UNAH
UNIVERSIDAD NACIONAL
AUTÓNOMA DE HONDURAS

Boletín Divulgativo

Carrera de Matemáticas

- Procesamiento de imágenes
- Estadística Robusta
- Modelación Matemática y Estadística
- Métodos para Análisis de Datos
- Matemática Pura

AGOSTO-2024

Presentación

Este documento fue desarrollado por el profesor David Méndez de la Carrera de Matemáticas de la UNAH, presenta artículos divulgativos y de investigación desarrollados por estudiantes del Seminario de Investigación de la Carrera de Matemáticas, curso desarrollado durante el segundo período académico del año 2024. Se abarca una temática variada: estimaciones robustas, modelos estadísticos y sus aplicaciones, análisis topológico de datos, machine learning, LLMs, programación funcional,...entre otros; en algunos de los trabajos se desarrolló una revisión bibliográfica de trabajos pertinentes y se resumió según lo comprendido por cada autor, en otros casos, se realizó experimentación original y se obtuvo resultados interesantes.

El objetivo principal de desarrollar este documento es que a futuro, en base a la experiencia obtenida y después de tener varias experiencias similares, se transforme en una revista científica de Matemáticas, cuestión que requiere de mucho trabajo por parte del equipo de profesores investigadores del programa y otros colaboradores externos; además de ser una muestra de que en la Carrera de Matemáticas se está desarrollando en los estudiantes un espíritu investigador. Cabe destacar que documentos similares se han desarrollado desde el programa de Maestría en Matemáticas de la UNAH.

Todas las revisiones bibliográficas y temas aquí presentados se encasillan dentro de las líneas de investigación de la UNAH, entre los temas prioritarios abarcados se encuentran: ciencia, cambio climático y vulnerabilidad, y productividad. Esto evidencia que la Carrera de Matemáticas está interesada en colaborar con las prioridades investigativas de la universidad y mantiene un compromiso con vincularse con la sociedad.

Se espera que a futuro se sigan desarrollando publicaciones similares, que a pesar de ser en su mayoría trabajos de divulgación, forman un inicio en la actividad investi-

gativa de la Carrera de Matemáticas.

Agosto del año 2024, Ciudad Universitaria

Tegucigalpa, M.D.C., Honduras

Edificio F1, Segundo Piso, Ciudad Universitaria
Tegucigalpa, M.D.C. Honduras.
https://matematica.unah.edu.hn/escuela/carreras/licenciaturas/
carrera.matematica@unah.edu.hn
Tel. 2216-3055

Contenido

ALGORITMOS DE SEGMENTACIÓN APLICADOS A IMÁGENES MÉDICAS

OLGA MICHELLE CARRASCO ANDINO

Resumen. En este artículo se estudian algoritmos de segmentación de imágenes, con un enfoque particular en la segmentación del cuerpo humano en imágenes médicas. El objetivo principal es desarrollar una base de datos precisa de imágenes segmentadas, la cual podrá ser utilizada para mejorar el diagnóstico médico. Se implementa y evalúa el algoritmo de Etiquetado de Componentes Conectados (CCL, por sus siglas en inglés) debido a su eficacia en la identificación y separación de estructuras anatómicas.

Abstract In this article, we study image segmentation algorithms, with a particular focus on human body segmentation in medical images. The main objective is to develop an accurate database of segmented images, which can be used to improve medical diagnosis. The Connected Component Labeling (CCL) algorithm is implemented and evaluated due to its effectiveness in identifying and separating anatomical structures.

1. Introducción

La segmentación de imágenes es uno de los pasos más importantes que conducen al análisis de datos de imágenes procesados, su objetivo principal es dividir una imagen en partes que tengan una fuerte correlación con objetos o áreas del mundo real contenidas en la imagen [1]. La segmentación puede entenderse como la realización de dos tareas esenciales: el reconocimiento y la delineación. El reconocimiento consiste en identificar los distintos objetos o regiones de interés dentro de la imagen. La delineación consiste en trazar los límites precisos de esos objetos o regiones para separarlos claramente del fondo y de otros elementos presentes en la imagen [2]. Esta combinación de tareas permite no solo reconocer qué está presente en una imagen, sino también definir con precisión dónde termina un objeto y comienza otro.

El desarrollo de algoritmos que segmentan imágenes surgió como consecuencia de la necesidad de contar con mecanismos automatizados de interpretación y análisis de imágenes. Es por ello, que la segmentación se ha convertido en un componente clave para el delineamiento de estructuras y otras regiones con el objetivo de asistir y automatizar ciertas tareas, siendo este el primer paso para la modelación tridimensional de una estructura determinada basado en su estructura real [2].

Fecha: Agosto 12, 2024.

Palabras y frases clave. segmentación de imágenes, segmentación del cuerpo humano, Etiquetado de Componentes Conectados, imágenes médicas.

Dado que el proceso de segmentación de imágenes juega un papel fundamental en el desarrollo de programas informáticos que utilizan tecnología de visión por computador, la aplicación de este proceso es muy amplia. Por ejemplo, los sistemas de videovigilancia utilizan la segmentación de imágenes para identificar objetos dentro de las grabaciones de vídeo [3]. Por otro lado, los profesionales del área de la salud recurren a la segmentación de imágenes cuando utilizan programas de imágenes médicas, ya que estos deben ser capaces de identificar elementos específicos del cuerpo humano. Mediante el proceso de segmentación de imágenes, se ha creado algoritmos de aprendizaje automático que pueden detectar tumores malignos en órganos específicos del cuerpo humano [4]. En relación con eso, la tecnología digital ha alterado realmente la forma en que los programas informáticos interactúan con el mundo físico, pues las fotografías ya no se limitan a sus formas físicas. La segmentación de imágenes tiene muchas aplicaciones en el mundo actual y seguramente se encontrarán nuevas aplicaciones para esta herramienta en un futuro próximo. Estos nuevos enfoques no solo mejoran la exactitud de la segmentación, sino que también reducen el tiempo y los costos asociados al análisis manual.

La segmentación de imágenes médicas es un desafío en el campo de la informática médica debido a la presencia de bajo contraste, ruido y detalles faltantes en las imágenes. Estas características dificultan la tarea de segmentar regiones basadas únicamente en la intensidad de los píxeles. Por lo cual nos ayudaremos del Algoritmo de Etiquetado de Componentes Conectados(CCL), una técnica que escanea la imagen para identificar y etiquetar regiones de píxeles conectados que comparten propiedades similares. Estas regiones, denominadas componentes conectados, pueden representar objetos, patrones o características relevantes en la imagen [5]. Por lo que, CCL es una herramienta versátil que se puede utilizar para resolver una variedad de problemas en el procesamiento de imágenes digitales. Es una técnica esencial para cualquier persona que trabaje en este campo.

Los resultados de este trabajo aportarán herramientas de preprocesado de imágenes para el proyecto de diagnóstico de cáncer de colon a partir de imágenes médicas con modelos de redes neuronales, llevado a cabo en colaboración con el Centro de Innovación en Realidad Extendida (CIRE) y el Centro de Innovación en Cómputo Científico (CICC). Este tema se encuentra dentro de los ejes prioritarios de investigación de la Universidad Nacional Autónoma de Honduras (UNAH), destacando su relevancia para el avance científico y tecnológico en el ámbito de la salud y la medicina. La UNAH apoya activamente esta línea de investigación debido a su potencial para mejorar las técnicas de diagnóstico y tratamiento del cáncer de colon.

2. Antecedentes

La segmentación de imágenes médicas ha avanzado considerablemente gracias a la participación de diversos investigadores y los desarrollos tecnológicos alcanzados. Geoffrey Hinton figura entre los pioneros destacados, reconocido por su trabajo seminal en redes neuronales profundas, las cuales han elevado la precisión y eficacia de los algoritmos de segmentación mediante el aprendizaje profundo [6, 7]. Sus contribuciones han sido fundamentales para el desarrollo de aplicaciones específicas en imágenes biomédicas, mejorando tanto la detección como la clasificación de estructuras de interés.

Un logro destacado en este campo es la arquitectura U-Net, desarrollada por Olaf Ronneberger y colaboradores en 2015, diseñada específicamente para mejorar la segmentación precisa de imágenes médicas [8]. Esta red convolucional ha sido ampliamente adoptada debido a su capacidad para capturar detalles finos y su eficiencia computacional en aplicaciones médicas críticas.
Igualmente, herramientas como SimpleITK, desarrolladas por Ziv Yaniv y su equipo en el 2013, han facilitado el procesamiento y la segmentación de imágenes médicas mediante una biblioteca robusta y accesible [9]. SimpleITK ha permitido a los investigadores y profesionales de la salud realizar análisis precisos, mejorando la interpretación clínica de imágenes complejas.

El avance hacia redes neuronales más profundas y complejas ha sido impulsado por trabajos como el de Kaiming He et al., quienes introdujeron las redes residuales (ResNet) en el año 2016, diseñadas para mitigar el problema de degradación del rendimiento a medida que aumenta la profundidad de la red [10].

Asimismo, en 1984 técnicas clásicas como el clustering difuso, desarrollado por James C. Bezdek, han encontrado aplicación en la segmentación de imágenes médicas para la detección y clasificación de regiones de interés [11]. Este enfoque ha proporcionado métodos efectivos para gestionar la variabilidad y complejidad de las imágenes médicas, mejorando la precisión y fiabilidad de los resultados.
En 1988, los modelos de contornos activos (snakes), propuestos por Kass, Witkin y Terzopoulos, han sido fundamentales en la delimitación precisa de estructuras anatómicas en imágenes médicas mediante la minimización de una energía definida [12].

Además de los avances tecnológicos y metodológicos mencionados, la segmentación de imágenes médicas ha evolucionado en áreas específicas como la neuroimagen y la radiología diagnóstica. Por ejemplo, técnicas avanzadas de segmentación basadas en redes neuronales recurrentes (RNNs) ha demostrado eficacia en la segmentación precisa de estructuras cerebrales en imágenes de resonancia magnética (MRI) [13]

En el ámbito de la medicina personalizada, la segmentación automática de imágenes es esencial al permitir la cuantificación precisa de biomarcadores en imágenes de medicina nuclear y PET-CT. Herramientas como las redes generativas adversarias (GANs) están siendo exploradas para mejorar la calidad de las segmentaciones y superar desafíos como el ruido y la variabilidad en imágenes clínicas [14].

Un avance innovador es la integración de la inteligencia artificial (IA) en sistemas de asistencia quirúrgica, donde la segmentación precisa en tiempo real de imágenes intraoperatorias es crucial para guiar intervenciones quirúrgicas mínimamente invasivas, no solo mejoran la precisión de la cirugía, sino que también reducen el tiempo de recuperación del paciente y los costos asociados con procedimientos prolongados [15]. En cuanto a los retos futuros, la interpretación precisa y reproducible de imágenes médicas sigue siendo un área de investigación activa. La integración de técnicas de aprendizaje semi-supervisado y de transferencia de aprendizaje se perfila como una dirección prometedora para mejorar la generalización de los modelos de segmentación a diferentes tipos de datos clínicos y escenarios de adquisición [16].

3. Segmentación en el Procesamiento de Imágenes

En el procesamiento de imágenes digitales y la visión por computadora , la segmentación de imágenes es el proceso de dividir una imagen digital en múltiples segmentos de imagen , también conocidos como regiones de imagen u objetos de imagen, que consisten en conjuntos de píxeles. La finalidad de la segmentación es simplificar o transformar la representación de una imagen para que sea más significativa y más sencilla de analizar [17].

Generalmente, la segmentación de imágenes se emplea para identificar objetos y límites (líneas, curvas, etc.) en las imágenes. En términos más específicos, la segmentación de imágenes consiste en asignar una etiqueta a cada píxel de una imagen, de modo que los píxeles con la misma etiqueta compartan determinadas características. El resultado de la segmentación de imágenes puede ser un conjunto de segmentos que abarcan toda la imagen en su totalidad, o un conjunto de contornos extraídos de la imagen.

Este problema general se descompone en problemas más específicos, dando lugar a ejemplos como:

- Segmentación por color.
- Segmentación por texturas.
- Superpíxel.
- Segmentación semántica.

Cada uno de estos problemas especializados asigna un significado particular a las categorías utilizadas para clasificar los píxeles. Uno de los casos más básicos de segmentación es la umbralización, que es una forma específica de segmentación por color con solo dos categorías: claro y oscuro. En este método, cada píxel se clasifica como claro u oscuro comparando su intensidad con una intensidad de referencia, denominada como umbral [1].

3.1. Aspectos Fundamentales de la Segmentación. Cada algoritmo de segmentación distingue una cierta cantidad de clases o categorías, y a cada una le asigna una etiqueta o identificador, un valor entero que el algoritmo utiliza para representar esa categoría. La clasificación mínima posible es la que distingue dos categorías, un caso comúnmente conocido como segmentación binaria [17].
Todos los algoritmos de segmentación trabajan sobre una imagen, que puede ser a color o en blanco y negro, y produce otra imagen artificial del mismo tamaño. En esta imagen resultante, cada píxel está marcado con una etiqueta que representa su categoría. Las áreas formadas por píxeles contiguos con la misma etiqueta se conocen como segmentos. Estos segmentos pueden visualizarse como piezas de un rompecabezas: no se superponen entre sí y juntos cubren la totalidad de la imagen [17].
La imagen resultante de una segmentación está optimizada para su procesamiento computacional, pero no necesariamente para ser visualizada directamente. Para mostrar los resultados de manera comprensible a un usuario, se requiere una etapa de visualización que asigne colores fácilmente distinguibles a cada etiqueta. En lugar de utilizar una imagen con etiquetas, el resultado de una segmentación puede

ser representado mediante contornos: líneas cerradas que delinean los bordes de los segmentos de forma arbitraria [1].

4. Etiquetado de Componentes Conectados

4.1. Definiciones y conceptos preliminares. En el análisis de imágenes y procesamiento de imágenes, la identificación y etiquetado de componentes conectados es una técnica fundamental. Para comprender este proceso, es esencial familiarizarse con varios conceptos clave [18].

Una imagen binaria B se puede obtener a partir de una imagen en escala de grises o en color mediante una operación que selecciona un subconjunto de los píxeles de la imagen como píxeles de primer plano, los píxeles de interés en una tarea de análisis de imágenes, dejando el resto como píxeles de fondo para ser ignorados. Los píxeles de una imagen binaria B son 0's y 1's; los 1's se utilizarán para indicar los píxeles de primer plano y los 0's, los píxeles de fondo [17].

Definición 4.1. (Píxel). Un píxel es la unidad más pequeña de una imagen digital, representada por una posición en una cuadrícula y un valor que define su color o intensidad. Cada píxel contribuye a formar la imagen completa en una matriz de coordenadas $[i, j]$.

Definición 4.2. (Vecinos). Dos píxeles son vecinos de 4 (N_4) si comparten un borde común. Para un píxel ubicado en la posición $[i, j]$, sus vecinos de 4 son $[i-1, j], [i+1, j], [i, j-1], [i, j+1]$ hace referencia a norte, sur, oeste, este respectivamente. Ver Figura 1. Además de los vecinos de 4, un píxel también tiene vecinos diagonales que comparten una esquina. Para un píxel en la posición $[i, j]$, sus vecinos de 8 (N_8) incluyen todos los vecinos de 4 más $[i-1, j-1], [i-1, j+1], [i+1, j-1], [i+1, j+1]$ hace referencia a noroeste, noreste, suroeste, sureste respectivamente. Ver Figura 2.

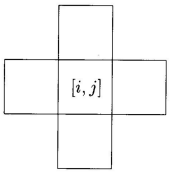


Figura 1. 4-vecinos [18].

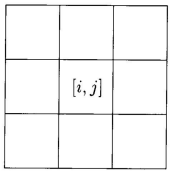
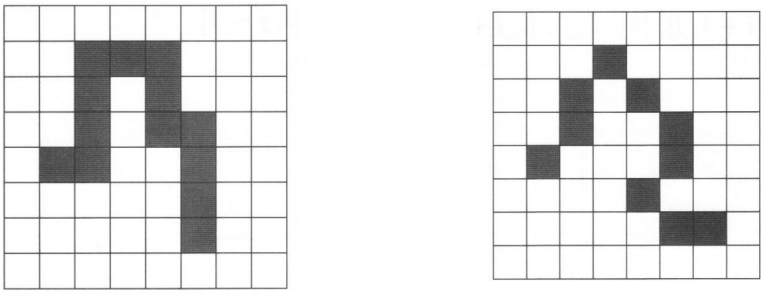


Figura 2. 8-vecinos [18].

Definición 4.3. (Conectividad-Vecinos). Un píxel está conectado en 4 a otros píxeles si estos están entre sus vecinos de 4. Un píxel está conectado en 8 a otros píxeles si estos están entre sus vecinos de 8.

Definición 4.4. (Camino). Un camino desde el píxel en $[i_0, j_0]$ hasta el píxel en $[i_n, j_n]$ es una secuencia de índices de píxeles $[i_0, j_0], [i_1, j_1], [i_2, j_2], \ldots, [i_n, j_n]$ tal que el píxel en $[i_k, j_k]$ es vecino del píxel en $[i_{k+1}, j_{k+1}]$ para todo k con $0 \leq k \leq n-1$. Si la relación de vecindad utiliza la conexión de 4, entonces el camino es un 4-camino; para la conexión de 8, el camino es un 8-camino. Ver Fiugura 3.



(a) 4-Caminos

(b) 8-Caminos

FIGURA 3. Ejemplo de 4-caminos y 8-caminos [18].

En la Figura 3.a) 4-camino utiliza la conexión de 4. Cada píxel comparte un borde común (norte, sur, este, oeste) con el siguiente. En la Figura 3.b) 8-camino utiliza la conexión de 8. Cada píxel comparte un borde común o una esquina (norte, sur, este, oeste, noreste, noroeste, sureste, suroeste) con el siguiente. Este ejemplo es importante para entender cómo los píxeles se agrupan y forman estructuras en la imagen.

Definición 4.5. (Primer Plano). El conjunto de todos los píxeles con valor 1 en una imagen se llama el primer plano y se denota por S.

Definición 4.6. (Conectividad). Un píxel $p \in S$ se dice que está conectado a $q \in S$ si hay un camino desde p hasta q que consiste únicamente en píxeles de S.

Nótese que la conectividad es una relación de equivalencia. Para cualquier tres píxeles p, q y r en S, tenemos las siguientes propiedades:

1. El píxel p está conectado consigo mismo (reflexividad).
2. Si p está conectado a q, entonces q está conectado a p (simetría).
3. Si p está conectado a q y q está conectado a r, entonces p está conectado a r (transitividad).

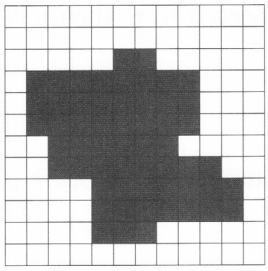
Definición 4.7. (Componentes Conectados). Un conjunto de píxeles en el cual cada píxel está conectado a todos los demás píxeles se llama un componente conectado.

Definición 4.8. (Fondo). El conjunto de todos los componentes conectados de $\bar{S}$ (el complemento de S) que tienen puntos en el borde de una imagen se llama fondo. Todos los otros componentes de $\bar{S}$ se llaman agujeros.

Definición 4.9. (Borde o Frontera). El borde de S es el conjunto de píxeles de S que tienen vecinos de 4 en $\bar{S}$. Normalmente, el borde se denota como S'.

Definición 4.10. (Interior). El interior es el conjunto de píxeles de S que no están en su borde. El interior de S es $(S - S')$.

Definición 4.11. (Rodear). La región T rodea a la región S (o S está dentro de T) si cualquier camino desde cualquier punto de S hasta el borde de la imagen debe intersectar T.



(a) Imagen Original

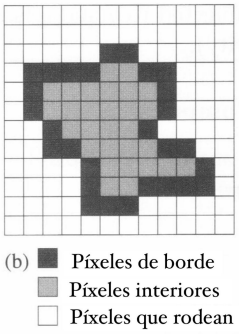


FIGURA 4. Una imagen binaria con su borde, interior y píxeles que la rodean. [18].

4.2. Fundamentos del Etiquetado de Componentes Conectados. Una vez definidos estos conceptos preliminares, procedemos a abordar el CCL. Este proceso es esencial en el análisis de imágenes digitales, ya que identifica y asigna etiquetas únicas a grupos de píxeles conectados, permitiendo un análisis detallado de la distribución y características de objetos en la imagen. En esta etapa, se busca no solo etiquetar cada grupo de píxeles contiguos, sino también comprender cómo se relacionan entre sí dentro de la imagen, facilitando así análisis más profundos y específicos sobre la distribución y características de los objetos o estructuras presentes en la imagen digital.

El CCL es una operación sobre imágenes binarias que se emplea para distinguir objetos (previamente binarizados) presentes en una imagen [17].

Los algoritmos CCL asignan un identificador único a cada grupo de píxeles conectados que comparten las mismas propiedades en una imagen. Este estudio se enfoca en algoritmos que, a partir de una imagen binaria (B), analizan las conectividades entre píxeles dentro de un vecindario de cuatro (N_4) u ocho vecinos (N_8). De esta manera, al final del proceso, dos píxeles, p y q, pertenecerán al mismo componente si ambos forman parte del primer plano o del fondo y existe un camino de píxeles

del mismo tipo que los conecte, cumpliendo la ecuación correspondiente, donde S es un subconjunto de píxeles de la imagen binaria B [19].

$$p \text{ conectado a } q \iff \exists\{s_i \in S \mid s_1 = p, s_{n+1} = q, s_{i+1} \in N(s_i), i = 1, \ldots, n\}$$

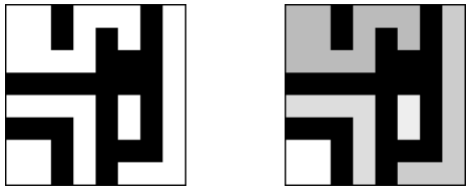
Esta definición establece claramente que p está conectado a q si existe una secuencia de píxeles en S donde cada píxel está conectado a su sucesor según la relación de vecindad definida por $N(s_i)$.

1	1	0	1	1	1	0	1
1	1	0	1	0	1	0	1
1	1	1	1	0	0	0	1
0	0	0	0	0	0	0	1
1	1	1	1	0	1	0	1
0	0	0	1	0	1	0	1
1	1	0	1	0	0	0	1
1	1	0	1	0	1	1	1

(a) Imagen Binaria

1	1	0	1	1	1	0	2
1	1	0	1	0	1	0	2
1	1	1	1	0	0	0	2
0	0	0	0	0	0	0	2
3	3	3	3	0	4	0	2
0	0	0	3	0	4	0	2
5	5	0	3	0	0	0	2
5	5	0	3	0	2	2	2

(b) Etiquetado de Componentes Conectados



(c) Imagen Binaria Etiquetada, ampliada para su visualización

FIGURA 5. Una imagen binaria con cinco componentes conectados del valor 1 [17].

En la figura 5.a) muestra una imagen binaria con cinco de dichos componentes conectados de 1's; estos componentes en realidad están conectados con respecto a la definición de ocho o cuatro vecindarios.

Una etiqueta es un símbolo que nombra de forma única una entidad. Si bien las etiquetas de caracteres son posibles, los números enteros positivos son más convenientes y se utilizan con mayor frecuencia para etiquetar los componentes conectados. La Figura 5.b) muestra el etiquetado de los componentes conectados de la imagen binaria de la Figura 5.a).

Existen diversos algoritmos para realizar el etiquetado de componentes conectados. Algunos algoritmos asumen que toda la imagen puede ser almacenada en memoria y emplean un enfoque simple y recursivo que opera en un componente a la vez, moviéndose por toda la imagen durante el proceso. Otros algoritmos están diseñados para imágenes más grandes que no caben completamente en memoria, operando únicamente en dos filas de la imagen simultáneamente. Además, existen algoritmos diseñados para sistemas altamente paralelos, que emplean estrategias de propagación paralela para mejorar la eficiencia y el rendimiento [17].

4.3. Algoritmos de Etiquetados de Componentes Conectados. Desde los años 60, se han dedicado considerables esfuerzos al avance y desarrollo de los algoritmos CCL. Varios criterios permiten clasificar los distintos procedimientos, como la eficiencia en el acceso a memoria y la forma de representación de la imagen. Sin embargo, es común representar la imagen como una matriz bidimensional y clasificar los algoritmos según el número de pasadas que realizan sobre la misma, entendiendo una pasada como la exploración de todos los píxeles sin importar el orden seguido. De este modo, podemos identificar algoritmos que (tomado de [19]):

4.3.1. Algoritmos de Una Pasada. Estos algoritmos recorren la imagen una sola vez, pero enfrentan desafíos como accesos irregulares y aleatorios a las estructuras de datos que almacenan la imagen o las etiquetas asignadas, lo cual dificulta predecir los tiempos de ejecución. Este método es rápido y fácil de implementar, basado en técnicas de recorrido de grafos de la teoría de grafos. En síntesis, una vez localizado el primer píxel de un componente conectado, se etiquetan todos los píxeles conectados de ese componente antes de pasar al siguiente píxel de la imagen. Es importante destacar que este proceso se refiere al enfoque de los algoritmos de una pasada en general. Sin embargo, no todos los algoritmos de una pasada necesariamente siguen este método exacto. Algunos pueden tener variaciones en cómo identifican y etiquetan los componentes conectados, aunque la idea principal de recorrer la imagen una sola vez se mantiene.

4.3.2. Algoritmos de Múltiples Pasadas. Estos algoritmos realizan varios pasadas de la imagen, típicamente en patrones alternos como de arriba abajo y de izquierda a derecha, y de abajo arriba y de derecha a izquierda. Esto asegura un acceso regular a la memoria. El tiempo de ejecución de estos procedimientos depende de la disposición específica de los píxeles en cada imagen, lo que dificulta establecer de antemano la duración del método en cada caso. Sin embargo, su implementación tanto en software como en hardware suele ser más sencilla en comparación con algoritmos de otros grupos. Esta versión aclara cómo estos algoritmos manejan el acceso a la memoria y cómo varía el tiempo de ejecución según la disposición de los píxeles en la imagen, destacando la simplicidad relativa de su implementación.

4.3.3. Algoritmos de dos Pasadas. El algoritmo realiza dos pasadas sobre la imagen: la primera para asignar etiquetas temporales y establecer equivalencias entre registros, y la segunda para sustituir cada etiqueta temporal por la más pequeña de su clase equivalente. Los datos de entrada pueden ser modificados directamente (con el riesgo de corrupción de datos) o la información de etiquetado puede mantenerse en una estructura de datos adicional.

4.4. Algoritmo clásico de etiquetado de componentes conectados en dos pasos con Union-Find. Se han desarrollado diversos algoritmos CCL que optimizan la complejidad espacial, la complejidad temporal y el procesamiento paralelo. En este artículo, analizaremos el algoritmo clásico diseñado por Rosenfeld y Pfaltz en 1966, el cual se basa en resultados de la teoría de grafos. Este algoritmo escanea la imagen de manera secuencial, de arriba a abajo y de izquierda a derecha, etiquetando cada píxel según las etiquetas de los píxeles vecinos. Posteriormente, se realiza una segunda pasada para corregir las inconsistencias de etiquetado causadas

por la forma de ciertos componentes. Las etiquetas equivalentes se almacenan en una estructura de datos "Union-Find" [20].

El propósito de la estructura de datos Union-Find es almacenar una colección de conjuntos disjuntos y proporcionar implementaciones eficientes para las operaciones de unión y búsqueda. La operación de unión fusiona dos conjuntos en uno solo, mientras que la operación de búsqueda determina a qué conjunto pertenece un elemento particular. Esta estructura de datos es ideal para manejar las etiquetas equivalentes durante la segmentación de imágenes en el algoritmo clásico de etiquetado de componentes conexos (CCL) de Rosenfeld y Pfaltz. En el contexto del CCL, cada conjunto representa un grupo de píxeles conectados que comparten la misma etiqueta. Los conjuntos se almacenan utilizando una estructura de árbol, donde cada nodo del árbol representa una etiqueta de píxel y apunta a su nodo principal. Inicialmente, cada píxel etiquetado es su propio conjunto y, por lo tanto, su propio nodo raíz en el árbol. La estructura de datos Union-Find permite manejar eficientemente las etiquetas equivalentes durante el proceso de etiquetado de componentes conexos, garantizando que las operaciones de unión y búsqueda se realicen en tiempo casi constante, lo que es esencial para el rendimiento del algoritmo [17].

4.4.1. Implementación del Algoritmo Clásico de CCL en dos pasos con Union-Find. La siguiente implementación del algoritmo está basado en [20].

Limitaremos nuestras entradas a imágenes binarias (en blanco y negro), donde cada píxel puede ser un píxel de primer plano (negro) o un píxel de fondo (blanco). Utilizaremos números enteros positivos para etiquetar los componentes de primer plano, mientras que los píxeles de fondo serán etiquetados con '0'. Por lo tanto, nuestro objetivo es que el algoritmo realice la siguiente operación:

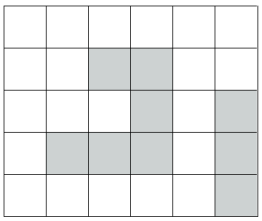


Imagen de Entrada

0	0	0	0	0	0
0	0	1	1	0	0
0	0	0	1	0	2
0	1	1	1	0	2
0	0	0	0	0	2

Salida con componentes etiquetados

FIGURA 6. Objetivo del Algoritmo CCL [20].

En la figura 6, la imagen de entrada es una imagen binaria (5 píxeles de alto y 6 píxeles de ancho) que contiene 2 componentes conectados.

Primer Pase. En la primera pasada, escaneamos la imagen píxel por píxel y examinamos los píxeles vecinos inmediatos de cada píxel. Si un píxel es adyacente a un píxel ya etiquetado, se considera que forma parte del mismo componente y, por lo tanto, debe recibir la misma etiqueta que su vecino. La selección de los píxeles vecinos que se consideran (la "conectividad") depende del propósito del análisis de

imagen y puede variar según el tipo de imagen y los objetivos específicos.
Las dos conectividades más comúnmente utilizadas son la conectividad 4 y la conectividad 8.

Al escanear la imagen fila por fila, de arriba a abajo, y dentro de cada fila de izquierda a derecha, solo necesitamos examinar los píxeles vecinos que están ubicados arriba y a la izquierda del píxel actual. Esto se debe a la dirección del escaneo: los píxeles a la derecha y debajo del píxel actual aún no se han procesado, por lo que no han sido etiquetados. Por lo tanto, nuestro núcleo de etiquetado (la metodología que utilizamos para escanear la imagen y considerar los píxeles vecinos) se configura de la siguiente manera:

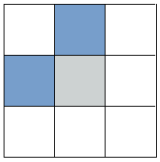


FIGURA 7. Nucleo de Etiquetado (4-conectividad) [20].

Una vez que se han recuperado las etiquetas de los vecinos relevantes, pueden surgir tres escenarios posibles:

1. El píxel no tiene vecinos etiquetados (es decir, todos los vecinos son píxeles de fondo). En este caso, el píxel es el primer miembro de una nueva forma y debe recibir una nueva etiqueta. Se utiliza un contador de etiquetas para asignar una etiqueta única y asegurar que no se repita ninguna etiqueta previamente usada.

2. El píxel tiene uno o más vecinos con la misma etiqueta. Esto indica que el píxel actual pertenece al mismo componente que sus vecinos, por lo que se le asigna la misma etiqueta que la de sus vecinos.

3. El píxel tiene vecinos con diferentes etiquetas. Esto significa que el píxel es parte del mismo componente que sus vecinos, pero se ha detectado una inconsistencia en el etiquetado, ya que el algoritmo no había identificado previamente que estos vecinos pertenecían al mismo componente. En este caso, se le asigna al píxel la etiqueta más pequeña de entre las etiquetas de sus vecinos, y la inconsistencia será corregida en la segunda pasada del algoritmo.

Manejo de inconsistencias. Algunos patrones específicos en la imagen pueden causar inconsistencias en el etiquetado, como se describe en el escenario 3 anterior, generando componentes con áreas de píxeles que tienen etiquetas distintas. Esto se debe a la dirección en la que escaneamos la imagen: los componentes que se conectan solo en el borde derecho pueden parecer como dos componentes distintos hasta que se detecta la sección de conexión en el borde derecho, lo que lleva a la asignación de etiquetas diferentes.

Para corregir estos errores de etiquetado, es importante registrar las equivalencias de etiquetas a medida que se encuentran durante la primera pasada. Estas equivalencias se registran porque en realidad deberían ser la misma etiqueta. Una vez que se han registrado estas equivalencias, podemos corregirlas mediante una segunda pasada de la imagen. Utilizamos la estructura de datos Union-Find (también conocida como Disjoint-Set) para almacenar y gestionar eficientemente las equivalencias de etiquetas. Esta estructura realiza un seguimiento de las etiquetas equivalentes y permite recuperar de manera rápida la etiqueta más baja (la "representativa") en cada conjunto de etiquetas equivalentes.

Así, si durante la primera pasada encontramos la siguiente situación:

0	1	0
3		

FIGURA 8. Inconsistencia, escenario 3 [20].

A mitad del primer paso, nos encontramos con un píxel cuyos vecinos tienen etiquetas diferentes. En nuestra estructura de datos de conjunto disjunto, registramos que las etiquetas 1 y 3 son equivalentes.

Segundo Pase. Utilizamos las equivalencias de etiquetas registradas para corregir cualquier inconsistencia detectada durante la primera pasada.
Realizamos un nuevo escaneo de la imagen, píxel por píxel. Para cada píxel etiquetado, comprobamos si su etiqueta tiene equivalencias en nuestra estructura de datos de conjuntos disjuntos. Si encontramos alguna equivalencia, actualizamos la etiqueta del píxel con la etiqueta más baja (o 'representativa') de su conjunto de equivalencias. Finalmente, aseguramos que todas las etiquetas sean reemplazadas por la etiqueta más baja dentro de su conjunto. Con esto, las etiquetas están uniformemente corregidas, ver Figura 9.

0	0	0	0	0	0
0	0	1	1	0	0
0	0	0	1	0	2
0	3	3	1	0	2
0	0	0	0	0	2

Resultado del primer pase

0	0	0	0	0	0
0	0	1	1	0	0
0	0	0	1	0	2
0	1	1	1	0	2
0	0	0	0	0	2

Resultado del segundo pase - finito

FIGURA 9. Resultados del CCL [20].

4.5. Análisis de Complejidad del Algoritmo de Etiquetado de Componentes Conectados utilizando Union-Find. La complejidad computacional del algoritmo de etiquetado de componentes conectados (CCL) utilizando la estructura de datos Union-Find en dos pasos se puede desglosar en función de las dos fases principales del proceso: la primera pasada para etiquetar y registrar equivalencias, y la segunda pasada para corregir las etiquetas.

4.5.1. Primera Pasada.

Operaciones Principales.

- Escaneo de la Imagen: Cada píxel de la imagen se examina una vez. Para una imagen de tamaño $m \times n$, esto toma $O(m \cdot n)$ tiempo, donde m es el número de filas y n es el número de columnas.
- Unión y Búsqueda en Union-Find: Para cada píxel, se realizan operaciones de unión y búsqueda para gestionar las equivalencias de etiquetas. La estructura Union-Find con compresión de caminos y unión por rango proporciona una complejidad reducida casi constante para estas operaciones, aproximadamente $O(\alpha(m \cdot n))$, donde α es la función inversa de Ackermann, que crece muy lentamente.

En el peor caso, el número de operaciones de unión y búsqueda que se realizan es proporcional al número de píxeles, es decir, $O(m \cdot n)$. Como estas operaciones son reducidas casi constantes, el tiempo total para esta fase se puede considerar como $O(m \cdot n \cdot \alpha(m \cdot n))$ [21, 22].

Complejidad Total de la Primera Pasada.

- Tiempo: $O(m \cdot n \cdot \alpha(m \cdot n))$, que se aproxima a $O(m \cdot n)$ para la mayoría de las aplicaciones prácticas debido al crecimiento lento de α.
- Espacio: $O(m \cdot n)$ para almacenar las etiquetas y las estructuras de datos Union-Find.

4.5.2. Segunda Pasada.

Operaciones Principales.

- Escaneo de la Imagen para Actualización de Etiquetas: Cada píxel es revisado nuevamente para actualizar su etiqueta a la etiqueta más baja en su conjunto de equivalencias. Esto toma $O(m \cdot n)$ tiempo.
- Búsqueda en Union-Find: La búsqueda para encontrar la etiqueta representativa de cada píxel también tiene una complejidad de $O(\alpha(m \cdot n))$ por operación, pero debido a que se realiza una sola vez por píxel, esto también se aproxima a $O(m \cdot n)$ en la práctica.

Complejidad Total de la Segunda Pasada:

- Tiempo: $O(m \cdot n)$ debido al escaneo y a las búsquedas en Union-Find.
- Espacio: $O(m \cdot n)$ para almacenar la imagen y las etiquetas actualizadas.

4.5.3. Complejidad Total Combinada. La complejidad computacional total del algoritmo de etiquetado de componentes conectados utilizando Union-Find en dos pasos es la suma de las complejidades de ambas pasadas:

- Tiempo Total: $O(m \cdot n \cdot \alpha(m \cdot n))$, en el peor caso teórico. En la práctica, se aproxima a $O(m \cdot n)$.
- Espacio Total: $O(m \cdot n)$ para almacenar las etiquetas y las estructuras de datos Union-Find.

La complejidad práctica del algoritmo de etiquetado de componentes conectados en dos pasos con Union-Find es $O(m \cdot n)$ tanto en tiempo como en espacio, lo que lo hace muy eficiente para imágenes de gran tamaño. Aunque el análisis teórico puede mostrar $O(m \cdot n \cdot \alpha(m \cdot n))$ debido a la función inversa de Ackermann, en la práctica, se considera casi constante [21, 22].

4.6. Aplicación del Algoritmo de Dos Pases con Union-Find para la Etiquetación de Componentes Conectados en Imágenes Binarias. Se implementó y se aplicó el algoritmo de dos pases para la etiquetación de componentes conectados en una imagen binaria. A continuación, se presenta la imagen original y la imagen etiquetada:

(a) Imagen Original (b) Imagen Etiquetada

FIGURA 10. Ejemplo 1. Elaboración propia

En la Figura 10.b), se pueden observar diferentes regiones conectadas que han sido etiquetadas con valores distintos. El algoritmo de dos pases se encarga de identificar y etiquetar estas regiones conectadas, facilitando así el análisis y procesamiento de la imagen.

En este segundo ejemplo, se utiliza nuevamente el algoritmo de dos pases para la etiquetación de componentes conectados en una imagen binaria. Se presentan tanto la imagen original como la imagen etiquetada:

(a) Imagen Original (b) Imagen Etiquetada

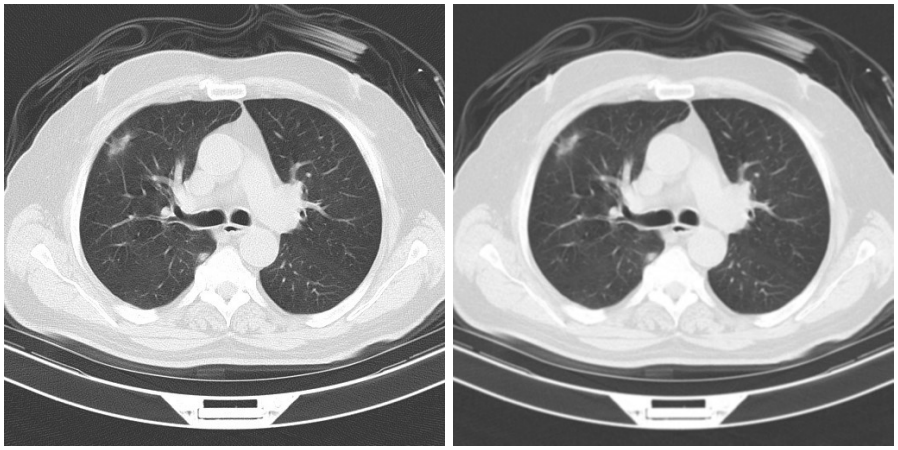
FIGURA 11. Ejemplo 2. Elaboracion Propia

En la Figura 11.b), se pueden distinguir varias áreas conectadas, cada una etiquetada con valores únicos. El algoritmo de dos pases ha logrado identificar y etiquetar de manera precisa estos componentes conectados, lo que simplifica su análisis y procesamiento.

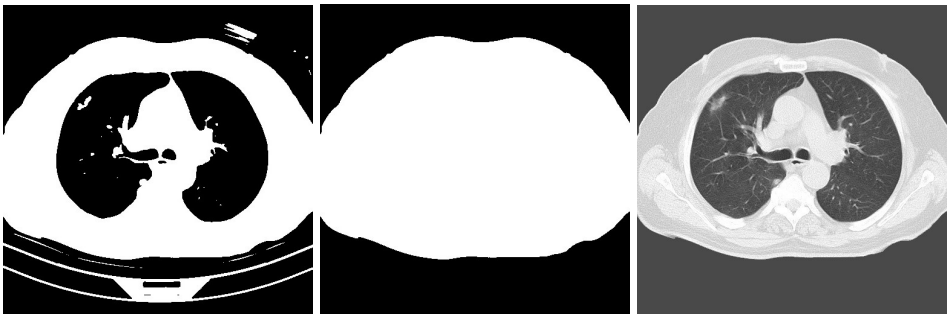
4.7. Procesamiento de Imágenes para Aislar Contornos Corporales y Eliminar Elementos Externos. El proceso de segmentación realizado en este ejemplo tiene como objetivo separar el cuerpo del paciente del fondo en una imagen en escala de grises. Primero, se aplica un filtro Gaussiano a la imagen para suavizarla, lo que ayuda a reducir el ruido y las variaciones menores de intensidad, preparando así la imagen para una segmentación más precisa. Luego, se utiliza el método de Otsu para realizar la segmentación, que determina automáticamente un valor de umbral que separa los píxeles de la imagen en dos categorías: fondo y objeto (en este caso, el cuerpo del paciente). Los píxeles con valores de intensidad por encima de este umbral se consideran parte del cuerpo, mientras que los píxeles por debajo se consideran parte del fondo. A partir del umbral calculado, se genera una imagen binaria en la que los píxeles pertenecientes al cuerpo se muestran en blanco (valor 255) y los píxeles del fondo se muestran en negro (valor 0), simplificando así la identificación de regiones de interés.

En la imagen binaria, se identifican todos los contornos presentes, y el contorno más grande, que corresponde al área del cuerpo, se extrae para crear una máscara que delimita la región del cuerpo. Usando este contorno principal, se genera una máscara que cubre únicamente el cuerpo del paciente, con el área del cuerpo en blanco y el resto en negro. Para asegurar que cualquier región brillante del fondo no se incluya erróneamente en la máscara del cuerpo, se revisan las áreas fuera del cuerpo. Se identifican las regiones oscuras del fondo, se calcula su valor medio y se reemplazan las áreas exteriores al cuerpo en la imagen original con este valor medio, eliminando visualmente el fondo. Finalmente, la imagen segmentada muestra el cuerpo del paciente sin las interferencias del fondo, con las áreas exteriores reemplazadas por un valor medio que no distrae del área de interés [1].

No se realiza explícitamente un etiquetado de componentes conectados. El enfoque aquí está más centrado en la segmentación y el aislamiento de una región de interés (el cuerpo) y la eliminación de detalles externos. El etiquetado de componentes conectados sería un paso adicional si se quisiera identificar y etiquetar regiones individuales conectadas dentro de la imagen segmentada.



(a) Imagen Original [23]. (b) Suavizado Gaussiano



(c) Binaria (d)Mascara (e) Segmentada

FIGURA 12. Procesamiento de Imágenes para Identificación de Contornos.

Este ejemplo ilustra cómo se puede procesar una imagen médica para eliminar características externas y destacar la región de interés, utilizando técnicas de procesamiento de imágenes en Python con OpenCV.

5. Conclusiones

Después de haber estudiado el proceso de segmentación de imágenes y el etiquetado de componentes conectados (CCL), junto con el uso del algoritmo Union-Find, es momento de sintetizar los hallazgos clave. A continuación, se presentan las conclusiones derivadas de este análisis, subrayando la importancia y el impacto de estas técnicas en el procesamiento de imágenes.

1. La segmentación de imágenes ayuda a reducir la cantidad de datos que necesitan ser procesados, almacenados y transmitidos. Al enfocarse en las regiones de interés, se optimiza el uso de recursos computacionales y de almacenamiento, lo que es especialmente importante en aplicaciones que manejan grandes volúmenes de datos de imagen. La segmentación de imágenes sigue siendo un área activa de investigación e innovación. Los avances en algoritmos, técnicas de preprocesamiento y tecnologías de hardware continúan mejorando su rendimiento y ampliando su aplicabilidad,

2. El CCL es eficiente para identificar regiones o componentes en una imagen que están conectadas. Esto es primordial en aplicaciones de procesamiento de imágenes, donde la segmentación y el análisis de regiones conectadas son importantes, incluso el algoritmo se adapta bien a imágenes de diferentes tamaños y resoluciones. Su rendimiento puede escalarse en función del tamaño de la imagen y del número de componentes conectados.

3. Como se observo en los ejemplos, el CCL se puede aplicar tanto a imágenes binarias como a imágenes en escala de grises, pero su eficacia es más pronunciada en imágenes binarias, donde el objetivo es identificar regiones de píxeles conectados. Para imágenes en escala de grises o en color, se requieren pasos adicionales de preprocesamiento o modificaciones en el algoritmo

4. Durante el estudio, se pudo determinar que el uso del algoritmo Union-Find en el contexto de CCL mejora la eficiencia del etiquetado de componentes conectados. La capacidad de unir y buscar componentes de manera rápida reduce el tiempo de ejecución, especialmente en imágenes grandes o complejas. Al igual, Union-Find reduce la complejidad computacional en comparación con otros métodos tradicionales, pero en aplicaciones especializadas, como la segmentación de imágenes médicas o la detección de objetos en entornos industriales, el uso de Union-Find puede requerir ajustes y calibraciones adicionales para abordar características específicas de las imágenes

5. Las futuras investigaciones podrían centrarse en mejorar aún más la eficiencia del etiquetado de componentes conectados, explorando nuevas técnicas de optimización y adaptando el algoritmo a diferentes tipos de imágenes y aplicaciones emergentes.

Referencias

1. Milan Sonka, Vaclav Hlavac, Roger Boyle *Image Proccessing, Analysis, and Machine Vision Fourth Edition*, 2015, 2008 Cengage Learning.
2. Walter Maximiliano Martínez Krawczuk, *Segmentacion de Imagenes Medicas*, Universidad Nacional de La Plata, 2008.

3. Natalia Gutierrez, Manuel Velazco *Diseño y simulación de un sistema de detección de movimiento basado en segmentación de imágenes utilizando los método de umbralización y sustracción de fondo aplicado a un sistema de videovigilancia*, Universidad de San Martin de Porres, 2019.
4. Carlos Abadía Cutillas *Detección de tumores en imagénes medicas mediante transformers* Universidad de Alicante, 2023.
5. Dr.Mrs. A.J.VYAVAHARE *Connected Component based Medical Image Segmentation* IJIREEICE, 2014.
6. Hinton, G. E., & Salakhutdinov, R. R *Reducing the dimensionality of data with neural networks. Science* 2006, 313(5786), 504-507.
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E.*Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems* 2012, 25, 1097-1105.
8. Ronneberger, O., Fischer, P., & Brox, T. *U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention* Springer, Cham 2015, (pp. 234-241).
9. Lowekamp, B. C., Chen, D. T., Ibanez, L., & Blezek, D *The design of SimpleITK. Frontiers in neuroinformatics,* 2013.
10. He, K., Zhang, X., Ren, S., & Sun, J. *Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition* 2016, (pp. 770-778).
11. Bezdek, J. C., Ehrlich, R., & Full, W. *FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences* 1984.
12. Kass, M., Witkin, A., & Terzopoulos, D. *Snakes: Active contour models. International journal of computer vision, 1* 1988, 321-331.
13. Brosch, T., & Tam, R. *3D convolutional neural networks for efficient and robust segmentation of anatomical structures in MRI: A comparison study with 3D U-Net. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* MICCAI, 2016.
14. Bowles, C., Taneja, S., & Farahani, K. *Applications of generative adversarial networks in medical imaging. IEEE Reviews in Biomedical Engineering* 2020, 13, 98-113.
15. Liu, F., Zhou, Z., & Jang, J*Deep learning for medical image segmentation: A review. Computers in Biology and Medicine* 2022, 144, 103698.
16. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L.*DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence* 2023, 40(4), 834-848.
17. Linda G. Shapiro & George C. Stockman *Computer Vision*, Prentice Hall, 2002.
18. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck *Machine Vision*, McGraw-Hil, 1995.
19. Elisa Calvo. *Un algoritmo en tiempo real para etiquetado de componentes conectados en imágenes*, Universidad de Sevilla. Sevilla España.
20. Jack Lawrence-Jones *Etiquetado de Componentes Conectados*, 2016.
21. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. *Introduction to Algorithms*, (3rd ed.). MIT Press. (2009) Sección sobre Union-Find.
22. Mehlhorn, K., & Sanders, P. *Algorithms and Data Structures: The Basic Toolbox*. Springer. (2009) Sección sobre Union-Find.
23. Kang Zhang, Xiaohong Liu, Jun Shen, et al. Jianxing He, Tianxin Lin, Weimin Li, Guangyu Wang. *Sistema de inteligencia artificial clínicamente aplicable para diagnóstico preciso, mediciones cuantitativas y pronóstico de neumonía por COVID-19 mediante tomografía computarizada*. (2020).

Departamento de matemáticas, Universidad Nacional Autónoma de Honduras

Dirección de correo electrónico: omcarrasco@unah.hn

COMPLEJOS SIMPLICIALES PARA EL ANÁLISIS TOPOLÓGICO DE DATOS

CHRISTIAN ARIEL PALACIOS HERNÁNDEZ

Resumen. Los complejos simpliciales son herramientas fundamentales en diversas áreas de las matemáticas y la ciencia de datos, incluido el análisis topológico de datos (TDA). Este artículo presenta una revisión del estado del arte sobre los complejos simpliciales, explorando su teoría matemática y sus aplicaciones en el TDA. Se revisan conceptos clave como los complejos simpliciales abstractos y geométricos, junto con sus propiedades relevantes. Además, se examina cómo estos conceptos se aplican en el TDA, destacando técnicas como la filtración, la homología persistente y los diagramas de código de barras. Finalmente, se analiza cómo el TDA complementa los enfoques tradicionales de análisis de datos, proporcionando una visión más profunda de conjuntos de datos complejos.

Abstract. Simplicial complexes are fundamental tools in various areas of mathematics and data science, including topological data analysis (TDA). This paper provides a state-of-the-art review on simplicial complexes, exploring their mathematical theory and applications in TDA. Key concepts such as abstract and geometric simplicial complexes, along with their relevant properties, are reviewed. Additionally, the paper examines how these concepts are applied in TDA, highlighting techniques such as filtration, persistent homology, and barcode diagrams. Finally, it discusses how TDA complements traditional data analysis approaches, providing a deeper insight into complex data sets.

1. Introducción

Los complejos simpliciales, estructuras matemáticas que generalizan la noción de triangulación de un espacio, han demostrado ser esenciales en la representación y análisis de espacios topológicos. Su origen se remonta a los orígenes de la topología algebraica, con los trabajos de Poincaré a finales del siglo XIX, donde se sentaron las bases de la topología combinatoria y se exploraron las primeras aplicaciones de los complejos simpliciales en la representación de espacios topológicos.

Los complejos simpliciales en la época moderna se han estado utilizado en diversas áreas de la ciencia y con múltiples aplicaciones, algunas de estas en la Neurociencia [2], la Ciencia de Redes [8] y la Física [6].

En el ámbito del Análisis Topológico de Datos (TDA), los complejos simpliciales han adquirido un papel fundamental [4]. El TDA es un campo que se encuentra en la intersección del análisis de datos, la topología algebraica, la geometría computacional, la informática, la estadística y otras áreas relacionadas [14]. El objetivo

Fecha: 13 de agosto de 2024.

Palabras y frases clave. complejos simpliciales, análisis topológico de datos, homología simplicial, topología combinatoria.

principal del TDA es utilizar ideas y resultados de la geometría y la topología para desarrollar herramientas para estudiar características cualitativas de los datos [14]. En este contexto, los complejos simpliciales ofrecen un marco matemático idóneo para representar y analizar la estructura topológica subyacente en conjuntos de datos complejos.

Este artículo se adentra en el estudio de los complejos simpliciales, explorando su rica teoría matemática y cómo pueden aplicarse en el TDA. Hay muchas maneras de representar un espacio topológico, siendo una de ellas una descomposición en piezas simples. Esta descomposición califica para ser llamada un complejo si las piezas son topológicamente simples y sus intersecciones comunes son piezas de menor dimensión del mismo tipo [4].

Si bien la aplicación de los complejos simpliciales en TDA ha sido ampliamente explorada, este documento se centra en los conceptos fundamentales relacionados con los complejos simpliciales, tanto abstractos como geométricos, y sus propiedades y consecuencias. Posteriormente, se aborda cómo estos conceptos se aplican en el TDA, destacando técnicas como la filtración de complejos, la homología persistente y los diagramas de código de barras. Finalmente, se examina cómo el TDA complementa los enfoques tradicionales de análisis de datos, proporcionando una visión más profunda y detallada de los conjuntos de datos complejos.

Además, la investigación de complejos simpliciales, al ser herramientas esenciales en el TDA, tiene el potencial de abordar problemas complejos en diversas áreas. En Honduras, esta metodología puede ser aplicada para modelar y analizar sistemas complejos como redes de infraestructura, datos epidemiológicos y patrones climáticos. Esto se alinea con las líneas prioritarias de investigación de la UNAH en desarrollo económico y social, población y condiciones de vida, y ambiente, biodiversidad y desarrollo. Al mejorar la comprensión de estos sistemas, se pueden desarrollar soluciones más efectivas a problemas nacionales de urgencia, como la optimización de recursos, la prevención de enfermedades y la mitigación del cambio climático. Asimismo, el desarrollo de software y herramientas computacionales para esta rama fomenta el desarrollo y aplicación de tecnologías de la información y comunicación, impulsando la innovación tecnológica en el país. En [2] pueden observarse algunas aplicaciones.

2. Antecedentes

Los complejos simpliciales, como estructuras matemáticas discretas que modelan espacios topológicos, han desempeñado un papel fundamental en el desarrollo de la topología algebraica y la geometría combinatoria. Sus orígenes se remontan a los trabajos de Bernhard Riemann en el siglo XIX, quien sentó las bases para el estudio de variedades y superficies mediante triangulaciones. Sin embargo, fue a principios del siglo XX cuando la teoría de los complejos simpliciales comenzó a formalizarse y adquirir relevancia gracias a las contribuciones de matemáticos como Henri Poincaré y Oswald Veblen [3].

Poincaré, en su trabajo “Analysis Situs”, introdujo ideas fundamentales sobre la conectividad y la estructura de los espacios topológicos, sentando las bases para el estudio de la homotopía e introducción al concepto de homología, siendo este último su aporte más significativo, que luego, posteriormente, dicho concepto desarrollaría

la homología simplicial. Esta última utiliza los complejos simpliciales, construidos a partir de símplices (la "unidad" básica de estos espacios), como herramienta algebraica para clasificar y estudiar las propiedades de los espacios topológicos mediante triangulaciones [15].

Veblen, por su parte, realizó importantes contribuciones a la definición rigurosa de los complejos simpliciales y sus propiedades. En su obra "Theory of Finite Projective Geometries", estableció una conexión fundamental entre los complejos simpliciales y la geometría proyectiva, demostrando que los complejos simpliciales podían utilizarse para representar espacios proyectivos y estudiar sus propiedades geométricas [21].

A lo largo del siglo XX, la teoría de los complejos simpliciales experimentó un gran desarrollo. Matemáticos como James W. Alexander y Solomon Lefschetz realizaron importantes avances en la topología algebraica, utilizando los complejos simpliciales como herramienta clave para estudiar la homología y la homotopía de espacios topológicos. Estos conceptos resultaron fundamentales para comprender la estructura y las propiedades de los complejos simpliciales, así como su relación con otros objetos matemáticos [11].

Con el auge de la computación en la segunda mitad del siglo XX, los complejos simpliciales encontraron un espacio natural en el desarrollo de algoritmos y estructuras de datos para la representación y manipulación de información geométrica. La capacidad de los complejos simpliciales para representar espacios topológicos de manera discreta y eficiente los convirtió en una herramienta indispensable en áreas como la geometría computacional, el diseño asistido por computadora y la visualización científica [4].

En las últimas décadas, la teoría de los complejos simpliciales ha continuado evolucionando y encontrando nuevas aplicaciones en diversas áreas de la matemática y otras disciplinas. En topología computacional, los complejos simpliciales se utilizan para modelar y analizar datos geométricos, permitiendo la reconstrucción de superficies, la simplificación de mallas y la detección de características topológicas en conjuntos de datos. En teoría de grafos, se emplean para estudiar redes y estructuras discretas, proporcionando herramientas para analizar la conectividad, la detección de comunidades y la identificación de patrones en grafos complejos. Además, los complejos simpliciales han encontrado aplicaciones en áreas como la física, la química y la biología, donde se utilizan para modelar estructuras moleculares, analizar redes neuronales y estudiar sistemas biológicos complejos [4].

En la actualidad, TDA ha emergido como una rama interdisciplinaria en rápido crecimiento, con aplicaciones en una amplia gama de áreas, desde la ciencia de datos y el aprendizaje automático hasta la biología y la física. Los complejos simpliciales siguen siendo una herramienta fundamental en este campo, proporcionando un marco matemático riguroso para analizar la forma y la estructura de los datos. En particular, la homología persistente, una técnica clave en el TDA, se basa en la construcción de complejos simpliciales filtrados para identificar características topológicas significativas en los datos a diferentes escalas [1].

3. Complejos simpliciales para el TDA

3.1. Definiciones previas. En esta sección, estableceremos los conceptos clave para estudiar los complejos simpliciales, comenzando con la independencia geométrica de puntos en el espacio euclidiano y los planos generados por estos puntos. También introduciremos la envolvente convexa, que delimita regiones en el espacio, y las nociones de encaje topológico y homotopía, esenciales para entender la equivalencia topológica. Estos conceptos son fundamentales para la construcción y el estudio de los complejos simpliciales en las secciones siguientes.

Definición 3.1. [17] Decimos que un conjunto $\{a_0, a_1, \ldots, a_n\}$ de $n+1$ puntos de $\mathbb{R}^N$ es geométricamente independiente si para cualesquiera $t_i \in \mathbb{R}$ tales que $\sum_{i=0}^{n} t_i = 0$ se tiene que:

$$\sum_{i=0}^{n} t_i a_i = 0 \Leftrightarrow t_0 = t_i = \cdots = t_n = 0.$$

Visto desde el punto de vista del álgebra lineal, de forma equivalente, puede decirse que el conjunto antes descrito es geométricamente independiente si y solo si el conjunto de vectores $\{a_1 - a_0, a_2 - a_0, \ldots, a_n - a_0\}$ son linealmente independientes [17].

Definición 3.2. [13] Dado un conjunto geométricamente independiente de puntos como en la definición anterior, un n-plano P generado por esos puntos es el conjunto de todos los puntos x de $\mathbb{R}^N$ tales que

$$x = \sum_{i=0}^{n} t_i a_i,$$

para algunos escalares t_i con $\sum t_i = 1$.

Dado que los a_i son geométricamente independientes, los t_i están determinados de forma única por x. Notemos que cada punto a_i pertenece al plano P.

El plano P también puede describirse como el conjunto de todos los puntos x tales que

$$x = a_0 + \sum_{i=1}^{n} t_i (a_i - a_0)$$

para algunos escalares $t_1, \ldots, t_n$. En esta forma, hablamos de P como el "plano que pasa por a_0 paralelo a los vectores $a_i - a_0$" [13].

Definición 3.3. [16] Sea $V = \{a_0, a_1, \ldots, a_n\}$ un conjunto con $n+1$ puntos en $\mathbb{R}^N$, se define la envolvente convexa como el conjunto

$$C(V) = \left\{ \sum_{i=0}^{n} t_i a_i \;\middle|\; t_i \geq 0, \sum_{i=0}^{n} t_i = 1 \right\}.$$

En otras palabras, la envolvente convexa de un conjunto de puntos es el conjunto convexo más pequeño que los contiene. Es decir, si tomamos dos puntos cualesquiera dentro de la envolvente convexa, el segmento que los une también estará contenido dentro de ella.

A continuación, exploraremos algunos conceptos topológicos clave que serán la base para analizar los complejos simpliciales. Estos conceptos ayudarán a comprender mejor la estructura y las relaciones entre espacios topológicos, lo cual es crucial para el estudio y manipulación de complejos simpliciales en el contexto del TDA.

Definición 3.4. [16] Sea $f : X \to Y$ una función continua e inyectiva, sea $Z = f(X)$ visto como un subespacio de Y, de manera que podemos hablar de la función $g : X \to Z \subset Y$. Si g^{-1} es continua, decimos que f es un encaje topológico (o simplemente encaje).

Otra forma de definir un encaje topológico es la de decir que si existe una función f desde un espacio topológico X a un espacio topológico Y tal que si f al restringir su rango a $f(X)$ es un homeomorfismo, entonces f es un encaje.

Definición 3.5. [16] Dada una función sobreyectiva $f : X \to Y$, se dice que es una función cociente si para todo conjunto V en Y, V es abierto en Y si y solo si $f^{-1}(V)$ es abierto en X.

Definición 3.6. [7] Si f y g son aplicaciones continuas del espacio X en el espacio Y, decimos que f es homotópica a g si existe una aplicación continua $F : X \times I \to Y$ tal que

$$F(x, 0) = f(x)$$

y

$$F(x, 1) = g(x)$$

Definición 3.7. [7] Dos espacios topológicos X e Y se dicen homotópicos (y se denota por $X \simeq Y$) si existen aplicaciones continuas $f : X \to Y$ y $g : Y \to X$ tales que $g \circ f$ es homotópica a 1_X y $f \circ g$ es homotópica a 1_Y.

Definición 3.8. [7] Un espacio topológico X se dice contráctil si es homotópico a un punto.

Dicho de otro modo, dos funciones continuas entre espacios topológicos son homótopas si una puede transformarse en la otra de forma continua. Un espacio es contráctil si puede encogerse continuamente hasta convertirse en un punto.

3.2. Estructura básica. En esta sección, estableceremos la teoría de complejos simpliciales, basándonos en los conceptos de independencia geométrica y envolvente convexa descritos anteriormente. Nos centraremos en los símplices, los componentes básicos de estos complejos, y en cómo se combinan para formar estructuras más complejas. Exploraremos tanto su representación geométrica como su estructura combinatoria abstracta, y presentaremos resultados fundamentales para entender las propiedades topológicas de los complejos simpliciales y su relación con espacios topológicos generales.

Definición 3.9. [17, 16] Dado un conjunto de puntos geométricamente independiente $\{a_0, \ldots, a_n\}$ de $\mathbb{R}^N$, definimos el n-símplice (o simplemente ímplice) σ generado por $a_0, \ldots, a_n$ y denotado como $\sigma = (a_0, \ldots, a_n)$, como el conjunto de puntos de $\mathbb{R}^N$ de la forma:

$$x = \sum_{i=0}^{n} t_i a_i,$$

donde

$$\sum_{i=0}^{n} t_i = 1 \text{ con } t_i \geq 0 \text{ para todo } i.$$

Definición 3.10. [16] El simplejo estandar denotado como $\Delta^n \subset \mathbb{R}^n$, se define como $\Delta^n = (e_0, e_1, \ldots, e_n)$, donde $\{e_i\}_{i=1}^n$ es la base estandar de $\mathbb{R}^n$ y $e_0 = 0$.

De manera intuitiva, podemos visualizar los símplices de baja dimensión de la siguiente manera: un 0-símplice es un punto, un 1-símplice es un segmento de línea, un 2-símplice es un triángulo sólido, un 3-símplice es un tetraedro sólido, y así sucesivamente. En general, un n-símplice es la generalización en n dimensiones de un triángulo (Figura 1).

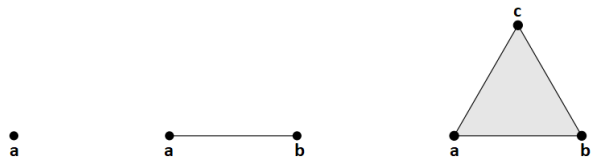


FIGURA 1. Un 0-simplice (a), un 1-simplice (a, b) y un 2-simplice (a, b, c), respectivamente [5].

Además, un n-símplice puede entenderse directamente como la envolvente convexa del conjunto de puntos dado, donde los a_i son los vértices. Si $z \in \sigma$, entonces las coordenadas de z se llaman coordenadas baricéntricas de z. Asimismo, podemos definir el interior de un símplice σ como los puntos tales que $t_i > 0$. Por su parte, la frontera de σ, denotada como $bd(\sigma)$ es $\sigma \setminus Int(\sigma)$ (Figura 2) [5].

FIGURA 2. La frontera de un 2-simplice (izq). El interior de un 2-simpice (der) [5].

Establezcamos ahora un resultado topológico importante sobre el n-símplice estándar.

Teorema 3.11. [12] Δ^n *es compacto.*

Este teorema es relevante porque establece que dicho conjunto es cerrado y acotado en el espacio euclidiano en el que está embebido. La compacidad es una propiedad topológica que tiene implicaciones importantes en la construcción de modelos matemáticos y en el TDA esto no es la excepción.

Definición 3.12. [17] Sea σ un n-símplice, entonces la dimensión de σ, denotada como $dim(\sigma)$, es n.

Es decir, la dimensión de un símplice simplemente es la cantidad de componentes que tiene, menos 1. Por definición, simplicidad y practicidad, $dim(\emptyset) = -1$.

Definición 3.13. [17] Sea σ un n-símplice. Una cara de σ es un símplice τ generado por un subconjunto de $\{a_0, \ldots, a_n\}$. Esto se denota como $\tau \leq \sigma$ (o $\tau \subseteq \sigma$). Por otra parte, una cara τ de σ se llama cara propia si $\tau \neq \sigma$, es decir, si τ es un símplice de dimensión estrictamente menor que n.

Así pues, un punto es una cara de su respectivo segmento si lo hay, un segmento una cara de su respectivo triángulo si lo hay, un triángulo una cara de su respectivo tetraedro si lo hay y así sucesivamente.

Ahora que hemos definido los símplices, es natural preguntarse cómo podemos establecer correspondencias entre ellos. Para ello, introducimos el concepto de función simplicial, que nos permite mapear un símplice en otro de manera que se respete su estructura combinatoria y geométrica.

Definición 3.14. [16] Una función entre dos símplices $f : (a_0, \ldots, a_n) \to (b_0, \ldots, b_m)$ es una función simplicial si para todo $i \in \{0, 1, \ldots, n\}$ se cumplen:

1. $f(a_i) = b_j$ para algún $j \in \{0, 1, \ldots, m\}$,
2. $f\left(\sum \lambda_i a_i\right) = \sum \lambda_i f(a_i)$.

[16] Dada una función simplicial f, se satisfacen las siguientes propiedades:

1. Si f es continua en el símplice, también es cerrada.
2. Si f es inyectiva, entonces es un encaje.
3. Si f es sobreyectiva, entonces es una función cociente.
4. Si f es biyectiva, entonces es un homeomorfismo.

Ahora, estableceremos un resultado fundamental que conecta los símplices arbitrarios con el símplice estándar. Este teorema nos permitirá aprovechar las propiedades del símplice estándar para estudiar símplices generales, simplificando en gran medida nuestro análisis:

Teorema 3.15. [12] *$\sigma = (a_0, a_1, \ldots, a_n)$ es homeomorfo a Δ^n.*

Este teorema es muy importante ya que nos permite trasladar propiedades y resultados del símplice estándar a símplices arbitrarios. En particular, la compacidad de Δ^n implica directamente la compacidad de cualquier n-símplice. Esta conexión entre símplices arbitrarios y el símplice estándar será fundamental en nuestro estudio de los complejos simpliciales y sus aplicaciones en el TDA.

Definición 3.16. [13] Un complejo simplicial geométrico K en $\mathbb{R}^N$ es una colección de símplices en $\mathbb{R}^N$ tal que:

1. Cada cara de un símplice de K está en K.
2. La intersección de dos símplices cualesquiera de K es una cara de cada uno de ellos.

En otras palabras, un complejo simplicial geométrico es una estructura formada por símplices que se "pegan" de manera consistente a lo largo de sus caras. Esta definición formaliza la idea intuitiva de construir objetos geométricos a partir de bloques de construcción más simples, como puntos, segmentos, triángulos y sus análogos en dimensiones superiores.

En el TDA, a menudo trabajamos con datos en forma de nube de puntos, es decir, un conjunto finito de puntos en un espacio métrico (como el espacio euclidiano $\mathbb{R}^n$). Para analizar la topología de estos datos, construimos un complejo simplicial a partir de ellos. Los vértices del complejo simplicial son la nube de puntos, y los símplices de dimensiones superiores se forman conectando los puntos cercanos.

Seguido a esto podemos encontrarnos con subestructuras como las siguientes:

Definición 3.17. [17, 13] Se dice que L es un subcomplejo simplicial de un complejo simplicial K si L es una subcolección de K que contiene todas las caras de sus elementos. Esto se denota como $L \subseteq K$.

Definición 3.18. [17, 13] Los p-esqueletos $K^{(p)}$ son los subcomplejos de K formados por los símplices de dimensión p o menor. En particular, $K^{(0)}$ son los vértices de K.

También podemos introducir una noción fundamental para caracterizar su complejidad:

Definición 3.19. [17] Para un complejo simplicial K, definimos su dimensión como la mayor de las dimensiones de sus símplices.

Es importante destacar que la dimensión de un complejo simplicial no se refiere a la dimensión del espacio en el que está embebido, sino a la complejidad de su estructura interna. Un complejo simplicial puede tener una dimensión baja, incluso si sus símplices están contenidos en un espacio de dimensión alta.

Definición 3.20. [17] El poliedro de un complejo simplicial (o espacio subyacente) K, denotado por $|K|$, se define como la unión de los símplices que forman K, i.e.,

$$|K| := \bigcup_{\sigma \in K} \sigma.$$

Además, se le puede dotar de una estructura de espacio topológico a $|K|$ al considerar la topología inducida por la usual de $\mathbb{R}^N$, siendo así un espacio compacto.

Ejemplo. El complejo simplicial

$$K = \{(a_1), (a_2), (a_3), (a_4), (a_5), (a_6), (a_7), (a_1, a_2), (a_2, a_3), (a_2, a_4), (a_3, a_4), (a_4, a_5), (a_5, a_6), (a_5, a_7), (a_6, a_7), (a_5, a_6, a_7)\},$$

(Figura 3).

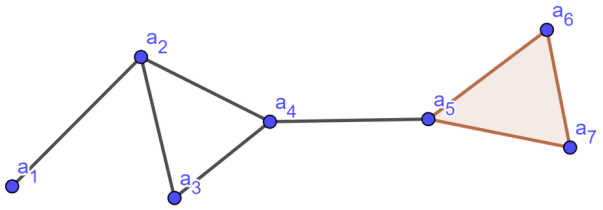


FIGURA 3. Poliedro $|K|$ asociado al complejo simplicial [17].

Definición 3.21. [12] Si existe un complejo simplicial K cuyo poliedro $|K|$ es homeomorfo a un espacio topológico X, entonces se dice que X es un espacio triangulable, y el complejo K se llama una triangulación de X.

La triangulación es una herramienta clave en topología computacional y TDA, ya que permite discretizar espacios complejos mediante estructuras simpliciales, lo cual facilita el desarrollo de algoritmos eficientes y potencialmente menos costosos desde el punto de vista computacional, capaces de analizar y extraer información topológica relevante de grandes conjuntos de datos. Algunos ejemplos de triangulación son: la n-esfera (S^n), el Toro ($\mathbb{T}^2$), la banda de Möbius, la botella de Klein y el cilíndro [20].

A continuación exploraremos algunas propiedades topológicas fundamentales de los poliedros asociados a complejos simpliciales. Estos resultados nos permitirán comprender cómo la estructura combinatoria de un complejo simplicial se relaciona con la topología de su poliedro.

Sea K un complejo simplicial y X un espacio topológico, entonces:

1. Si L es un subcomplejo (simplicial) de K, $|L|$ es un subespacio cerrado de $|K|$. En particular, si $\sigma \in K$, entonces σ es un subespacio cerrado de $|K|$.
2. Una función $f : |K| \to X$ es continua si y solo si $f|_\sigma$ es continua para cada $\sigma \in K$.
3. $|K|$ es Hausdorff.
4. Si K es finito, entonces $|K|$ es compacto. Recíprocamente, si un subconjunto A de $|K|$ es compacto, entonces $A \subset |K_0|$ para algún subcomplejo finito K_0 de K [13].

Hasta ahora, hemos estudiado los complejos simpliciales como objetos individuales. Sin embargo, es fundamental comprender cómo se relacionan entre sí. Para ello,

introduciremos la noción de aplicación simplicial, que nos permitirá establecer correspondencias entre los vértices y símplices de diferentes complejos simpliciales, preservando su estructura combinatoria.

Definición 3.22. [17] Sean K y L dos complejos simpliciales, decimos que $f : K \to L$ es una aplicación simplicial si es una aplicación entre sus conjuntos de vértices $f : K^{(0)} \to L^{(0)}$ de tal modo que para todo símplice $\sigma = (a_0, \ldots, a_n) \in K$, los vértices $f(a_0), \ldots, f(a_n)$ pertenecen a un mismo símplice de L.

De forma natural podemos también definir la composición de aplicaciones simpliciales, que de hecho es una aplicación simplicial. Asimismo, si existe otra aplicación simplicial $g : L \to K$ tal que $g \circ f = 1_K$ y $f \circ g = 1_L$, entonces f es un isomorfismo simplicial, y K y L son simplicialmente isomorfos [20].

También, toda aplicación simplicial $f : K \to L$ induce una aplicación continua entre sus poliedros asociados $|f| : |K| \to |L|$, definida como sigue:

$$|f|(x) := \sum_{i=0}^{n} t_i f(a_i)$$

para cada $x = \sum_{i=0}^{n} t_i a_i$, con $\sigma \in K$ [17].

Además de la definición geométrica, existe una definición abstracta de complejo simplicial que se centra en su estructura combinatoria.

Definición 3.23. [9] Decimos que la colección S de conjuntos finitos y no vacíos es un complejo simplicial abstracto si para cualquier elemento A de S todo subconjunto no vacío de A también está en S.

Así pues, un elemento U perteneciente a un complejo simplicial S se denomina símplice. Si U tiene los elementos $a_0, a_1, \ldots, a_n$, usaremos la notación de llaves: $\{a_0, a_1, \ldots, a_n\}$. Esta notación distingue los complejos simpliciales abstractos de los geométricos, donde se utiliza la notación de paréntesis $(a_0, a_1, \ldots, a_n)$. La dimensión de este símplice U es igual al número de sus elementos menos uno. Cualquier subconjunto no vacío de U se considera una cara de U. La dimensión del complejo simplicial S es la mayor dimensión de sus símplices. El conjunto de vértices V de S (también denotado como V_S) se define como la unión de los elementos de S que tienen un solo elemento. Las notaciones $v \in V$ y $\{v\} \in S$ hacen referencia al mismo elemento. Además, una subcolección de S que también sea un complejo simplicial se denomina subcomplejo simplicial de S.

Así observamos que esta nueva definición abstracta de complejo simplicial es equivalente con la definición geométrica que presentamos al inicio, así como con las propiedades que establecimos.

Ejemplo. La colección $S = \{\{1, 2, 3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{3, 4\}, \{1\}, \{2\}, \{3\}, \{4\}\}$ es un complejo simplicial abstracto. En la Figura 4 puede verse su poliedro asociado.

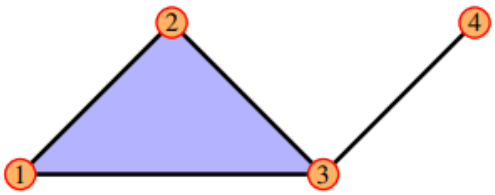


FIGURA 4. Complejo simplicial abstracto [9].

De manera similar, una aplicación simplicial abstracta entre complejos simpliciales abstractos $f : S \to L$ se define como una función entre sus vértices que preserva la estructura simplicial, es decir, si un conjunto de vértices forma un símplice en S, sus imágenes bajo f también forman un símplice en L. En consecuencia, dos complejos simpliciales abstractos S y L son simplicialmente isomorfos si existe una biyección entre sus vértices que preserva esta estructura simplicial.

Ahora que hemos establecido la equivalencia entre complejos simpliciales abstractos y geométricos, podemos introducir el concepto de símplice ordenado. Un símplice ordenado es una generalización de un simple estándar, que tiene en cuenta el orden específico de sus vértices. Este concepto es crucial para la definición de la noción de orientación en un complejo simplicial, ya que la orientación proporciona una manera de distinguir entre los símplices que están orientados de manera diferente pero que, en esencia, podrían representar la misma entidad topológica. La orientación de los símplices permite la construcción de grupos de cadenas, que son estructuras algebraicas que capturan la información sobre los símplices en términos de combinaciones lineales.

Definición 3.24. [17] Un símplice ordenado (orientado) es un símplice cuyos vértices están dotados de un orden total. Denotamos un símplice ordenado como $[a_0, a_1, \ldots, a_n]$, donde el orden de los vértices es $a_0 < a_1 < \ldots < a_n$ (Figura 5).

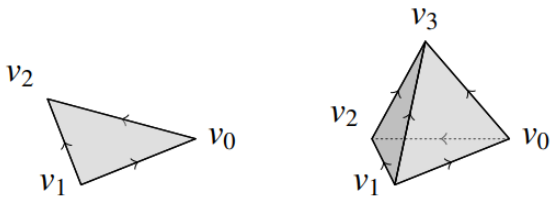


FIGURA 5. Un 2-simplejo ordenado (izq.), un 3-simplejo ordenado (der.) [9].

La noción de orientación también es fundamental para el estudio de la homología simplicial, una herramienta clave en la topología algebraica que nos ayuda a entender las características topológicas de los espacios. La orientación afecta la manera

en que se definen las fronteras y las cadenas, influyendo en cómo se calculan los grupos de homología y, por ende, en la interpretación de las propiedades topológicas del complejo. Por lo tanto, una comprensión profunda del concepto de símplice ordenado no solo facilita la definición precisa de la orientación, sino que también proporciona una base sólida para el análisis de estructuras topológicas más complejas en estudios posteriores.

Podemos afirmar que dos símplices ordenados $[a_0, a_1, \ldots, a_n]$ y $[b_0, b_1, \ldots, b_n]$ son equivalentes si los vértices b_i son una permutación par de los vértices a_i. En caso contrario, se dice que los símplices ordenados tienen orientaciones opuestas.

Definición 3.25. [13] Si S es un complejo simplicial, sea V el conjunto de vértices de S. Sea $\mathcal{S}$ la colección de todos los subconjuntos $\{a_0, \ldots, a_n\}$ de V tales que los vértices $a_0, \ldots, a_n$ forman un símplice de S. La colección $\mathcal{S}$ se llama el esquema de vértices de S.

La colección $\mathcal{S}$ es un ejemplo particular de un complejo simplicial abstracto. De hecho, es el ejemplo crucial, lo mostramos en el siguiente teorema:

Teorema 3.26. [13, 17] *Todo complejo simplicial abstracto es isomorfo al esquema de vértices de algún complejo simplicial geométrico. Además, dos complejos simpliciales (abstractos o geométricos) son isomorfos si y solo si sus esquemas de vértices son isomorfos.*

En otras palabras, dado un complejo simplicial abstracto S, podemos construir un complejo simplicial geométrico K cuyos vértices y símplices corresponden exactamente a los de S. Este complejo geométrico K se llama la realización geométrica de S, y es único salvo isomorfismo.

Por lo anterior, en la práctica, podemos trabajar indistintamente con complejos simpliciales abstractos o geométricos. Cuando nos refiramos al poliedro asociado a un complejo simplicial abstracto, estaremos haciendo alusión al poliedro correspondiente a su realización geométrica. Asimismo, a partir de este momento, utilizaremos de manera intercambiable las notaciones $\sigma = (a_0, \ldots, a_n)$ y $\{a_0, \ldots, a_n\}$ para representar los elementos (símplices) de un complejo simplicial abstracto.

Definición 3.27. [1] Sea $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ un recubrimiento abierto de un espacio topológico X. El nervio de $\mathcal{U}$, denotado por $\mathcal{N}(\mathcal{U})$, es el complejo simplicial abstracto cuyo conjunto de vértices es A, y donde una familia $\{\alpha_0, \ldots, \alpha_k\}$ define un k-símplice si y sólo si $U_{\alpha_0} \cap \ldots \cap U_{\alpha_k} \neq \emptyset$.

Esta es una construcción extremadamente útil en teoría de homotopía. La razón principal en nuestro caso es el siguiente teorema, que proporciona criterios que garantizan que $\mathcal{N}(\mathcal{U})$ es homotópicamente equivalente al espacio subyacente X.

Teorema 3.28. [1] *Sea $\mathcal{U}$ un recubrimiento abierto numerable de un espacio topológico X. Si para todo $S \subseteq A$, la intersección $\bigcap_{s \in S} U_s$ es contráctil o vacía, entonces el nervio $\mathcal{N}(\mathcal{U})$ es homotópicamente equivalente a X.*

3.3. Principales complejos para el TDA. El TDA se apoya en varias construcciones de complejos simpliciales que capturan la estructura topológica de los datos. Hasta aquí hemos visto teoría relevante acerca de los complejos simpliciales. A continuación, presentamos algunas de las principales construcciones utilizadas en la topología computacional.

Definición 3.29. [1] Dado un espacio métrico (X, d), un subconjunto $V \subseteq X$ y un número real $\epsilon > 0$, el complejo de Čech, denotado por $\check{C}(V, \epsilon)$, se define como el nervio del recubrimiento $\{B_\epsilon(v)\}_{v \in V}$, donde $B_\epsilon(v)$ denota la bola abierta de radio ϵ centrada en v. Es decir, $\check{C}(V, \epsilon)$ es el complejo simplicial abstracto cuyos vértices son los puntos de V y donde un conjunto de $k+1$ vértices $\{v_0, \ldots, v_k\}$ forma un k-símplice si y sólo si la intersección de las bolas correspondientes $B_\epsilon(v_0) \cap \ldots \cap B_\epsilon(v_k)$ es no vacía.

Definición 3.30. [10] Sea X un espacio métrico. El complejo de Vietoris-Rips, denotado como, $R_\epsilon(X)$ a escala ϵ se define como el complejo simplicial abstracto cuyos símplices son todos los subconjuntos finitos $\sigma = \{x_0, \ldots, x_k\} \subseteq X$ que satisfacen

$$d(x_i, x_j) \leq \epsilon \quad \text{para todo } x_i, x_j \in \sigma.$$

El complejo de Čech puede ser computacionalmente costoso, especialmente cuando se maneja una gran cantidad de datos, debido a la necesidad de calcular y verificar la intersección de bolas abiertas para todos los subconjuntos posibles de puntos. En comparación, el complejo de Vietoris-Rips suele ser menos costoso, ya que solo requiere verificar distancias entre pares de puntos en lugar de calcular intersecciones de bolas, lo que lo hace más manejable para grandes conjuntos de datos.

Estas dos construcciones están estrechamente relacionadas y se puede establecer una conexión formal entre ellas. Este vínculo se resume en el siguiente resultado:

Teorema 3.31. [1]

$$\check{C}(X, \epsilon) \subseteq VR(X, 2\epsilon) \subseteq \check{C}(X, 2\epsilon).$$

Esta relación es útil para entender cómo las diferentes escalas de análisis afectan la estructura topológica capturada por estos complejos.

Además de los complejos de Vietoris-Rips y Čech, existen otras construcciones relevantes en TDA. El diagrama de Voronoi y su correspondiente complejo de Delaunay son fundamentales para estudiar la estructura espacial de los puntos en $\mathbb{R}^N$.

El diagrama de Voronoi divide el espacio en regiones, cada una asociada a un punto específico del conjunto dado. Cada punto dentro de una célula está más cerca de su punto de referencia que de cualquier otro punto de referencia.

Definición 3.32. [17] Sean $X \subset \mathbb{R}^N$ nuestra nube de puntos y $\lambda \in X$ un punto de referencia (representante), llamamos célula de Voronoi asociada a λ a

$$V_\lambda = \{x \in \mathbb{R}^N : d(x, \lambda) \leq d(x, \lambda') \ \forall \lambda' \in X\}.$$

Definición 3.33. [17] El nervio asociado al diagrama de Voronoi es un complejo simplicial que recibe el nombre de complejo de Delaunay y se denota como

$$D(X) = \left\{ \sigma \subseteq X : \bigcap_{\lambda \in \sigma} V_\lambda \neq \emptyset \right\}.$$

Por lo tanto, apoyándonos en el Teorema del nervio, intersectando estos subespacios podemos elaborar un complejo simplicial conservando el tipo de homotopía, pues estas intersecciones, $R_\epsilon(\lambda) = V_\lambda \cap B_\epsilon(\lambda)$ para todo $\lambda \in X$, forman un recubrimiento de nuestra nube de puntos X. Veamos cómo construir dicho complejo:

Definición 3.34. [17] Dado $\epsilon > 0$, definimos el complejo Alfa como el nervio del recubrimiento dado por $\{R_\epsilon(\lambda)\}_{\lambda \in X}$.

$$A_\epsilon(X) = \left\{ \sigma \subset X : \bigcap_{\lambda \in \sigma} R_\epsilon(\lambda) \neq \emptyset \right\}.$$

3.4. Aplicaciones al TDA. Aunque la discusión sobre filtraciones, homología persistente y su aplicación en el TDA no forma parte de los objetivos principales de esta investigación, es importante abordar estos conceptos para proporcionar un contexto más amplio y permitir una comprensión más profunda de las aplicaciones prácticas de los complejos simpliciales, es por esto que esta sección será muy breve, pero relevante. A continuación, exploraremos cómo los complejos simpliciales se utilizan en TDA, centrándonos en la técnica de filtraciones y en la homología persistente.

Definición 3.35. [17] Una filtración de un complejo simplicial K, denotada por $\mathcal{F} = \{K_i\}_{i \in I}$ con I totalmente ordenado, es una familia de subcomplejos de K tal que se verifica que $K_i \subseteq K_j$ para todo $i \leq j$ y $\emptyset$, $K \in F$.

Esta técnica es esencial en TDA, ya que permite estudiar la evolución de las características topológicas a través de diferentes escalas. La idea central es construir una serie de complejos simpliciales, cada uno correspondiente a un parámetro diferente.

Los métodos de Čech, Alpha y Vietoris-Rips nos permiten construir filtraciones de tipo finito con parámetro real. Por cuestiones de economía computacional e interpretabilidad, en la práctica no se suele tomar la filtración completa sino que se escoge un M máximo y se impone que $K_\epsilon = K_M$ para todo $\epsilon \geq M$.

El uso de filtraciones lleva a la homología persistente, una herramienta clave en TDA. La homología persistente estudia cómo cambian las características topológicas (como componentes conexas, agujeros y cavidades) a medida que varía el parámetro de la filtración. Esto proporciona información sobre la importancia y la relevancia de estas características a través de diferentes escalas.

Para representar la información de la homología persistente, se utilizan diagramas de persistencia, siendo el código de barras uno de los más comunes. Este diagrama muestra la persistencia de características topológicas a lo largo de diferentes escalas de filtración, donde la longitud de cada barra indica la duración de la característica en el tiempo.

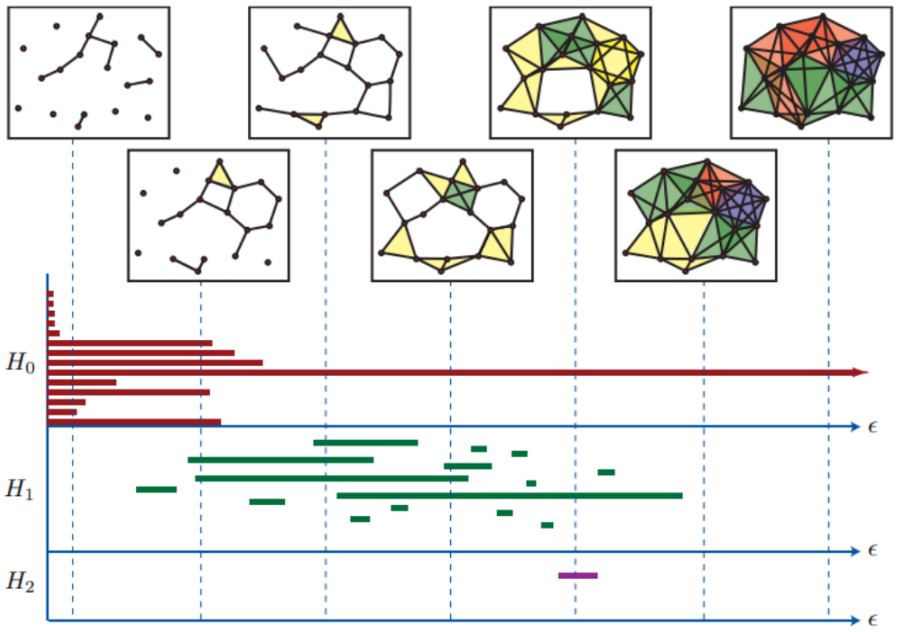


FIGURA 6. Ejemplo de filtración con su código de barras asociado [17].

Ejemplo. Imaginemos que tenemos una nube de puntos que representa las posiciones de diversas estaciones meteorológicas en una región. Queremos analizar la estructura topológica de estos datos para identificar características significativas como áreas densamente conectadas (componentes conexas) y posibles patrones de vacíos o huecos (agujeros y cavidades) en los datos.

En la figura 6 se muestra un ejemplo de una filtración de un complejo simplicial construido a partir de esta nube de puntos y su correspondiente diagrama de código de barras.

En la parte superior de la figura, se observa cómo evoluciona el complejo simplicial a medida que aumenta el parámetro ϵ. Cada subcomplejo K_i se muestra en un recuadro separado. A medida que el parámetro crece, se agregan más simplices al complejo, cambiando su topología.

En la parte inferior de la figura, el diagrama de barras muestra la persistencia de características topológicas en diferentes escalas de filtración. Las barras rojas en H_0 indican componentes conexas, las verdes en H_1 representan agujeros, y la morada en H_2 una cavidad. La longitud de cada barra refleja la duración de la característica, lo que ayuda a identificar las más significativas en los datos.

Las características topológicas, representadas en diagramas de persistencia como el código de barras, se integran con el análisis de datos tradicional. Esto enriquece la interpretación de los datos y proporciona una base sólida para tomar decisiones y desarrollar modelos predictivos o descriptivos.

4. Conclusiones

Hemos examinado la aplicación de complejos simpliciales para analizar datos complejos. Estos complejos proporcionan una herramienta poderosa para representar y entender la topología subyacente en conjuntos de datos, revelando patrones y estructuras que no se capturan con métodos tradicionales.

Los complejos simpliciales permiten construir representaciones geométricas y abstractas de datos multidimensionales, lo que es crucial para el análisis en campos como la biología computacional, la ciencia de materiales y la ingeniería. En virtualización científica, por ejemplo, los complejos simpliciales se utilizan para modelar y visualizar datos de simulaciones científicas complejas, facilitando el análisis de fenómenos físicos [18]. En la ciencia de materiales, permiten estudiar la topología de materiales porosos y su relación con propiedades físicas [19]. Estas aplicaciones subrayan la versatilidad y la utilidad de los complejos simpliciales.

Aunque presentan desafíos como la alta carga computacional y la necesidad de conocimientos especializados para interpretar los resultados, los complejos simpliciales son una herramienta valiosa que complementa y enriquece los métodos tradicionales de análisis de datos y es muy recomendado optimizar los algoritmos de construcción y análisis, y explorar su integración con técnicas como el aprendizaje automático para expandir sus aplicaciones y mejorar su eficiencia.

Referencias

1. Gunnar Carlsson, *Topology and data*, Bulletin of the American Mathematical Society **46** (2009), no. 2, 255–308.
2. Moo K. Chung, Hyekyoung Lee, Hernando Ombao, and Victor Solo, *Exact topological inference of the resting-state brain networks in twins*, Network Neuroscience **3** (2019), no. 3, 648–670.
3. Jean Dieudonné, *A history of algebraic and differential topology 1900-1960*, Birkhäuser, 1989.
4. Herbert Edelsbrunner and John L. Harer, *Computational topology: An introduction*, American Mathematical Society, 2010.
5. Juan Camilo Gutiérrez Díaz, *Introducción al análisis topológico de datos*, 2023.
6. James B. Hartle, *Simplicial quantum gravity*, Journal of Mathematical Physics **26** (1985), no. 7, 1730–1739.
7. Allen Hatcher, *Algebraic topology*, Cambridge university press, 2002.
8. Danijela Horak, Slobodan Maletic, and Milan Rajkovic, *Persistent homology of complex networks*, Journal of Statistical Mechanics: Theory and Experiment **2009** (2009), no. 03, P03034.
9. Benjamín Alfonso Itza-Ortíz, Federico Menendez-Conde Lara, Erika Elizabeth Rodríguez-Torres, Margarita Tetlalmatzi-Montiel, and Rafael Villarroel-Flores, *Un método topológico para el análisis de complejidad de series de tiempo*, Publicación Semestral Pádi **9** (2021), no. 17, 103–107.
10. Karla Saraí Jiménez-Martínez and Daniel Trejo-Medina, *Topología algebraica para el análisis de datos*, Investigación y Desarrollo (2016), 1–22.
11. Charles Richard Francis Maunder, *Algebraic topology*, Dover Publications, 1996.
12. Jose Cristobal Molina Cortes, *Conceptos de homología simplicial*, 2016.
13. James R. Munkres, *Elements of algebraic topology*, Addison-Wesley Publishing Company, 1984.
14. Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington, *A roadmap for the computation of persistent homology*, EPJ Data Science **6** (2017), no. 1, 17.
15. Henri Poincaré, *Analysis situs*, `https://www.maths.ed.ac.uk/~v1ranick/papers/poincare2009.pdf`, 2009, Original work published 1895.
16. Jesús Manuel Pérez Angulo, *Análisis topológico de datos: robusticidad y análisis de sensibilidad de algoritmos*, Master's thesis, Centro de Investigación en Matemáticas, 12 2016.
17. Luis Sánchez Cano, *Introducción al análisis topológico de datos*, 07 2022, Trabajo de Fin de Grado, Universidad de Salamanca.

18. Julien Tierny, *Topological data analysis for scientific visualization*, Springer, 2018.
19. Gonzalo Turiel García, *Revisión de software para análisis topológico de datos*, junio 2024.
20. Laura Daniela Vargas Araujo, *Introducción a la homología simplicial y algunas consecuencias*, 2017.
21. Oswald Veblen and John Wesley Young, *Projective geometry*, Ginn and Company, 1910.

Carrera de matemática, Universidad Nacional Autónoma de Honduras, Tegucigalpa

Dirección de correo electrónico: christian.palacios@unah.hn

ANÁLISIS PREDICTIVO EN DATOS DE TEMPERATURA EN HONDURAS MEDIANTE MODELOS ESTADÍSTICOS Y MACHINE LEARNING

NATHALYE NICOL DERAS DURON

Resumen. Este estudio aborda el desafío de pronosticar la variación de la temperatura en Honduras a través de un análisis de series temporales, utilizando datos proporcionados por el Instituto Hondureño de Ciencias de la Tierra (IHCIT) de la UNAH. Los datos de la temperatura promedio se recopilan diariamente por la estación meteorológica ubicada en la UNAH. Para lograr la mejor precisión en la estimación y predicción de la serie temporal, se emplea un enfoque basado en modelos de redes neuronales, comparando los resultados con los modelos estadísticos ARIMA. Al implementar estos modelos, se busca resolver la complejidad inherente en la predicción climática, permitiendo el análisis de los datos históricos y actuales de la temperatura con el fin de identificar patrones, tendencias y proyecciones futuras. Los resultados obtenidos en esta investigación permitirán establecer la factibilidad de implementar esta metodología en diversas estaciones meteorológicas para obtener una visión más precisa del comportamiento de la temperatura en la región hondureña. Estos hallazgos pueden proporcionar la formulación de estrategias para la adaptación y mitigación del cambio climático en Honduras, lo que representa una contribución valiosa para abordar este desafío crucial a nivel regional.

ABSTRACT. This study tackles the challenge of forecasting the temperature changes in Honduras via a time series analysis, using data provided by IHCIT, UNAH. The average temperature data is measured daily by the meteorological station located in UNAH. To get the the most accuracy from the time series estimation and prediction, the focus is based in artificial neural network models, comparing them with ARIMA statistic models. By using these models, we try to solve the the complexity of climatic forecasting, allowing historic and current data analysis with the objective of identifying patterns, trends and future predictions. The results obtained in this research will stablish the factibility of implementing this methodology in diverse meteorological stations to get a more precise overview of temperature behaviour in Honduras. This discoveries will provide for a formulation of strategies for the mitigation of climate change in Honduras, which represents a valuable contribution to approach this crucial challenge at a regional level.

.

1. Introducción

La predicción con series de tiempo tiene importancia crucial en varios temas de la ciencia e ingeniería [1].

Fecha: Agosto de 2024.

Palabras y frases clave. Machine Learning, Redes Neuronales, Series Temporales, Modelos ARIMA, Cambio Climático.

Hasta el momento, se han llevado acabo muchas investigaciones sobre el modelamiento y predicción de series de tiempo con estacionariedad. El resultado de estas investigaciones es el desarrollo de redes neuronales, que se han utilizado desde la década de los 80 [2]. El cambio climático tiene impacto en sectores de importancia fundamental en Honduras, como la agricultura. Algunos impactos relevantes incluye la erosión de la tierra, el aumento en la temperatura, el aumento de valores climáticos extremos, y cambios en la calidad de los cultivos [3, 4].
En este trabajo, se estará usando como parámetro indicador de cambio climático la temperatura y sus variaciones desde el año 1979 hasta 2023, con el objeto de estimar las fluctuaciones de este parámetro, y predecir el mismo a futuro, por medio de modelos estadísticos ARIMA, y Machine Learning por medio de redes neuronales.

Los modelos ARIMA son comunmente utilizados para predecir parámetros de cambio climático como temperatura, y precipitación con el objetivo de aportar en aplicaciones de ingeniería ambiental y civil [5, 6]. Y debido a la estacionariedad de los datos que fueron proporcionados por el IHCIT, serán los modelos de comparación ideales respecto a los modelos de redes neuronales.

Las redes neuronales son un sistema computacional de aprendizaje supervisado muy popular que simula técnicas de aprendizaje de organismos biológicos [7] , que dan como resultado un ajuste altamente preciso, lo cual será aplicado para estudiar y proponer soluciones a la creciente problemática del cambio climático en Honduras, siendo este tema uno de los ejes prioritarios de investigación en la UNAH.

2. Antecedentes

El inicio de las redes neuronales artificiales llegó en 1943 cuando Warren McCulloch y Walter Pitts desarrollaron el primer modelo de redes neuronales, publicando un artículo sobre como las neuronas funcionan [6]. Sus redes se basaban en elementos simples y los resultados de su trabajo fueron funciones lógicas simples [8].

Posteriormente, en 1958, Rosenblatt llevó a cabo un trabajo sobre perceptrones [11]. El perceptron era un dispositivo electrónico que se construyó de acuerdo a principios biológicos y mostraba habilidades de aprendizaje, así que Rosenblatt, en 1962 publicó un libro sobre neuro-computación [10].

Luego de un periodo de entusiasmo, el progreso de las redes neuronales y la neuro-computación se vio estancada por un periodo de aproximadamente diez años, iniciado en 1969, cuando Minsky y Papert escribieron un libro sobre Perceptrones, parte de una campaña de desacreditación a la investigación de las redes neuronales, mostrando una cantidad importante de problemas fundamentales y sobre las limitaciones de los Perceptrones.

A pesar del estancamiento, varios eventos en 1980 renovaron el interés en el tema, uno de ellos siendo la publicación de Kohonen, quien introdujo la red neuronal artificial, que es comunmente llamada Mapa o red de Kohonen.
Más adelante, en 1982, Hopfield publica un artículo describiendo las redes neuronales artificiales. Para 1985, el institito americano de Física inició lo que ahora se lleva a cabo como una reunión anual: Redes neuronales para computación. En 1987, la primera conferencia abierta sobre redes neuronales se llevó a cabo, y a día de hoy, progresos significativos se han llevado a cabo en el área, atrayendo atención

y financiamiento para continuar las investigaciones. A día de hoy, las discusiones sobre redes neuronales ocurren en todas partes [8].

3. El uso de redes neuronales en el pronóstico usando series de tiempo

3.1. Preliminares.

3.1.1. Definición. Se llama series de tiempo a un conjunto de observaciones sobre valores que toma una variable (cuantitativa) en diferentes momentos del tiempo. Los datos se pueden comportar de diferentes formas a través del tiempo, puede que se presente una tendencia, un ciclo; no tener una forma definida o aleatoria, variaciones estacionales (anual, semestral, etc).
Las observaciones de una serie de tiempo serán denotadas por $Y_1, Y_2, ..., Y_T$, donde Y_t es el valor tomado por el proceso en el instante t. Los modelos de series de tiempo tienen un enfoque netamente predictivo y en ellos los pronósticos se elaborarán sólo con base al comportamiento pasado de la variable de interés [13].

3.1.2. Componentes de una serie de tiempo. Se dice que una serie de tiempo puede descomponerse en cuatro componentes que no son directamente observables, de los cuales únicamente se pueden obtener estimaciones. Estas cuatro componentes son:

- Tendencia: Representa el comportamiento predominante de la serie. Esta puede ser definida vagamente como el cambio de la media a lo largo de un extenso período de tiempo.
- Ciclo: Caracterizado por oscilaciones alrededor de la tendencia con una larga duración, y sus factores no son claros. Por ejemplo, fenómenos climáticos, que tienen ciclos que duran varios años.
- Estacionalidad: Es un movimiento periódico que se producen dentro de un periodo corto y conocido. Este componente está determinado, por ejemplo, por factores institucionales y climáticos.
- Aleatoriedad: Son movimientos erráticos que no siguen un patrón específico y que obedecen a causas diversas. Este componente es prácticamente impredecible. Este comportamiento representan todos los tipos de movimientos de una serie de tiempo que no son tendencia, variaciones estacionales ni fluctuaciones cíclicas [13].

3.1.3. Deep Learning. Deep learning es una forma de "machine learning"que permite que las computadoras aprendan a partir de la experiencia y entiendan el mundo en términos de una jerarquía de conceptos. Como la computadora recopila información de experiencia, no es necesario que haya un operador humano de la computadora que especifique el conocimiento requerido por la computadora.
La jerarquía de conceptos, permite que la computadora aprenda conceptos complicados construyéndolos a partir de conceptos más sencillos [14].

3.1.4. Modelos estadísticos ARIMA. ARIMA es una reconocida familia de modelos de series de tiempo que se originó por su uso en economía. Esta familia de modelos, capaz de predecir puntos futuros en un conjunto de datos de series de tiempo, son valoradas por sus características estadísticas, su capacidad de implementar un rango de modelos exponenciales que aportan suavidad, y por la integración del método de Box-Jenkins durante la fase de entrenamiento del modelo.

Antes de comprender el modelo ARIMA, es crucial entender el operador de retrasos B, el cual es un dispositivo notacional bastante útil cuando se trata de lidiar con retrasos en una sucesión. Para una serie de tiempo Y_t, la serie con retraso se denotará por $BY_t = Y_{t-1}$, y similarmente $B_kY_t = Y_{t-k}$ [16]. Los modelos ARIMA contienen tres parámetros (p, d, q) y se pueden representar mediante:

$$(1 - B)^d Y_t = \mu + \Phi(B)(1 - B)^d Y_t = \mu + \Phi(B)Z(t) + \Theta(B)\epsilon_t. \tag{3.1}$$

3.2. Redes Neuronales. Una red neuronal es un tipo de inteligencia artificial que intenta imitar la forma en que el cerebro humano funciona. En lugar de utilizar un modelo digital, donde todos los cálculos manipulan ceros y unos, una red neuronal funciona creando conexiones entre los elementos procesados, lo cual, es el equivalente computacional de las neuronas.

Las redes neuronales son particularmente efectivas prediciendo eventos cuando las redes tienen una base de datos grande previo a las situaciones a predecir. También, son un conjunto de algoritmos modelados holgadamente en el cerebro humano, diseñados para reconocer patrones. Interpretan datos sensoriales a través de percepción de máquina, etiquetando o agrupando los datos de entrada. Los patrones reconocidos son numéricos, contenidos en vectores, hacia los cuales, todos los datos reales, sean estos imágenes, sonido, texto o series de tiempo, deben ser traducidos.

En su forma más sencilla, un cerebro biológico es una colección muy grande de neuronas. Cada neurona tiene signos químicos y eléctricos de entradas a través de sus numerosas dendritas y transmite las señales de salida mediante un axón.

Los axones entran en contacto con otras neuronas en uniones especializadas llamadas sinapsis, donde pasan sus señales de salida hacia otras neuronas para repetir el mismo proceso millones de veces.

Inspirándose en el cerebro, una red neuronal artificial es una colección de unidades conexas, también llamadas neuronas. La conexión entre las neuronas puede transportar señales entre ellas. Cada conexión transporta un valor de número real, el cual determina la fuerza de la señal [15].

Más aún, los modelos de aprendizaje profundo tienen la ventaja de ser capaces de descubrir automáticamente estructuras intrincadas en datos de altas dimensiones, una tarea que comúnmente requiere ser trabajada manualmente en el caso de bosque aleatorio. Los modelos de aprendizaje profundo se construyen utilizando múltiples capas de neuronas artificiales o nodos, diseñadas para replicar la estructura y funcionamiento del cerebro humano. Procesando los datos y creando patrones para la toma de decisiones, estos modelos son la clave tecnológica detrás de aplicaciones avanzadas, incluyendo el procesado de lenguaje natural, reconocimiento de voz, visión de computadora, bioinformática, entre otras.

3.2.1. Funciones de activación. El aprendizaje profundo se basa en redes neuronales artificiales con representación en capacidades de aprendizaje. Las arquitecturas del aprendizaje profundo como ser las redes neuronales profundas, las redes neuronales recurrentes y las redes neuronales convolucionales han sido aplicadas a través de varios campos, consistentemente demostrando una exactitud notable en tareas impulsadas por el reconocimiento de patrones en los datos. Comparativamente, métodos como bosque aleatorio, aunque son más sencillas de entrenar e interpretar, pueden fallar con tareas que requieran interacciones de alto nivel, y en áreas de

predicción de sucesiones, donde los métodos de aprendizaje profundo sobresalen [16].

Las redes neuronales se organizan en múltiples capas y cada capa se compone de un número de nodos interconectados que tienen funciones de activación asociadas a ellos. Se le proporciona los datos a la red por medio de la capa de entrada, la cual se comunica con otras capas y procesan los datos de la entrada con la ayuda de un sistema de conexiones pesadas. Estos datos procesados se obtienen posteriormente por medio de la capa de salida.

3.2.2. ¿Por qué las redes neuronales requieren funciones de activación? Dado que las redes neuronales son una red de múltiples capas de neuronas que consisten de nodos que se utilizan para la clasificación y predicción de datos ingresados como entrada de la red. Existe una capa de entrada, y una o múltiples capas ocultas y múltiples capas. Todas las capas tienen nodos y cada nodo tiene un peso que es tomado en consideración al procesar información de una capa a la siguiente.

Si una función de activación no se usa en una red neuronal, luego la señal de salida seria simplemente una una función lineal. Aunque una ecuación lineal es sencilla y simple de resolver, su complejidad es limitada y no tienen la habilidad de aprender y reconocer mapeos intricados de los datos. Una red neuronal sin función de activación sería equivalente a un modelo de regresión lineal con desempeño y potencia limitados la mayor parte del tiempo.

Es deseable que una red neuronal no solo aprenda y calcule, si no que, también desempeñe tareas más complicadas que modelar, como ser, el procesado de imágenes, video, audio, texto, etc. Esta es la razón por la cual usar funciones de activación y técnicas como el deep learning, que comprendan conjuntos de datos no lineales, o de altas dimensiones, donde el modelo tenga múltiples capas ocultas. El modelo antes mencionado, es la red neuronal densa, o también llamada multicapa. Otro modelo significativo en el aprendizaje profundo es la red neuronal recurrente (RNR). Las RNR son particularmente efectivas cuando se trabaja con datos secuenciales, y su particularidad es que mantienen una forma de memoria utilizando su propia salida como la entrada del siguiente paso. Esto hace que las RNR sean especialmente eficaces para tareas como el análisis de series de tiempo y procesado de lenguaje natural, donde el orden de los datos es significativo.

3.2.3. Arquitecturas utilizadas. Existen varios tipos de redes neuronales artificiales. Este tipo de redes son implementadas basadas en las operaciones matemáticas y un conjunto de parámetros requeridos para determinar la salida [15]. Para efectos de este estudio, aparte de los modelos estadísticos ARIMA, se utilizaron redes neuronales multicapa, y redes neuronales recurrentes, hablaremos de estas últimas a continuación.

Red Neuronal Multicapa: La red neuronal multicapa, o también llamada red neuronal densa, es el tipo más básico de red neuronal artificial, no hay circulación en la red. Los datos viajan en una misma dirección. Luego, una red neuronal multicapa es una función no lineal de entradas, lo cual es la composición de funciones de sus neuronas.

Una gran variedad de posibles topologías puede plantearse, bajo la única restricción que las conexiones de los grafos sean acíclicas. Sin embargo, la mayoría de aplicaciones utiliza redes neuronales multicapa, cuya arquitectura se muestra en la figura 1.

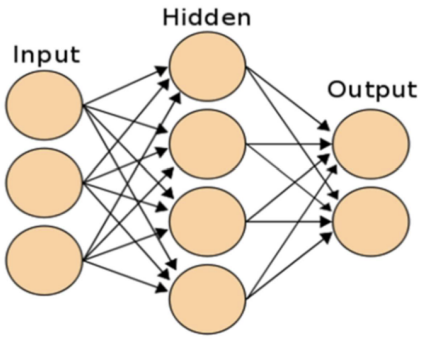


FIGURA 1. Red Neuronal Multicapa [18].

La red neuronal usada en la sección de experimentación, tiene una sola capa oculta, la ecuación que describe la red neuronal recurrente con una capa oculta es:

$$y_k = \tilde{g}\left(\sum_{j=1}^{M} w_{kj}^{(2)} \cdot g\left(\sum_{i=1}^{D} w_{ji}^{(1)} \cdot x_i + w_{j0}^{(1)}\right) + w_{k0}^{2()}\right), k = 1, ..., K, \tag{3.2}$$

donde $\tilde{g}$ y g son funciones de activación. En el caso de este estudio, se utilizó una función de activación de unidad lineal rectificada, también llamada ReLU.

Por otro lado, la red neuronal recurrente es la arquitectura más general de redes neuronales, cuyos grafos de conexión muestran ciclos. En dicho grafo, existe al menos un camino, seguido de las conexiones, que conducen de regreso a la neurona inicial, dicho camino es llamado ciclo.

Como la salida de la neurona no puede ser una función en si misma, una arquitectura de este tipo, requiere que el tiempo sea tomado en cuenta de forma explícita: la salida de la neurona no puede ser una función en el mismo instante de tiempo, pero puede ser función de su valor anterior.

En la actualidad, la basta mayoría de aplicaciones de redes neuronales se implementan como sistemas digitales, por lo cual, los sistemas discretos son el marco referencial para investigar usando redes neuronales recurrentes, las cuales se describen matemáticamente como ecuaciones recurrentes, que son los equivalentes discretos de las ecuaciones diferenciales continuas [17]. La arquitectura de una red neuronal recurrente se muestra en la figura 2.

3.3. Resultados y Experimentación. Se trabajó con datos reales, los cuales fueron proporcionados por el IHCIT en la UNAH, los cuales son recopilados como un promedio diario. A continuación, se muestra la serie de tiempo con los datos, los cuales, comprenden los años 1979 a 2023. Se tomaron los años 1979 a 2022 como entrenamiento para los modelos estadísticos y para las redes neuronales, usándose el año restante, 2023 para predicción.

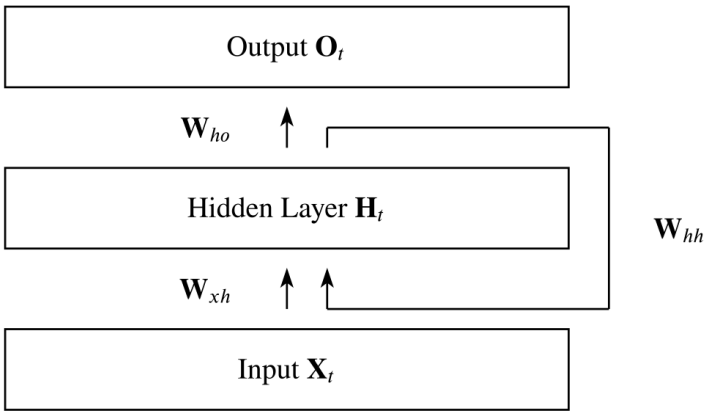


FIGURA 2. Red Neuronal Recurrente, [19].

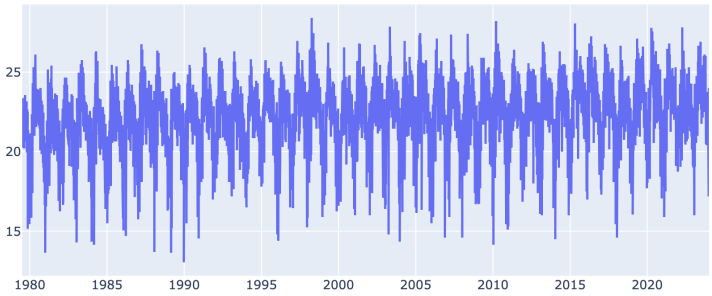


FIGURA 3. Datos reales de temperatura diaria

Los resultados de estimación obtenidos usando ARIMA (3,1,3) como fue sugerido por AutoArima en R, de acuerdo a la figura 4 es notorio que los modelos ARIMA se adaptan muy acertadamente a la tarea de estimación, incluso en valores que exceden los promedios de temperatura baja, y alta.

A pesar de los resultados sobresalientes en estimación, es válido afirmar que el modelo ARIMA falla cuando se trata de predecir, pues no se adapta a la estacionariedad de los datos, ni a los valores más bajos o altos de temperatura. Esto se evidencia en la figura 5.

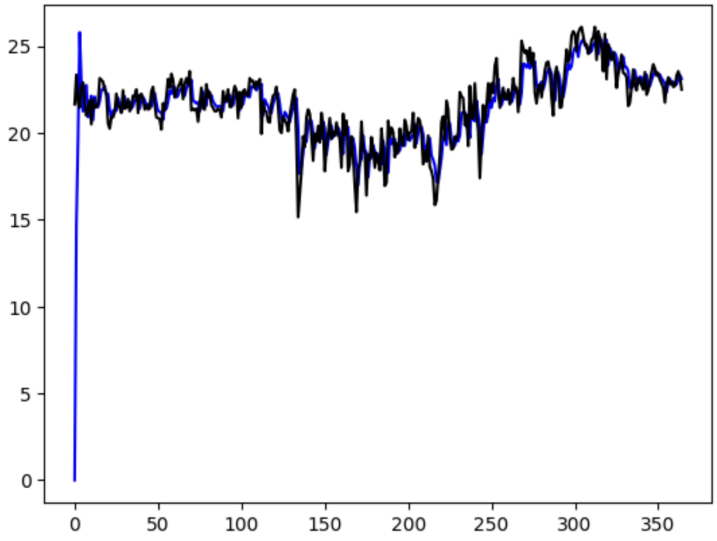


FIGURA 4. Estimación usando ARIMA (3,1,3)

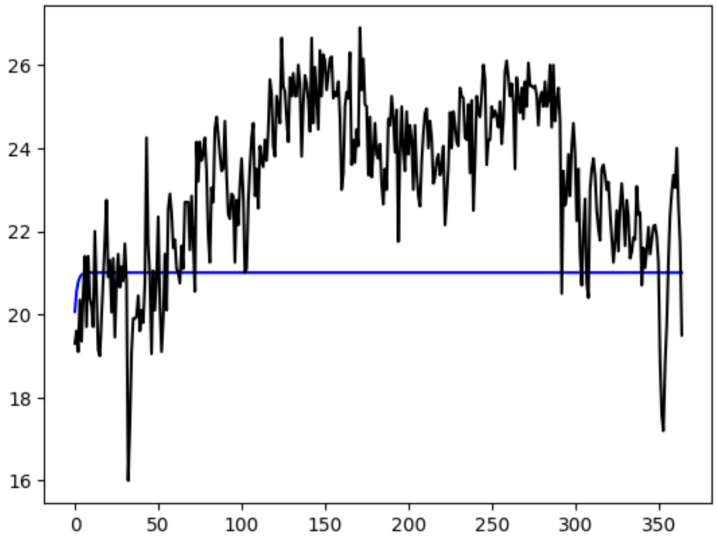


FIGURA 5. Predicción del año 2023 usando ARIMA (3,1,3)

La red neuronal densa, evidentemente muestra ventajas en el aprendizaje de la estacionariedad observada en los datos de temperatura, comparado con el modelo ARIMA visto anteriormente, y más aún, muestra habilidades prometedoras en predicción, replicando muy bien el comportamiento de los datos. El azul representa los datos reales, el verde la estimación realizada por la red neuronal, y el naranja es la predicción hecha por la misma.

La estimación usando la red neuronal recurrente es incluso mejor que los dos trabajados anteriormente.

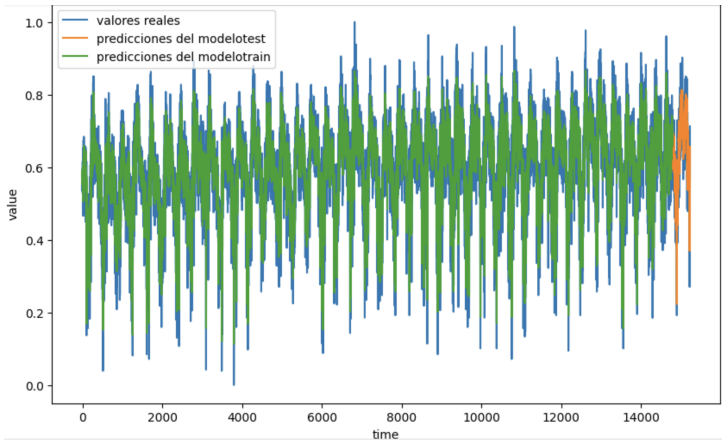


FIGURA 6. Estimación y predicción usando una red neuronal multicapa)

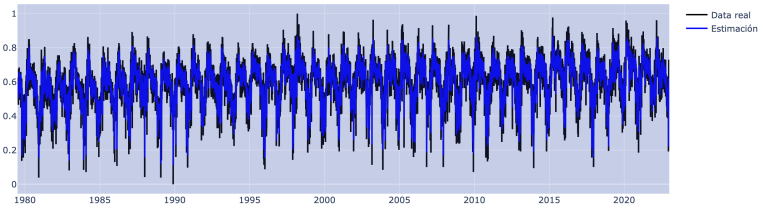


FIGURA 7. Estimación usando una red neuronal recurrente, periodo de entrenamiento comprendido de 1979 a 2022

Resultados de predicción usando una red neuronal recurrente

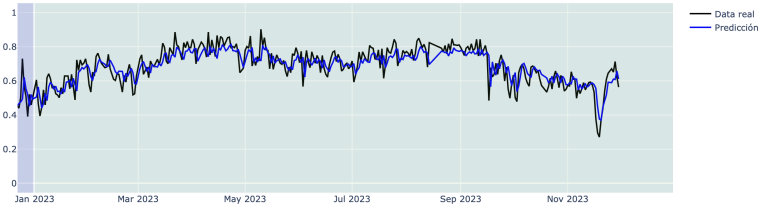


FIGURA 8. Predicción del año 2023 mediante una red neuronal recurrente

Cabe mencionar, con este último gráfico, que los valores máximos y mínimos son mejor estimados por la red neuronal recurrente, lo cual fue mencionado teóricamente en la sección 3.2.

4. CONCLUSIONES

- El modelo ARIMA, muestra muy buenos valores de estimación, pero discrepa en predicción.
- La red neuronal densa presenta resultados muy acertados tanto de estimación como de predicción, siendo una opción muy factible para predecir la estacionariedad del parámetro temperatura.

- La red neuronal recurrente presenta los mejores resultados dentro de los modelos utilizados para estimación y predicción por igual.
- Durante el desarrollo de este trabajo, surgieron algunas propuestas de investigación sin abordar a nivel de país, una de ellas siendo la implementación de estos modelos y metodologías con datos a nivel nacional, o bien, considerando más variables, creando así un modelo bivariado que correlacione variables meteorológicas, dando así una idea más clara del comportamiento de la temperatura como variable en Honduras.

Referencias

1. J.G. De Gooijer, R.J. Hyndman. *25 Years of Time Series Forecasting* Int. J. Forecasting, vol. 22, no 3, (2006).
2. I. Khandelwal, U. Satija, R. Adhikari *Forecasting Seasonal Time Series with Functional Link Artificial Neural Network*, 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, (2015).
3. D. L. Smith & J. J. Almaraz *Climate change and crop production: contributions, impacts, and adaptations, Canadian Journal of Plant Pathology*, 26:3, 253-266, DOI: 10.1080/07060660409507142 (2004)
4. H. Díaz-Ambrona, C. Gregorio, R. Gigena, and C. O. Mendoza. *Climate change impacts on maize and dry bean yields of smallholder farmers in Honduras* Iberoamerican Journal of Development Studies 2.1 (2013).
5. T. Dimri, S. Ahmad & M. Sharif. *Time series analysis of climate variables using seasonal ARIMA approach* J Earth Syst Sci, DOI: 10.1007/s12040-020-01408-x, (2020).
6. Y. Lai, and D. A. Dzombak. *Use of the autoregressive integrated moving average (ARIMA) model to forecast near-term regional temperature and precipitation.* Weather and Forecasting 35.3 (2020).
7. C. C. Aggarwal. *Neural Networks and Deep Learning*, Springer, Suiza, (2018).
8. Macukow, Bohdan. *Neural networks–state of art, brief history, basic models and architecture* Computer Information Systems and Industrial Management: 15th IFIP TC8 International Conference, CISIM 2016, Vilnius, Lithuania, Proceedings 15. Springer International Publishing, (2016).
9. McCulloch, W.S., Pitts, W. *A logical calculus of the ideas immanent in nervous activity* Bulletin of Mathematical Biophysics 5, 115–133. https://doi.org/10.1007/BF02478259 (1943)
10. Rosenblatt, F., Principles of Neurodynamics. Spartan Books, Washington (1962)
11. Rosenblatt, F. *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychol. Rev. 65(6), 386–408 (1958)
12. Murillo, Joaquín, Alvaro Trejos, and PATRICIA CARVAJAL OLAYA. *Estudio del pronóstico de la demanda de energía eléctrica, utilizando modelos de series de tiempo.* Scientia et technica 3.23 (2003).
13. Hurtado, Carlos, and G. Ríos. *Series de tiempo.* Santiago: Universidad de Chile (2008).
14. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning* MIT press (2016).
15. Islam, Mohaiminul, Guorong Chen, and Shangzhu Jin. An overview of neural network. American Journal of Neural Networks and Applications 5.1 (2019): 7-11.
16. Ospina, Raydonal, et al. *An overview of forecast analysis with ARIMA models during the COVID-19 pandemic: Methodology and case study in Brazil* Mathematics 11.14 (2023): 3069.
17. Dreyfus, G., and G. Dreyfus. *Neural networks: an overview. Neural Networks: Methodology and Applications* (2005): 1-83
18. Kokate, Pranali & Pancholi, Sidharth & Joshi, Amit. *Classification of Upper Arm Movements from EEG signals using Machine Learning with ICA Analysis.* (2021)
19. Schmidt, Robin M. *Recurrent neural networks (rnns): A gentle introduction and overview.* arXiv preprint arXiv:1912.05911 (2019).

Escuela de Matemática, Facultad de Ciencias, Universidad Nacional Autónoma de Honduras

Dirección de correo electrónico: nathalye.deras@unah.hn

SERIES TEMPORALES APLICANDO MEDIAS MÓVILES AL IPC PARA PREDECIR INFLACIÓN

IRVIN SAID CANALES ORDOÑEZ

Resumen. Este escrito narra paso a paso el desafío de pronosticar los niveles de inflación en el índice de precio del consumidor en los distintos productos en la canasta básica en Honduras, a través de un análisis de series temporales, utilizando datos proporcionados por la "Secretaría Ejecutiva del Consejo Monetario Centroamericano" (SECMC por sus siglas en español). Al implementar el modelo de medias móviles mejor conocido como: MA, que es el modelo básico en el que se busca comparar datos anteriores y actuales y ver en base a los resultados ¿De cuanto es el nivel porcentual de inflación en el territorio nacional? Cabe resaltar que existe la posibilidad también (aunque sea poco probable) de que tengamos 0 % de inflación, es decir, tendríamos estabilidad, e incluso en lugar de inflación podríamos tener rebajas o reducción en los precios.

Abstract. This paper narrates step by step the challenge of forecasting inflation levels in the consumer price index in the different products in the basic basket in Honduras through a time series analysis, using data provided by the "Executive Secretariat of the Central American Monetary Council" (SECMC by its acronym in Spanish). By implementing the moving average model, better known as: MA, which is the basic model, which seeks to compare previous and current data and see, based on the results, how much is the percentage level of inflation in the National territory? It should be noted that there is also the possibility (although it is unlikely) that we will have 0 % inflation, that is, we would have stability, and even instead of inflation we could have discounts or reductions in prices.

Fecha: 10 de Junio del 2024.
Palabras y frases clave. Series Temporales, Inflación, Estabilidad, Medias Móviles, Reducción.

1. Introducción

Ésta es una investigación ligada a la Escuela de Matemáticas en el área de Estadística y Economía de la Universidad Nacional Autónoma de Honduras y está dirigida a toda persona en cualquier ámbito (profesional o no) que tenga el interés de saber ¿Qué tanto se ven afectados sus ingresos económicos con respecto a la inflación en la economía actual del país? Ya que es un tema importante que afecta a cualquiera sin ver distinciones de clases sociales.
La inflación al consumidor es el aumento generalizado y sostenido de los precios de los bienes y servicios más representativos del consumo de los hogares de un país. Si bien hay distintas medidas de inflación, cuando hablamos de este fenómeno hacemos referencia a la que tiene que ver con el incremento de los precios que enfrentan los consumidores por sus compras habituales [2].

Este trabajo pretende ser una breve introducción al estudio de las series temporales, las cuales poseen una gran importancia en el campo de la Economía y la Estadística dada la abundancia de este tipo de observaciones; de hecho, dichas series constituyen la mayor parte del material estadístico con el que trabajan los economistas. Pero, ¿qué es una serie temporal? Sin duda es una pregunta un tanto compleja ya que su respuesta es demasiado extensa, pero en pocas palabras, una serie temporal es una sucesión de observaciones de una variable realizadas a intervalos regulares de tiempo [1]. Es claro que al avanzar en nuestra investigación indagaremos un poco mas en la definición formal y además de incluir los tipos de series y sus respectivos modelos matemáticos. El objetivo fundamental en el estudio de las series temporales es el conocimiento del comportamiento de una variable a través del tiempo y; a partir de dicho conocimiento y bajo el supuesto de que no van a producirse cambios estructurales, poder realizar predicciones, es decir, determinar qué valor tomará la variable objeto de estudio en uno o más períodos de tiempo situados en el futuro mediante la aplicación de un determinado modelo calculado previamente [1, 3].
Es importante recalcar que el "objetivo principal" de esta investigación es poder aplicar el modelo de medias móviles denominado como: MA al IPC con una base de datos proporcionada por la Secretaría Ejecutiva del Consejo Monetario Centroamericano; apoyándonos en el lenguaje de programación de R y con eso poder predecir si en los próximos meses a corto plazo (de 1 a 5 meses aproximadamente) habrá inflación en el IPC de Honduras.

Una observación importante es que en los tipos de series temporales están: las univariadas y las multivariadas [3] y este trabajo investigativo se basa únicamente en el caso univariante que tiene como variable independiente los índices de inflación del país ya que en el análisis multivariante considera simultáneamente series temporales múltiples. En general, es mucho más complicado que el análisis de series de tiempo univariante, especialmente cuando el número de series consideradas es grande.
Entre los distintos tipos de medias móviles que se pueden construir nos vamos a referir a dos tipos: centradas y asimétricas. El primer tipo se utiliza para la representación de la tendencia, mientras que el segundo lo aplicaremos para la predicción en modelos con media constante. Entre dichos tipos encontramos modelos como el ya mencionado MA(q) (de orden q), Alisado Exponencial Simple, están también los modelos ARMA que son una combinación entre los modelos de medias móviles MA y los modelos auto-regresivos AR [4]. En base a lo antes mencionado el modelo MA es una herramienta bastante útil para el análisis de datos y predicciones a futuro.

2. antecedentes

El análisis científico de las series de tiempo tiene una larga historia. De hecho, ha comenzado en 1664 cuando Sir Isaac Newton descompuso una señal luminosa (o serie de

tiempo) en sus componentes en diferentes frecuencias haciendo pasar la señal por un prisma de vidrio, sin embargo su trabajo estaba ligado al campo de el cálculo diferencial y tras muchos años de estudio y con el desarrollo de la estadística como ciencia en 1894 M. I. Pupin inventó el filtro de onda eléctrica. Este aparato ampliaba considerablemente el dominio de las frecuencias sobre las cuales la serie de tiempo podría ser analizada. La potencia de una señal eléctrica podía ser medida entonces en un intervalo de bandas de frecuencias [5].

El desarrollo teórico del análisis de series temporales comenzaron con el estudio de los procesos estocásticos. La primera aplicación a datos puede atribuirse al trabajo de Udny Yule y J. Walker entre 1920 y 1930. El modelo auto-regresivo (conocido hasta esa fecha como un sistema de cálculo) pudo ser descrito matemáticamente, con un buen grado de aproximación como una ecuación diferencial estocástica de segundo orden. Normalmente el tiempo discreto de la ecuación diferencial estocástica es que sea de primer orden; pero al aumentar el grado de la ecuación se obtuvo una mejor estimación, lo cual hoy en día es conocido como modelo auto-regresivo de segundo orden (denotado como AR(2)) [5].
Las series temporales son resultado de la confluencia de cinco líneas de trabajo en distintos campos científicos. En primer lugar, tenemos los estudios de series astronómicas y climáticas que dio lugar a la teoría de procesos estocásticos estacionarios desarrollados por los matemáticos Kolmogorov, Wiener y Cramer alrededor de 1910. En segundo lugar, con el fin de refinar la predicción en series de producción y ventas en la década de los sesenta, se desarrollaron los métodos de alisado, aprovechando la aparición de los primeros ordenadores. Se continuó con la teoría de predicción y control de sistemas de lineales, impulsada por la ingeniería aeronáutica y espacial. Una década después entró en juego la teoría de procesos no estacionarios y no lineales, desarrollada sobre todo por economistas y estadísticos. Y finalmente, el último campo, corresponde con los modelos multivariantes (en este trabajo desarrollamos modelos univariantes) y los métodos de reducción de la dimensión en sistemas dinámicos que siguen en desarrollo [6].

Una clase alternativa, y en muchas formas complementaria de otros métodos fueron los llamados modelos de promedios móviles de orden "q" (denotados como MA(q)), los cuales tienen propiedades bastantes diferentes a las de los modelos auto-regresivos, pero todavía pueden mostrarnos una forma de comportamiento "pseudo" periódico. El estudio sistemático de las propiedades estadísticas de los modelos de promedios móviles fue realizado por Herman Wold en 1949 [5]. En 1970 George Box y Gwilym Jenkins desarrollaron el modelo ARMA, el cual es una combinación entre los modelos de auto-regresión (AR) y medias móviles (MA), el libro que publicaron llamado "Time Series Analysis" [5, 7] marcó un hito en el análisis de series temporales al presentar una metodología unificada para estudiar series estacionarias y no estacionarias y aplicar estos modelos para casos prácticos, además de su gran contribución en el desarrollo de los fundamentos estadísticos de los sistemas de control.

Actualmente, se están desarrollando sistemas basados en Machine Learning y Deep Learning. De hecho en 2020, un grupo de científicos del MIT publicaron un artículo [8], en el que presentaban una herramienta que permite manejar complicados sistemas de inteligencia artificial para pronosticar la evolución de la bolsa, el tiempo o la probabilidad de desarrollar enfermedades. La nueva herramienta se llama tspDB (base de datos de predicción de series temporales) y permite obtener un nuevo algoritmo de predicción sobre una base de datos de series temporales ya existente. La base del nuevo algoritmo es otro anterior llamado SSA (análisis del espectro singular) que fue creado por el mismo equipo. Este potente algoritmo es capaz de corregir y sustituir los valores que faltan dentro de una serie de datos y pronosticar series temporales únicas. Este equipo propuso una solución para aplicación de estas predicciones sobre conjuntos de datos multivariantes a través de

apilar las distintas matrices de datos únicas y tratarlas como si fuera una sola matriz con mayor dimensión (MSSA) sobre la que se aplica el anterior algoritmo mencionado (SSA). Los investigadores aseguran que el sistema resultante es realmente efectivo en predicciones basadas en datos de series temporales multivariantes. De acuerdo al investigador principal, Devavrat Shah, aunque los datos de las series temporales sean cada vez más complejos, este algoritmo puede captar eficazmente cualquier estructura de las series temporales, permitiéndoles encontrar la lente adecuada para observar la complejidad del modelo de datos. Actualmente, el equipo está trabajando en mejorar la funcionalidad y la facilidad de uso del sistema y se está buscando nuevos algoritmos que sean compatibles. Uno de sus objetivos es poder permitir añadir cambios puntuales en el comportamiento de las series de datos y que el sistema los detecta automáticamente para incluirlos en sus predicciones, para ello necesitaron todos los modelos que anteriormente hemos mencionado (AR, MA, ARMA, ARIMA, Suavizado Exponencial etc.) [6, 8].

En 2019 la Universidad Nacional Autónoma de Puebla aplico el modelo de medias móviles (MA) como alternativas de inversión en un índice accionario, el propósito era encontrar la estrategia adecuada incorporando análisis técnico en los precios de activos financieros al momento de invertir en los índices accionarios, ha sido una labor de investigación por mucho tiempo con el fin de maximizar las ganancias en los mercados financieros que se encuentran en constante movimiento o tendencia [9] .

Es de conocimiento público que la expansión de las ciencias de la computación combinado con la estadística han permitido el desarrollo y las aplicaciones del análisis de datos utilizando las series temporales y sus distintos modelos, los cuales son aplicados (como hemos podido ver a lo largo de toda la historia) en cualquier campo profesional. De hecho, no parece tener un límite el uso de esta herramienta estadística que sin duda motiva a muchas personas a lo largo del tiempo a adaptarse e innovar para descubrir nuevos beneficios y avances en el análisis de datos.

3. Análisis y Conceptos Preliminares De Las Series De Tiempo

3.1. Series de Tiempo. En esta sección exploraremos los conceptos clave para el análisis de series temporales que serán fundamentales para nuestro trabajo. Destacaremos la importancia de la estacionariedad que nos será de utilidad para entender los datos que mas adelante analizaremos.

Definición 3.1. Se llama Series de Tiempo a un conjunto de observaciones sobre valores que toma una variable (cuantitativa) en diferentes momentos del tiempo. Las observaciones de una serie de tiempo serán denotadas por:

$$Y_1, Y_2, ..., Y_n \text{ donde } Y_i \ (\text{con } i = 1, 2, 3, ..., n) \tag{3.1}$$

es el valor tomado por el proceso en el instante i [10].

3.1.1. Proceso Estocástico. Un proceso estocástico es una sucesión de variables aleatorias ordenadas y equidistantes cronológicamente referida o varias características de una unidad observable en diferentes momentos.

Definición 3.2. Un proceso estocástico Y_t es estacionario cuando las propiedades estadísticas de cualquier sucesión finita $Y_1, ..., Y_n$ son semejantes a las de la sucesión Y_{n+h} para $h \in \mathbb{Z}$ [3].

3.1.2. Estacionariedad. Se refiere que los datos analizados poseen patrones que se repiten de manera periódica a lo largo de toda una serie temporal.

Definición 3.3. Una serie es estacionaria cuando es estable a lo largo del tiempo, es decir cuando la media y varianza son constantes y no cambian en el tiempo. Esto se

refleja gráficamente en que los valores de la serie tienden a oscilar alrededor de una media constante y la variabilidad con respecto a esa media también permanece constante en el tiempo. Además, una serie temporal $\{Y_n, n \in \mathbb{N}\}$ se considera débilmente estacionaria si su media, varianza y covarianza son constantes, en la condición de estacionariedad débil, asumimos que los primeros dos momentos de Y_n son finitos [4].

3.2. Variaciones Estacionales. Son movimientos ascendentes y descendentes respecto de la tendencia que se consuman en el término de un año y se repiten anualmente, estas variaciones suelen identificarse con base en datos mensuales o trimestrales [1].

Observación 3.3.1. Si la serie no cumple la definición descrita anteriormente entonces es no estacionaria, en ese caso podemos corregir el problema ya que en la mayoría de los modelos de series de tiempo necesitamos que la serie sea estacionaria y para ello necesitamos lo siguiente:

- Ajuste por diferencial: calcula las diferencias entre observaciones consecutivas.
- Cálculo de log o raíz cuadrada: para estabilizar la varianza no constante.
- Prueba de raíz unitaria: esta prueba se usa para descubrir la primera diferencia o regresión que se debe usar para hacerla estacionaria. En la prueba Kwiatkowski-Phillips-Schmidt-Shin (KPSS), los valores pequeños de p sugieren que se requiere una diferenciación (o varias) [4, 14].

3.3. Propiedades.

- La estacionariedad débil no implica una estacionariedad estricta.
- Una secuencia independiente e idénticamente distribuida es estrictamente estacionaria [12].

3.3.1. Función de Auto-correlación y Correlación. Dichas funciones nos ayudan a evaluar el grado de dependencia en los datos y además marca la relación entre las variables en un determinado tiempo y uno(s) anterior a este.

Definición 3.4. El Coeficiente de Correlación entre dos variables aleatorias X e Y es definido como: $\rho_{x,y}$ que no es más que la covarianza de ambas variables sobre la raíz cuadrada del resultado de multiplicar la varianza de cada una de las variables y recordemos que Var(X)= E$(X - \mu_x)^2$ (de manera análoga para Y) además note que μ_x y μ_y son las medias de X e Y respectivamente. Este coeficiente mide la fuerza de la dependencia lineal entre X e Y, las dos variables aleatorias no están correlacionadas si $\rho_{x,y} = 0$ [11].

Definición 3.5. Considere una serie estacionaria r_t. Cuando la dependencia lineal entre r_t y sus valores pasados r_{t-l} son de interés, el concepto de correlación está generalizado a la auto correlación, es decir los coeficientes de correlación entre r_t y r_{t-l} se llama el auto correlación de retraso l de r_t y comúnmente se denota por ρ_l [11] y que bajo las normas de estacionariedad sigue siendo una función de l definida como:

$$\rho_l = \frac{\text{Cov}(r_t, r_{t-l})}{\text{Var}(r_t)}. \tag{3.2}$$

3.3.2. Ruido Blanco. El ruido blanco es una señal aleatoria que describe fluctuaciones aleatorias sin ninguna estructura o patrón discernible. Es una herramienta esencial en el análisis de series temporales, ya que su objetivo principal es modelar la variabilidad aleatoria en los datos. Además, ayuda a distinguir entre la señal (patrones de interés) y el ruido (fluctuaciones aleatorias) en un conjunto de datos observados, lo que facilita el análisis y la interpretación de los datos [12].

Definición 3.6. Una serie temporal Y_t se llama ruido blanco si $\{Y_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con media y varianza finitas.

Proposición 3.7. Sea Y_t una serie de ruido blanco, entonces dicha serie es estacionaria

3.3.3. Criterio De Comparación de Series. La búsqueda de un método que permita identificar el grado de similitud entre dos o más series de tiempo es importante de realizar ya que en la práctica, es frecuente enfrentar situaciones en las que se requiere comparar dos o mas series temporales, o analizar un gran número de estas series para separarlas en grupos tan homogéneos como sea posible. Normalmente se generan bases de datos temporales que luego se estudian para identificar concordancias y disimilitudes [13].

Definición 3.8. Sean S_1 y S_2 dos series de tiempo observadas en los instantes $t_1, ..., t_p$ ($con\ p \in \mathbb{N}$).Decimos que S_1 y S_2 son similares en comportamiento, si en cualquier período de tiempo observado aumentan o disminuyen de forma simultánea con una misma tasa de crecimiento. En otras palabras, si el valor de correlación entre ambas series se aproxima a 1 [13].

3.3.4. Comprobación del Modelo. Un modelo ajustado debe examinarse cuidadosamente para verificar posibles deficiencias. Si el modelo es adecuado, entonces la serie de residuos debería comportarse como un ruido blanco. El ACF (prueba de auto-correlación) de los residuos puede usarse para verificar la cercanía de $\hat{a}_t$ a un ruido blanco [15].

3.4. Modelo AR. El modelo AR es el llamado modelo auto-regresivo en el cual la variable de estudio en un período de tiempo t es explicada por las observaciones de ella misma.

Definición 3.9. Sea r_t una serie de tiempo, el modelo AR de orden 1 es define como

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t \tag{3.3}$$

donde $\{a_t\}$ se asume ser una serie de ruido blanco con media cero y varianza σ_a^2 y $\phi_0 + \phi_1 r_{t-1}$ se conoce como el polinomio de retraso [4].

Este modelo tiene la misma forma que la conocida regresión lineal simple. modelo en el que r_t es la variable dependiente y r_{t-1} es la variable explicativa. En las lecturas sobre series temporales anteriores hemos visto que el modelo se denomina autor-regresivo AR (modelo de orden 1) o simplemente modelo AR(1). Este modelo simple también es ampliamente utilizado en el modelado de volatilidad estocástica cuando r_t se reemplaza por su volatilidad logarítmica [11].

Una generalización sencilla del modelo AR(1) es el modelo AR(p) denotado:

$$r_t = \phi_0 + \phi_1 r_{t-1} + \cdots + \phi_p r_{t-p} + a_t, \tag{3.4}$$

donde p es un número entero no negativo y a_t se define como en la ecuación de orden 1. Este modelo dice que los valores p pasados r_{t-i} $(i = 1, 2, ..., p)$ determinan conjuntamente la expectativa condicional de r_t dados los datos pasados. El modelo AR(p) tiene la misma forma que un modelo de regresión lineal múltiple con valores rezagados que sirven como variables aclaratorias o de explicaciones de datos [11].

Una representación gráfica de la ACF de un modelo AR(p) estacionario debería mostrar un efecto que tiende a reducir la amplitud de las oscilaciones (asemejando a un oscilador armónico) en otras palabras, la ACF mostraría una mezcla de amortiguación de senos y cosenos con un decaimiento exponencial que dependerá de la naturaleza de las raíces características.

3.4.1. Función de Auto-correlación Parcial (PACF). La PACF de una serie temporal estacionaria es función de su ACF y es una herramienta útil para determinar el orden p de un modelo AR. Una forma sencilla pero eficaz de introducir la PACF es considerar los siguientes modelos AR en orden consecutivo:

$$r_t = \phi_{0,1} + \phi_{1,1} r_{t-1} + e_{1,t},$$

$$r_t = \phi_{0,2} + \phi_{1,2} r_{t-1} + \phi_{2,2} r_{t-2} + e_{2,t},$$

$$\vdots$$

donde $\phi_{0,j}$, $\phi_{i,j}$, y $\{e_{j,t}\}$ son respectivamente el término constante, el coeficiente de r_{t-i} y el término de error de un modelo AR(j). Estos modelos están en forma de una regresión lineal múltiple y pueden ser estimados por el método de mínimos cuadrados. La estimación $\hat{\phi}_{1,1}$ de la primera ecuación se llama la PACF de muestra de lag-1 de r_t. La estimación $\hat{\phi}_{2,2}$ de la segunda ecuación es la PACF de muestra de lag-2 de r_t, y así sucesivamente [11].

3.4.2. Criterio de Información o Comparación. Existen varios criterios de Comparación para determinar el orden p de un modelo. Todos ellos se basan en la verosimilitud [11, 4]. El criterio de información de Akaike (AIC) se define como:

$$\text{AIC} = -\frac{2}{T}\ln(\text{verosimilitud}) + \frac{2}{T} \times (\text{número de parámetros}),$$

donde la función de verosimilitud se evalúa en las estimaciones de máxima verosimilitud y T es el tamaño de la muestra. El primer término del AIC mide la bondad de ajuste del modelo a los datos, mientras que el segundo término se llama función de penalización del criterio porque penaliza un modelo candidato por el número de parámetros utilizados [11]. Diferentes funciones de penalización resultan en diferentes criterios de información. Otro criterio de función comúnmente utilizado es el criterio de información Schwarz-Bayesiano (BIC) [11]

$$\text{BIC}(P) = \ln(\tilde{\sigma}_P^2) + \ln(T)\frac{p}{T}.$$

La penalización por cada parámetro utilizado es de 2 para AIC y $\ln(T)$ para BIC. Por lo tanto, en comparación con AIC, BIC tiende a seleccionar un modelo más bajo cuando el tamaño de la muestra es moderado o grande.

3.5. Modelo de Medias Móviles MA. Es el promedio de los n valores de datos más recientes de una serie de tiempo, a medida de que se dispone del nuevo valor de un dato de una serie de tiempo, la nueva observación reemplaza a la antigua en la serie de n valores como base para determinar el nuevo promedio.

El promedio móvil sirve para pronosticar valores de datos del siguiente periodo de tiempo de la serie, es decir el t_{i+1} pero no los datos de periodos más distantes a futuro. Es decir que realiza predicciones a futuro de manera inmediata y no tan extenso en el tiempo. Este procedimiento sirve para promediar el componente irregular de los datos mas recientes de una serie de tiempo [1].

Definición 3.10. Las variaciones irregulares son variaciones erráticas respecto de la tendencia que no puedan atribuirse a las influencias cíclicas o estaciónales [1].

Hay varias formas de introducir modelos MA. Un enfoque es tratar el modelo como una simple extensión de series de ruido blanco. Otro enfoque es tratar el modelo como un modelo AR de orden infinito con algunas restricciones de parámetros [11]. Adoptamos el segundo enfoque. No hay una razón particular sólo la simplicidad, para asumir a priori que el orden de un modelo AR es finito.
Podemos considerar al menos en teoría un modelo AR con orden infinito como:

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + \cdots + a_t. \tag{3.5}$$

Sin embargo, dicho modelo AR no es realista porque tiene un número infinito de parámetros. Una forma de hacer que el modelo sea práctico es asumir que los coeficientes ϕ_i satisfacen algunas restricciones para que estén determinados por un número finito de parámetros. Un caso especial de esta idea es:

$$r_t = \phi_0 - \theta_1 r_{t-1} - \theta_2 r_{t-2} - \theta_3 r_{t-3} - \cdots + a_t, \tag{3.6}$$

Donde los coeficientes dependen de un solo parámetro θ_1 a través de $\phi_i = -\theta_1^i$ para $i \geq 1$. Para que el modelo en la Ecuación (3.5) sea estacionario, θ_1 debe ser menor que 1 en valor absoluto; de lo contrario, θ_1^i y la serie divergirán, es decir que tendremos datos incomprensibles que se extenderán al infinito. Dado que $|\theta_1| < 1$, tenemos $\theta_1^i \to 0$ a medida que $i \to \infty$. Por lo tanto, la contribución de r_{t-i} a r_t decae exponencialmente a medida que i aumenta. Esto es razonable ya que la dependencia de una serie estacionaria r_t en su valor rezagado r_{t-i}, si la hay, debería decaer con el tiempo [11].

Definición 3.11. La forma general de un modelo MA(1) es

$$r_t = c_0 + a_t - \theta_1 a_{t-1} \tag{3.7}$$

donde c_0 es una constante y $\{a_t\}$ es una serie de ruido blanco.

Un modelo MA(q) es de la forma

$$r_t = c_0 + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{3.8}$$

donde $q > 0$ [1, 11, 4].

3.5.1. Suavizado Exponencial. Es un método de pronóstico basado en el uso de promedios móviles ponderados que consiste en añadir un valor de ponderación para cada dato, no son promedios en los que se ponderan por igual los valores de datos precedentes, la base de ponderación es exponencial por lo que se concede la mayor ponderación al valor correspondiente al periodo inmediatamente anterior al periodo de pronóstico y las ponderaciones decrecen exponencialmente para los valores de datos de periodos anteriores.
El siguiente modelo sirve para representar la determinación de ponderaciones exponencialmente decrecientes. Sea α una constante de suavización para hacer pronósticos, inicialmente se requiere de un valor "semilla" y dicho modelo se denota:

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t). \tag{3.9}$$

Donde $\hat{Y}_{t+1}$ es el pronóstico para el siguiente periodo, $\hat{Y}_t$ es el pronóstico del periodo más reciente, α es la constante de suavización entre 0 y 1 y $\hat{Y}_t$ es el valor real para el periodo mas reciente [1].

3.6. Modelo ARMA. En algunas aplicaciones, los modelos AR o MA discutidos en las secciones anteriores se vuelven tediosos porque puede ser necesario un modelo de alto orden con muchos parámetros para describir adecuadamente la estructura dinámica de los datos. Para superar esta dificultad, se introducen los modelos auto-regresivos de medias móviles (ARMA) [11]. Básicamente, un modelo ARMA combina las ideas de los modelos AR y MA en una forma compacta de manera que el número de parámetros utilizados se mantenga pequeño, logrando la parsimonia en la parametrización. En esta sección, estudiamos el modelo ARMA(1,1) más simple.

Definición 3.12. Una serie temporal r_t sigue un modelo ARMA(1,1) si satisface la siguiente ecuación:

$$r_t - \phi_1 r_{t-1} = \phi_0 + a_t - \theta_1 a_{t-1} \tag{3.10}$$

Donde $\{a_t\}$ es una serie de ruido blanco. El lado izquierdo de la ecuación (3.10) es el componente AR del modelo y el lado derecho da el componente MA. El término constante es ϕ_0. Para que este modelo sea significativo, necesitamos $\phi_1 = \theta_1$; de lo contrario, habría una cancelación en la ecuación y el proceso se reduciría a una serie de ruido blanco [11].

Definición 3.13. Un modelo ARMA(p,q) general tiene la forma:

$$r_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i}, \tag{3.11}$$

donde $\{a_t\}$ es una serie de ruido blanco y p y q son enteros no negativos. Los modelos AR y MA son casos especiales del modelo ARMA(p,q).

3.6.1. Identificación del Orden. El ACF y el PACF no son informativos para determinar el orden de un modelo ARMA [11]. La función que nos ayudará a identificar el orden es el "Likelihood Ratio Test" [16]. El Likelihood Ratio Test (LRT) es una prueba estadística utilizada para comparar la adecuación de dos modelos estadísticos, evaluando si un modelo más complejo es significativamente mejor que un modelo más simple al comparar la verosimilitud de los datos entre ambos modelos.

3.6.2. Series temporales estacionarias de tendencia. Un modelo estrechamente relacionado que exhibe una tendencia lineal es el modelo de series temporales estacionarias de tendencia [12].

Definición 3.14. Sea r_t una serie temporal estacionaria, un modelo de series temporales estacionarias de tendencia está dado por:

$$p_t = \beta_0 + \beta_1 t + r_t, \tag{3.12}$$

donde p_t crece linealmente en el tiempo con una tasa β_1.

Definición 3.15. La serie estacionaria de tendencia puede transformarse en una estacionaria eliminando la tendencia temporal mediante un análisis de regresión lineal simple.

3.6.3. Modelos Generales No Estacionarios de Raíz Unitaria. Considera un modelo ARMA. Si se extiende el modelo permitiendo que el polinomio AR tenga 1 como raíz característica, entonces el modelo se convierte en el conocido modelo auto-regresivo integrado de media móvil (ARIMA). Se dice que un modelo ARIMA es no estacionario debido a raíz unitaria porque su polinomio AR tiene una raíz unitaria. Un enfoque convencional para manejar la no estacionariedad debido a la raíz unitaria es utilizar la diferenciación.

3.6.4. Diferenciación. Algunas series temporales no son estacionarias debido a tendencias o efectos estacionales [5]. Las series no estacionarias pueden transformarse en series estacionarias mediante la diferenciación. Una vez diferenciadas, podemos ajustar un proceso ARMA en ellas. Estos procesos son conocidos como procesos auto-regresivos integrados de media móvil (o ARIMA), ya que la serie diferenciada necesita ser sumada o integrada para recuperar la serie original [16].

Definición 3.16. Una serie temporal y_t se dice que es un proceso ARIMA(p, 1, q) si la serie de cambios $c_t = y_t - y_{t-1} = (1 - B)y_t$ sigue un modelo ARMA(p,q) estacionario. Donde $c_t = y_t - y_{t-1}$ se refiere como la primera serie diferenciada de y_t.

Definición 3.17. Una serie temporal y_t puede contener múltiples raíces unitarias y necesita ser diferenciada varias veces para volverse estacionaria. Además, si s_t sigue un modelo ARMA(p,q), entonces y_t es un proceso ARIMA(p, 2, q) [17].

3.6.5. Prueba de Raíz Unitaria. Para probar si una serie temporal p_t es estacionaria, empleamos los modelos:

$$p_t = \phi_1 p_{t-1} + e_t \tag{3.13}$$

$$p_t = \phi_0 + \phi_1 p_{t-1} + e_t \tag{3.14}$$

donde e_t denota el término de error, y consideramos la hipótesis nula $H_0 : \phi_1 = 1$ frente a la hipótesis alternativa $H_a : \phi_1 < 1$. Este es el conocido problema de prueba de raíz unitaria [16]. Una estadística de prueba conveniente es la razón t de la estimación de mínimos cuadrados (LS) de ϕ_1 bajo la hipótesis nula. Para la Ecuación (3.12), el método de mínimos cuadrados (LS) proporciona

$$\hat{\phi}_1 = \frac{\sum_{t=1}^{T} p_{t-1} p_t}{\sum_{t=1}^{T} p_{t-1}^2}, \quad \hat{\sigma}_e^2 = \frac{\sum_{t=1}^{T} (p_t - \hat{\phi}_1 p_{t-1})^2}{T - 1}, \tag{3.15}$$

donde $p_0 = 0$ y T es el tamaño de la muestra. La razón t es

$$DF \equiv \text{t ratio} = \frac{\hat{\phi}_1 - 1}{\text{std}(\hat{\phi}_1)} = \frac{\sum_{t=1}^{T} p_{t-1} e_t}{\hat{\sigma}_e \sqrt{\sum_{t=1}^{T} p_{t-1}^2}}, \tag{3.16}$$

que comúnmente se conoce como la prueba de Dickey-Fuller (DF). Si $\{e_t\}$ es una serie de ruido blanco con momentos finitos de orden ligeramente mayor que 2, entonces la estadística DF converge a una función de la movilidad Browniana estándar a medida que $T \to \infty$ [18].

3.6.6. Series Temporales Estacionales. En algunas aplicaciones, la estacionalidad es de importancia secundaria y se elimina de los datos, lo que resulta en una serie temporal ajustada estacionalmente que luego se utiliza para hacer inferencias [11].

Definición 3.18. El procedimiento para eliminar la estacionalidad de una serie temporal se denomina ajuste estacional [11].

En otras aplicaciones como la predicción, la estacionalidad es tan importante como otras características de los datos y debe manejarse en consecuencia. Dado que la predicción es un objetivo importante para esta investigación, nos centramos y discutimos algunos modelos que son útiles en el modelado de series temporales estacionales.

3.7. Experimento.

3.7.1. Procesamiento De Datos. Nuestro experimento comienza con el análisis exhaustivo de la base de datos proporcionada por la Secretaría Ejecutiva del Consejo Monetario Centroamericano, en ella tenemos los resultados de la variación en el índice de precio al consumidor desde Febrero de 1990 hasta Abril de 2024, lo primero que notamos fue que los datos poseen 2 variables (fecha y variación) lo que hicimos fue tomar únicamente la variable de variación ya que al momento de importar nuestros datos a R procederemos a transformar los datos en una serie de tiempo y al hacer eso automáticamente en los datos aparecerá la fecha por defecto pero sin ser una variable que pueda ser manipulada.

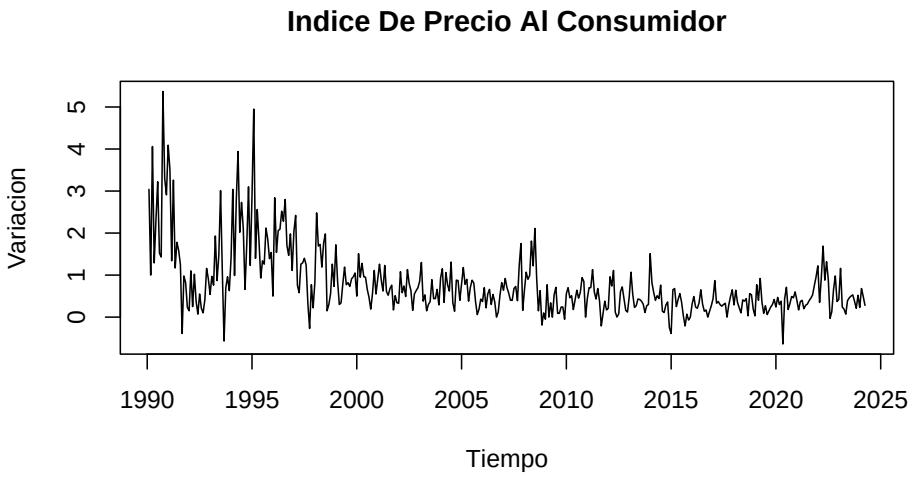


FIGURA 1. Comportamiento de la variación en los precios con respecto al tiempo.

Se realizó una visualización de la serie de tiempo con el objetivo de ver patrones, tendencias, y por supuesto ver el comportamiento de los datos através del tiempo con el fin de determinar si era necesario hacer algún test de estacionariedad, en este caso la prueba visual fue determinante como se mostró en la figura 1 .

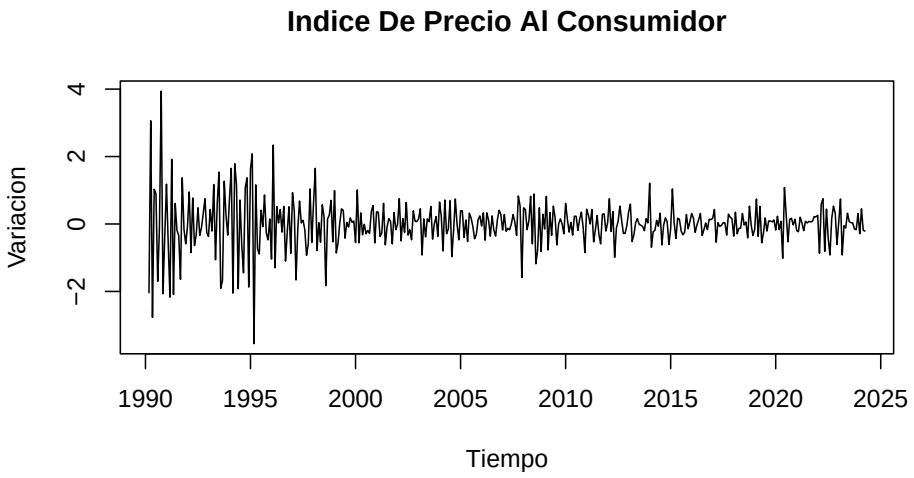


FIGURA 2. Comportamiento de la variación en los precios con respecto al tiempo con una diferencia.

Con la ayuda del gráfico mostrado en la figura 1 la serie parece tener un buen comportamiento, sin embargo los datos parecen extenderse en los años 90's hasta el año 2000. Así que para asegurarnos que la serie es estacionaria o no procedemos hacer la prueba de Dickye- Fuller aumentada [16]. Una vez realizada la prueba se observa el valor obtenido (p) y se compara con el valor de significancia para determinar si se acepta o se rechaza la hipótesis nula, es decir si $p \leq 0.05$ se rechaza la hipótesis nula H_0 lo que nos indica que la serie es estacionaria. En nuestro caso $p = 0.01$ así que rechaza la hipótesis nula y nos confirma que la serie es estacionaria por ende no necesita ninguna hacer ninguna diferencia. Sin embargo por la varianza inestable que presentan los datos desde el 90 hasta el 2000 se decidió realizar la prueba de KPSS (Kwiatkowski-Phillips-Schmidt-Shin) que nos indica el número de diferencias necesarias para obtener estacionariedad y desde luego los datos arrojaron que necesitamos una diferencia y el resultado mejoró considerablemente como se puede ha podido apreciar en la figura 2.

Luego procedemos a aplicar el modelo de medias móviles MA(q), pero antes necesitamos realizar la prueba ACF (auto-correlación) para determinar el valor de q [11]. Una vez realizada la prueba arrojó un valor para q entre 10 y 19, sin embargo dicha prueba no arroja un valor exacto y no es suficiente para determinar el mejor modelo que se ajuste a los datos, entonces se hizo la prueba Auto ARIMA que consiste en realizar un análisis exhaustivo para determinar el mejor modelo ARIMA y la prueba sugirió un modelo SARIMA con una diferencia y 2 medias móviles. Finalmente luego de realizar todo el análisis respectivo para el modelo se tomó la decisión de tomar un rango para q que va desde $q = 2$, hasta $q = 30$ y luego de ir realizando prueba tras prueba y ver que los resultados no varían significativamente cuando el valor de q en la prueba realizada es muy cercano al siguiente

(es decir $q = 2$ y $q = 3$ por ejemplo) por ende nos quedamos con 3 modelos propuestos como los mejores candidatos y el objetivo será ver cual de los 3 se ajusta mejor a la serie para poder tener mejores predicciones y los candidatos fueron MA(2), MA(19) y MA(30). Para ello comenzamos particionando la serie en 2, una que va desde Febrero de 1990 hasta Diciembre del 2023 esto con el objetivo de comparar los datos que ya tenemos de la serie original y los de predicción que nos proporcionarán los modelos y los resultados para los meses de Enero hasta Abril de 2024 fueron:

Fecha	Real.	MA(2) Pred.	MA(19) Pred.	MA(30) Pred.
Enero-2024	0.23	0.4246	0.1889	0.4029
Febrero-2024	0.68	0.4180	0.4642	0.4226
Marzo-2024	0.49	0.4180	0.5434	0.5879
Abril-2024	0.28	0.4180	0.4478	0.4677

Tabla 1. IPC

Note que en todos los casos nos acercamos pero ninguno fue tan preciso como para llegar al valor real aunque hay que destacar el modelo MA(19) en el mes de enero, pero también el modelo MA(2) en el mes de marzo estuvo bastante cerca. Entonces ¿Cómo saber que modelo es el mejor de los 3? para ello haremos uso de la raíz cuadrada del error cuadrático medio (RMSE) y también del error cuadrático medio absoluto (MAE) y en base a los resultados se decidirá cuál es el mejor modelo.

Como mencionamos anteriormente particionamos la serie en dos. Una llamada Serie Training que va desde Febrero del 90 hasta Diciembre del 2023 y en ella se observará en cual caso tendremos un mejor ajuste, por otro lado tenemos la serie Test y en ella se observará en que caso tendremos una mejor predicción.

Pruebas Realizadas	MA(2)	MA(19)	MA(30)
RMSE Serie Training	0.5884992	0.5266832	0.5027334
RMSE Serie Test	0.1807986	0.1407712	0.1877541
MAE Serie Training	0.3909988	0.3820614	0.3645991
MAE Serie Test	0.1666574	0.1195393	0.1790185

Tabla 2. Métricas De Evaluación De Modelos

Como podemos observar en la tabla de métricas, calculamos el RMSE y el MAE para las series Training y Test y ambas métricas funcionan así: “Entre mas bajo sea el número observado el modelo es mejor” , sin embargo no siempre es así, en nuestro caso el RMSE de la serie Training dictamina un mejor ajuste del modelo propuesto con respecto a la serie original (Llamada IPC.ts) y como podemos observar el modelo MA(30) es el que mejor se ajusta a los datos ¡Pero! el hecho que se ajuste mejor no necesariamente es mejor, de hecho en algunos casos podríamos tener “sobreajuste” eso sucede cuando en lugar de que el modelo se adapte a la serie pasa lo contrario que la serie se adapte al modelo, normalmente ocurre cuando tenemos un modelo pesado como en este caso que el MA(30) tiene 30 medias móviles son muchas más de las que sugirió la prueba de auto-correlación

ACF lo que lo convierte en un modelo poco favorable.
Por otro lado tenemos el RMSE de la serie test el cual nos indica una mejor predicción y como podemos observar este nos dice que el modelo que mejor predijo los valores de inflación con respecto a la serie original fue el MA(19).

De manera análoga se hace el mismo análisis para el MAE y notamos que al igual que el RMSE el modelo que tiene un mejor ajuste es el MA(30), sin embargo el modelo que mejor predice al igual que en el caso anterior es el MA(19) por ende de los tres candidatos es el mejor modelo para este experimento.
En el siguiente gráfico se muestran las proyecciones de predicción que arrojó el modelo MA(19) al futuro inmediato y como se puede observar los puntos azules denotan una alza mínima pero lo suficientemente confiable ya que se adapta a los datos de la serie original siguiendo el mismo patrón lo que demuestra que su ajuste es bueno

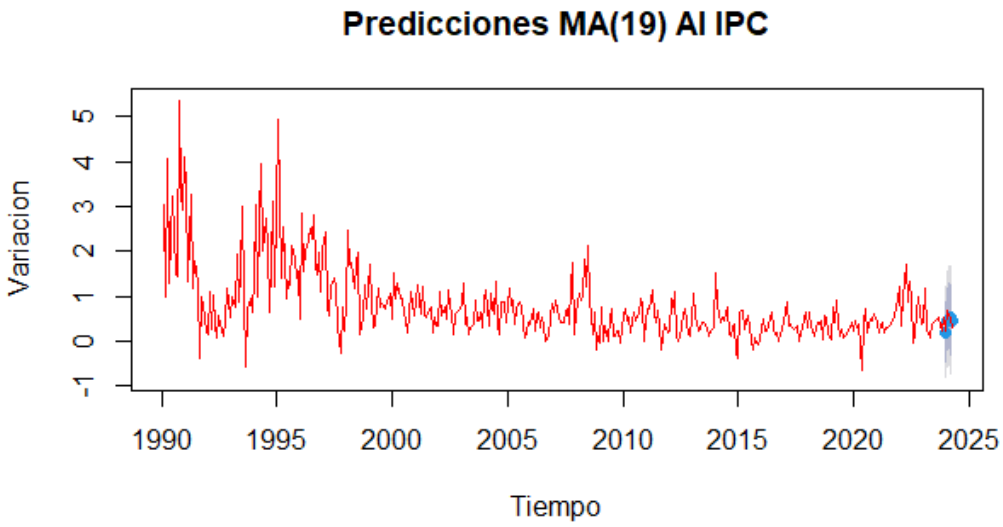


Figura 3. Predicciones A Futuro.

4. Conclusiones

- El modelo de medias móviles es eficiente para predecir a futuro inmediato, sin embargo para predecir niveles de inflación en Honduras necesitaríamos mas variables como por ejemplo: La variación del precio de los combustibles ya que en un país de tercer mundo, es una variable de mucha importancia en el IPC y así poder tener mejores y más precisos resultados.
- Aunque con 19 medias móviles obtuvimos mejores resultados que los otros 2 modelos no deja de ser un modelo que es algo pesado y normalmente se acostumbra que aunque sea un poco menos significativo es mejor quedarse con el modelo más simple.
- Ya que no es el modelo más simple de medias móviles habría que indagar más a fondo y comprobar si los resultados agregando nuevas variables mejoran o no. Si no habría que ver que modelo de los mencionados en este informe de series temporales

se ajusta mejor a la serie y posiblemente podríamos tener mejores resultados de predicción

- Como trabajos a futuro se puede usar siempre el modelo medias móviles pero para evitar tener tantas, se puede aplicar el suavizado exponencial. Es un modelo que no sirve para predecir pero si para hacer más simple un modelo
- Analizar varias variables que estén co-relacionadas como dijimos anteriormente una posible variable extra seria el precio de los combustibles, la variación del dolar, cantidad de remezas etc. Un método que podría ser más eficaz sería usar una red neuronal para obtener resultados más precisos

Referencias

1. Luis C. Toledo Vega, *Modelos de Series de Tiempo*, Universidad Autónoma del Estado de México, 2015.
2. Paul Samuelson y William Nordhaus, *Macroeconomía con aplicaciones a Latinoamérica*, 19ed, McGraw-Hill interamericana Editores, México, 2010.
3. José A. Mauricio, *análisis de series temporales*, Universidad Complutense de Madrid, Madrid, 2007
4. Abelardo Monsalve y Pedro Harmath, *Introducción al análisis con series de tiempo con aplicaciones a la econometría y finanzas*, Universidad Centro occidental Lisandro Alvarado, Merida - Venezuela, 2015
5. Juan C. Abril, *Análisis De La Evolución De Las Técnicas De Series De Tiempo. Un Enfoque Unificado*, Instituto Interamericano de Estadística, Universidad Nacional de Tucumán, Argentina, 2011.
6. David G. Ramos y M. Isabel Galiano, *Análisis de Series Temporales*, Universidad Politécnica de Madrid, Madrid - España 2022.
7. Robert H. Shumway and David S. Stoffer, *Time Series Analisys and Its Applications With R Examples*, 4ed, Davis CA, Pittsburgh, 2010.
8. Anish Agarwal, Abdullah Alomar, and Devavrat Shah, *tspDB: Time Series Predict DB*, Massachusetts Institute of Technology, Cambridge, MA, USA, 2021.
9. Edmundo M. Sánchez, Lágrima de María Montiel, Gerardo A. Rosas, *Medias Móviles, Como Alternativas De Inversión En Un Índice Accionario*, Benemérita Universidad Autónoma de Puebla, México, 2019.
10. Gonzalo Rios y Carlos Hurtado, *Series De Tiempo*, Universidad De Chile, Noviembre, 2008
11. RUEY. S TSAY, *Analysis Of Financial Of Series*, 2ed, Wiley Interscience, University Of Chicago, 2005
12. Peter J. Brockwell Richard A. Davis, *Introduction to Time Series and Forecasting* Vol. 2, Editorial Board 1996
13. Laura A. Rodriguez y Silvia M.Ojeda, *Medidas De Disimilitud En Series Temporales*, Universidad Nacional de Córdoba, 2017
14. Pável Vidal Alejandro, Eduardo Hernández Roque, Carlos Pérez Soto, Mercedes García Armenteros, Guillermo Gil Gómez y Maria de los Ángeles Llorente *Estacionariedad y Estacionalidad En Series De La Economía Cubana*, Banco Central de Cuba, 2002.
15. Ramón, M. Dolors García. *Valor actual del modelo de Von Thünen y dos comprobaciones empíricas.* Revista de geografia 1976.
16. Cheung, Yin-Wong, and Kon S. Lai. *Practitioners corner: Lag Order and Critical Values of a Modified Dickey-Fuller Test.* Oxford Bulletin of Economics and Statistics 57.3 1995.
17. Yaglom, A. M. (1955).*The correlation theory of processes whose n-th difference constitute a stationary process*, 1995
18. Stock, James H. *Unit roots, structural breaks and trends.* Handbook of econometrics 4 1994.

FUNDAMENTOS DE LLMS Y GENERACIÓN AUMENTADA: APLICACIÓN PRÁCTICA A INFORMACIÓN UNIVERSITARIA

FABRICIO MURILLO

Resumen. Los recientes avances en el Procesamiento del Lenguaje Natural (NLP: "Natural Language Processing") han impulsado el desarrollo de Modelos de Lenguaje de Gran Escala (LLMs: "Large Language Models") que muestran resultados notables en diversas tareas. Esta investigación tiene como objetivo explorar los fundamentos de los LLMs y examinar la técnica de Generación Aumentada con Recuperador (RAG: "Retriever-Augmented Generation") como un método para mejorar la precisión y relevancia de las respuestas generadas. Se presenta un estudio teórico de estas tecnologías, seguido de un caso de estudio utilizando fuentes de información de la Universidad Nacional Autónoma de Honduras (UNAH) para desarrollar un chatbot informativo. Este enfoque busca comprender cómo los LLMs y RAG pueden aplicarse para abordar desafíos específicos en el contexto universitario, mejorando el acceso y la distribución de información.

Abstract. Recent advances in Natural Language Processing (NLP) have driven the development of Large Language Models (LLMs) that demonstrate remarkable results across various tasks. This research aims to explore the fundamentals of LLMs and examine the technique of Retriever-Augmented Generation (RAG) as a method to enhance the accuracy and relevance of generated responses. A theoretical study of these technologies is presented, followed by a case of study using information sources from the National Autonomous University of Honduras (UNAH) to develop an informative chatbot. This approach seeks to understand how LLMs and RAG can be applied to address specific challenges in the university context, improving access to and distribution of information.

1. Introducción

En la Universidad Nacional Autónoma de Honduras (UNAH), la accesibilidad a información precisa y estructurada representa un desafío significativo para estudiantes, profesores, personal administrativo y la comunidad en general. La dispersión y desorganización de los datos académicos y administrativos conllevan pérdidas de tiempo sustanciales y limitan la eficiencia en las actividades universitarias y la interacción con la comunidad. Estos problemas afectan la toma de decisiones informadas y eficaces, ya que la información a menudo no es clara ni accesible.

Recientes avances en el campo del Procesamiento del Lenguaje Natural (NLP: "Natural Language Processing") han permitido el desarrollo de modelos de lenguaje de gran escala (LLMs: "Large Language Models"), los cuales han mostrado resultados impresionantes en diversas aplicaciones [1]. Modelos como GPT-3 [2],

Fecha: 5 de Junio del 2024 y en revisión desde el 10 de junio del 2024.
Palabras y frases clave. NLP, LLMs, RAG, chatbot.

entrenados con vastas cantidades de datos, tienen el potencial de transformar significativamente la manera en que se consulta la información. Además, la técnica RAG ha emergido como una técnica prometedora para mejorar la precisión y relevancia de las respuestas generadas por los LLMs.

El objetivo principal de este estudio es abordar los fundamentos de los LLMs y explorar la técnica RAG como un medio para mejorar la calidad y contextualización de las respuestas generadas. Se busca comprender cómo estas tecnologías pueden aplicarse para afrontar los desafíos de acceso a la información en entornos universitarios. Para ello, se realizará un estudio teórico de los LLMs y RAG, seguido de un caso práctico en la UNAH, donde se desarrollará un prototipo de chatbot informativo basado en estas tecnologías.

Implementar un chatbot basado en tecnologías avanzadas de LLMs y RAG facilita el acceso a datos actualizados y relevantes por parte de estudiantes y personal universitario, optimizando al mismo tiempo los procesos administrativos internos y reduciendo la carga de trabajo del personal de atención al cliente. Trabajar en el desarrollo de este chatbot responde a la necesidad urgente de sistemas que impulsen la rápida y eficaz distribución de información dentro de la UNAH. Estas tecnologías tienen el potencial de resolver problemas específicos, como la falta de claridad y desactualización en la información académica y administrativa, permitiendo la toma de decisiones más informadas y eficaces. Además, al ser pioneros en la adaptación de LLMs y RAG en el contexto educativo hondureño, este estudio proporciona un marco replicable que puede ser adoptado por otras instituciones educativas que enfrentan problemas similares.

La necesidad de mejorar la accesibilidad a la información en la UNAH se enmarca en la respuesta a las prioridades nacionales relacionadas con la productividad, e innovación y tecnología. Al desarrollar un chatbot basado en LLM, esta investigación no solo posiciona a la UNAH como líder en la adopción de tecnologías avanzadas para promover la excelencia académica y administrativa, sino que también fortalece su capacidad para competir en un mundo globalizado. De esta manera, la investigación se alinea específicamente con las líneas prioritarias de investigación establecidas por la UNAH, satisfaciendo las demandas crecientes de un entorno globalizado y tecnológicamente avanzado.

2. Antecedentes

Los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés) representan un avance significativo en la evolución del Procesamiento del Lenguaje Natural (NLP). Su desarrollo se basa en décadas de investigación en lingüística computacional y aprendizaje automático, culminando en sistemas capaces de generar texto coherente.

El camino hacia los LLMs comenzó con modelos basados en reglas y gramáticas formales. Noam Chomsky introdujo la idea de gramáticas en la década de 1950, sentando las bases para el procesamiento computacional del lenguaje [3]. Sin embargo, estos modelos carecían de la flexibilidad necesaria para capturar la complejidad del lenguaje natural.

En los años 80 y 90, los modelos estadísticos, como los n-gramas, ganaron popularidad por su capacidad para predecir palabras basándose en contextos limitados [4]. Aunque simples, estos modelos demostraron ser sorprendentemente efectivos en muchas tareas de NLP.

El verdadero salto cualitativo llegó con la introducción de las redes neuronales para el modelado del lenguaje. En 2003, Bengio et al. propusieron el primer modelo de lenguaje neuronal, superando significativamente a los modelos n-gram [5]. Este trabajo sentó las bases para los futuros desarrollos en el campo.

La era moderna de los LLMs comenzó con la introducción de las arquitecturas de "transformers" por Vaswani et al. en 2017 [1]. Esta arquitectura, basada en el mecanismo de atención, permitió entrenar modelos mucho más grandes y eficientes, capaces de capturar dependencias a largo plazo en el texto.

El primer LLM que captó la atención general fue GPT ("Generative Pre-trained Transformer"), introducido por OpenAI en 2018 [7]. GPT demostró que el pre-entrenamiento a gran escala en datos de texto no etiquetados podía producir modelos con capacidades sorprendentes en una variedad de tareas de NLP.

BERT ("Bidirectional Encoder Representations from Transformers"), presentado por Google en 2018, introdujo el concepto de pre-entrenamiento bidireccional, mejorando significativamente el rendimiento en tareas de comprensión del lenguaje [9].

En 2020, OpenAI lanzó GPT-3, un modelo con 175 mil millones de parámetros que demostró capacidades impresionantes y dando pie al futuro producto conocido como ChatGPT de OpenAI [2]. Más recientemente, modelos como GPT-4 y Llama han elevado aún más el estándar, demostrando un rendimiento excepcional en numerosas tareas y dominios, acercándose a la experiencia humana en campos como la codificación, la medicina y el derecho [11, 12].

A medida que los LLMs se hicieron más sofisticados, también se volvieron evidentes sus limitaciones, particularmente en cuanto a la precisión factual y la actualización de conocimientos. Estas deficiencias impulsaron el desarrollo de técnicas para enriquecer los LLMs con información externa, siendo el "fine-tuning" o ajuste fino una de las más prometedoras.

El "fine-tuning" se puede conceptualizar como un proceso de "educación continua" para los LLMs. Así como un profesional actualiza sus conocimientos en áreas específicas, el "fine-tuning" permite a un modelo pre-entrenado especializarse en dominios o tareas particulares. Este proceso implica exponer al modelo a un conjunto de datos más pequeño pero altamente relevante para la tarea objetivo, permitiéndole ajustar sus parámetros internos de manera más precisa [8].

La eficacia del "fine-tuning" radica en su capacidad para aprovechar el conocimiento general adquirido durante el pre-entrenamiento, adaptándolo a contextos específicos, mejorando el rendimiento en tareas particulares. Su ventaja radica en el potencial para mantener los modelos actualizados, en campos donde el conocimiento evoluciona rápidamente, como la medicina o la tecnología, el "fine-tuning" permite incorporar nueva información sin necesidad de reentrenar completamente el modelo.

No obstante, "fine-tunning" no está exento de desafíos, uno de ellos es el sobreajuste. Este fenómeno puede llevar al modelo a memorizar los datos de entrenamiento en lugar de aprender patrones generalizables, limitando su eficacia en nuevos escenarios [15]. Asimismo, surge el reto del "olvido catastrófico", donde el LLM, al adaptarse a una nueva tarea, puede perder parte de su conocimiento general previamente adquirido [10]. Este desafío obliga a encontrar un equilibrio entre la adaptación a nuevas tareas y la retención del conocimiento base.

A pesar de sus beneficios, el "fine-tuning" presenta limitaciones en cuanto a la actualización continua y robusta de su conocimiento. La necesidad de ajustar el LLM para incorporar nueva información puede ser muy costosa en términos de tiempo y en recursos de computo. Además, no resuelve completamente el problema de la precisión factual, ya que el modelo sigue dependiendo principalmente de la información con la que fue entrenado originalmente.

Estas limitaciones motivaron a los investigadores a buscar enfoques diferentes para abordar de manera más efectiva los desafíos de actualización de conocimientos. En este contexto, surgió una técnica innovadora que prometía superar algunas de las limitaciones inherentes al "fine-tuning": la Generación Aumentada con Recuperación (RAG).

En 2020, Lewis et al. [13] introdujeron RAG como una solución innovadora. Esta técnica introduce un modelo para recuperar información junto a un LLM, permitiendo al modelo acceder a información externa durante la generación de texto. Esta técnica aborda varios problemas clave de los LLMs tradicionales, incluyendo la actualización de conocimientos, la trazabilidad y la mejora de la precisión factual.

La idea fundamental detrás de RAG es permitir que el modelo "consulte" fuentes externas de información en tiempo real, similar a cómo un humano podría buscar información adicional al abordar una tarea. Esto proporciona al modelo la capacidad de acceder a conocimientos actualizados sin necesidad de reentrenamiento constante, ofreciendo así una solución más flexible y adaptable a la actualización de conocimientos.

Un estudio de enero del 2024 de Microsoft ha explorado en profundidad las diferencias entre RAG y "fine-tuning", proponiendo un proceso completo para ambas técnicas y evaluando sus ventajas y desventajas en múltiples LLMs populares, incluyendo Llama2-13B, GPT-3.5 y GPT-4 [16].

El estudio reveló que tanto RAG como "fine-tuning" son efectivos para mejorar el rendimiento de los LLMs. RAG demostró ser particularmente útil cuando los datos son contextualmente relevantes, mientras que el "fine-tuning" fue eficaz para enseñar al modelo nuevas habilidades específicas del dominio. Notablemente, se observó un aumento acumulativo en la precisión cuando se combinaron ambas técnicas.

La aplicación de LLMs en dominios específicos representa un área de investigación emergente y prometedora. El estudio de Microsoft demostró cómo estas tecnologías pueden adaptarse para abordar desafíos específicos en el contexto agrícola, mejorando el acceso y la distribución de información [16].

Nuestra investigación se basa en este trabajo pionero, pero se enfoca específicamente en el uso de RAG para abordar el problema de acceso a la información en el contexto universitario de la UNAH. Debido a limitaciones de recursos, nuestro estudio no incluye "fine-tuning", centrándose en su lugar en la utilización de la técnica RAG para mejorar la precisión y relevancia de las respuestas generadas por el LLM.

Este enfoque tiene implicaciones significativas para el desarrollo de copilotos de inteligencia artificial (IA) en diversas industrias. Los copilotos de IA, impulsados por LLMs, están transformando la forma en que las organizaciones operan e interactúan con su entorno, proporcionando asistencia invaluable en el procesamiento de datos y la toma de decisiones en campos como la salud, la manufactura y las finanzas.

3. Fundamentos de los LLMs, RAG y un caso de estudio

3.1. Fundamentos de los LLMs. Los LLMs, representan uno de los avances más significativos en el campo del NLP en la última década. Estos modelos son sistemas de inteligencia artificial diseñados para comprender y generar lenguaje humano de una manera que se asemeja notablemente a cómo lo haríamos nosotros.

Un LLM es un sistema que ha sido entrenado con vastas cantidades de texto, aprendiendo patrones y estructuras del lenguaje a un nivel que antes parecía inalcanzable. Imagine una persona que ha leído millones de libros, artículos y conversaciones, y que puede utilizar todo ese conocimiento para entender y producir texto en una variedad de contextos. Así funcionan los LLMs, pero a una escala mucho mayor y con una velocidad de procesamiento incomparable con la capacidad humana.

El término "gran escala" en LLM se refiere tanto a la cantidad de datos con los que estos modelos son entrenados como al número de parámetros que contienen. Los parámetros son, en términos simples, los "conocimientos" del modelo, representados como valores numéricos que el modelo ajusta durante su entrenamiento. Modelos usados en Chatgpt, por ejemplo, contienen millones de parámetros, una escala que hace solo unos años parecía imposible de lograr [2].

La habilidad de los LLMs para entender y generar texto se basa en su capacidad para predecir la próxima palabra en una secuencia dada. Este proceso, aparentemente simple, es el resultado de un entrenamiento complejo que permite al modelo capturar las sutilezas del lenguaje, incluyendo contexto, gramática, e incluso ciertos aspectos del significado de las oraciones.

3.1.1. Arquitectura Transformer. Para entender mejor cómo los LLMs procesan el lenguaje, es importante examinar la arquitectura que hace posible su funcionamiento: el Transformer. Introducida por Vaswani et al. en 2017, la arquitectura Transformer revolucionó el campo del NLP al proporcionar un mecanismo eficiente para manejar secuencias largas de texto [1].

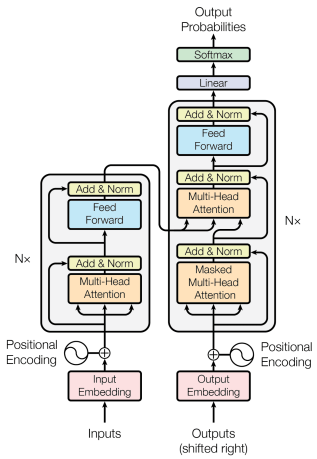


Figura 1. Arquitectura del Transformer, tomada de *Attention is All You Need* de Vaswani et al. [1].

La Figura 1 muestra una representación simplificada de la arquitectura Transformer. En este diagrama, podemos observar los dos componentes principales: el codificador ("encoder") y el decodificador ("decoder"). Imaginemos el Transformer como una máquina de traducción avanzada; el codificador sería la parte que "lee y entiende" el texto original, mientras que el decodificador sería la parte que "escribe" la traducción.

El codificador es responsable de procesar el texto de entrada y convertirlo en una representación que captura su significado y contexto. Está compuesto por varias capas, cada una con dos partes principales:

- Una capa de atención de múltiples cabezas: Esta capa permite al modelo enfocarse en diferentes partes de la entrada simultáneamente. Es como si el modelo pudiera leer una frase varias veces, cada vez prestando atención a diferentes palabras y sus relaciones.
- Una red neuronal de procesamiento: Esta capa toma la información de la capa de atención y la procesa aún más, permitiendo al modelo entender aspectos más complejos del texto.

El decodificador, por su parte, toma la información procesada por el codificador y la utiliza para generar el texto de salida. También está compuesto por varias capas, incluyendo:

- Una capa de atención que solo mira las palabras ya generadas: Esto asegura que el modelo solo use información que ya ha "escrito" para decidir qué escribir a continuación, similar a cómo escribimos una frase palabra por palabra.
- Una capa de atención que se conecta con la salida del codificador: Esta capa permite al decodificador usar la información del texto original procesado por el codificador.
- Una red neuronal de procesamiento: Similar a la del codificador, esta capa ayuda a generar el texto final.

El corazón de la arquitectura Transformer, y lo que la hace tan poderosa, es el mecanismo de atención. Este mecanismo permite al modelo "enfocarse" en diferentes partes del texto cuando procesa cada palabra, similar a cómo los humanos prestamos más atención a ciertas palabras que a otras al leer una frase.

Para entender mejor cómo funciona la atención, consideremos un ejemplo práctico. Imaginemos la frase "El gato que está sobre la mesa es negro". Cuando el modelo procesa esta frase, el mecanismo de atención funcionaría de la siguiente manera:

1. Al procesar "El", el modelo no tiene mucho contexto, así que presta atención por igual a esta única palabra.
2. Para "gato", la atención se centra principalmente en "El" y "gato", estableciendo la relación entre estas palabras.
3. Al llegar a "que", el modelo presta más atención a "gato", entendiendo que "que" se refiere al gato.
4. Para "está", la atención se distribuye entre "gato" y "que", manteniendo la conexión con el sujeto de la oración.
5. En "sobre", el modelo atiende principalmente a "está" y "gato", comprendiendo la acción y quién la realiza.
6. Al procesar "la", la atención se centra en "sobre", anticipando que vendrá un objeto.

7. Para “mesa”, el foco está en “sobre” y “la”, completando la idea de la ubicación del gato.
8. Al llegar a “es”, el modelo vuelve a prestar más atención a “gato”, preparándose para una descripción del sujeto principal.
9. Finalmente, para “negro”, la atención se distribuye entre “gato” y “es”, entendiendo que esta palabra describe al gato.

Este proceso permite al modelo capturar la estructura de la oración y las relaciones entre las palabras. Por ejemplo, entiende que “negro” se refiere al “gato” y no a la “mesa”, a pesar de que “mesa” está más cerca en la secuencia.

Es importante destacar que este ejemplo es una simplificación. En realidad, los modelos Transformer utilizan múltiples “cabezas” de atención en paralelo, cada una capaz de capturar diferentes tipos de relaciones en el texto. Esto es como si varios lectores leyeran la misma frase, cada uno fijándose en aspectos diferentes, y luego combinaran sus observaciones. Esta capacidad permite a los Transformers “entender” matices complejos del lenguaje, como la gramática, el contexto y, hasta cierto punto, el significado.

La habilidad de los Transformers para procesar texto de esta manera tan sofisticada es lo que permite a los LLMs generar respuestas coherentes y contextualmente apropiadas. Pueden “entender” una pregunta y generar una respuesta que parece haber sido escrita por un humano.

Sin embargo, es crucial recordar que, a pesar de su impresionante rendimiento, estos modelos no “entienden” el texto en el sentido humano. Son herramientas estadísticas muy avanzadas que han aprendido patrones del lenguaje, pero carecen de una verdadera comprensión de lo que están procesando. Por ejemplo, un LLM podría generar una respuesta gramaticalmente perfecta y aparentemente lógica sobre gatos voladores, sin “saber” realmente que los gatos no pueden volar.

3.1.2. Funcionamiento Detallado de los LLMs. Para comprender cómo los LLMs procesan y generan texto, es necesario examinar en detalle los pasos clave de su funcionamiento. Esto implica varias etapas fundamentales: tokenización, “embeddings”, y la predicción de la siguiente palabra.

La tokenización es el proceso inicial y crucial en el funcionamiento de un LLM. Consiste en dividir el texto de entrada en unidades más pequeñas y manejables llamadas “tokens”. La tokenización es fundamental porque permite al modelo trabajar con unidades discretas de significado, en lugar de procesar el texto como una secuencia continua de caracteres. Un “token” puede representar una palabra completa, una parte de una palabra, o incluso un carácter individual, dependiendo del diseño específico del modelo y del idioma que se está procesando.

Para ilustrar este proceso, consideremos la frase “La estudiante universitaria lee un libro interesante”. Un proceso de tokenización simple podría dividir esta frase de la siguiente manera: [“La”, “estudiante”, “universitaria”, “lee”, “un”, “libro”, “interesante”]. No obstante, la tokenización puede ser más sofisticada, especialmente para palabras menos comunes o en idiomas con estructuras complejas. Por ejemplo, la palabra “universitaria” podría dividirse en “tokens” más pequeños como [“universi”, “taria”], permitiendo al modelo manejar variaciones de palabras que no ha visto exactamente durante su entrenamiento [18].

La importancia de la tokenización radica en su capacidad para mejorar la eficiencia del procesamiento del texto y permitir al modelo manejar un vocabulario extendido. Al dividir palabras en subunidades, el modelo puede entender y generar

palabras nuevas o poco comunes basándose en sus componentes familiares. Esto es particularmente útil en idiomas con palabras compuestas.

Una vez que el texto se ha dividido en "tokens", el siguiente paso crucial es convertirlos en representaciones numéricas que el modelo pueda procesar. Estas representaciones se denominan "embeddings". Un "embedding" es esencialmente un vector, es decir, una lista ordenada de números, que representa un "token" en un espacio multidimensional. La idea fundamental detrás de los "embeddings" es que "tokens" con significados o usos similares tendrán representaciones numéricas similares en este espacio multidimensional [18].

Para comprender mejor este concepto, imaginemos un espacio simplificado de solo tres dimensiones. Los "embeddings" podrían verse así:

- "estudiante": [0.35, 0.55, 0.75]
- "alumna": [0.3, 0.6, 0.7]
- "libro": [0.05, 0.9, 0.3]

En este ejemplo simplificado, podemos observar que los "embeddings" de "estudiante" y "alumna" son más similares entre sí que con el "embedding" de "libro". Esto refleja la similitud semántica entre "estudiante" y "alumna", que son conceptos estrechamente relacionados.

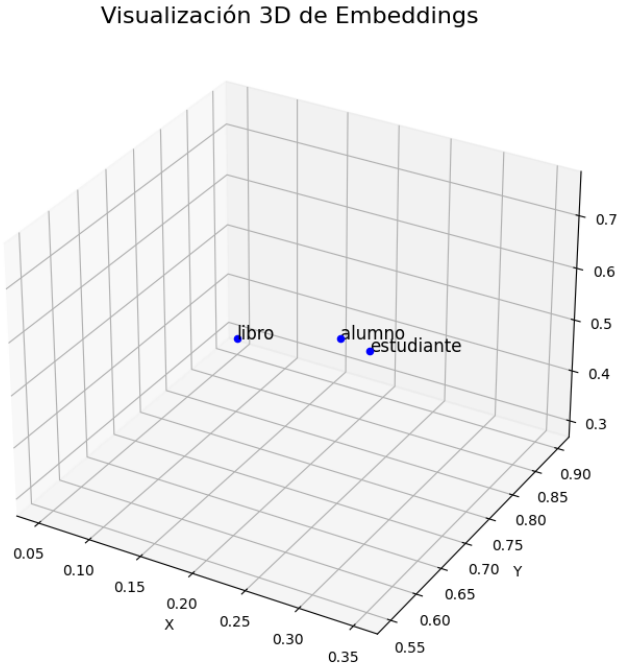


FIGURA 2. Visualización tridimensional de "embeddings", ilustrando cómo las palabras sémanticamente similares están más cercanas en el espacio.

En la Figura 2 el espacio es tridimensional, sin embargo, es importante destacar que en la práctica, estos "embeddings" tienen cientos o incluso miles de dimensiones,

lo que permite capturar relaciones muy sutiles y complejas entre palabras y conceptos. Esta representación multidimensional es lo que permite a los LLMs capturar y utilizar las complejas relaciones semánticas y sintácticas del lenguaje.

Con los "tokens" convertidos en "embeddings", el LLM puede iniciar el proceso central de su funcionamiento: la predicción de la siguiente palabra o "token" en una secuencia. Este proceso se puede visualizar como una serie de decisiones probabilísticas basadas en el contexto proporcionado. Para cada posición en la secuencia de texto, el modelo utiliza los "tokens" anteriores como contexto para predecir el siguiente "token" más probable.

Consideremos un ejemplo concreto. Supongamos que el modelo ha procesado la frase incompleta "La estudiante lee un". Para predecir la siguiente palabra, el modelo podría asignar probabilidades a diferentes opciones:

- "libro": 0.6
- "artículo": 0.2
- "texto": 0.1
- "periódico": 0.05
- otras palabras: 0.05

en este caso, el modelo seleccionaría "libro" como la palabra más probable para continuar la frase, basándose en los patrones que ha aprendido durante su entrenamiento.

Es crucial entender que este proceso no es determinista. Dependiendo de la configuración del modelo y los parámetros de generación, se puede introducir cierta aleatoriedad en las selecciones. Esta aleatoriedad controlada es lo que permite a los LLMs generar respuestas variadas y creativas, evitando la producción de texto repetitivo o predecible [19].

Cuando un LLM genera texto, repite este proceso de predicción múltiples veces. Cada palabra o "token" predicho se añade al contexto y se utiliza para predecir el siguiente elemento en la secuencia. Este proceso continúa hasta que se alcanza un criterio de parada predefinido, como un número máximo de palabras o un "token" especial que indica el fin de la secuencia.

Para ilustrar este proceso de generación iterativa, consideremos un ejemplo donde se le pide al modelo que complete la frase "El estudiante universitario está":

- Contexto: "El estudiante universitario está", predicción: "estudiando".
- Contexto: "El estudiante universitario está estudiando", predicción: "para".
- Contexto: "El estudiante universitario está estudiando para", predicción: "sus".
- Contexto: "El estudiante universitario está estudiando para sus", predicción: "exámenes".

Este proceso continúa hasta que el modelo genera una frase o párrafo completo que satisface los criterios de longitud o completitud establecidos. En cada paso de este proceso, el modelo está considerando todo el contexto anterior para tomar su decisión. Esto permite generar texto que mantiene coherencia y relevancia a lo largo de secuencias de texto.

La capacidad de los LLMs para mantener la coherencia en secuencias largas se debe en gran parte a la arquitectura Transformer y su mecanismo de atención. A diferencia de modelos previos que tenían dificultades para manejar dependencias a largo plazo en el texto, los Transformers pueden "atender" eficientemente a cualquier parte del contexto previo, sin importar cuán lejana esté.

Esta atención a largo plazo permite al modelo capturar y utilizar información relevante de manera más efectiva. Por ejemplo, si al principio de un texto se menciona un personaje o un tema específico, el modelo puede referirse a esa información incluso muchas oraciones después, manteniendo así la coherencia temática y narrativa [1]. Sin embargo, no poseen una comprensión real del significado detrás de las palabras que generan. Están, en esencia, realizando predicciones estadísticas basadas en patrones observados en los datos con que fueron entrenados. Esta limitación puede llevar a varios desafíos:

- Generación de información incorrecta: Los LLMs pueden producir afirmaciones que suenan plausibles pero son factualmente incorrectas, un fenómeno conocido como "alucinaciones".
- Sensibilidad al contexto inicial: Pequeños cambios en el "prompt" inicial pueden llevar a resultados muy diferentes, lo que puede hacer que el comportamiento del modelo sea a veces impredecible.

Estas limitaciones son áreas activas de investigación en el campo de la inteligencia artificial y el procesamiento del lenguaje natural. Una de estas técnicas prometedoras es RAG, la cual combina la capacidad de generación de texto de los LLMs con un sistema de recuperación de información, permitiendo al modelo acceder a fuentes externas de conocimiento durante el proceso de generación. Esto puede mejorar significativamente la precisión factual y la actualidad de la información proporcionada por el modelo.

3.2. RAG. Introducida por Lewis et al. en 2020 [13], RAG combina dos elementos fundamentales: LLMs y sistemas de recuperación de información. En términos simples, RAG permite a los sistemas "pensar" de una manera más parecida a cómo lo hacemos los humanos cuando buscamos información. Imagine a un estudiante universitario que necesita información sobre cómo cambiar de carrera de la universidad. Un LLM tradicional podría proporcionar información general basada en su entrenamiento, pero esta información podría estar desactualizada o no ser específica para su universidad. Aquí es donde RAG marca la diferencia. En lugar de depender únicamente del conocimiento con el que fue entrenado, un sistema RAG buscaría activamente en fuentes actualizadas, como el sitio web de la universidad o bases de datos administrativas recientes, para proporcionar información precisa y relevante sobre el proceso actual de cambio de carrera.

La importancia de RAG radica en su capacidad para superar limitaciones críticas de los LLMs tradicionales, permitiendo acceder a información actualizada en tiempo real, mejorando significativamente la precisión y relevancia de las respuestas [16].

En el contexto universitario, la aplicación de RAG podría revolucionar la forma en que los estudiantes acceden a la información. Ya sea para consultar sobre cambios en los planes de estudio, averiguar sobre nuevas oportunidades de becas, o entender los pasos para transferirse a otro campus, podría proporcionar respuestas precisas y actualizadas, adaptadas a las políticas específicas de cada institución.

3.2.1. Componentes de RAG. La arquitectura se compone de dos elementos principales: el Recuperador y el Generador. Estos componentes trabajan en conjunto para producir respuestas informadas y contextualizadas. El Recuperador actúa como un bibliotecario digital extremadamente eficiente. Su función es buscar y seleccionar información relevante de una base de conocimientos externa. En el contexto universitario, esta base de conocimientos podría incluir reglamentos académicos,

descripciones de cursos, anuncios recientes de la administración, y otros documentos relevantes.

Por ejemplo, si un estudiante pregunta sobre el proceso para cambiar de campus, el Recuperador buscaría en los documentos más recientes de la universidad relacionados con transferencias entre campus. Podría acceder a formularios de solicitud y quizá cuestionarios de preguntas frecuentes.

El Generador, por su parte, es similar a un asesor estudiantil experto. Toma la información proporcionada por el Recuperador y la utiliza para crear una respuesta coherente y apropiada. No solo debe entender la información recuperada, sino también integrarla de manera fluida con su conocimiento base para producir una respuesta que sea relevante, precisa y fácil de entender para el estudiante [21].

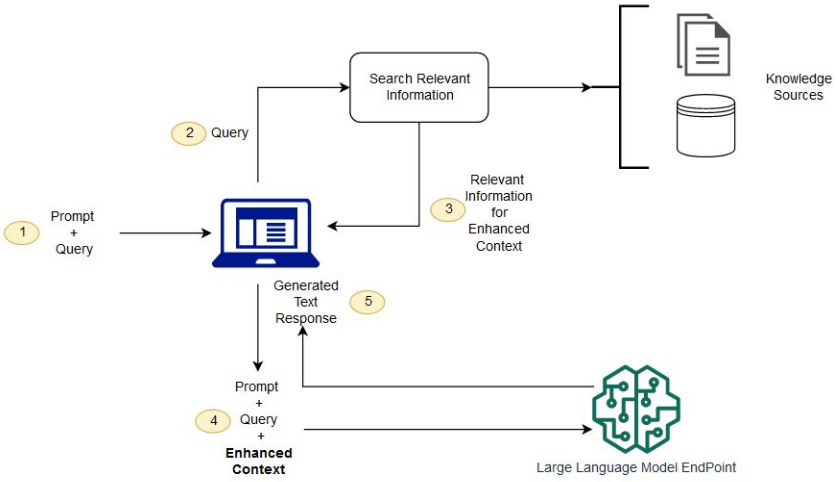


FIGURA 3. Arquitectura básica de RAG [22].

La Figura 3 ilustra cómo el sistema RAG trabaja para responder a consultas, mostrando el flujo de información desde las bases de datos hasta la respuesta final generada. El proceso de funcionamiento puede describirse como una serie de pasos interconectados y sofisticados. Inicialmente, el sistema recibe una consulta del usuario. El Recuperador analiza esta consulta para identificar los conceptos clave y realiza una búsqueda exhaustiva en la base de conocimientos. Esta búsqueda no se limita a una simple coincidencia de palabras clave, sino que emplea técnicas avanzadas de procesamiento del lenguaje natural para comprender el contexto y la intención detrás de la consulta [23]. Cada proceso se puede definir así:

- Recepción de la consulta: El sistema recibe la pregunta o solicitud del usuario.
- Análisis de la consulta: El Recuperador analiza la consulta para identificar los conceptos clave y la intención del usuario.
- Recuperación de información: Se realiza una búsqueda exhaustiva en la base de conocimientos externa para encontrar información relevante.
- Selección de información: De los resultados obtenidos, se seleccionan los textos más relevantes y correlacionados a la búsqueda.

- Contextualización: La información seleccionada se combina con la consulta original para crear un contexto enriquecido.
- Generación de respuesta: El Generador utiliza el contexto enriquecido para producir una respuesta coherente y relevante.
- Refinamiento y presentación: La respuesta generada se refina y se presenta al usuario en un formato apropiado.

3.2.2. Aspectos Técnicos de RAG. Desde una perspectiva técnica, RAG incorpora conceptos avanzados de álgebra lineal y aprendizaje automático. Un aspecto crucial es la representación vectorial de textos, que permite una comparación eficiente entre la consulta y los documentos en la base de conocimientos [23].

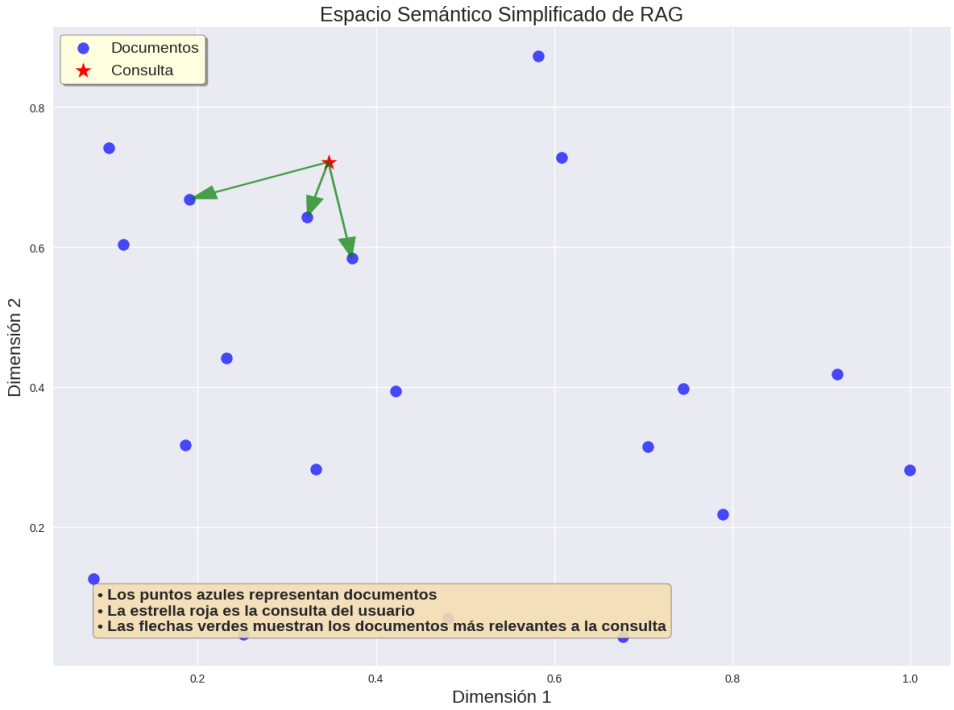


FIGURA 4. Representación vectorial en el espacio semántico de RAG.

La Figura 4 ilustra cómo las consultas y los documentos se representan como vectores en un espacio semántico. El espacio de la figura está simplificado a dos dimensiones, sin embargo, este puede ser de altas dimensiones. En este espacio, la similitud entre vectores se calcula típicamente utilizando la similitud del coseno:

$$\text{similitud}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|},$$

donde $\mathbf{a}$ y $\mathbf{b}$ son vectores que representan la consulta y un documento, respectivamente. Esta representación vectorial permite al sistema RAG realizar búsquedas eficientes en grandes bases de datos de documentos, identificando rápidamente los

más relevantes para una consulta dada. Sin embargo, existen diferentes métodos para determinar la similaridad [24].

Las ventajas del uso de RAG sobre los LLMs tradicionales son múltiples y significativas tal y como lo presentaron Balaguer et al. [16]. Estas incluyen una mayor precisión en las respuestas, mayor transparencia y verificabilidad, capacidad para manejar consultas complejas, y adaptabilidad a diferentes dominios de conocimiento sin necesidad de un reentrenamiento extensivo del modelo base. El principal desafío se encuentra en obtener, procesar, transformar y gestionar la data que servirá para nutrir la base de datos. Sin embargo, abren la puerta a implementar LLMs con RAG para responder a las necesidades de sectores donde la IA aún no ha llegado, cómo ser la agricultura, servicios de salud o de atención al cliente.

3.3. Caso de estudio: Construcción de un Chatbot Universitario. La Universidad Nacional Autónoma de Honduras (UNAH) enfrenta desafíos significativos en cuanto a la accesibilidad y distribución eficiente de información. Estudiantes, profesores y personal administrativo a menudo se encuentran con dificultades para acceder a datos precisos y actualizados sobre procesos académicos, administrativos y servicios universitarios. Esta situación no solo genera pérdidas de tiempo considerables, sino que también impacta negativamente en la toma de decisiones y en la eficacia general de las operaciones universitarias.

Para abordar esta problemática, proponemos el desarrollo de un chatbot informativo basado en RAG. Este sistema busca transformar la manera en que la comunidad universitaria accede y utiliza la información institucional, proporcionando respuestas precisas, contextualizadas y actualizadas a una amplia gama de consultas.

La UNAH, como muchas instituciones de educación superior, se enfrenta a una serie de desafíos en la gestión y distribución de información. Para abordar estos desafíos, proponemos el desarrollo de un chatbot informativo basado en RAG, diseñado específicamente para el contexto de la UNAH. Este sistema combinará la potencia de los LLMs con una base de conocimientos estructurada y actualizada de la universidad.

El chatbot propuesto funcionará de la siguiente manera:

- Recopilación de datos: Crear una base de conocimientos utilizando técnicas de "web scraping", extracción de datos de documentos PDF y otras fuentes oficiales de la UNAH. Esta base de datos incluirá información sobre programas académicos, procesos administrativos, servicios estudiantiles, calendarios, reglamentos y otros aspectos relevantes de la vida universitaria.
- Implementación de RAG: Se utilizará un modelo LLM de código abierto (como LLaMA o similares) como base para el sistema. La técnica RAG se implementará para permitir que el modelo acceda y utilice la base de conocimientos de la UNAH en tiempo real durante la generación de respuestas.
- Interfaz de usuario: Se desarrollará una interfaz de chat intuitiva y accesible.
- Procesamiento de consultas: Cuando un usuario realice una pregunta, el sistema RAG analizará la consulta, recuperará la información relevante de la base de conocimientos y generará una respuesta contextualizada y precisa.

La arquitectura propuesta para el chatbot informativo de la UNAH se compone de los siguientes elementos principales:

- Interfaz de usuario: Una capa que permite a los usuarios interactuar con el chatbot a través de una interfaz de chat.
- Módulo de NLP: Responsable de interpretar las consultas de los usuarios y prepararlas para el sistema RAG.
- Sistema RAG:
  1. Recuperador: Busca y selecciona la información más relevante de la base de conocimientos de la UNAH.
  2. Generador: Utiliza un LLM para generar respuestas coherentes y contextualizadas basadas en la información recuperada.
- Base de conocimientos: Una base de datos estructurada que contiene toda la información relevante de la UNAH, organizada de manera que facilite la recuperación eficiente.

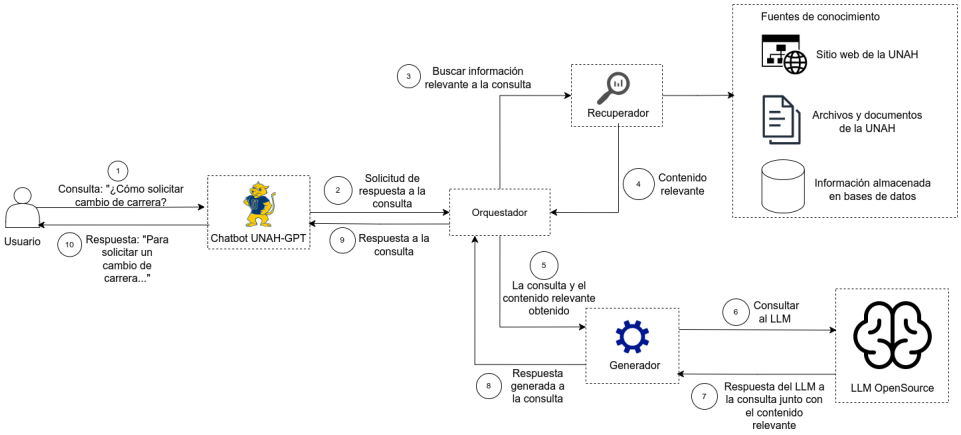


FIGURA 5. Arquitectura del chatbot informativo basado en RAG para la UNAH

La Figura 5 ilustra ilustra el concepto de arquitectura propuesto para el prototipo de chatbot informativo de la UNAH, mostrando los componentes básicos y su interacción.

Esta propuesta se plantea como un paso inicial en la exploración de cómo la tecnología RAG podría aplicarse en el contexto de la UNAH. Los objetivos principales incluyen:

- Demostrar la viabilidad de utilizar RAG para mejorar el acceso a la información universitaria.
- Identificar desafíos técnicos y operativos en la implementación de esta tecnología en un entorno universitario.
- Proporcionar una base para futuros desarrollos y estudios más detallados.

Es importante señalar que este prototipo serviría como prueba de concepto y no se pretende que sea un sistema completo o listo para producción. La implementación y evaluación de este prototipo podrían formar la base de un trabajo de grado más extenso, permitiendo un análisis profundo de las implicaciones técnicas y prácticas de implementar un sistema RAG en la UNAH.

Esta propuesta de prototipo representa un primer paso hacia la exploración de soluciones innovadoras basadas en inteligencia artificial para los desafíos de gestión

de información en la UNAH. Al desarrollar y evaluar este prototipo, se espera generar conocimentos valiosos que puedan guiar futuras investigaciones y desarrollos en este campo, potencialmente abriendo el camino para una transformación significativa en cómo la universidad maneja y distribuye información a su comunidad.

4. Conclusiones

El estudio de los Modelos de LLMs y RAG revela un panorama transformador en el campo del NLP, con implicaciones significativas para diversos sectores, incluyendo la educación superior. A lo largo de esta investigación, hemos explorado los fundamentos de estas tecnologías y su potencial aplicación en el contexto universitario, específicamente en la Universidad Nacional Autónoma de Honduras (UNAH).

Los LLMs representan un avance revolucionario en la capacidad de las máquinas para comprender y generar lenguaje humano. La arquitectura Transformer, con su innovador mecanismo de atención, ha permitido el desarrollo de modelos cada vez más sofisticados y capaces. Sin embargo, también hemos identificado limitaciones importantes, como la dependencia de conocimientos estáticos y la posibilidad de generar información incorrecta o desactualizada.

Es en este contexto donde la técnica RAG emerge como una solución prometedora. Al combinar la potencia generativa de los LLMs con sistemas de recuperación de información en tiempo real, RAG ofrece una vía para superar muchas de las limitaciones inherentes a los LLMs tradicionales. Esta técnica no solo mejora la precisión y relevancia de las respuestas generadas, sino que también permite una actualización continua del conocimiento sin necesidad de reentrenar constantemente el modelo base.

La aplicación de RAG en el contexto universitario, como se propone en nuestro caso de estudio para la UNAH, ilustra el potencial transformador de esta tecnología. Un chatbot informativo basado en RAG podría revolucionar la forma en que los estudiantes, profesores y personal administrativo acceden a la información institucional. Este enfoque promete no solo mejorar la eficiencia en la distribución de información, sino también democratizar el acceso al conocimiento dentro de la comunidad universitaria.

Más allá del ámbito académico, la combinación de LLMs y RAG abre nuevas posibilidades para abordar desafíos de información en diversos sectores. Como se ha visto en estudios recientes aplicados a la agricultura, estas tecnologías tienen el potencial de llevar conocimientos especializados a comunidades que tradicionalmente han tenido acceso limitado a la información. En países en desarrollo como Honduras, donde el uso de aplicaciones de mensajería está ampliamente extendido, un sistema basado en RAG podría proporcionar acceso a información crucial a través de interfaces familiares y accesibles.

El estudio de los LLMs y RAG no solo nos proporciona herramientas poderosas para el procesamiento del lenguaje natural, sino que también nos invita a repensar fundamentalmente cómo gestionamos y accedemos a la información en la era digital. A medida que continuamos explorando y refinando estas tecnologías, es crucial mantener un enfoque equilibrado que aproveche sus beneficios mientras aborda de manera proactiva sus limitaciones y desafíos éticos.

El caso de estudio propuesto para la UNAH representa un paso inicial en la exploración de cómo estas tecnologías pueden aplicarse en un contexto universitario

específico. Su implementación y evaluación no solo beneficiarían directamente a la comunidad de la UNAH, sino que también proporcionarían valiosas lecciones para la adopción más amplia de estas tecnologías en el sector educativo y más allá.

Referencias

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention Is All You Need*, arXiv:1706.03762, 2023.
2. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language Models are Few-Shot Learners*, arXiv:2005.14165, 2020.
3. N. Chomsky, *Syntactic Structures*, Mouton, The Hague, 1957.
4. P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, *Class-based n-gram models of natural language*, Computational Linguistics, 18(4):467-479, 1992.
5. Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, *A Neural Probabilistic Language Model*, Journal of Machine Learning Research, 3:1137-1155, 2003.
6. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention Is All You Need*, Advances in Neural Information Processing Systems, 30:5998-6008, 2017.
7. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving Language Understanding by Generative Pre-Training*, mikecaptain.com, 2018.
8. C. Jeong, *Domain-specialized LLM: Financial fine-tuning and utilization method using Mistral 7B*, Journal of Intelligence and Information Systems, 30(1):93-120, 2024.
9. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805, 2018.
10. J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, *Overcoming catastrophic forgetting in neural networks*, Proceedings of the National Academy of Sciences, 114(13):3521-3526, 2017.
11. OpenAI, *GPT-4 Technical Report*, arXiv:2303.08774, 2023.
12. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, *LLaMA: Open and Efficient Foundation Language Models*, arXiv:2302.13971, 2023.
13. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv:2005.11401, 2020.
14. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, Journal of Machine Learning Research, 21(140):1-67, 2020.
15. J. Howard and S. Ruder, *Universal Language Model Fine-tuning for Text Classification*, arXiv:1801.06146, 2018.
16. A. Balaguer, V. Benara, R. Cunha, R. Estevão, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski, and R. Chandra, *RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture*, arXiv:2401.08406, 2024.
17. B. Silva, L. Nunes, V. Estevão, R. Aski, and R. Chandra, *GPT-4 as an agronomist assistant? answering agriculture exams using large language models.*, arXiv:2310.06225, 2023.
18. S. Wolfram, *What Is ChatGPT Doing ... and Why Does It Work?*, Wolfram Media Inc., 2023.
19. M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, *Is Temperature the Creativity Parameter of Large Language Models?*, arXiv:2405.00492, 2024.
20. K. Martineau, *What is retrieval-augmented generation?*, IBM Research, 2024. URL: `https://research.ibm.com/blog/retrieval-augmented-generation-RAG`.
21. R. Mandal, *Enhance Your Applications with Retrieval-Augmented Generation (RAG) Architecture*, Learn | Hevo, 2024. URL: `https://hevodata.com/learn/what-is-rag-architecture/`.

22. AWS, *¿Qué es RAG?: explicación de la IA de generación aumentada por recuperación*, Amazon Web Services, Inc., URL: `https://aws.amazon.com/es/what-is/retrieval-augmented-generation/`, (n.d.).
23. Pinecone, *Choosing an embedding model*, Pinecone, URL: `https://www.pinecone.io/learn/series/rag/embedding-models-rundown/`, (n.d.).
24. R. Schwaber-Cohen, *Vector similarity explained*, Pinecone, URL: `https://www.pinecone.io/learn/vector-similarity/`, (n.d.).

Escuela de Matemática y Ciencias de la Computación, Universidad Nacional Autónoma de Honduras, Tegucigalpa, Honduras
Email address: `fabriciomurillo24@gmail.com`

ESTIMACIÓN ROBUSTA DE PARÁMETROS DE UNA DISTRIBUCIÓN GAMMA BASADO EN LA TRANSFORMACIÓN INTEGRAL DE PROBABILIDAD

ULISES ARIEL OBANDO REYES

Resumen. En la estadística, la estimación robusta es una metodología utilizada para obtener estimaciones más fiables y menos sensibles a las variaciones de los datos. A menudo, en un conjunto de datos, se presentan valores atípicos que, bajo métodos estadísticos clásicos, tienen una gran capacidad para distorsionar las estimaciones; a diferencia de los métodos robustos que están diseñados para ser menos afectados por estos valores y proporcionar resultados más representativos. En este artículo, se construirá un estimador para los parámetros de una distribución Gamma mediante el método de transformación integral de probabilidad y se comparará con el método clásico de máxima verosimilitud.

Abstract. In statistics, robust estimation is a methodology used to obtain more reliable and less sensitive estimates in the presence of data variations. Often, datasets contain outliers that, under classical statistical methods, have a significant potential to distort the estimates. In contrast, robust methods are designed to be less affected by these outliers and to provide more representative results. In this article, an estimator for the parameters of a Gamma distribution will be constructed using the probability integral transformation method and compared with the classical maximum likelihood method.

1. Introducción

Los métodos estadísticos están basados en una serie de suposiciones ideales. Cuando analizamos situaciones del mundo real, no siempre se presentan datos que se adapten a distribuciones que nos faciliten el análisis, o pueden existir valores que se alejen mucho de las observaciones obtenidas en el mismo estudio. En tales situaciones, los métodos clásicos pueden producir resultados engañosos o erróneos, ya que las estimaciones se ven afectadas por valores atípicos. Aquí es donde entra la estimación robusta, empleando métodos que son resistentes a las desviaciones de las suposiciones del modelo. Estos métodos intentan minimizar el efecto de estos datos atípicos, proporcionando así estimaciones que reflejen mejor la tendencia central o los parámetros relacionados con el conjunto de datos [18].

El enfoque robusto para los modelos estadísticos y el análisis de datos busca derivar métodos que generen estimaciones de parámetros confiables, pruebas asociadas e intervalos de confianza, sin importar si los datos siguen una distribución dada o si los datos siguen una distribución aproximada. Una ventaja de los métodos robustos

Fecha: 12 de agosto de 2024.
Palabras y frases clave. Estimación robusta, Distribución Gamma, Transformación integral de probabilidad, Estadística.

es que se ajustan mejor a la mayoría de los datos. Si los datos no contienen valores atípicos, los resultados del método robusto son aproximadamente iguales a los del método clásico; mientras que, si los datos contienen una pequeña proporción de valores atípicos, el método robusto se ajusta bien tanto a los valores atípicos como a los no atípicos.

La estimación robusta es muy importante en estadística porque permite obtener resultados precisos y fiables incluso si en los datos contamos con valores atípicos o que no cumplen con las suposiciones del modelo tradicional. Dentro de esta área tenemos técnicas como la regresión robusta, que son fundamentales para minimizar el impacto de valores atípicos. Al ser una herramienta muy útil para realizar estimaciones, la estimación robusta tiene muchas áreas de aplicación, entre las cuales se pueden mencionar economía [1], finanzas, medicina [3], biología, meteorología [2], etc.

Por otra parte, el método de estimación robusta de transformación integral de probabilidad es una técnica estadística que se utiliza para evaluar y mejorar la robustez de los modelos de probabilidad. Se basa en la transformación de los datos originales utilizando la función de distribución acumulada del modelo teórico, con el fin de obtener mayor precisión y fiabilidad en las inferencias estadísticas [19,20].

La distribución Gamma puede ser aplicada en el estudio de tiempos de espera en la Teoría de Colas, en la meteorología para ajustar los datos de las precipitaciones, también para el análisis del tiempo que pasa hasta que ocurre un evento, para el ajuste de distribución de la renta y en el modelado de los datos de consumo de productos a granel [8]. Esta investigación puede ser utilizada para ampliar el conocimiento en las líneas de investigación que se ha planteado la Universidad Nacional Autónoma de Honduras, en el eje temático de la ciencia.

2. Antecedentes

Es importante conocer la función Gamma que da origen a la distribución Gamma. Según Fuss (1843), el surgimiento de la función Gamma se da de la interacción entre Leonhard Euler y Christian Goldbach. La función Gamma habría surgido de la intersección de dos desarrollos matemáticos. El primero era el problema de interpolación de los números factoriales, que se había hecho muy popular durante el siglo XVIII, mientras que el segundo era el cálculo integral [4]. Bernoulli y Euler enviaron cartas a Goldbach donde presentaban algunas expresiones para calcular factoriales de números racionales. Fue en 1738 que Euler publicó con más detalles en un artículo titulado "Sobre las progresiones trascendentes o aquellas cuyos términos generales no pueden ser dados algebraicamente", cuyo nombre original es “De progressionibus transcendentibus seu quarum termini generales algebraice dari nequeunt” [5]. Posteriormente, Euler usó los resultados de su trabajo de interpolación [6] y lo transformó para poder realizarlo mediante el cálculo de una integral. En ese tiempo, esta integral de Euler no era vista como una función, sino como una fórmula para obtener los valores de los factoriales de números no naturales, y fue Adrien-Marie Legendre quien estableció el valor de la integral como la definición de una función, en .^{Ex}ercices de Calcul Intégral", Vol. 1 (1811). Realizó diversos cambios de variables para llegar a la hoy conocida función Gamma. Se puede conocer más sobre la función Gamma en [7].

La distribución Gamma es una de las distribuciones continuas más conocidas, aunque no es muy utilizada. Es importante ya que muchas distribuciones utilizadas en la práctica son casos particulares de la distribución Gamma, como la distribución Exponencial y la distribución chi-cuadrado [8]. La distribución Gamma aparece en múltiples ocasiones como modelo para el cálculo de probabilidades y para el ajuste de datos reales. En la práctica, la distribución Gamma es el modelo de referencia para variables continuas y positivas, como pueden ser los flujos de agua, consumos de productos a granel, rentas, recogidas de residuos urbanos, y tantos otros. Algunos trabajos donde se aplica la distribución Gamma se pueden ver en [9,10,11]. Uno de los principales desarrolladores de esta distribución fue Karl Pearson, quien trabajó con distribuciones continuas para modelar una amplia gama de datos en diversos campos. Pearson desarrolló el Sistema de Distribuciones de Pearson, que comprende distintos tipos enumerados de I al VII, y en 1895 presentó este sistema [12], que posteriormente en 1901 y 1916 amplió en dos publicaciones [13,14].

El método de mínimos cuadrados era considerado el procedimiento para generar estimadores de parámetros en modelos estadísticos, pero se ha demostrado que este método no utiliza toda la información que puede proporcionar una muestra. Debido a esto, se ha optado por otros métodos como el de máxima verosimilitud, aunque en presencia de valores atípicos deja de ser tan preciso. Por esto, el desarrollo de métodos de estimación robusta se volvió algo muy importante durante el siglo XX y comienzos del siglo XXI. En 1935, Ronald Fisher concibió la idea de robustez en términos de sensibilidad de la prueba t, bajo el supuesto de normalidad. Pero fue hasta 1953 cuando Box introdujo el término robusto para denotar un procedimiento que es insensible a las desviaciones de los supuestos en los cuales está basada la prueba [15]. Peter Huber publicó un artículo en el año 1964 donde habla de cómo diseñar procedimientos robustos. También uno de sus grandes aportes es su libro Robust Statistic"[16]. Otros libros relevantes que se pueden consultar son [17,18].

El Método de Transformación Integral de Probabilidad ha sido desarrollado mayormente por Peter Huber y Frank Hampel. Han publicado artículos donde se introdujeron los conceptos y se sentaron las bases teóricas de este método, desarrollándose posteriormente en otras publicaciones. En [19,20] se puede encontrar información sobre cómo la Transformación Integral de Probabilidad puede ser aplicada para evaluar y mejorar la robustez de los estimadores.

La función Gamma, cuyos orígenes se remontan al trabajo de Euler y su posterior formalización por Legendre, proporciona la base matemática fundamental para la distribución Gamma. Esta distribución es crucial en diversos contextos prácticos debido a su capacidad para modelar variables continuas y positivas, como los flujos de agua y el consumo de productos. En el ámbito de la estimación de parámetros, se ha reconocido que los métodos tradicionales como el de mínimos cuadrados no son siempre suficientes, lo que ha llevado al desarrollo de métodos de estimación robusta, una línea de investigación que se ha fortalecido desde las contribuciones de Fisher y se ha formalizado con los trabajos de Huber y Hampel [16,18]. Es en este contexto que el Método de Transformación Integral de Probabilidad, desarrollado por Huber y Hampel, cobra relevancia, al ofrecer una forma de evaluar y mejorar la robustez de los estimadores. La presente investigación busca sintetizar estos conceptos y avances, aplicando la Transformación Integral de Probabilidad para obtener estimaciones robustas de los parámetros de una distribución Gamma, aportando

así al campo de la estadística robusta y mejorando la precisión y fiabilidad de los modelos estadísticos en aplicaciones prácticas.

3. Estimación robusta de parámetros

3.1. Fundamentos teóricos. Las siguientes definiciones y proposiciones se encuentran en [7,8].

Primeramente, tenemos la definición de la función Gamma.

Definición 3.1.1 Se define la función Gamma como $\Gamma : (0, \infty) \to \mathbb{R}$ donde

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

Proposición 3.1.1 Para cualquier $\alpha > 1$ tenemos que

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

Corolario 3.1.1 Para cualquier $n \in \mathbb{N}$ tenemos que

$$\Gamma(n) = (n - 1)!$$

La proposición 1.1 y el corolario 1.1 son propiedades de la función Gamma que nos ayudan a relacionar el valor de la función evaluada en un número α en términos de la función evaluada en $\alpha - 1$, además se establece el resultado de la función Gamma cuando se calcula para un número natural que simplifica el cálculo en esos casos particulares.

Definición 3.1.2 Diremos que X es una variable aleatoria que sigue una distribución Gamma con parámetros $\alpha > 0$ y $\beta > 0$, si su función de densidad es

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{si } x > 0 \\ 0 & \text{otro caso} \end{cases}$$

El parámetro α es conocido como el "parámetro de forma", mientras que β se conoce como el "parámetro de escala". En la figura 1, vemos las funciones de densidad de la distribución Gamma para $\beta = 1$ y distintos valores de α.

Proposición 3.1.2 Dada X una variable aleatoria tal que $X \sim \Gamma(\alpha, \beta)$ entonces $aX \sim \Gamma(\alpha, a\beta)$ siendo $a > 0$ un número real positivo.

Ahora, veamos la forma de la función de distribución o función de probabilidad acumulada.

Definición 3.1.3 Sea $X \sim \Gamma(\alpha, \beta)$, la función de distribución asociada a X es,

$$F_x(t) = P(X \leq t) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^t x^{\alpha-1} e^{-x/\beta} dx$$

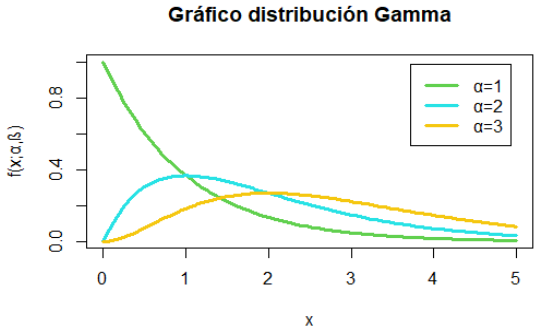


FIGURA 1. Función de densidad de la Gamma con $\beta = 1$

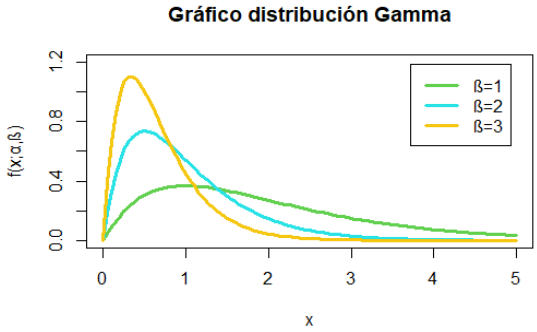


FIGURA 2. Función de densidad de la Gamma con $\alpha = 2$ y distintos valores de β

La esperanza y la varianza para una variable aleatoria que sigue una distribución Gamma está dada por

Proposición 3.1.3 Sea X una variable aleatoria tal que $X \sim \Gamma(\alpha, \beta)$ entonces $E(X) = \alpha\beta$.

Proposición 3.1.4 Sea X una variable aleatoria tal que $X \sim \Gamma(\alpha, \beta)$ entonces $Var(X) = \alpha\beta^2$.

Las proposiciones anteriores son resultados que se obtienen del cálculo de la esperanza y la varianza de una variable aleatoria que sigue una distribución Gamma, en [7] se muestra el procedimiento para obtener estos valores.

La función generatriz de momentos es una variable aleatoria X es $E[e^{tX}]$, entonces

Proposición 3.1.5 Sea X una variable aleatoria tal que $X \sim \Gamma(\alpha, \beta)$ entonces su función generatriz de momentos es,

$$M_x(t) = E[e^{tX}] = \frac{1}{(1-t\beta)^{\alpha}}$$

Proposición 3.1.6 Sean $X_1, X_2, ..., X_n$ variables aleatorias independientes tales que $X_i \sim \Gamma(\alpha_i, \beta)$ siendo $i = 1, 2, ..., n$. Entonces,

$$Y = \sum_{i=1}^{n} X_i \sim \Gamma\Big(\sum_{i=1}^{n} \alpha_i, \beta\Big)$$

Mediante la proposición 2.5 se encuentra la distribución de una variable aleatoria Y, que está compuesta por la suma de variables aleatorias que siguen una distribución Gamma, este resultado nos sirve para determinar los parámetros de Y.

Las siguientes tres distribuciones son casos particulares de la distribución Gamma, donde se plantean los parámetros que tienen según cada distribución, con esto podemos analizar otras distribuciones a través de la distribución Gamma.

Definición 3.1.4 (Distribución Exponencial) Sea X una variable aleatoria, decimos que sigue una distribución exponencial si su función de densidad es de la forma,

$$f_x(x) = \frac{1}{\beta} e^{-x/\beta} \qquad , x > 0$$

La distribución exponencial es un caso particular de una distribución Gamma con $\alpha = 1$, es decir, si $X \sim Exp(\beta)$ entonces $X \sim \Gamma(1, \beta)$ [8].

A continuación, veremos una propiedad interesante de la distribución exponencial.

Proposición 3.1.7 La suma de exponenciales independientes es una distribución Gamma.

Definición 3.1.5 (Distribución de Erlang) Sea X una variable aleatoria, decimos que sigue una distribución de Erlang si su función de densidad es de la forma

$$f_X(x) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & \text{otro caso} \end{cases}$$

siendo $\beta > 0$ y α un número natural.

La distribución de Erlang es otro caso particular de una distribución Gamma siendo el parámetro α un valor entero y positivo. Esta distribución representa el tiempo de espera hasta que ocurre el α-ésimo evento [8].

Definición 3.1.6 (Distribución chi-cuadrado) Una variable aleatoria X sigue una distribución χ^2 con n grados de libertad si su función de densidad es de la

siguiente forma

$$f_X(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} & x \geq 0 \\ 0 & \text{otro caso} \end{cases}$$

La distribución χ^2 también es un caso específico de la distribución Gamma tomando como $\alpha = n/2$ y $\beta = 2$. Con esos valores de los parámetros tenemos la distribución χ^2 con n grados de libertad, es decir, si $X \sim \Gamma(n/2, 2)$ entonces $X \sim \chi^2(n)$ [8].

Existe una relación entre la distribución de Poisson y la distribución Gamma,

Definición 3.1.7 Sea X una variable aleatoria, decimos que sigue una distribución de Poisson de parámetro $\lambda \in (0, +\infty)$ si su función de probabilidad es de la siguiente forma [8]

$$p(x, \lambda) = P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{con} \quad x \in \{0, 1, 2, 3, \ldots\}$$

y su función de distribución es

$$P[X \leq x] = \sum_{k=0}^{x} \frac{1}{k!} \left(\frac{1}{\beta}\right)^k e^{-\frac{1}{\beta}}$$

Ahora definiremos un proceso de Poisson,

Definición 3.1.8 Un proceso de Poisson es la aparición aleatoria de sucesos a lo largo de un tiempo cumpliendo:

- El número de sucesos que ocurren en intervalos de tiempo disjuntos son variables aleatorias independientes.
- El número medio de sucesos por unidad de tiempo, λ, se mantiene constante a lo largo del tiempo.
- En un intervalo de tiempo de longitud diferencial $[t, t + \Delta t]$ sólo se puede producir a lo sumo un suceso [8].

Proposición 3.1.8 El número de veces que ocurre un evento en un intervalo de tiempo $(0, t)$ sigue una distribución de Poisson de parámetro λt con $\lambda \in (0, +\infty)$ si, y solo si, el tiempo que transcurre hasta el n-ésimo evento sigue una distribución de Erlang de parámetros n y λ con $\lambda \in (0, +\infty)$ [8].

Como método clásico, se eligió el método de máxima verosimilitud para estimar los parámetros de la distribución Gamma.

La función de logverosimilitud [21], para $\Gamma(a, b)$ es

$$\log L(\alpha, \beta) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i,$$

y las ecuaciones de verosimilitud son

$$0 = \frac{\partial}{\partial \alpha} \log L(\alpha, \beta) = -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - n \log \beta + \sum_{i=1}^{n} \log x_i,$$

$$0 = \frac{\partial}{\partial \beta} \log L(\alpha, \beta) = \frac{1}{\beta}(-n\alpha + \frac{1}{\beta}\sum_{i=1}^{n} x_i),$$

resultando así

$$\beta = \frac{\bar{x}}{\alpha},$$

y para α el estimador es la solución de la siguiente ecuación

$$-n\phi(\alpha) - n\log\bar{x} + n\log\alpha + \sum_{i=1}^{n}\log x_i = 0,$$

donde

$$\phi(\alpha) = \frac{d}{d\alpha}\log\Gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \int_0^\infty (\frac{e^{-t}}{t} - \frac{e^{-\alpha t}}{1-e^{-t}})dt,$$

Este valor puede aproximarse con el uso de métodos numéricos [21].

3.2. Estimación Robusta. La estimación robusta es un conjunto de técnicas estadísticas diseñadas para generar estimaciones que son insensibles a pequeñas desviaciones de los supuestos del modelo, especialmente a la presencia de valores atípicos en los datos. En lugar de ser influenciados de manera significativa por estos valores atípicos, los estimadores robustos proporcionan resultados que reflejan mejor la estructura de los datos [18].

Un ejemplo para evidenciar la importancia de la estimación robusta es el siguiente donde se tiene un conjunto de datos que presenta un dato que se aleja demasiado del resto y este valor genera que la media esté muy distante de la mayoría de datos.

Ejemplo: [18] Considere el siguiente conjunto de datos

2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90
3.03	3.03	3.10	3.37	3.40	3.40	3.40	3.50
3.60	3.70	3.70	3.70	3.70	3.77	5.28	28.95

El valor 28.95 se aleja bastante del resto de datos, es por ello que será considerado como un valor atípico. Una conjetura sobre este valor es que a la hora de escribirlo se colocó mal el punto decimal y debió ser 2.895.

Es común que para evitar este tipo de problemas se elimine este valor o que sea reemplazado por otro valor que sea generado a partir del resto y que tenga más "sentido". Pero no siempre realizar cualquiera de estos procedimientos es correcto, ya que se podría estar eliminando o cambiando un valor que nos esté aportando una información distinta e importante sobre los datos. Aquí es donde entra la estimación robusta, pues busca que las estimaciones sean relativamente buenas sin importar si el conjunto de datos tiene un porcentaje de valores atípicos o que no los tenga [18].

Entonces, se calcula la media de los datos con y sin el valor atípico y los resultados son $\bar{x} = 4.28$ y $\bar{x} = 3.21$ respectivamente. Ahora bien, se propone otra forma de encontrar un valor que este cerca de "la mitad"de los datos, esta forma es el cálculo de la mediana, que es un estimador robusto. Con la media se obtiene que $\hat{x} = 3.38$ con el valor atípico y si se reemplaza por un valor más cercano al resto puede bajar hasta $\hat{x} = 3.23$.

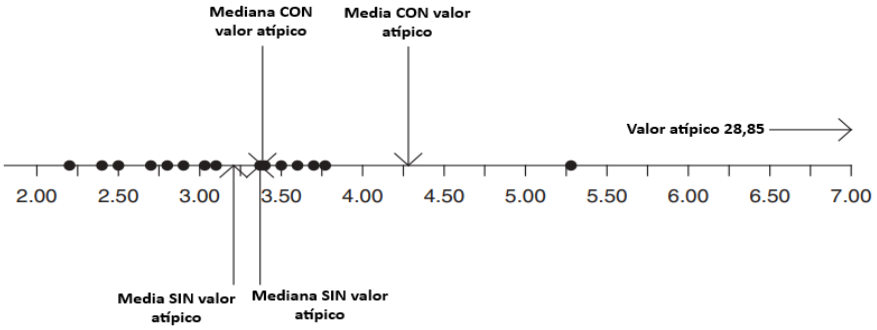


FIGURA 3. Valores del ejemplo graficados sobre la recta real donde se ubican los distintos valores de medias y medianas en ambos casos [18].

3.3. Transformación integral de probabilidad. Como método robusto se estimarán los parámetros de una distribución Gamma mediante el estimador M basado en la transformación integral de probabilidad, primeramente se presenta el método de manera general para cualquier distribución.

Sea $F_\theta, \theta \in \Theta$ con Θ un conjunto abierto en $\mathbb{R}^p$, una familia de funciones de distribución discretas o continuas y $p_\theta(k) = P_\theta(X = K)$.

Supongamos que estamos interesados en estimar θ, entonces la transformación integral de probabilidad $T(X)$ está definida por $T(X) = F_\theta(X) - Vp_\theta(X)$, donde V es una variable aleatoria que se distribuye uniforme $[0, 1]$ ($V \sim U[0, 1]$) que es independiente a X. Veamos que, cuando X es continua, esta transformación se reduce a $T(X) = F_\theta(X)$.

El siguiente teorema establece la propiedad muy conocida de que $T(X)$ tiene una distribución estándar uniforme, a partir de este se obtiene el estimador M para los parámetros de una distribución [22].

Teorema 3.2.1 Si X es una variable aleatoria continua o discreta y T es la transformación integral de probabilidad, entonces $T(X) \sim U[0, 1]$.

Conociendo la distribución de esta transformación de X, podemos calcular los momentos: $E_\theta((F_\theta(X) - Vp_\theta(X)^k) = 1/(k + 1)$, donde $E_\theta(f(x))$ es la esperanza de $f(x)$ cuando $X \sim F_\theta$.

Definamos

$$\Psi^{'}(x, \theta) = (\psi_1^{'}(x, \theta), ..., \psi_p^{'}(x, \theta)),$$

donde

$$\psi_k^{'}(x, \theta) = E_\theta((F_\theta(X) - Vp_\theta(X)^k | X = x) - 1/(k + 1),$$

cuando X es continua, la definición de $\Psi_k^{'}$ es:

$$\Psi_k^{'}(x,\theta) = F_\theta^k(X) - 1/(k+1),$$

con esto, se define el estimador M, $\hat{\theta_n}$ de θ como la solución de

$$\sum_{i=1}^{n} \Psi_k^{'}(X_i,\theta) = 0.$$

Este trabajo consiste en estimar los parámetros de una distribución Gamma mediante el método de transformación integral de probabilidad, para ello necesitamos encontrar el estimador M definido anteriormente cuando la variable aleatoria X sigue una distribución Gamma, entonces sustituyendo en $\Psi_k^{'}$ tenemos

$$\Psi^{'}(x,\alpha,\beta) = (F_x(x) - 1/2, F_x^2(x) - 1/3),$$

el estimador M es la solución de

$$\sum_{i=1}^{n} \Psi^{'}(x_i,\alpha,\beta) = 0,$$

sustituyendo obtenemos el siguiente sistema

$$\begin{cases} \sum_{i=1}^{n}(F_x(x_i) - 1/2) = 0 \\ \sum_{i=1}^{n}(F_x^2(x_i) - 1/3) = 0 \end{cases}$$

donde F_x es la función de distribución de la variable aleatoria X.

Se resuelve el sistema para las variables α y β,

$$\begin{cases} \sum_{i=1}^{n}((\frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^{x_i} x^{\alpha-1}e^{-x/\beta}dx) - 1/2) = 0 \\ \sum_{i=1}^{n}((\frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^{x_i} x^{\alpha-1}e^{-x/\beta}dx)^2 - 1/3) = 0 \end{cases}$$

Estos valores pueden aproximarse con el uso de métodos numéricos.

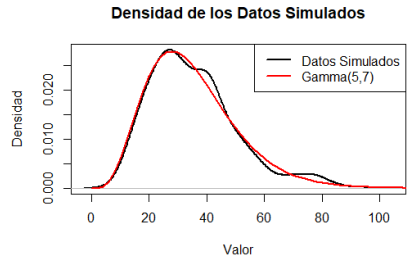
3.4. Comparación de métodos. En esta sección compararemos un método clásico; el método de máxima verosimilitud y un método robusto; basado en la transformación integral de probabilidad.

Primeramente, se extrae una muestra de tamaño 500 que siga una distribución Gamma con parámetros $\alpha = 5$, $\beta = 7$. Con esto probaremos los métodos sin la presencia de valores atípicos. En la figura 4, se muestra la densidad y un histograma de los datos de la muestra extraída.

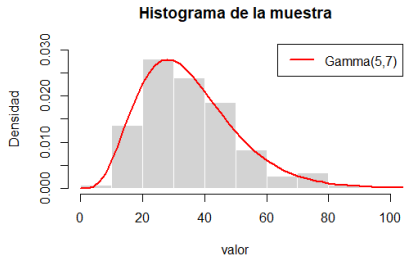
Seguidamente se aplica el método de máxima verosimilitud para estimar los parámetros de la muestra, se obtiene como resultado $\bar{\alpha} = 5.6185$ y $\bar{\beta} = 6.2965$.

Al aplicar el método robusto se obtiene $\bar{\alpha} = 5.5614$ y $\bar{\beta} = 6.3227$.

En la figura 5, se muestra el histograma de la muestra extraída junto a la función de densidad de la Gamma con párametros $\alpha = 5$ y $\beta = 7$ (Gamma(5,7)), la función densidad de la Gamma donde los párametros son estimados por el método de máxima verosimilitud (Gamma MLE) y la función de densidad de la Gamma donde los parámetros son estimados por el método robusto (Gamma Robusta).



(a) Densidad de los datos de la muestra junto a la densidad de una Gamma(5,7)



(b) Histograma de la muestra junto a la densidad de una Gamma(5,7)

FIGURA 4. Gráficos de la muestra

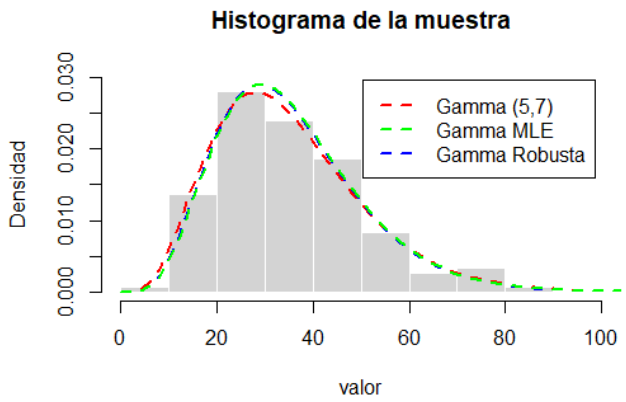


FIGURA 5. Histograma de la muestra junto a la densidad de una Gamma(5,7), la densidad de la Gamma MLE y la Gamma Robusta.

A modo de resumen se muestra el siguiente cuadro, se incluye el error cuadrático medio para cada método.

	Máxima Verosimilitud	**Método Robusto**
α	5.6185	5.5614
β	6.2965	6.3227
ECM	0.8278	0.6726

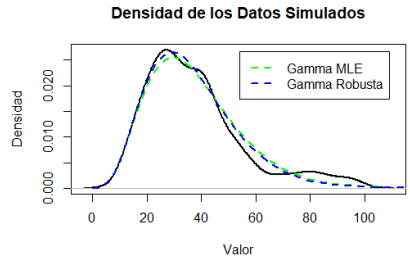
CUADRO 1. Comparación de parámetros estimados y error cuadrático medio entre el método de máxima verosimilitud y el método robusto.

Como se menciona en este trabajo, los métodos clásicos son sensibles a los valores atípicos, a diferencia de los métodos robustos que nos generan mejores estimaciones

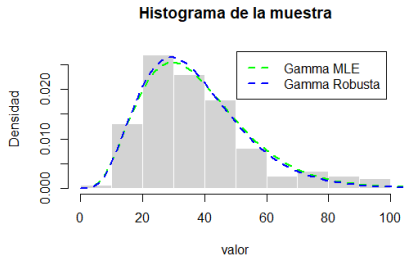
con presencia de valores atípicos, a continuación, se insertarán valores atípicos en la muestra obtenida anteriormente y se aplicarán ambos métodos nuevamente.

En el método de máxima verosimilitud se obtiene $\bar{\alpha} = 4.7187$ y $\bar{\beta} = 7.9410$.

Mientras que con el método robusto se obtiene $\bar{\alpha} = 4.9250$ y $\bar{\beta} = 7.4292$.



(a) Densidad de los datos de la muestra contaminada junto a la densidad de la Gamma MLE y Gamma Robusta



(b) Histograma de la muestra junto a la densidad de la Gamma MLE y Gamma Robusta

FIGURA 6. Gráficos de la muestra contaminada

En la figura 6, se muestra la densidad y el histograma de la muestra contaminada junto a la función densidad de la Gamma donde los párametros son estimados por el método de máxima verosimilitud (Gamma MLE) y la función de densidad de la Gamma donde los parámetros son estimados por el método robusto (Gamma Robusta). El siguiente cuadro muestra los valores de α, β y el error cuadrático medio para cada método.

	Máxima Verosimilitud	**Método Robusto**
α	4.7186	4.9250
β	7.9409	7.4292
ECM	208.2549	3.8954

CUADRO 2. Comparación de parámetros estimados y error cuadrático medio entre el método de máxima verosimilitud y el método robusto cuando la muestra contiene valores atípicos.

Se puede apreciar que sin presencia de valores atípicos ambos métodos generan buenas estimaciones, pero cuando hay valores atípicos el método robusto funciona mucho mejor que el método clásico.

4. CONCLUSIONES

A continuación, se resume el trabajo realizado y los hallazgos observados durante la experimentación:

1. Los métodos robustos son diseñados para ser insensibles a la presencia de valores atípicos en los datos. A diferencia de los métodos tradicionales, como el de máxima verosimilitud, que pueden ser significativamente afectados por

estos valores extremos, los métodos robustos proporcionan estimaciones más fiables y representativas de la verdadera estructura de los datos.

2. En lugar de eliminar o ajustar los valores atípicos, lo que podría resultar en la pérdida de información importante, los métodos robustos permiten mantener todos los datos en el análisis. Esto asegura que ninguna información potencialmente valiosa se descarte prematuramente, lo que es crucial en estudios donde cada observación puede aportar al entendimiento del fenómeno estudiado.

3. Se utilizó una muestra de tamaño 500 de una distribución Gamma con parámetros $\alpha = 5$ y $\beta = 7$ sin valores atípicos para probar el método de máxima verosimilitud y el método robusto basando en la transformación integral de probabilidad, se busca que las estimaciones de estos métodos sean precisas, por lo que valores cercanos a 5 y 7 para α y β ,respectivamente, nos indican que los métodos de estimación son adecuados para los datos seleccionados. Como resultados en el método de máxima verosimilitud se obtuvo $\bar{\alpha} = 5.6185$ y $\bar{\beta} = 6.2965$, y para el método robusto se obtuvo $\bar{\alpha} = 5.5614$ y $\bar{\beta} = 6.3227$, como podemos ver el método robusto genera valores mas cercanos a 5 y a 7, que es un buen indicador de que el método es mas eficiente.

4. El error cuadrático medio (ECM) es una medida que nos ayuda a cuantificar la diferencia promedio al cuadrado entre los valores predichos por el modelo y los valores observados o reales, por lo que nos conviene obtener un ECM bajo para garantizar que los datos estimados estén bastante cerca de los valores reales, como podemos ver, el ECM para ambos métodos cuando no hay valores atípicos es bastante bajo con valores de 0.8278 y 0.6726, con esto podemos concluir que ambos métodos son muy precisos en la estimación.

5. Como se esperaba, cuando hay presencia de valores atípicos el método robusto es más preciso que el método de máxima verosimilitud, en cuanto a los valores estimados de los parámetros vemos que son bastante cercanos a 5 y 7, recordando que la muestra fue contaminada con valores atípicos y no se espera que la nueva muestra siga la misma distribución que la muestra anterior, por lo que una forma correcta de comparar estos métodos es a través del ECM, donde se logra una diferencia notoría con valores de 208.2549 para el método de máxima verosimilitud y 3.8954 para el método robusto.

Como trabajo futuro se puede aplicar este método robusto a datos reales en áreas donde se utiliza la distribución Gamma como en la meteorología, finanzas, ingeniería, salud, etc.

Referencias

1. Agostini, C. A., Hojman, D., Román, A. & Valenzuela, L. (2016). Segregación residencial de ingresos en el Gran Santiago, 1992-2002: una estimación robusta. Eure (Santiago), 42(127), 159-184.
2. Rodríguez Gómez, E. A. (2016). Control no-lineal de un generador eólico usando un sistema de estimación robusta del viento.
3. Julián, A., Pablo Kontaxis, S., & Gil Herrando, E. Estimación robusta de la diferencia del tiempo de tránsito del pulso sanguíneo a partir de señales fotopletismográficas.
4. P. H. Fuss (1843). Correspondance Mathématique et Physique de Quelques Célèbres Géomètries du XVIII. Saint-Pétersbourg: Acad. Imperiale des Sci., pág. 713 (vid. pág. 1)

5. L. Euler (1738). “De progressionibus transcendentibus seu quarum termini generales algebraice dari nequeunt”. Comm. Acad. Sci. Petropolitanae 5, págs. 36-57 (vid. pág. 2).
6. P. J. Davis (1959). “Leonhard Euler’s integral: A historical profile of the gamma function”. Amer. Math. Monthly 66, págs. 849-869. MR: 0106810 (vid. págs. 1-3, 58).
7. P. Díaz & R. Labarca (2019). “Función Gamma: Propiedades clásicas e introducción a su dinámica”. IMPA, Brasil.
8. R. Troncoso de la Cuesta (2021). “La distribución Gamma”. Universidad de Santiago de Compostela
9. Lechuga, M. L. (1998). La distribución Gamma como modelo para analizar la distribución de la renta: una aplicación a la EPF 1990-91. Revista de Estudios Regionales, 1, 161-186.
10. Llanos Arias, J. (1994). Ajuste de la distribución Gamma a datos de precipitación pluvial.
11. Villazón Gómez, J. A., Noris Noris, P., & Martín Gutiérrez, G. (2021). Determinación de la precipitación efectiva en áreas agropecuarias de la provincia de Holguín. Idesia (Arica), 39(2), 85-90. R. Troncoso de la Cuesta (2021). “La distribución Gamma”. Universidad de Santiago de Compostela
12. Pearson, Karl (1895). Çontributions to the mathematical theory of evolution, II: Skew variation in homogeneous material".
13. Pearson, Karl (1901). "Mathematical contributions to the theory of evolution, X: Supplement to a memoir on skew variation".
14. Pearson, Karl (1916). "Mathematical contributions to the theory of evolution, XIX: Second supplement to a memoir on skew variation"
15. Ortíz, M.C.L (1988). Algunos métodos robustos en el análisis de varianza. Tesis de licenciatura en estadística, Facultad de Ciencias, Universidad Veracruzana.
16. Huber, P.J.(1981). Robust Statistics. John Wiley and Sons.
17. Tiku, M.L., Tan, W.Y and Balakrishnam, N. (1986) Robust Inference. Marcel Dekker Inc.
18. Ricardo A. Maronna, R. Douglas Martin, Victor J. Yohai and Matías Salibián-Barrera (2019). Robust Statistics, theory and methods. John Wiley & Son
19. P.J. Huber (1981). Robust Estimation of Location and Scale Using the Median and the Interquartile Range
20. F.P. Hampel, P. Rousseeuw, & E.A. Ronchetti (1986). The Hampel-Rousseeuw Estimator for Multivariate Location and Scale
21. A. Redchuk, D. Soria, (2000). Estimación Máximo Verosímil del Parámetro de Forma de la Distribución Gamma. Revista de la Escuela de Perfeccionamiento en Investigación Operativa
22. M. Valdora, V. Yohai, (2019). M estimators based on the probability integral transformation with applications to count data. Universidad de Buenos Aires.

Carrera de matemática, Universidad Nacional Autónoma de Honduras, Tegucigalpa

Dirección de correo electrónico: ulises.obando@unah.hn

UNA BREVE DESCOMPOSICIÓN DE FIBONACCI DE POLINOMIOS SIMÉTRICOS TETRANACCI

JOSUÉ ANTONIO ZÚNIGA GALO

RESUMEN. La descomposición de Fibonacci en el contexto de los polinomios simétricos Tetranacci es una área de estudio que combina conceptos fundamentales de la teoría de números y álgebra polinómica. Este trabajo explora la relación entre la famosa sucesión de Fibonacci y los polinomios simétricos Tetranacci. A través de esta investigación, se espera ampliar nuestra comprensión de los polinomios simétricos Tetranacci y su conexión con los números de Fibonacci.

ABSTRACT The Fibonacci decomposition in the context of symmetric Tetranacci polynomials is an area of study that combines fundamental concepts from number theory and polynomial algebra. This work explores the relationship between the famous Fibonacci sequence and symmetric Tetranacci polynomials. Through this research, we aim to expand our understanding of symmetric Tetranacci polynomials and their connection to Fibonacci numbers.

1. INTRODUCCIÓN

La sucesión de Fibonacci, está definida por la relación de recurrencia $F_{n+1} = F_n + F_{n-1}$ con $n \geq 1$ y $F_0 = 0$ y $F_1 = 1$. Esta sucesión fue descrita en Europa por Leonardo de Pisa, matemático italiano del siglo $XIII$ también conocido como Fibonacci. Tiene numerosas aplicaciones en ciencias de la computación, matemática, biología y teoría de juegos [4, 5]. Sin embargo, su relación con los polinomios simétricos Tetranacci, que son una generalización de los polinomios de Fibonacci $F_n(x)$, estos son una sucesión de polinomios definidos por:

$$F_n(x) = xF_{n-1}(x) + F_{n-2}(x)$$

con $F_0(x) = 0$ y $F_1(x) = 1$. Los polinomios simétricos Tetranacci consideran cuatro términos consecutivos en lugar de dos, la palabra "Tetranacci" proviene de dos partes: "Tetra", del griego "tettares", que significa cuatro e indica que la sucesión involucra cuatro términos anteriores, y "nacci", de "Fibonacci". Los polinomios simétricos son polinomios en n variables que permanecen invariantes bajo cualquier permutación de estas variables. En otras palabras, si intercambiamos cualquier par de variables en el polinomio, el resultado sigue siendo el mismo polinomio. Estos polinomios han recibido menos atención hasta ahora [3]. Una área de interés es la relación entre polinomios simétricos y sucesiones generalizadas como la Tetranacci.

Palabras y frases clave. Polinomios, Fibonacci, Tribonacci, Tetranacci, descomposición y simetría.

Los polinomios simétricos Tetranacci, definidos recursivamente por la ecuación:

$$\xi_{n+2} = \zeta\xi_n - \xi_{n-2} + \eta(\xi_{n+1} + \xi_{n-1}),$$

donde ζ y η representan coeficientes complejos, estos polinomios presentan una rica estructura algebraica que merece ser explorada más a fondo. La simetría aquí no se refiere a la simetría clásica de polinomios, sino a la estructura balanceada de la relación recursiva. Cada término de la secuencia depende de una combinación de términos anteriores y posteriores de una manera que puede presentar patrones simétricos [5].

El objetivo de este trabajo es investigar la descomposición de Fibonacci de los polinomios simétricos Tetranacci, proponer y ejemplificar algunos problemas basados en el tema. A través de esta investigación, esperamos no solo ampliar nuestra comprensión de los polinomios Tetranacci simétricos y su conexión con la secuencia de Fibonacci, sino también descubrir nuevas propiedades y aplicaciones de estos polinomios en distintas áreas.

Definimos formalmente los polinomios Tetranacci simétricos y presentamos la estrategia básica para encontrar su expresión en forma cerrada. Primero, haremos un breve desarrollo de los polinomios de Tribonacci, seguidamente, introducimos los llamados polinomios Tetranacci básicos y discutimos algunas de sus propiedades. Finalmente, demostramos que ciertos polinomios Fibonacci generalizados también obedecen la fórmula de recurrencia Tetranacci[5].

El estudio de la sucesión y los polinomios simétricos de Tetranacci muestra una importancia significativa para la ciencia, así como áreas de interés por parte de la Universidad Nacional Autónoma de Honduras, ya sea en estudios del área de salud, biología, ingeniería y ciencias de la computación.

2. Antecedentes

El origen del estudio de la secuencia de Fibonacci comienza con un simple problema, conocido como el problema de los conejos. Se plantea de la siguiente manera:

1. Se coloca una pareja de conejos en un campo.
2. Cada pareja de conejos produce una nueva pareja cada mes a partir del segundo mes de vida.
3. El objetivo es calcular cuántas parejas de conejos habrá después de un año.

En el primer mes, hay 1 pareja de conejos (la original). En el segundo mes, sigue habiendo 1 pareja (ya que los conejos no se reproducen hasta el segundo mes). En el tercer mes, hay 2 parejas (la pareja original más una nueva pareja nacida). En el cuarto mes, hay 3 parejas (la pareja original produce otra nueva pareja, además de la pareja nacida el mes anterior que aún no se reproduce). En el quinto mes, hay 5 parejas (la pareja original y la nueva pareja nacida en el tercer mes producen cada una, una nueva pareja) y así sucesivamente [4].

Este patrón origina la sucesión mencionada en la introducción. A partir de esta sucesión, se derivan los polinomios de Fibonacci. De manera similar, así también, el estudio de una sucesión Tribonacci, que consiste en la suma de tres términos

anteriores de la sucesión, esto da lugar a los polinomios Tribonacci, y es así como de manera parecida se desarrolla la sucesión Tetranacci. De esta forma, se desarrollan también los polinomios Tetranacci [1, 2, 6].

Históricamente, el estudio de las sucesiones de Fibonacci y Tetranacci se centraba en sus propiedades numéricas y combinatorias. Sin embargo, en décadas recientes ha habido un interés creciente en explorar sus extensiones polinomiales y aplicaciones en álgebra abstracta. Investigadores como Koshy han ampliado nuestra comprensión de estas sucesiones, mientras que estudios recientes han comenzado a vincular estos conceptos con polinomios simétricos y otras estructuras algebraicas complejas [1].

Diversos estudios han abordado muchos temas diferentes relacionados con la sucesión de Fibonacci, sin embargo, nuestro objetivo es centrado en los polinomios simétricos Tetranacci, la combinación específica de estos conceptos en forma de polinomios simétricos Tetranacci y su descomposición de Fibonacci es un área emergente con mucho potencial por explorar. Estudios recientes han comenzado a revelar las complejas interacciones entre estas estructuras y sus aplicaciones en teoría de números y combinatoria [1, 3, 5].

En tiempos más recientes, investigadores como Koshy han trabajado en los polinomios de Fibonacci, mientras que Leumer ha explorado la descomposición de Fibonacci de polinomios simétricos Tetranacci. Estas investigaciones han revelado nuevas propiedades y aplicaciones de estos polinomios, abriendo nuevos caminos para la investigación [1, 5].

A pesar de los avances significativos en el estudio de polinomios simétricos y sucesiones de Fibonacci y Tetranacci, aún existen numerosas áreas que no han sido suficientemente exploradas. En particular, la descomposición de Fibonacci de polinomios simétricos Tetranacci, es un área relativamente nueva que merece atención adicional para entender mejor sus propiedades y aplicaciones. Existe una asociación dedicada a estudios y soluciones de problemas relacionados con Fibonacci, que lleva por nombre Asociación de Fibonacci, donde publican investigaciones y problemas mediante su revista “The Fibonacci Quarterly”. Esta asociación es una organización de Canadá, reúne a investigadores y entusiastas de todo el mundo, proporcionando un foro para la discusión y difusión de descubrimientos relacionados con la sucesión de Fibonacci [6].

3. Las sucesiones recursivas Tribonacci y Tetranacci

La sucesión de Fibonacci, mencionada previamente en los antecedentes, sirve como base para la comprensión de secuencias recursivas y sus aplicaciones en diversas áreas matemáticas y científicas. A partir del problema clásico de los conejos, hemos aprendido cómo una simple regla de crecimiento puede generar una sucesión con propiedades matemáticas profundas.

En esta sección, extendemos el concepto de sucesiones recursivas para introducir los polinomios simétricos Tetranacci. Estos polinomios se construyen siguiendo una regla de recurrencia similar a la de Fibonacci, pero considerando cuatro términos anteriores en lugar de dos.

Para ilustrar la transición desde la sucesión de Fibonacci a los polinomios Tetranacci, presentamos la imagen que representa el problema de los conejos explicado en los antecedentes:

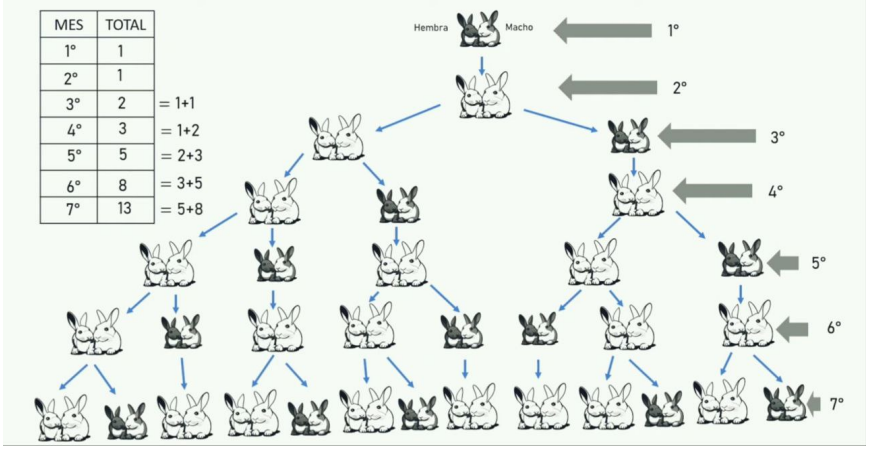


FIGURA 1. El problema de los conejos [9].

La sucesión Tribonacci es una secuencia numérica que sigue un patrón específico de recursión y que se relaciona estrechamente con la secuencia más conocida de Fibonacci. La sucesión Tribonacci es una generalización de la secuencia de Fibonacci, donde en lugar de sumar dos términos anteriores, se suman tres.

Ambas secuencias comparten propiedades estructurales y pueden ser estudiadas usando técnicas similares, veamos a continuación la definición de la sucesión de Tribonacci:

Definición 3.1. La sucesión Tribonacci es una sucesión numérica donde cada término se obtiene sumando los tres términos anteriores. Los primeros términos de la sucesión tribonacci son:

$$T_0 = 0, \quad T_1 = 0, \quad T_2 = 1.$$

A partir de estos, se define recursivamente como:

$$T_n = T_{n-1} + T_{n-2} + T_{n-3} \quad \text{para } n \geq 3.$$

Los primeros términos de la sucesión son:

$$0, 0, 1, 1, 2, 4, 7, 13, 24, 44, \ldots$$

.

En la siguiente definición nos permite calcular cualquier término T_n de manera eficiente la cual es una herramienta poderosa para analizar, manipular y aplicar esta sucesión en diversas áreas matemáticas y prácticas.

Proposición 1. [9] La función generatriz para la sucesión tribonacci es:

$$F(x) = \frac{x^3}{1 - x - x^2 - x^3}. \tag{3.1}$$

Ejemplo 3.2. Expansión de la función generatriz de la sucesión Tribonacci para $n = 2$.

Solución:

Consideramos la expansión del denominador usando la serie geométrica:

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n \quad \text{donde} \quad z = x + x^2 + x^3$$

Sustituimos z y expandiendo:

$$\frac{x^3}{1 - (x + x^2 + x^3)} = x^3 \sum_{n=0}^{\infty} (x + x^2 + x^3)^n$$

Desarrollemos algunos términos de $(x + x^2 + x^3)^n$:

Para $n = 0:$ $\quad (x + x^2 + x^3)^0 = 1$
El término es $x^3 \cdot 1 = x^3$.

Para $n = 1$:
El término es: $x^4 + x^5 + x^6$.

Para $n = 2$:
El término es: $x^5 + 2x^6 + 3x^7 + 2x^8 + x^9$

De esta manera, podemos encontrar en los coeficientes de cada variable un número de la sucesión tribonacci,

$$F(x) = x^3 + x^4 + 2x^5 + 3x^6 + 3x^7 + 2x^8 + x^9 + \cdots$$

que corresponden a los términos de la sucesión tribonacci.

Ejemplo 3.3. (Relación con la Geometría y la Naturaleza:)
Al igual que la secuencia de Fibonacci, la sucesión tribonacci aparece en varios contextos geométricos y naturales, como en el crecimiento de algunas plantas y en la disposición de ciertos objetos en la naturaleza.

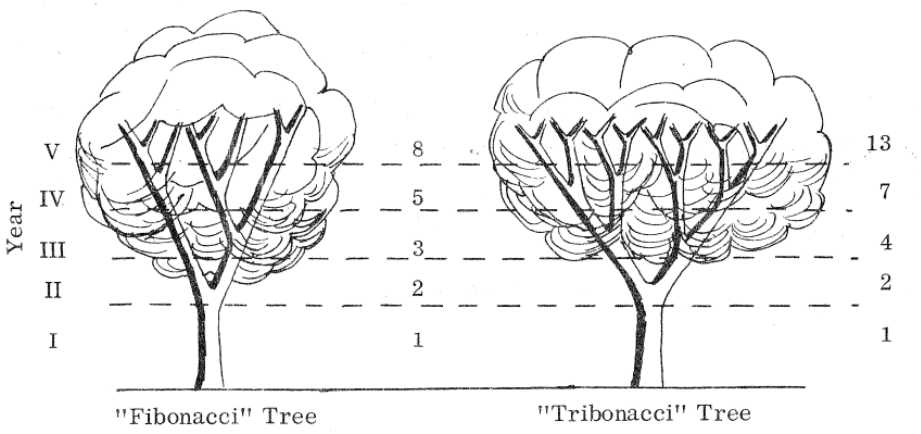


FIGURA 2. Crecimiento de las ramas de los arboles "siguen el patrón de una sucesión de Fibonacci y Tribonacci" [6].

A continuación, analizaremos la sucesión de Tetranacci, que amplía aún más este concepto, sumando los cuatro términos anteriores para generar cada nuevo término. Esta sucesión posee propiedades únicas que la hacen particularmente interesante para el estudio de polinomios simétricos.

Definición 3.4. [10] Sea T_n una sucesión Tetranacci definida por la relación de recurrencia:

$$T_n + T_{n+1} + T_{n+2} + T_{n+3} = T_{n+4} \tag{1.0}$$

con valores iniciales $T_0 = 0, T_1 = 0, T_2 = 0, T_3 = 1$.

Estos son los primeros términos de la sucesión: $0, 0, 0, 1, 1, 2, 4, 8, 15, 29, 56, 108, 208, 401, 773, \cdots$

n	T_n	Expresión	Resultado
4	T_4	$T_3 + T_2 + T_1 + T_0 = 1 + 0 + 0 + 0$	1
5	T_5	$T_4 + T_3 + T_2 + T_1 = 1 + 1 + 0 + 0$	2
6	T_6	$T_5 + T_4 + T_3 + T_2 = 2 + 1 + 1 + 0$	4
7	T_7	$T_6 + T_5 + T_4 + T_3 = 4 + 2 + 1 + 1$	8
8	T_8	$T_7 + T_6 + T_5 + T_4 = 8 + 4 + 2 + 1$	15
9	T_9	$T_8 + T_7 + T_6 + T_5 = 15 + 8 + 4 + 2$	29
10	T_{10}	$T_9 + T_8 + T_7 + T_6 = 29 + 15 + 8 + 4$	56

CUADRO 1. Cálculo de algunos términos de la sucesión Tetranacci.

Al igual que con la función generatriz Tribonacci 3.1, la función generatriz Tetranacci se utiliza para representar la sucesión en una forma compacta y facilitar el análisis de sus propiedades. La técnica de expansión en series de potencias y la manipulación algebraica aplicada a la función generatriz Tribonacci también se aplica de manera análoga a la función generatriz Tetranacci.

Proposición 2. La función generatriz sin factoriales está dada por

$$T(x) := \frac{x}{1 - x - x^2 - x^3 - x^4} = \sum_{n=0}^{\infty} T_n x^n$$

debido a la relación de recurrencia.

La siguiente definición se basa en la idea de la sucesión de Tetranacci conocida por su utilidad en diversas aplicaciones matemáticas y computacionales.

Definición 3.5. [10] La matriz A que define la relación de recurrencia de la secuencia Tetranacci está dada por:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Esta matriz permite expresar la relación de recurrencia de la sucesión Tetranacci 1.0, en forma matricial .

Veamos la construcción de la matriz:
La sucesión Tetranacci definida por la relación de recurrencia 1.0, para expresar esta relación en términos de una matriz, consideramos un vector columna que contiene los términos de la sucesión en orden descendente:

$$T_n = \begin{pmatrix} T_n \\ T_{n-1} \\ T_{n-2} \\ T_{n-3} \end{pmatrix}$$

Cuando multiplicamos este vector columna T_n por una matriz específica A, obtenemos el siguiente vector columna T_{n+1}:

$$T_{n+1} = AT_n$$

Para construir la matriz A, observamos cómo cada término de la sucesión Tetranacci depende de los cuatro términos anteriores. La matriz A se define de tal manera que cuando la multiplicamos por T_n, obtenemos un vector columna que representa los términos $(T_{n+1}, T_n, T_{n-1}, T_{n-2})$.

Vamos a explicar por qué esta matriz es adecuada:

1. Primera fila: El primer término T_{n+1} es T_n, por lo tanto, en la primera fila de A, ponemos 1 en la segunda columna.
2. Segunda fila: El segundo término T_n es T_{n-1}, por lo tanto, en la segunda fila de A, ponemos 1 en la tercera columna.

3. Tercera fila: El tercer término T_{n-1} es T_{n-2}, por lo tanto, en la tercera fila de A, ponemos 1 en la cuarta columna.

4. Cuarta fila: Para el cuarto término T_{n-2}, sumamos todos los términos anteriores $T_n + T_{n-1} + T_{n-2} + T_{n-3}$, por lo tanto, en la cuarta fila de A, ponemos 1 en todas las columnas.

Esta matriz A se utiliza para calcular los términos sucesivos de la sucesión Tetranacci, multiplicamos la matriz A por el vector columna T_n:

$$T_{n+1} = AT_n$$

Esto nos da un método sistemático para calcular cada término sucesivo de la sucesión Tetranacci utilizando la matriz A.

Ejemplo 3.6. Si tenemos los valores iniciales:

$$T_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Para encontrar T_4:

$$T_4 = AT_3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

De esta manera hemos calculado T_4

Calculamos T_5:

$$T_5 = AT_4 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

Así, hemos calculado T_5 utilizando la matriz Tetranacci.

La siguiente definición es una generalización de la famosa sucesión de Fibonacci basada en polinomios.

Definición 3.7. [8] Los polinomios de Fibonacci generalizados F_n como (para $n \geq 1$)

$$F_{n+1} = xF_n + yF_{n-1}, F_0 = 0, F_1 = 1. \tag{3.2}$$

Ampliando el rango de recursión en la Ec. 3.2 de dos a tres se obtienen los números Tribonacci o polinomios Tribonacci, dependiendo de los coeficientes y suponiendo valores iniciales adecuadamente elegidos [5]. Posteriormente, la primera noción de números Tetranacci, donde el siguiente elemento de la secuencia se forma con los cuatro anteriores, de esta forma se definen los polinomios Tetranacci.

Proposición 3. [5] La forma más genérica de lo que llamamos en adelante polinomios Tetranacci t_n (para $n \geq 0$) se define como:

$$t_{n+2} = x_1 t_{n+1} + x_0 t_n + x_{-1} t_{n-1} + x_{-2} t_{n-2} \tag{1.2}$$

con algunos valores iniciales $t_{-2}, \ldots, t_1$ y coeficientes dados $x_1, \ldots, x_{-2}$. En contraste con la Ec. 3.2, nos enfocamos en la situación especial donde $x_{-2} = -1$ y $x_1 = x_{-1}$ (ver Ec. 3.3 abajo), pero aún con x_0, x_1 genéricos.

3.1. Polinomio simétrico Tetranacci. Las siguientes definiciones, lemas, proposiciones y teoremas están basadas en [5].

Definición 3.8. El polinomio Tetranacci simétrico ξ_j se define recursivamente por

$$\xi_{j+2} = \zeta \xi_j - \xi_{j-2} + \eta(\xi_{j+1} + \xi_{j-1}), \quad j \in \mathbb{Z} \tag{3.3}$$

en términos de sus valores iniciales $\xi_i = g_i(\zeta, \eta) \in \mathbb{C}$ para $i = -2, \ldots, 1$ y coeficientes complejos ζ, η.

Aunque los valores iniciales pueden o no depender de ζ y/o η, siempre utilizamos la notación abreviada $g_{-2}, \ldots, g_1$ y ξ_j respectivamente, en lugar de mencionar explícitamente esta dependencia. Para ilustrar, los primeros términos ξ_j son

$\xi_2 = -g_{-2} + \eta g_{-1} + \zeta g_0 + \eta g_1,$
$\xi_3 = -\eta g_{-2} + \zeta g_1 - g_{-1} + \eta(-g_{-2} + \eta g_{-1} + \zeta g_0 + \eta g_1 + g_0)$
$\xi_4 = \zeta(-g_{-2} + \eta g_{-1} + \zeta g_0 + \eta g_1) - g_0 + \eta(-\eta g_{-2} + \zeta g_1 - g_{-1} + \eta(-g_{-2} + \eta g_{-1} + \zeta g_0 + \eta g_1) + g_0 + g_1)$

De forma más clara, podemos ver el polinomio simétrico Tetranacci tal como estamos acostumbrados a ver un polinomio con variables conocidas, por ejemplo:

Para $j = 2$ en la ecuación 3.3, tenemos:

$$\xi_4 = \zeta \xi_2 - \xi_0 + \eta(\xi_3 + \xi_1)$$

$$\xi_0 = w, \quad \xi_1 = x, \quad \xi_2 = y, \quad \xi_3 = z, \quad \xi_4 = P(w, x, y, z), \quad \zeta = 2 \quad \text{y} \quad \eta = 3$$

Sustituyendo las variables:

$$P(w, x, y, z) = 2y - w + 3(z + x)$$

Los términos siguientes se derivan de la Ecuación 3.3. Alternativamente, también podemos recurrir a la función generadora, a la que nos dirigimos a continuación.

Proposición 4. La función generadora $E(t) = \sum_{n=0}^{\infty} \xi_n t^n$ de los polinomios Tetranacci simétricos se lee:

$$E(t) = \frac{g_1 t + g_0(1-\eta t) + g_{-1}(\eta t^2 - t^3) - g_{-2}t^2}{1 - \eta t - \zeta t^2 - \eta t^3 + t^4} \tag{3.4}$$

Ejemplo 3.9. Expansión de la función generatriz de los polinomios Tetranacci simétricos para $n = 2$.

Utilizamos la expansión en serie geométrica para el denominador:

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n$$

donde $z = \eta t + \zeta t^2 + \eta t^3 - t^4$.

Sustituimos z y expandiendo en la serie geométrica para $n = 0$, $n = 1$, y $n = 2$:

Para $n = 0$:

$$g_1 t + g_0(1-\eta t) + g_{-1}(\eta t^2 - t^3) - g_{-2}t^2$$

.

Para $n = 1$:

$$(g_1 t + g_0(1-\eta t) + g_{-1}(\eta t^2 - t^3) - g_{-2}t^2)(\eta t + \zeta t^2 + \eta t^3 - t^4)$$

Para $n = 2$:

$$= \eta^2 t^2 + 2\eta\zeta t^3 + (\eta^2 + \zeta^2)t^4 - 2\eta t^5 - 2\zeta t^6 - \eta^2 t^6 + t^8$$

Ahora, sumando los términos:

$$E(t) = (g_1+\eta)t + (g_0 - \eta g_0 + \zeta)t^2 + (\eta g_0 + g_{-1}\eta - g_1)t^3 - (g_{-1}t^4 + \eta^2 t^2 + 2\eta\zeta t^3 + (\eta^2+\zeta^2)t^4)$$

De esta manera la expansión de la función generatriz permite encontrar los coeficientes de cada variable en la serie de potencias, lo que corresponde a los términos de los polinomios Tetranacci simétricos.

Definición 3.10. Los polinomios básicos de Tetranacci $T_i(j)$ $(i = -2, \ldots, 1)$ satisfacen la Ecuación 3.3 para $j \in \mathbb{Z}$ y sus valores iniciales se resumen en:

$$T_i(j) = \delta_{ij}, \quad i, j = -2, \ldots, 1 \tag{3.1}$$

Aquí, δ_{ij} denota el Delta de Kronecker y llamamos a la Ecuación 3.1 la propiedad selectiva de $T_i(j)$.

Corolario 1. Cualquier polinomio simétrico de Tetranacci ξ_j puede escribirse como:

$$\xi_j = \sum_{i=-2}^{1} g_i T_i(j), \quad j \in \mathbb{Z} \tag{2.8}$$

para $\eta, \zeta \in \mathbb{C}$ genéricos y valores iniciales complejos $\xi_i = g_i$, $i = -2, \ldots, 1$.

El siguiente lemma proporciona una forma para expresar los polinomios simétricos Tetranacci en términos de funciones exponenciales complejas. Esta representación puede simplificar el análisis y la aplicación de estos polinomios en diversas áreas.

Lema 3.11. *Cualquier polinomio simétrico Tetranacci puede expresarse como*

$$\xi_j = Ae^{i\theta_1 j} + Be^{-i\theta_1 j} + Ce^{i\theta_2 j} + De^{-i\theta_2 j}, \tag{2.22}$$

siempre que $S_1 \neq \pm S_2$ *y* $S_1^2, S_2^2 \neq 4$*, donde*

$$S_{1,2} = \frac{\eta \pm \sqrt{\eta^2 + 4(\zeta + 2)}}{2}.$$

En la ecuación (2.22), introducimos $\theta_{1,2} \in \mathbb{C}$ *definidos por* $2\cos(\theta_{1,2}) := S_{1,2}$*. Los coeficientes* A, B, C, D *están determinados por los valores iniciales* $\xi_i = g_i$*,* $i = -2, \ldots, 1$*.*

Parámetros:

- ξ_j: Los términos del polinomio Tetranacci.
- A, B, C, D: Coeficientes que dependen de los valores iniciales ξ_i.
- θ_1, θ_2: Ángulos complejos definidos mediante $S_{1,2}$, que a su vez dependen de los coeficientes η y ζ.
- $S_{1,2}$: Expresiones que dependen de η y ζ.

Para que la fórmula sea válida, se deben cumplir las siguientes condiciones:

1. $S_1 \neq \pm S_2$
2. $S_1^2 \neq 4$
3. $S_2^2 \neq 4$

Estas condiciones aseguran que los términos exponenciales no se colapsen en términos redundantes "necesitamos asegurar que cada término en la combinación lineal ξ_j sea único y contribuya de manera independiente a la solución, cumpliendo las condiciones", manteniendo la independencia lineal de los componentes.

Ejemplo 3.12. Vamos a ilustrar el uso de este lema con un ejemplo sencillo. Supongamos que tenemos los siguientes valores iniciales y parámetros:

$$\xi_{-2} = 1, \quad \xi_{-1} = 2, \quad \xi_0 = 3, \quad \xi_1 = 5$$

Dado: $\eta = 1$ y $\zeta = -1{,}5$.

Podemos calcular $S_{1,2}$ de la siguiente manera:

$$S_{1,2} = \frac{\eta \pm \sqrt{\eta^2 + 4(\zeta + 2)}}{2}$$

$$S_{1,2} = \frac{1 \pm \sqrt{1^2 + 4(-1{,}5 + 2)}}{2}$$

$$= \frac{1 \pm \sqrt{1 + 4 \times 0{,}5}}{2} = \frac{1 \pm \sqrt{3}}{2}$$

$$S_1 = \frac{1 + \sqrt{3}}{2}, \quad S_2 = \frac{1 - \sqrt{3}}{2}$$

$S_1 \approx 1{,}366$ $S_2 \approx -0{,}366$
Cálculo de θ_1 y θ_2:

$$\theta_1 = \cos^{-1}\left(\frac{S_1}{2}\right) = \cos^{-1}\left(\frac{1{,}366}{2}\right) = \cos^{-1}(0{,}683)$$

$$\theta_2 = \cos^{-1}\left(\frac{S_2}{2}\right) = \cos^{-1}\left(\frac{-0{,}366}{2}\right) = \cos^{-1}(-0{,}183)$$

$$\theta_1 \approx 0{,}821$$
$$\theta_2 \approx 1{,}755$$

Por tanto, los valores iniciales cumplen para que la fórmula 2.22 sea válida.

Para finalizar este trabajo, presentamos una aplicación relacionada con los polinomios simétricos Tetranacci. Los polinomios simétricos Tetranacci pueden utilizarse para obtener los valores propios de las matrices hamiltonianas de Toeplitz (matriz cuadrada en la que cada elemento en una diagonal descendente de izquierda a derecha es constante y cada elemento en una misma diagonal son iguales). Esto se hace resolviendo la ecuación característica de la matriz, que es un polinomio cuyos coeficientes están relacionados con los elementos de la matriz. Al determinar los valores propios de la matriz hamiltoniana, los polinomios simétricos Tetranacci permiten identificar los estados estacionarios del sistema [5].

4. Conclusiones

- En conclusión, este estudio ha destacado la relevancia de la sucesión de Fibonacci y ha explorado los polinomios simétricos Tetranacci, estableciendo su relación con dicha sucesión. Hemos presentado definiciones, proposiciones, lemas, teoremas y ejemplos que facilitan la comprensión y el desarrollo de esta teoría. El análisis hecho es significativo, ya que proporcionan una mejor comprensión de comportamientos y relaciones en física, química, biología, matemáticas, entre otras disciplinas. Este estudio no solo amplía nuestro conocimiento sobre los polinomios simétricos Tetranacci, sino que también sugiere nuevas direcciones para investigaciones futuras.
- Se ha demostrado que los polinomios simétricos Tetranacci, al ser una extensión natural de los polinomios de Fibonacci, poseen características algebraicas que amplían el conocimiento previo sobre secuencias recursivas. La función generatriz y la matriz de recurrencia asociada proporcionan herramientas efectivas para el análisis y cálculo de estos polinomios.
- En resumen, la investigación confirma que los polinomios simétricos Tetranacci ofrecen una perspectiva valiosa sobre las secuencias polinomiales, abriendo nuevas avenidas para aplicaciones y estudios en teoría de números y álgebra.
- Continuar investigando en esta área es crucial para ampliar nuestro conocimiento y comprensión sobre los polinomios simétricos Tetranacci y sus aplicaciones. Para trabajos futuros, se sugiere profundizar en la relación entre los polinomios simétricos tetranacci y otros polinomios recursivos. También sería beneficioso desarrollar métodos numéricos eficientes para calcular estos polinomios y sus aplicaciones en problemas concretos de la física y la biología. Los hallazgos futuros no solo podrán enriquecer la teoría matemática existente, sino que también tienen el potencial de influir en otras áreas de la ciencia y la ingeniería, proporcionando nuevas herramientas y enfoques para resolver problemas complejos.

Referencias

[1] V. Hoggatt Jr. and M. Bicknell, Generalized Fibonacci Polynomials, Fibonacci Q. 11(5), 457–465 (1973). Available at: https://www.fq.math.ca/Scanned/11-5/hoggatt.pdf.

[2] DeTemple, D. W. *Tetranacci Numbers*. The Fibonacci Quarterly, 1983.

[3] Macdonald, I. G. *Symmetric Functions and Hall Polynomials*. Oxford University Press, 1998.

[4] Koshy, T. *Fibonacci and Lucas Numbers with Applications*. Wiley-Interscience, 2001.

[5] Leumer, N. G. *The Fibonacci Decomposition of Symmetric Tetranacci Polynomials*. arXiv:2208.10527v2 [math-ph], 19 Aug 2022. Disponible en: https://arxiv.org/html/2208.10527v2

[6] M. Feinberg. Fibonacci-Tribonacci, Fibonacci Q. Available at: https://www.fq.math.ca/Scanned/1-3/feinberg.pdf.

[7] *The Fibonacci Association*. Available at: https://www.fq.math.ca/.

[8] V. Hoggatt Jr. and C. T. Long, Divisibility properties of generalized Fibonacci Polynomials, Fibonacci Q. 12(2), 113-120 (1974), https://www.fq.math.ca/Scanned/12-2/hoggatt1.pdf.

[9] Canal de Youtube "'Áuero Educa", El problema de los conejos Origen de la sucesión de Fibonacci.https://www.youtube.com/watch?v=goI3xO70LwEt=366s

[10] M. E. Waddill, The Tetranacci sequence and generalizations, *Fibonacci Quarterly*, 30(1), 9-19 (1992), https://www.fq.math.ca/Scanned/30-1/waddill.pdf.

[11] Mansi N. Zaveri, Jayant K. Patel, Generalized Tetranacci Sequence and Its Period, *2nd International Conference on Multidisciplinary Research Practice (ICMRP)*, 2015, https://www.rsisinternational.org/2ICMRP2015/81-84.pdf.

Departamento de Matemáticas, Universidad Nacional Autónoma de Honduras

Dirección actual: Tegucigalpa, Honduras

Dirección de correo electrónico: josuezuniga@unah.hn

APLICACIÓN DE MODELOS GARCH EN LA ESTIMACIÓN DEL VALOR EN RIESGO

JOSE LEONEL MARTINEZ

Resumen. El modelo GARCH (modelo autorregresivo condicional heterocedástico generalizado) es un modelo estadístico para series de tiempo que describe la varianza del error en función de los errores al cuadrado pasados y de las variaciones. En la presente investigación, se hará la aplicación de modelos GARCH y sus extensiones (EGARCH y TGARCH) para la estimación del Valor en Riesgo (VaR) del índice bursátil del precio del café. Se emplean diferentes distribuciones para los retornos, específicamente la distribución normal y la t-Student, y se utilizan tres tamaños de muestra históricas distintas, (500, 1000 y 1500 observaciones). Además, se realizo la selección del mejor modelo, mediante criterios de información, para medir la efectividad de los modelos de volatilidad en la predicción del VaR, proporcionando una comparación empírica de su desempeño y ofreciendo recomendaciones para su implementación práctica en la gestión del riesgo financiero.

ABSTRACT. The GARCH model (generalised autoregressive heterocedastico conditional autoregressive model) is a statistical model for time series describing a time series. co generalized autoregressive conditional autoregressive model) is a statistical model for time series that describes the error variance as a function of the past squared errors and of the variance of the time series. cations. In the present research, the application will be made on GARCH models and their extensions (EGARCH and extensions (EGARCH and TGARCH) for the estimation of the Value at Risk (VaR) of the index (VaR) of the stock market index of the price of coffee. Different distributions are used. The distributions for the returns, specifically the normal and the t-Student distributions, are used, and three different historical sample sizes are used, (500, 1000, and 1500 observations). In addition, the selection of the best model, by means of informa- tion criteria, is carried out to select the best model to be used. tion criteria, to measure the effectiveness of the volatility models in predicting the volatility in VaR prediction, providing an empirical comparison of their performance and offering recommendations on their performance and offering recommendations for practical implementation in the management of VaR. financial risk management.

1. introducción

El histórico Acuerdo de Capital de Basilea de 1988 marcó el primer paso significativo hacia un sistema financiero más seguro y sólido. Este acuerdo permitió a los bancos utilizar modelos de Valor en Riesgo (VaR) internos como base para determinar el capital necesario para cubrir el riesgo de mercado. Como resultado, el VaR ha sido oficialmente promovido como una práctica óptima de gestión del riesgo. Anteriormente, la solidez financiera era un concepto vago, pero ahora puede

Fecha: 28 mayo, 2024 y, en forma revisada, junio 10, 2024.
Palabras y frases clave. Modelos GARCH, Valor en Riesgo (VaR),Activos, Volatilidad.

ser medida en términos de la probabilidad de insolvencia, gracias a este avance en la medición de la estabilidad financiera. Esta mejora ha proporcionado a instituciones financieras y reguladores una visión más clara y precisa de los riesgos involucrados [1].

En el ámbito de la gestión de riesgos financieros, el Valor en Riesgo (VaR) se ha consolidado como una herramienta esencial para cuantificar y gestionar la exposición al riesgo de mercado de una cartera de inversiones. El VaR estima la máxima pérdida esperada de una cartera en un horizonte temporal determinado, bajo condiciones normales de mercado y para un nivel de confianza específico [2]. Esta medida ha ganado popularidad debido a su simplicidad y capacidad para sintetizar el riesgo en un solo número comprensible tanto para gestores de riesgos como para reguladores [5].

Sin embargo, se han desarrollado diversos enfoques para estimar el VaR, como los métodos tradicionales, que presentan limitaciones al no capturar adecuadamente la dinámica cambiante de la volatilidad en los mercados financieros, por eso destacámos los modelos de volatilidad condicional como los modelos GARCH (Generalized Autoregressive Conditional Heteroskedasticity) y sus extensiones, tales como EGARCH (Exponential GARCH) y TGARCH (Threshold GARCH). Estos modelos permiten capturar la dinámica temporal de la volatilidad de los activos financieros, ofreciendo una estimación más precisa y realista del riesgo[6].

Esta investigación tiene como objetivo principal analizar la efectividad de los modelos GARCH, incluyendo sus extensiones EGARCH y TGARCH, en la estimación del VaR para los índice bursátil representativos de mercados financieros como ser el commodities que incluye el precio del café. Al analizar este índice permite evaluar la aplicabilidad y eficacia de los modelos GARCH y sus extensiones en la estimación del VaR en diferentes contextos financieros, que presentan características y comportamientos de volatilidad distintos. Además, se consideran distintas distribuciones para los retornos, como la distribución normal y la t-Student, para analizar su impacto en la estimación del VaR. Para asegurar la robustez de los resultados, se utilizarán tres tamaños de muestra históricas diferentes (500, 1000 y 1500 observaciones), con el fin de estimar el VaR de un día con los niveles de confianza del 95 % y al 99 % para el índice mencionado anteriormente [7].

En la presente investigación se contribuye al conocimiento existente, así como al conocimiento personal, en el campo de la gestión del riesgo financiero de varias maneras. Se explora y compara la precisión de los modelos GARCH, EGARCH y TGARCH para capturar la asimetría y la dependencia en los retornos del índice del precio del café cotizado en la bolsa de Nueva York, en la estimación del VaR, bajo diferentes supuestos de distribución de retornos para evaluar su sensibilidad a la distribución de los datos. Además, se examina el impacto del tamaño de la muestra en la precisión de las estimaciones del VaR y se evalúa su desempeño mediante criterios de información como ser AIC, BIC, HQC, AICc. El objetivo es proporcionar una comprensión más profunda de las ventajas y limitaciones de estos modelos y configuraciones en diversos contextos financieros, así como ofrecer recomendaciones para la implementación práctica de estos modelos en la estimación del VaR.

Y seguidamente se abordará una revisión detallada de la teoría subyacente a los modelos GARCH y sus extensiones, y se presentarán aplicación empírica utilizando datos históricos del precio del cafe, y se discutirán los resultados. Finalmente, se ofrecerán conclusiones y recomendaciones para futuros estudios e implementaciones en el campo de la gestión del riesgo financiero.

2. Antecedentes

Los modelos GARCH (Heterocedasticidad Condicional Autorregresiva Generalizada) siguen siendo una herramienta fundamental para la estimación del Valor en Riesgo VaR de activos financieros.

El Valor en Riesgo (VaR) ha sido ampliamente utilizado en la gestión de riesgos financieros desde su introducción en la década de 1990. JP Morgan popularizó el VaR con su sistema de medición de riesgos "RiskMetrics", el cual utilizaba supuestos simplificados sobre la volatilidad y las distribuciones de retornos. Desde entonces, el VaR se ha convertido en un estándar en la industria financiera, no solo por su capacidad para cuantificar el riesgo de mercado, esto significa que el VaR estima cuánto puede perderse en una inversión o cartera en un período específico con un cierto nivel de probabilidad (o confianza) en circunstancias normales de mercado, sino también por su adopción por parte de los reguladores financieros como el Comité de Basilea [9].

Aunque el concepto de VaR es muy sencillo, su cálculo no es fácil. Las metodologías desarrolladas inicialmente para calcular el VaR, se clasifican en tres categorías: paramétrico, no paramétrico y simulación Montecarlo.

- El enfoque de varianza-covarianza, también llamado método Paramétrico. Según Hull y White en mil novecientos noventa y ocho, este método asume que los rendimientos de la cartera siguen una distribución normal requiriendo el cálculo de la volatilidad y las correlaciones entre los activos de la cartera [10].
- La Simulación Histórica también llamado método no paramétrico, lo cual no depende de ninguna distribución de probabilidad específica, y no es necesario estimar volatilidades ni correlaciones, lo que lo hace más flexible en cuanto a los supuestos sobre la distribución de los rendimientos [10].
- La simulación Montecarlo, que es un método semiparamétrico, lo cual permite estimaciones tanto paramétricas como no paramétricas, proporciona una descripción más realista del riesgo ajustando una distribución empírica de los retornos, lo que puede captar mejor las características reales de los datos financieros. [18].

Como es bien sabido, todas estas metodologías, habitualmente denominadas modelos estándar, presentan numerosas deficiencias, que han llevado al desarrollo de nuevas propuestas [1].

Por eso los siguientes estudios han trabajado para superar problemas en la medición del riesgo usando distribuciones condicionales, que ajustan mejor las características de los datos financieros, algunos estudios importantes son como el Bollerslev en el año 1986 utilizó la distribución T-Student y Angelidis en el 2004 la distribución de errores generalizada (GED), que tiene colas pesadas, lo que ayuda a capturar eventos extremos, algunas de las ventajas de estas distribuciones es que son ampliamente utilizadas porque presentan un mejor desempeño en la captura del exceso de curtosis. Aunque tiene sus limitaciones a pesar de sus ventajas, ambas

distribuciones GED y T-Student son simétricas y, por lo tanto, no pueden captar el componente de asimetría en los retornos financieros, que es una característica importante de las innovaciones en los datos.

Luego están los modelos de heterocedasticidad condicional, estos modelos fueron propuestos por Engle en 1982 y generalizados por Bollerslev en 1986, se utilizan para modelar la estructura autorregresiva en la varianza condicional, es decir, la varianza de un periodo depende de la varianza de periodos anteriores. Estos modelos ofrecen una buena primera aproximación a los hechos estilizados en los datos financieros, como la volatilidad que cambia con el tiempo. Los modelos GARCH son adecuados para describir ciertos patrones en los datos financieros, tales como Clúster de Volatilidad que significa los periodos de alta y baja volatilidad agrupadas y el Comportamiento de Cola Gruesa, que significa la mayor probabilidad de eventos extremos en comparación con la distribución normal [10].

Pero según Gallant en el año 1997 y Andersen en el año 2002, los modelos GARCH muestran una alta persistencia, lo que significa que la volatilidad tiende a permanecer alta o baja durante periodos prolongados. Estos modelos no son los más adecuados para manejar eventos raros en los retornos de los activos, como grandes caídas o subidas repentinas, lo que nos pueden llevar a una sobreestimación del riesgo de mercado real cuando se enfrentan a saltos o eventos raros en los retornos de los activos, debido a esto se han desarrollado modelos GARCH con propiedades adicionales, como la asimetría, (EGARCH, TGARCH) lo que significa que los impactos diferenciales de choques positivos y negativos son de igual magnitud en la volatilidad, es decir, las subidas y bajadas en los precios de los activos no afectan la volatilidad de la misma manera, y segun Cheong señala que la asimetría es crucial en el análisis del riesgo, especialmente cuando se gestionan inversiones en posiciones largas (comprar activos esperando que suban de valor) y cortas (vender activos esperando que bajen de valor) durante un período determinado, y algo importante es que la asimetría ayuda a entender mejor los comportamientos de las colas inferiores y superiores de la distribución de retornos, lo que es fundamental para evaluar el riesgo en ambos extremos del espectro [11].

Luego Nelson en el año 1991, hizo un importante avance al desarrollar el modelo exponencial GARCH (EGARCH) para analizar los efectos asimétricos en la volatilidad causados por noticias del mercado, así en este modelo, los choques negativos (malas noticias) y positivos (buenas noticias) tienen impactos diferentes en la volatilidad. Específicamente, los choques negativos tienden a aumentar la volatilidad más que los choques positivos de igual magnitud, y también existe la posibilidad que modelo EGARCH incluye un apalancamiento, lo que significa que los choques negativos aumentan la volatilidad mientras que los choques positivos la disminuyen, y así esta característica refleja cómo las malas noticias tienden a tener un mayor impacto en la volatilidad de los mercados financieros que las buenas noticias [12]. Y luego Zakoian en 1994, desarrollo otro modelo GARCH asimétrico, el TGARCH que también consideran la asimetría en los efectos de los choques de mercado sobre la volatilidad.

El estudio de Hung en 2011, destaca la importancia de considerar colas pesadas en el cálculo del VaR al 99 % para posiciones largas, las colas pesadas se refieren a la mayor probabilidad de eventos extremos en la distribución de retornos, lo que es crucial para una estimación precisa del VaR en escenarios de alto riesgo, y así estas investigaciones empíricas han demostrado que incorporar características como

colas pesadas, asimetría y choques dinámicos, al combinar diferentes técnicas de estimación, mejora la eficiencia y precisión de los modelos de VaR en la evaluación del riesgo de mercado [13].

Por último, se ha utilizado el backtesting, una técnica crucial para validar los modelos de VaR. El backtesting evalúa la precisión y conservadurismo de estos modelos al comparar las ganancias y pérdidas diarias reales e hipotéticas con las medidas de VaR generadas por el modelo, según lo definido por el Comité de Supervisión Bancaria de Basilea (BCBS, 2019). Esto permite verificar si el modelo predice adecuadamente el riesgo de pérdidas. Kupiec propuso la aproximación más utilizada en backtesting, conocida como la "prueba de cobertura incondicional". Esta prueba se centra en verificar si el número real de excepciones (días en que las pérdidas exceden el VaR estimado) coincide con el número esperado de excepciones. La "tasa de cobertura teórica", es la proporción de días en que se espera que las pérdidas excedan el VaR según el nivel de confianza del modelo (por ejemplo, 1 % para un VaR al 99 %). Según Laporta 2018, la prueba de Kupiec sigue un proceso de Bernoulli con parámetro p. Este parámetro p representa la tasa teórica de cobertura del VaR. Por ejemplo, si el VaR se calcula para un nivel de confianza del 99 %, entonces p sería 0.01 (1 %). En este contexto, la variable de prueba toma el valor de 1 si la pérdida en un día determinado excede el VaR estimado (indicando una excepción) y 0 en caso contrario [10].

ESTIMACIÓN DEL VAR CON MODELOS GARCH

2.1. Conceptos preliminares. En esta sección, exploraremos conceptos clave como los rendimientos de activos, series de tiempo, estacionariedad, y la prueba Portmanteau.

La mayoría de los estudios financieros utilizan los rendimientos en lugar de los precios de los activos por dos razones principales. En primer lugar, para los inversores promedio, la rentabilidad de un activo proporciona un resumen completo y sin escala de la oportunidad de inversión. En segundo lugar, las series de rendimientos son más fáciles de manejar que las series de precios, ya que presentan propiedades estadísticas más atractivas.

Definamos a P_t el precio de un activo en el índice temporal t . Suponiendo que nuestro activo no paga dividendos [15].

Definición 2.1. El logaritmo natural del rendimiento simple para un período desde la fecha t - 1 hasta t de un activo que se compone en forma continua se determina como [8]:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \approx ln(\frac{P_t}{P_{t-1}}) = ln(P_t) - ln(P_{t-1}).$$

2.1.1. Series de tiempo. Una serie de tiempo o serie temporal es una colección de observaciones $\{r_1, x_2, ..., r_T\}$ que registran el valor de una variable de interés en determinado período de tiempo. La colección $\{r_t\}_{t=1}^{T}$ se suele considerar como una muestra particular de un proceso estocástico $\{r_t\}_{-\infty}^{\infty}$ heredando sus propiedades. Cuando solo una variable varía con el tiempo estamos ante una serie temporal univariante, si se tienen múltiples variables se denomina serie temporal multivariante. En el análisis de series de tiempo el objetivo es extraer parámetros relevantes o características de ella para luego generar un modelo matemático adecuado que describa el proceso y poder hacer un análisis predictivo.

Definición 2.2. Diremos que una serie de tiempo $\{r_t\}_{i=1}^{T}$ es estrictamente estacionaria si

$$(r_{t1}, r_{t2}, ..., r_{tn}) \ \underline{d} \ (r_{t1+k}, r_{t2+k}, ..., r_{tn+k}).$$

Es decir, si la distribución conjunta es invariante bajo translaciones de tiempo. Es difícil verificar esta propiedad en la práctica por lo que se maneja el siguiente concepto.

Observación: $\underline{d}$ quiere decir idéntica a la distribución.

Definición 2.3. Una serie de tiempo $\{r_t\}_{i=1}^{T}$ se dice débilmente estacionaria si la media de $\{r_t\}_{i=1}^{T}$ y la covarianza entre r_t y r_{t-l} son invariantes en el tiempo, es decir,

$$E(r_t) = \mu, \qquad cov(r_t, r_{t-l}) = \gamma_l,$$

donde la covarianza es una función que depende únicamente de l [8].

En resumen, las dos definiciones anteriores nos indican que una serie temporal $\{r_t\}$ es estrictamente estacionaria si sus propiedades estadísticas son invariantes bajo traslaciones en el tiempo. Esto implica que todas sus distribuciones de probabilidad conjuntas son constantes a lo largo del tiempo. Por otro lado, si una serie $\{r_t\}$ tiene sus dos primeros momentos (media y varianza) finitos, entonces es débilmente estacionaria. Sin embargo, el recíproco, es decir, que una serie débilmente estacionaria sea estrictamente estacionaria, no siempre es cierto a menos que $\{r_t\}$ se distribuya normalmente.

Para fines prácticos, se supone que $\{r_t\}$ es débilmente estacionaria, lo cual simplifica el análisis y permite utilizar ciertas propiedades estadísticas. Es importante mencionar que si la serie es débilmente estacionaria, la covarianza entre $\{t_t\}$ y $\{r_{t-1}\}$ depende solo del retraso l, la cual llamaremos $lag - l$.
La covarianza tiene las siguientes propiedades:

- Varianza de $lag - l$: Es la covarianza en la $lag - l$, es γ_0, que es simplemente la varianza de $\{r_t\}$, es decir que $\gamma_0 = var(r_t)$.
- Simetría: La covarianza es simétrica con respecto al lag, es decir $\gamma_{-l} = \gamma_l$ [15].

Definición 2.4. El coeficiente de correlación entre dos variables x, y está definido como

$$\rho_{x,y} = \frac{cov(x,y)}{(var(x)var(y))^{\frac{1}{2}}}$$

Este coeficiente mide la fuerza de dependencia lineal entre x y y. Se puede probar que $-1 < \rho < 1$, $\rho_{x,y} = \rho_{y,x}$ [4].

Algo importante que hay que mencionar es que si la correlación es cero ($\rho_{x,y} = 0$), esto significa que no existe correlación entre dos variables aleatorias x e y, lo que equivale que no hay una relación lineal entre las dos variables. Ahora Si x e y son variables aleatorias normales y su correlación es cero, entonces x e y son independientes. Esto implica que el valor de una de las variables no afecta al valor de la otra. Pero el coeficiente de correlación tiene una suma importancia en las series de tiempo, porque si una serie de tiempo r_t es débilmente estacionaria, la dependencia entre r_t y sus valores pasados es crucial. Esta dependencia ayuda a entender y modelar la estructura y el comportamiento de la serie a lo largo del tiempo.

2.1.2. Prueba de Portmanteau. En finanzas se necesita probar conjuntamente que varias autocorrelaciones de r_t con r_{t-l} son cero, para ello utilizamos el estadístico Q.

$$Q^*(m) = T\sum_{l=1}^{m} \hat{\rho l}^2,$$

con hipótesis nula $H_0 : \rho_1 = ... = \rho_m = 0$ versus hipótesis alternativa $H_a : \rho_i \neq 0$ para $i = 1, 2, 3, ..., m$ donde $Q^*(m) \sim X_i^2$ con m grados de libertad.

En la práctica la selección de m puede afectar el rendimiento del estadśtico $Q(m)$, comúnmente se utiliza $m \approx ln(T)$ donde T es el tamaño de la muestra de r_t.

Definición 2.5. Una serie de tiempo a_t es llamado ruido blanco si $\{a_t\}_{t=1}^T$ es una sucesión de variables aleatorias idénticamente distribuidas con media μ y varianza $\sigma^2 > 0$. En particular supondremos que a_t se distribuye normalmente $N(0, \sigma^2)$, caso que se conoce como ruido blanco Gaussiano [8].

2.2. Valor en riesgo (VaR). Es una medida estadística utilizada para cuantificar el riesgo de pérdida en una cartera de inversiones.

Definición 2.6. Dado un nivel de confianza $\alpha \in (0, 1)$. El VaR de nuestra cartera en el nivel de confidencia α viene dado por el menor número l tal que la probabilidad de que la pérdida L supere l no sea mayor que $(1 - \alpha)$. Formalmente,

$$VaR_\alpha = inf\{l \in \mathbb{R} : P(L > l) \leq 1 - \alpha\} = inf\{l \in \mathbb{R} : F_L(l) \geq \alpha\}.$$

Donde F_L es el cuantil de la distribución de las pérdidas L que correspondiente al nivel de confianza α.

En términos probabilísticos, el VaR es simplemente un cuantil de la distribución de pérdidas [14]. Es importante destacar que el VaR se basa en tres componentes clave:

- Horizonte Temporal: El período durante el cual se estima la posible pérdida. Comúnmente se utilizan horizontes de un día, una semana, 10 días o un mes.
- Nivel de Confianza: La probabilidad con la que se asegura que la pérdida no excederá el VaR estimado. Los niveles de confianza comunes son 95 %, 99 % y 99.9 %.
- Magnitud de la Pérdida: El monto de la pérdida que no se espera superar más allá del nivel de confianza durante el horizonte temporal.

2.3. Modelos GARCH, sus Extensiones y volatilidad. En esta sección explicaremos los conceptos clave, como ser volatilidad, y los modelos GARCH, EGARCH, TGARCH. La volatilidad es conocida como una medida de la variabilidad de los precios de los activos financieros. Modelar la volatilidad es crucial para la estimación precisa del VaR.

2.3.1. Volatilidad. Es una medida de variación o dispersión de una variable de interés a lo largo del tiempo, y se define como la varianza condicional de la serie.

Cuando el precio del activo varía regularmente en plazos cortos, se dice que tiene alta volatilidad, si el precio se mantiene casi constante entonces tiene baja volatilidad, estos períodos de variabilidad indican heterocedasticidad en el proceso, es importante observar que la volatilidad no es posible de forma directa, sin embargo, tiene las siguientes características:

- Existen agrupamientos de volatilidad.
- Evoluciona con el tiempo de manera continua.
- No diverge al infinito, se puede decir que la volatilidad es estacionaria.
- Parece reaccionar de manera diferente a un gran aumento de precios o una gran caída de precios [16].

El modelo ARCH (Engle 1982) es el primer modelo que proporciona un marco de referencia sistemático para modelar la volatilidad. La idea básica de los modelos ARCH es que el ruido r_t de un activo es serialmente no correlacionado, pero dependiente, la dependencia de r_t puede ser descrita por una función cuadrática de sus valores rezagados [3].

Los modelos GARCH se generalizan a partir de los modelos ARCH, a pesar que el modelo ARCH es un modelo simple, usualmente requiere muchos parámetros para describir adecuadamente el proceso de la volatilidad de un conjunto de retornos. Por eso es necesario introducir un nuevo modelo heteroscedástico generalizado.

2.3.2. GARCH (Generalized Autoregressive Conditional Heteroskedasticity). Propuesto por Bollerslev (1986), captura la persistencia en la volatilidad mediante la incorporación de términos autoregresivos y de media móvil en la varianza condicional. Un modelo GARCH de parámetros p y q o GARCH(p, q), se puede representar como:

$$r_t = \sigma_t \epsilon_t \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

donde $\{\epsilon_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con media cero y varianza 1, usualmente se asume que ϵ_t se distribuye normal o t-student estándar, $\alpha_0 > 0$, y $\alpha_i \geq 0$, para $i \in \{1, 2, ..., p\}$ garantizando que la varianza condicional σ_t^2 sea positiva [3] .

2.3.3. EGARCH (Exponential GARCH). Introducido por Nelson (1991), modela la volatilidad logarítmica y captura la asimetría en los shocks de mercado. En particular, esta extensión fue creada para tener en cuenta los efectos asimétricos entre los rendimientos positivos y negativos de los activos, considerando la innovación ponderada,

$$g(\epsilon_t) = \theta \epsilon_t + \gamma[|\epsilon_t| - E(|\epsilon_t|)],$$

donde θ y γ son constantes. Pero ϵ_t y $|\epsilon_t| - E(|\epsilon_t|)$ son idénticamente distribuido con media cero con distribuciones continuas, así $E[g(\epsilon_t)] = 0$ La asimetría de $g(\epsilon_t)$ puede verse fácilmente reescribiéndola como,

$$g(\epsilon_t) = \begin{cases} (\theta + \gamma)\epsilon_t - \gamma E(|\epsilon_t|) & si \quad \epsilon_t \geq 0, \\ \\ (\theta - \gamma)\epsilon_t - \gamma E(|\epsilon_t|) & si \quad \epsilon_t < 0. \end{cases}$$

2.3.4. TGARCH (Threshold GARCH). Propuesto por Zakoian (1994), permite que la volatilidad reaccione de manera diferente a shocks positivos y negativos, reflejando sobre la volatilidad. Un modelo TGARCH(m, s) asume la forma

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{s} (\alpha_i + \gamma_i N_{t-i}) a_{t-i}^2 + \sum_{j=1}^{m} \beta_j \sigma_{t-j}^2,$$

donde N_{t-i} es un indicador de negativo a_{t-i}, asi que

$$N_{t-i} = \begin{cases} 1 & si \quad a_{t-i} < 0, \\ \\ 0 & si \quad a_{t-i} \leq 0, \end{cases}$$

y α_i, γ_i, y β_j son parámetros no negativos que cumplen condiciones similares a las de los modelos GARCH [15].

2.4. Distribuciones de Retornos de Activos. En esta sección explicaremos las definiciones de los retornos de distribución normal y t-student, ambas son ampliamente utilizadas en modelos financieros por su simplicidad y propiedades matemáticas favorables, subestiman la probabilidad de eventos extremos debido a que los retornos financieros muestran colas más gruesas y picos más altos de lo esperado, y la distribución t-Student, con sus grados de libertad ajustables, se adapta mejor a estas características empíricas [3].

Definición 2.7. Distribución Normal: En los modelos GARCH, se suelen asumir que los errores ϵ_t siguen una distribución normal con media cero y varianza condicional $\sigma_t^2 : \epsilon_t \sim N(0, \sigma_t^2)$, esto implica que $\epsilon = \sigma_t z_t$, donde z_t son variables aleatorias independientes e idénticamente distribuidas con una distribución normal estándar $N(0, 1)$.

Definición 2.8. Distribución t-Student: Los errores σ_t se modelan de la siguiente manera, $\epsilon_t = \sigma_t z_t$, donde z_t son variables aleatorias independientes e idénticamente distribuidas, que siguen una distribucion t de student estándar con v grados de libertad [15].

2.5. Aplicación a series financieras. En esta sección se realizarán cálculos numéricos y la modelación de la serie de tiempo. El análisis de los rendimientos del activo incluirá la generación de histogramas y la estimación de funciones de densidad. A partir de estas funciones de densidad, se determinará qué tipo de distribución se ajusta mejor a los datos. Se llevó a cabo la modelación de la volatilidad financiera en el retorno del precio de las acciones del café, cotizadas en la bolsa de Nueva York con la etiqueta KC=F. Los datos se obtuvieron de Yahoo Finance, con información diaria desde el 15-07-12 al 15-07-24, obteniéndose así una base de datos de 3017 observaciones. Nuestra variable de interés de la base de datos fue $KC = F\$Adjusted$, que representa el precio al que se cerró al final del día.

En esta sección las referencias utilizadas son, [3], [8], [15], [17].

2.5.1. Estadística descriptiva. A continuación, se presentan los datos más relevantes de nuestra serie de rendimientos del activo. Se incluyen los estadísticos descriptivos de los rendimientos logarítmicos diarios para el periodo comprendido entre el 15 de julio de 2012 y el 15 de julio de 2024.

CUADRO 1. Momentos estadísticos para los retornos del precio café.

Media	D.S	Min	Varianza	Maximun	No.Obs	Asimetría	Curtosis	Prob.JB
0.001	0.02120	-0.0902	0.00045	0.117	$3,017$	0.2825	23.736	$2.2e-16$

La media positiva y la desviación estándar moderada sugieren una ligera tendencia alcista en el precio del café con una volatilidad diaria del 2.12 %, y los valores extremos (mínimo y máximo) indican que hay tanto riesgos significativos de caídas (hasta el 9.02 %) como oportunidades para ganancias grandes (hasta el 11.7 %), también tenemos una asimetría positiva que indica una ligera inclinación a los rendimientos positivos, y la alta curtosis indica la presencia de eventos extremos, que es crucial para evaluar el riesgo. Por ultimo tenemos la prueba de Jarque-Bera que sugiere que los rendimientos no siguen una distribución normal, reforzando la necesidad de modelos que puedan capturar mejor la naturaleza de los datos financieros, como los modelos GARCH con distribución t de Student.

CUADRO 2. Percentiles y Cuantiles

10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
−0.024	−0.0155	−0.00997	−0.0051	0.000336	0.0040	0.0092	0.0156	0.0262

Los percentiles inferiores (10 %, 20 %, 30 %) muestran rendimientos negativos, sugiriendo que hay una mayor frecuencia de días con pérdidas menores (más del 1 % y hasta 2.4 %). Los percentiles superiores (70 %, 80 %, 90 %) muestran rendimientos positivos, sugiriendo que hay una mayor frecuencia de días con ganancias menores (menos del 1 % y hasta 2.62 %).Y el porcentil del 10 % de los días han tenido caídas superiores al 2.4 %, lo que destaca el riesgo de caídas significativas en el precio del café y el porcentil del 90 % muestran ganancias superiores al 2.62 %, indicando oportunidades de rendimientos altos en ciertos días. En conclusión, la interpretación de los percentiles muestra que la distribución de los rendimientos logarítmicos está generalmente centrada en valores cercanos a cero. Sin embargo, esta distribución también revela tanto riesgos significativos de caídas como oportunidades considerables de ganancias en ciertos días.

En la Figura 1 siguiente se representa la evolución del índice a lo largo del tiempo, y a pesar de la volatilidad diaria, es posible identificar una tendencia ascendente a lo largo de varios años, esto podría indicar un aumento general en los precios del café.

FIGURA 1. Serie diaria del índice precio del café

En la Figura 2 se muestran los rendimientos diarios que proporcionan una visión detallada de cómo los precios del café cambian día a día, permitiendo a los analistas e inversores evaluar la volatilidad y el riesgo del mercado.

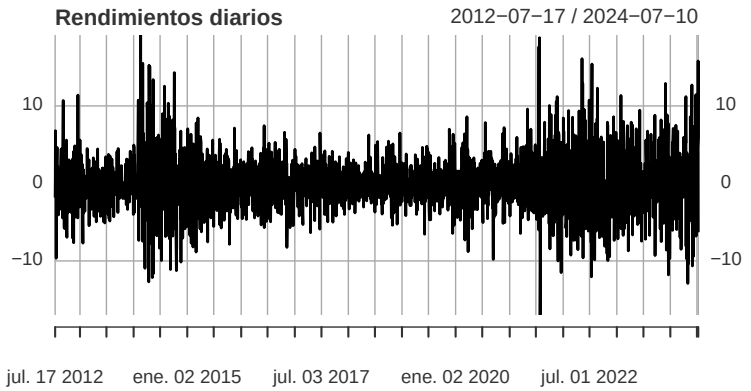


FIGURA 2. Rendimientos diarios del precio del café.

2.5.2. Histograma, la función de densidad empírica, ACF y PACF. Examinamos la serie con los correlogramas para identificar el tipo de proceso ARIMA que podría representar a la serie, por otra parte antes de hacer una estimación es importante también examinar la serie que estamos utilizando, tanto en niveles como en primeras diferencias para ver si esta serie es estacionaria o no, lo podemos hacer a través de correlogramas y básicamente es a través de la función de autocorrelación y la función de autocorrelación parcial. Si apreciamos la función de autocorrelación en la Figura 4, vemos que va cayendo lentamente y nunca muere lo cual es un patrón muy característico en series no estacionarias. También podemos observar en la Figura 3, el histograma y la función de densidad y podemos apreciar que se asemeja a una distribución normal pero tiene una pequeña asimetría positiva

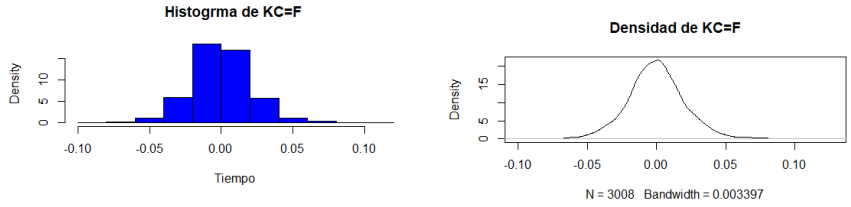


FIGURA 3. Histograma y función de densidad de los retornos observados.

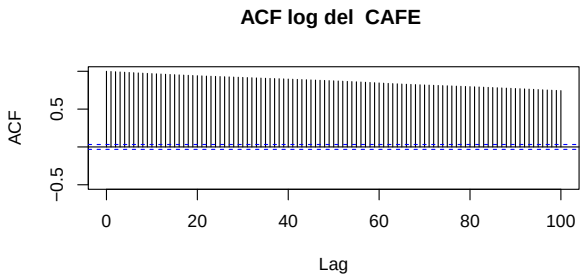


FIGURA 4. ACF

Sin embargo cuando graficamos ACF y PACF en primera diferencias podemos apreciar que rápidamente caen y los coeficientes de correlación se vuelven igual a cero.

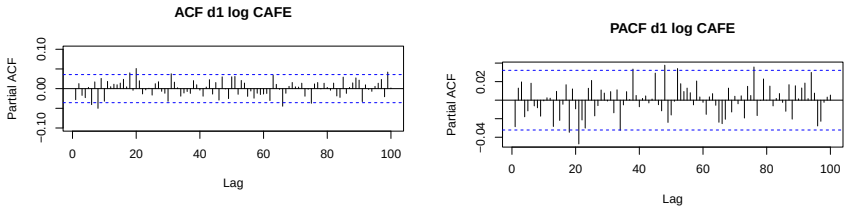


FIGURA 5. ACF Y PACF en primera diferencia.

2.5.3. Análisis de raíces unitarias o pruebas de estacionariedad. Realizamos las pruebas formales para verificar si una serie es estacionaria o no, tanto en niveles como en diferencias, que es a travéz de las pruebas de raíces unitarias.

1. Prueba de la raíz unitaria de Dickey-Fuller aumentada (ur.df).
2. Prueba de raíz unitaria de Phillips y Perron (ur.pp).
3. Prueba de raíz unitaria de Kwiatkowski et al (ur.kpss).

En el siguiente cuadro mostramos los resultados de las pruebas de raíz unitaria con un nivel de significancia 0.01^{***}.

CUADRO 3. Pruebas de raíz unitaria para el precio del café (log).

Prueba	Niveles	Primeras diferencia
ur.df	-2.088	-18.3783^{***}
ur.pp	-1.2748	-56.4486^{***}
ur.kpss	-1.3382	0.0696^{**}

Los resultados del Cuadro 3 indican que la serie de precios del café en logaritmos no es estacionaria en niveles, pero se vuelve estacionaria después de tomar la primera diferencia.

2.5.4. Ajuste del modelo autoregresivo. Para estimar un modelo ARIMA, podemos utilizar la metodología de Ljung-Box u otros métodos alternativos, especificando los órdenes correspondientes. Otra opción es emplear el paquete seasonal de R, que selecciona el mejor modelo ARIMA y sugiere posibles valores atípicos. Una función más rápida es el auto.arima, que selecciona o recomienda el modelo más adecuado. En nuestro caso, auto.arima recomendó un modelo ARIMA(0,0,1), que no tiene términos autorregresivos, pero si incluye un término de media móvil y la variable es estacionaria después de una diferencia [8].

Es crucial extraer los residuos de este modelo y someterlos a un análisis exhaustivo. Inicialmente, realizamos pruebas de normalidad para los residuos utilizando los tests de Portmanteau ($p - valor = 2.2 \times 10^{-16}$) y Shapiro-Wilk ($p - valor = 2.2 \times 10^{-16}$), ambos tests indican que los valores $p - valores$ son significativamente menores al 5 %, lo que nos permite concluir que los retornos no siguen una distribución normal. Debido a falta de normalidad, es necesario considerar una estimación ARCH. Aplicamos pruebas de heterocedasticidad autoregresiva y condicional con rezagos de 1 y 2, respectivamente. En ambos casos, los $p - valores$ fueron menores al 5 % y al 10 %, lo que nos lleva a rechazar la hipótesis nula de homocedasticidad. Esto sugiere que en nuestro modelo ARIMA existe y tiene un problema de heterocedasticidad que es autoregresivo y condicional. Ante esta situación, debemos proceder a estimar modelos de las familias GARCH para manejar adecuadamente la heterocedasticidad presente. Luego graficamos los residuos al cuadrado.

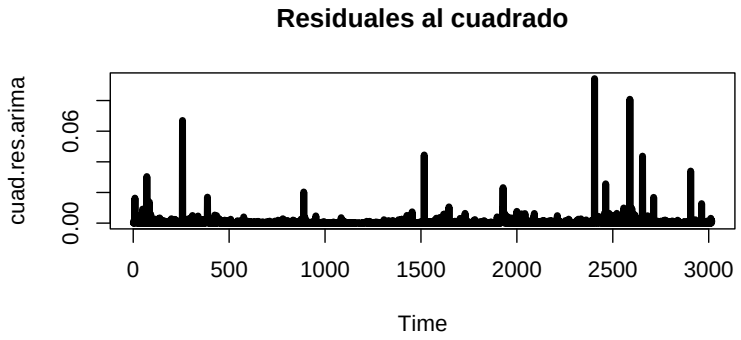


FIGURA 6. Residuos al cuadrado.

En la Figura 6, observamos períodos de alta variabilidad seguidos de períodos de tranquilidad, y viceversa. Este patrón es característico de la heterocedasticidad condicional en series de tiempo. Debido a esta característica, debemos analizar nuestra serie de tiempo utilizando la familia de modelos GARCH, considerando tanto los modelos GARCH simétricos como los asimétricos. El proceso GARCH es válido cuando los residuos al cuadrado están correlacionados y los gráficos ACF y PACF indican claramente una correlación significativa.

2.5.5. GARCH simémetrico. Estimamos el modelo GARCH estándar, con dos tipos de distribución normal y t-student, con diferentes ordenes y vamos a mostrar el mejor modelo que se ajusta obteniendo los siguientes resultados. Con un nivel de significancia 0.01^{***} y 0.05^{**}.

	Dist.Normal	Dist.t-student	Dist.t-student
Parametro	GARCH(1,1)	GARCH(1,0)	GARCH(1,1)
μ	0.001092^{**}	0.001318^{***}	0.001398^{***}
ϕ_1	0.7665^{***}	0.7249^{***}	0.7631^{***}
θ_1	-0.8040^{***}	-0.7608^{***}	-0.8087^{***}
ω	0.000002	0.00053^{***}	0.000003
α_1	0.0181^{***}	0.2196^{***}	0.0294^{***}
β_1	0.9794^{***}		0.9651^{***}
η		3.05438^{***}	3.7133^{***}

CUADRO 4. Estimaciones simétricas: Familias GARCH.

En el modelo GARCH(1,1) con especificación de distribución normal para las innovaciones, se observó que presentaba problemas de volatilidad condicional no estacionaria ($\alpha_1 + \beta_1 \approx 1$).

Tanto GARCH(1,1) como GARCH(1,0) con una especificación de distribución t de student capturan adecuadamente la volatilidad de los retornos del café, pero la persistencia de la volatilidad (valores altos de β_1) indica que los choques a la volatilidad tienen efectos prolongados y la condición del test ARCH con 3,5,7 rezagos resultaron significativos lo cual significa que se ha resuelto el problema de heterocedasticidad condicional.

2.5.6. GARCH asimétricos. EN los modelos TGARCH, EGARCH asimétricos el supuesto fundamental es considerar el efecto de malas noticias, persistencia de choques o volatilidades diferenciables. Obteniendo los siguientes resultados. Con un nivel de significancia 0.01^{***} y 0.05^{**}.

	Dist.t-student	Dist.t-student
Parametro	TGARCH(1,1)	EGARCH(1,1)
μ	0.001226^{***}	0.001256^{***}
ϕ_1	0.756985^{***}	0.760497^{***}
θ_1	-0.796329^{***}	-0.801910^{***}
ω	0.000199^{*}	-0.068580^{***}
α_1	0.063679^{***}	-0.036694^{***}
β_1	0.946135^{***}	0.991257^{***}
η_1	0.394699^{***}	
γ_1		0.101368^{***}

CUADRO 5. Estimaciones asimétricas: Familias GARCH.

En el modelo TGARCH(1,1) existe efecto de asimetría o malas noticias que incrementan la volatilidad condicional de forma positiva y estadísticamente significativa η_1. La volatilidad condicional fue de tipo no estacionaria ($\alpha_1 + \beta_1 > 1$).

En el modelo EGARCH(1,1) todos los parámetros fueron estadísticamente significativos y con varianza condicional estacionaria ($\alpha_1 + \beta_1 < 1$). También podemos ver que $\alpha_1 < 0$, significa que los choques positivos (buenas noticias) generan menor volatilidad en comparación con las malas noticias, y γ_1 muestra un efecto asimétrico en la volatilidad con respecto a choques negativos y positivos.

2.5.7. Selección del mejor modelo. Comparar y seleccionar modelos basados en criterios de información es una práctica estándar en el análisis de series de tiempo y otros contextos estadísticos. Al ajustar varios modelos y utilizar criterios como AIC, BIC, AICc y HQC, se puede identificar el modelo que ofrece el mejor equilibrio entre ajuste y complejidad. Esto es crucial para evitar el sobreajuste y mejorar la capacidad predictiva. Seleccionaremos el modelo con los valores más bajos en estos criterios de información, ya que se considera el mejor en términos de balance entre ajuste y simplicidad.

	GARCH simétricos			GARCH simétricos	
	Norm	t-student	t-student	t-student	t-student
IC	GARCH(1,1)	GARCH(1,0)	GARCH(1,1)	TGARCH(1,1)	EGARCH(1,1)
AIC	-4.6506	-4.9050	-4.9934	-5.0133	-5.0137
BIC	-4.6386	-4.8930	-4.9795	-4.9973	-4.9977
AICc	-4.6506	-4.9050	-4.9934	-5.0133	-5.0137
HQC	-4.6463	-4.9007	-4.9884	-5.0075	-5.0079

CUADRO 6. Criterios de Información (IC).

Una vez elegido el mejor modelo a través de los criterios de información, y verificando que cumple todas las restricciones, incluidos los signos de la varianza condicional, determinamos que el modelo EGARCH es el más adecuado entre las alternativas simétricas y asimétricas. Este tipo de modelo es frecuentemente recomendado en la literatura para series financieras. Con el mejor modelo EGARCH seleccionado, procedemos a realizar pronósticos de los retornos del precio del café [8].

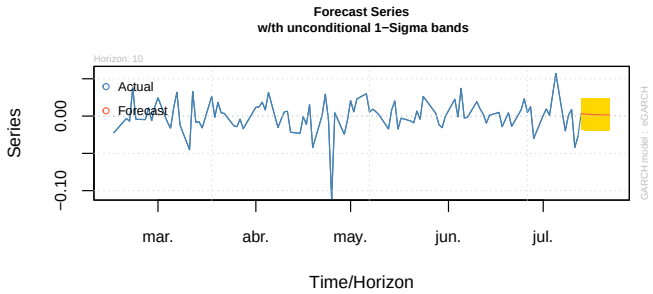


FIGURA 7. Predicción de la variabilidad de los precios en 10 días hacia adelante.

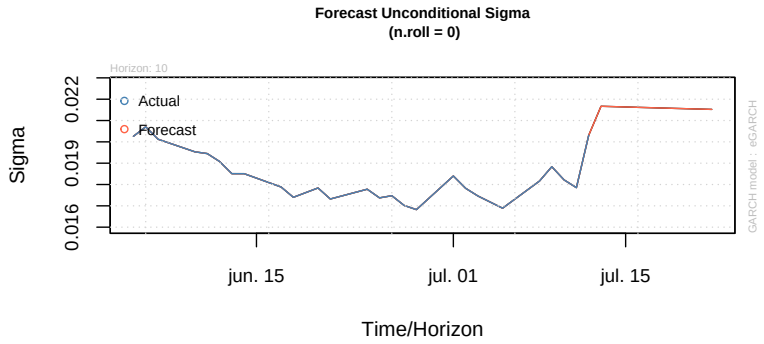


FIGURA 8. Predicción de la volatilidad incondicional en 10 días hacia adelante.

En la Figura 9 se presenta la volatilidad condicional histórica, medida por la desviación estándar de la ecuación de varianza y en comparación con la volatilidad incondicional.

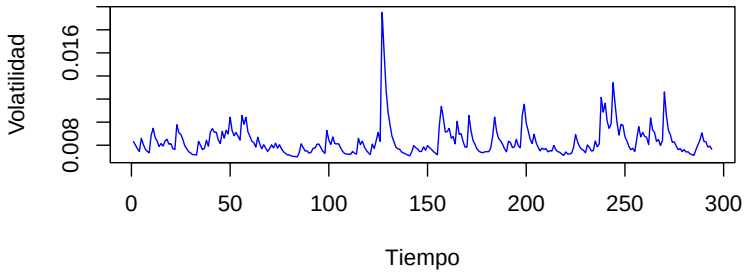


FIGURA 9. Volatilidad histórica condicional, modelo EGARCH.

2.6. Calculo del VaR. Se calculó el VaR a un día con diferentes niveles de significancia y tamaños de muestra. Es importante destacar que el modelo EGARCH, con un intervalo de confianza del 99 % y una muestra de 1000 datos, indicó la mayor pérdida esperada en comparación con los demás tamaños de muestra.

	VaR 95 %			Var 99 %	
Muestra	500	1000	1500	500	1000
GARCH(1,1)	-0.038142	-0.021701	-0.032179	-0.053945	-0.050137
TGARCH(1,1)	-0.036410	-0.037956	-0.035688	-0.051495	-0.049973
EGARCH	-0.035642	-0.039655	-0.039901	-0.050409	-0.058921

CUADRO 7. VaR con diferentes tamaños de muestra.

En la Figura 10 se muestra el VaR al 1 % de significancia con el mejor modelo seleccionado EGARCH, donde la línea rojo significa el escenario adverso y la línea verde es el escenario a favor [10].

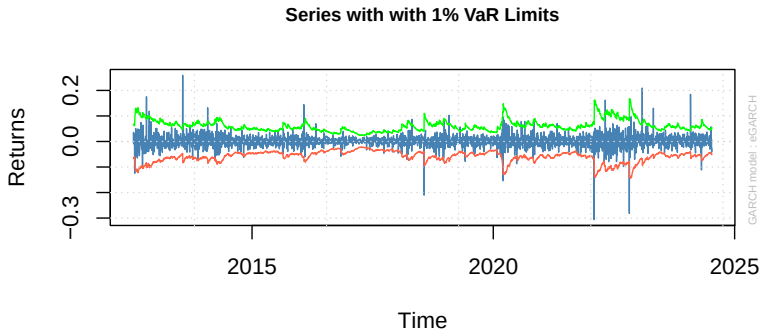


FIGURA 10. VaR al 1 %, modelo EGARCH.

3. CONCLUSIONES

A continuación, se resume el trabajo realizado y los hallazgos observados durante la experimentación.

- En el contexto del análisis financiero, especialmente en modelos como los GARCH o cualquier modelo que requiera estimaciones precisas de la volatilidad y otras características del rendimiento, una cantidad insuficiente de datos puede llevar a estimaciones imprecisas, lo que a su vez afecta a la toma de decisiones. Por lo tanto, contar con suficientes observaciones es crucial para la validez del análisis.
- En la investigación se concluye que, al estimar un modelo univariado (ARIMA) que presenta problemas de heterocedasticidad condicional autorregresiva (ARCH), es necesario abordar este problema mediante la inclusión de una estimación simultánea de una ecuación de la media y una ecuación de la varianza condicional, como los modelos GARCH(p, q). La literatura sugiere dos enfoques para esto, las familias estándar de GARCH o versiones simétricas, y las versiones asimétricas
- Uno de los aspectos más interesantes del análisis de series de tiempo es la predicción de futuros valores de los rendimientos. Sin embargo, es crucial considerar que la precisión de estas predicciones disminuye a medida que nos alejamos de los datos reales. En el contexto del índice del precio del café, esto significa que, aunque los modelos GARCH pueden ofrecer estimaciones útiles a corto plazo, las predicciones a largo plazo son cada vez menos precisas debido a la creciente incertidumbre.
- Con base en la evidencia empírica se encontró que la especificación de los modelos asimétricos EGARCH(1,1) con presencia significativa de efecto apalancamiento, es decir las malas noticias generan un aumento a ala volatilidad de los precios.

- Resulta impactante la facilidad para calcular el VaR del índice del precio del café, ya que el modelo GARCH univariado proporciona todos los resultados necesarios. En este caso, lo más importante es obtener una estimación precisa de la volatilidad de los rendimientos.
- R es una poderosa herramienta para la modelación de series financieras, gracias a paquetes como rmgarch, rugarch, seasonal, fgarch y tseries, los cuales son de libre descarga.

Como trabajo futuro, se propone explorar especificaciones alternativas de la familia de modelos GARCH simétricos y asimétricos, utilizando distribuciones como la t de Student asimétrica, la distribución de errores generalizados, y la distribución de errores generalizados asimétricos, para comparar resultados.

.

Referencias

1. Jorion, Philippe. *Value at risk: the new benchmark for managing financial risk.* McGraw-Hill, 2007.
2. Jeřábek, Tomáš. *The Efficiency of GARCH Models in Realizing Value at Risk Estimates.* Acta VŠFS-ekonomické studie a analýzy 14.1 (2020): 32-50.
3. Francq, Christian, and Jean-Michel Zakoian. *GARCH models: structure, statistical inference and financial applications.* John Wiley & Sons, 2019.
4. P. Jorion. *Value at Risk. The new benchmark for controlling derivatives risk.* McGraw Hill,
5. Cerović Smolović, Julija, Milena Lipovina-Božović, and Saša Vujošević. *GARCH models in value at risk estimation: empirical evidence from the Montenegrin stock exchange.* Economic research-Ekonomska istraživanja 30.1 (2017): 477-498
6. Bob, Ngoga Kirabo. *Estimación del valor en riesgo. Un enfoque garch-evt-copula.* Mathematiska institutionen (2013): 1-41.
7. Angelidis, Timotheos, Alexandros Benos, and Stavros Degiannakis. *The use of GARCH models in VaR estimation.*
8. González Escamilla, Jesús. *Modelos GARCH multivariados aplicados al cálculo del VaR.* 2020, https://doi.org/10.24275/uami.r494vk45.
9. P. Zangari, *An improved methodology for measuring VAR, RiskMetrics Monitor.* Reuters/JP Morgan, 1996.
10. Barbosa Camargo, María Ines, Alejandra Salazar Sarmiento, and Kelly Jhohana Peñaloza Gómez. *Valoración de riesgo mediante modelos GARCH y simulación Montecarlo: evidencia del mercado accionario colombiano.* Semestre económico 22.53 (2019): 53-75.
11. Gallant, Ronald, Hsieh, David y Tauchen, George. *Estimation of stochastic volatility models with diagnostics.* En: Journal of Econometrics, N°. 81, (1997). p. 159-192.
12. Nelson, Daniel. *Conditional heteroscedasticity in asset returns: a new approach* En: Econometrica, N°. 59, (1991). p. 347-370.
13. Su, Jung-Bin y Hung, Jui-Cheng. *Empirical analysis of jump dynamics, heavy-tails and skewness on value-at-risk estimation.* En: Economic Modelling, N°. 28, (2011). p. 1117-1130.
14. McNeil, Alexander J., Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition.* Princeton university press, 2015.
15. Tsay, Ruey S. *Analysis of financial time series.* John Wiley and Sons, Chicago, Illinois, 2002.
16. Daniel Vasquez. *Analisis De Series De Tiempo Con Metodos Estadisticos Y MACHINE LEARNING* Boletín Divulgativo - Escuela de Matemática y Ciencias de la Computación. 2024.p.6-24
17. ALI, G. *EGARCH, GJR-GARCH, TGARCH, AVGARCH, NGARCH, IGARCH and APARCH models for pathogens at marine recreational sites.* Journal of Statistical and Econometric Methods, (2013), 57-73.
18. Romero, Pilar Abad, Sonia Benito Muela, and CARMEN LÓPEZ Martin. *A comprehensive review of value at risk methodologies.* Documentos De Trabajo 711.1 (2013).

Departamento de Matemáticas, Universidad Nacional Autónoma de Honduras
Dirección actual: Departamento de matemáticas, Universidad Nacional Autónoma de Honduras
Dirección de correo electrónico: jlmartinez@unah.hn

PROGRAMACIÓN FUNCIONAL, UNA APLICACIÓN DE LA TEORÍA DE CATEGORÍAS

JARED MONTECINOS

Resumen. La teoría de categorías ha sido tomada como algo confuso u oscuro en la comunidad de matemática, ha sido hasta llamada un "sinsentido abstracto". En este breve artículo, nos proponemos a mostrar una aplicación en la programación funcional con el lenguaje de programación Haskell

Abstract. Category Theory has been deemed as something confusing or dark in the mathematics community, it has even been called "abstract nonsense". In this short article, we propose ourselves to show an application to functional programming with the Haskell programming language

1. Introducción

La teoría de categoría ha sido tomada primordialmente como una herramienta en matemáticas, pero en realidad es un área bastante sólida que ayuda a conectar, o entender distintas áreas de las matemáticas que al principio parecen disconexas. Con una simple y elegante base, construida sobre los objetos respectivos y morfismos, lo que podemos entender como los elementos más primitivos de las matemáticas, la teoría de categoría puede aplicarse a cualquier área de las matemáticas, sea teoría de conjuntos, topología, geometría diferencia, álgebra abstracta entre otros. [6]
La teoría de categorías da una "gramática" al lenguaje de las matemáticas, lo cual nos dá resultados increíbles donde esta ya no sólo se aplica y crece en la matemática, sino también en otras ciencias, como la física, la neurología [13] o la lingüística [3].
Para este artículo, nosotros nos enfocaremos en su aplicación para las ciencias de la computación en la programación funcional, y especificamente con el lenguaje de programación Haskell. Un concepto clave en la teoría de categoría, es el de composición de funciones, algo que es bastante visible en los diagramas que se usan para estudiarla, este concepto aplicado a la programación, al estudio de polimorfismo entre otros elementos, más el cálculo lambda, nos da la programación funcional.
La base principal de la teoría de categorías en la programación es el tratamiento de los tipos(tipos de dato) como objetos y las funciones como morfismos, esto le da una flexibilidad enorme y ventajas a los lenguajes de programación funcional sobre los lenguajes de tipo imperativo en aspectos de ahorro de memoria o complejidad computacional.

En el contexto de la UNAH, dadas las aplicaciones industriales de la programación funcional, es realmente importante para el sector empresarial del país en el contexto de la globalización.

Palabras y frases clave. Teoría de categoría, Haskell, Programación funcional, Cálculo lambda, Computación.

2. Antecedentes

2.1. **Teoría de Categoría.** Para esta sección, nos basaremos en [8].
La teoría de categoría surge de una manera compleja y fascinante, tanto así que cubrir su origen con detalles es meritorio de un artículo entero, así que aquí cubriremos una entrada con detalles generales. Las categorías, los functores y las transformaciones naturales aparecen por primera vez en un artículo de Eilenberg y McLane (1942), pero restringido a teoría de grupos. Fué la generalización de estos conceptos lo que los llevó a la teoría de categorías.

Al principio fué más un lenguaje util que una teoría completa, mostrando resultados satisfactorios en algebra homológica o topología algebraica, esto cambió hasta 1957 cuando Groethendieck escribe un artículo donde construye y define teorías más generales para luego aplicarlas a la topología algebraica.

En resumidas cuentas, Grothendieck mostró cómo desarrollar parte del álgebra homológica en un entorno abstracto basado en categorías. Esto permitió ver una categoría específica de estructuras, por ejemplo, una categoría de haces sobre un espacio topológico, como un representante de una categoría abstracta de cierto tipo (por ejemplo, una categoría abeliana). De esta manera, se hizo evidente cómo aplicar métodos como el álgebra homológica a otros campos, como la geometría algebraica.

De igual forma Kan en 1958 aplica categorías a límites y colímites, brindando importantes resultados en teoría de homotopías.

En la década de los 60's gracias a los esfuerzos de Freyd y Lawvere, los teóricos de las categorías reconocieron la importancia del concepto de funtores adjuntos. La existencia de adjuntos no solo permite definir categorías abstractas, sino que, como se mencionó anteriormente, muchos teoremas e incluso teorías en diversos campos son equivalentes a la existencia de funtores específicos entre categorías particulares. Para principios de la década de 1970, el concepto de funtores adjuntos se consideraba central en la teoría de categorías.

Estos avances permitieron que la teoría de categorías se consolidara como un campo de investigación autónomo, dando lugar a la "teoría de categorías pura". Experimentó un rápido crecimiento como disciplina y en sus aplicaciones, principalmente en sus áreas de origen (topología algebraica y álgebra homológica), pero también en geometría algebraica y álgebra universal (impulsada por la tesis doctoral de Lawvere).

Dicha tesis es un hito histórico, ya que Lawvere propuso la categoría de categorías como base para la teoría de categorías, la teoría de conjuntos y, por tanto, toda la matemática. También planteó el uso de categorías para estudiar los aspectos lógicos de las matemáticas.

A lo largo de la década de 1960, Lawvere sentó las bases para un enfoque completamente nuevo de la lógica y los fundamentos de las matemáticas. Sus logros incluyen:

- Axiomatizar la categoría de conjuntos (1964) y la categoría de categorías (1966).
- Caracterizar las categorías cerradas cartesianas y mostrar sus conexiones con los sistemas lógicos y diversas paradojas lógicas (1969).
- Demostrar que los cuantificadores y los esquemas de comprensión podrían capturarse como funtores adjuntos a operaciones elementales dadas (1966, 1969, 1970, 1971).

- Defender que los funtores adjuntos deberían desempeñar un papel fundamental a nivel fundacional a través de la noción de "doctrinas categóricas" (1969).

Todo este trabajo culminó en un concepto: el de topos. Un topo es como una categoría en los conjuntos con una estructura lógica para cubrir los elementos básicos de las matemáticas, y generalizado como espacio topológico, provee una conexión entre la lógica y la geometría.

Los topos son un concepto importantísimo en matemática, se les considera "Las Algebras de Lie del siglo XXI" teniendo aplicaciones no sólo en áreas de la matemática, sino también en la metamatemática, siendo importantes para la construcción de modelos en el intuicionismo y el constructivismo.

La teoría de categorías sin dudas, trae mucha riqueza a las matemáticas pero no sólo a estas, sino que ha encontrado multiplicidad de aplicaciones como serlo en teoría cuántica de campos y lo que trataremos en este documento, la teoría computacional.

2.2. **Programación funcional.** Para esta sección nos basaremos en [14]
El antecesor de la programación funcional es el cálculo Lambda creado por Alonzo Church, donde en su tesis desea definir un cálculo que pueda expresar el comportamiento de las funciones. Esto es una herramienta muy útil en matemáticas cuando queremos estudiar la composición de funciones, pero a su vez en ciencias de la computación por todas las aplicaciones que giran alrededor de este concepto.

A la hora de diseñar un lenguaje de programación, hay múltiples elementos en semántica que hay que tomar en cuenta, y es de hecho alrededor del análisis de esta semántica, sobretodo en el análisis de funciones polimórficas donde la teoría de la categoría toma importancia en la programación.

El primer lenguaje de programación en destacar fué LISP(List Processing) creado por John McArthy en la década de los 50's para el manejo de la computación simbólica, algo clave para la inteligencia artificial, dentro de los aspectos claves de Lisp y los más importantes de la programación funcional en general se encuentran: el uso de funciones de primera clase, la recursividad , la homoiconocidad y sobretodo su cambio de paradigma de ser un lenguaje imperativo a uno basado en el uso de funciones y composición de funciones.

Luego de LISP, el siguiente gran lenguaje fué ML que surgió cercanamente al tema de nuestro artículo, Haskell. ML (Meta Language) fué desarrollado por Robert Milner como un meta-lenguaje para probar teoremas, pero rápidamente tuvo su evolución hacia un lenguaje de programación completo, además de las características de LISP, tiene un fuerte sistema basado en tipos, y tipos de dato inmutables.

Haskell, diseñado por Haskell Curry en los 80's es el primer lenguaje puramente funcional, con todas las características mencionadas y dos nuevas: la evaluación perezosa(una expresión no es evaluada hasta que se necesite) y el concepto de mónadas, donde se maneja enormemente la esencia de la programación funcional.

Actualmente, más allá de los entornos académicos la programación funcional ha sido integrada a la industria por sus múltiples ventajas en el manejo de memoria, siendo untilizada por compañías de todo tipo [11].

3. Cuerpo del trabajo

3.1. **Nociones básicas de teoría de la categoría.** En esta sección, primero trabajaremos con definiciones y ejemplos de teoría de la categoría,seguiremos las definiciones y la construcción de [12]

Para una categoría C tenemos:

- Sea Ob(C) una clase, los elementos de esta clase son llamados objetos.
- Para A,B ∈ Ob(C); Mor(A,B) o C(A,B) denota el conjunto morfismo de A a B.
- Para todo A ∈ Ob(C),

$$1_A : A \longrightarrow A$$

denota el morfismo unidad.

- Para morfismos $f : A \longrightarrow B$ y $g : B \longrightarrow C$, el único morfismo de composición de f y g es:

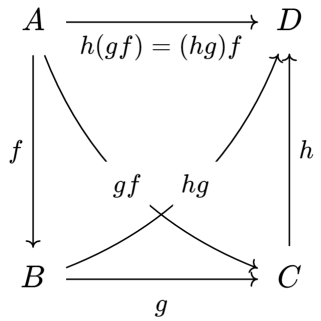
$$g \circ f : A \longrightarrow C$$

y tiene que satisfacer las siguientes condiciones:

(1) Para $f : A \longrightarrow B$ $g : B \longrightarrow C$ y $h : C \longrightarrow D$

$$h(gf) = (hg)f$$

o utilizando un diagrama:



decimos que este diagrama debe tener asociatividad.

(2) Ley de la identidad: Para cualquier morfismo $f : A \longrightarrow B$, la ecuación:

$$f.1_A = 1_B.f = f$$

debe ser cumplida. Con estas cuatro estructuras, cualquier categoría C se denota por:

$$C \sim (Ob(C), Mor(C), Mor(C) \times Mor(C) \overset{\circ}{\to} Mor(C); Condiciones)$$

3.2. **Ejemplos.**

(1) C=$Conjuntos$; Categoría de los conjuntos
$Ob(C)$: Clase de los conjuntos
$Mor(C)$: Conjuntos de funciones en conjuntos
$Composición$: Composición de funciones

(2) $C = Grp$; Categoría de los grupos
$Ob(C)$: Clase de los grupos
$Mor(C)$: Conjunto de los homomorfismos de grupos
Composición: Composición de homomorfismos

(3) $C = Top$; Categoría de los espacios topológicos
$Ob(C)$: Clase de los espacios topológicos
$Mor(C)$: Conjunto de funciones continuas entre espacios topológicos
Composición: Composición de funciones continuas

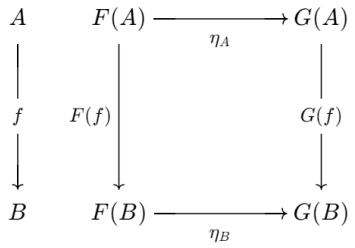
(4) $C = hTop$; Categoría de homotopías
$Ob(C)$: Clase de los espacios topológicos
$Mor(C)$: Clase de homotopías de funciones continuas
Composición: Composición de clases de homotopías

3.3. **Functores y transformaciones naturales.** Un functor es un mapeo que conserva la estructura entre categorías de la misma forma que un homomorfismo es un mapeo que conserva estructuras entre grafos.
Sean C y D las categorías, se cumple:

- Para un objeto A de C
$$F(id_A) = id_{F(A)}$$
- Para una composición $g \circ f$ de C
$$F(g \circ f) = F(g) \circ F(f)$$

La función F es un functor de la Categoría C a la D. Este functor es denotado por
$$F : C \to D$$
mientras tenemos que si, F y G son functores de la categoría C a D.

- Para cualquier A objeto de C
$$\eta_A : F(A) \to G(A)$$
es un morfismo de D.
- Sea f un morfismo de A a B en la categoría C. El diagrama,



es conmutativo. Si las condiciones son satisfechas, el morfismo
$$\eta : F \to G$$
es una transformación natural del functor F al functor G

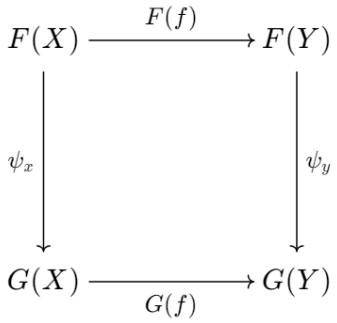


FIGURE 1. Diagrama conmutativo de la transformación natural

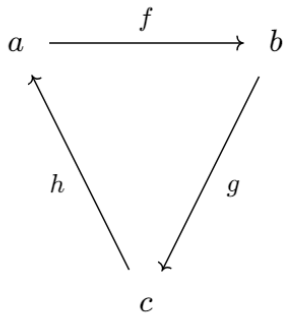
Para trabajar con una definición más precisa de transformación natural y posteriormente al de mónada junto a su aplicación en Haskell, nos basaremos en [4].

Dados dos functores, F, G entre las categorías C,D, una transformación natural. $\psi : F \to G$ es una colección de morfismos en D. Los morfismos deben satisfacer las siguientes condiciones:

(1) Para cada objeto x en C, hay un morfismo $\psi_x : F(x) \to G(x)$, llamado el componente de ψ en x.
(2) Los componentes deben satisfacer que para cada morfismo $f : X \to Y$ en C, donde tenemos $\psi_y \circ F(f) = G(f) \circ \psi_x$.

Es decir, una transformación natural puede ser vista cómo una manera de transformar un functor en otro. Además, si cada ψ_x es biyectivo, es decir que cada ψ es un isomorfismo natural, y que F y G son isomórficos. Note que mientras la notación da la idea de ψ siendo una función, F y G no son conjuntos, simplemente nos dice que ψ transforma un functor F en otro functor G.

3.4. **Ejemplo.** Para hacer el concepto de transformaciones naturales más claro, utilizaremos un ejemplo sencillo.
Sea C la categoría definida por el diagrama



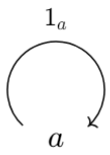


FIGURE 2. Diagrama conmutativo de la identidad

Sea F un functor $F : C \to C$ con $F(a) = F(b) = F(c) = ayF(f) = F(g0 = F(h) = 1_a$, y sea I el functor identidad en C que mapea cada objeto a si mismo, y cada morfismo a si mismo.

Un endofunctor T [10] lo entenderemos un functor F que mapea a la categoría C, con sus respectivos morfismos $Mor(C)$ Objetos $Ob(C)$, a sí misma. Es decir:
$T : C \to C$

Sea $\phi : I \to F$ una transformación natural con componentes:

$$\phi_a = h \circ g \circ f = 1_a$$
$$\phi_b = h \circ g$$
$$\phi_c = g$$

Notemos que cada para $x \in C$ existe un morfismo ϕ_x, entonces la primera ley es satisfecha. Entonces, examinamos los morfismos y mostramos que:

$$\phi_b \circ I(f) = h \circ g \circ f = (h \circ g \circ f) \circ 1_A = F(f) \circ \phi_a$$
$$\phi_c \circ I(g) = h \circ g = 1_a \circ (h \circ g) = F(g) \circ \phi_b$$
$$\phi_a \circ I(h) = h \circ f \circ g \circ h = F(h) \circ \phi_c$$

Así que F es una transformación natural , transformando C en una subcategoría.

Algo importante para la definición de Mónada es la de composición de transformaciones naturales, que la trabajaremos de la siguiente forma:
si $\alpha : F \to G$ y $\beta : G \to H$ son transformaciones naturales entre functores $F, G, H : C \to D$, podemos construir una transformación natural $\beta \circ \alpha : F \to H$, por componente definiendo $(\beta \circ \alpha)_x = \beta_x \alpha_x$. Ahora, dada una transformación natural $\xi : F \to G$ entre functores $F, G : C \to D$ y los functores $H_1 : B \to C$, $H_2 : D \to E$, podemos definir las transformaciones naturales $H_1\xi : H_1F \to H_1G$ y $\xi H_2 : FH_2 \to GH_2$ como

$$(H_1\xi)_x = H_1\xi_x$$
$$(\xi H_2)_x = \xi_{H2(x)}$$

Con esto ya definido, podemos hablar del concepto de mónada.

3.5. **Mónadas.** Una mónada en una categoría C consiste de un endofunctor $T : C \to C$ junto con dos transformaciones naturales:

- $\eta : 1_C \to T$, donde 1_C es el functor identidad en C
- $\mu : T \circ T \to T$

Sujeto a los criterios:
$\mu \circ T\mu = \mu \circ \mu T$

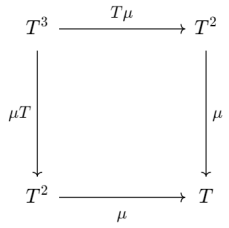


FIGURE 3. Diagrama de conmutatividad de la mónada, para más detalles [6].

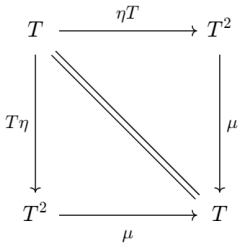


FIGURE 4. Diagrama equivalente con la doble linea del endofunctor

$\mu \circ T\eta = \mu \circ \eta T = 1_T$ donde 1_T es el morfismo identidad de T a T.

3.6. **Teoría de categorías y programación funcional.** Para los aspectos de esta sección, nos basaremos en [9]. La concepción matemática de los algoritmos computacionales se identifica como un factor clave evaluado por su capacidad de ahorrar tiempo y garantizar procesos estables en la programación. Diferentes algoritmos se obtienen mediante diversos métodos, como la encriptación asimétrica en la programación paralela. Estas características fundamentales de los lenguajes de programación deben ser compatibles con el pensamiento matemático para construir algoritmos sólidos. En este sentido, el lenguaje más adecuado corresponde a los lenguajes de programación funcional. Esto permite minimizar el número de líneas de código, el tiempo de procesamiento de los algoritmos y los errores derivados del compilador a un nivel aceptable.

La computación imperativa, que incluye los lenguajes de programación orientados a objetos, se centra precisamente en el resultado mostrado por el compilador. Cualquier salida adicional a este resultado se denomina efecto secundario. De hecho, estos efectos secundarios pueden ser factores que dificultan alcanzar el resultado deseado. En contraste, los lenguajes de programación funcional tienden a enfocarse en los cálculos. En conclusión, los lenguajes de programación funcional pueden tener propiedades más sólidas en comparación con otros lenguajes de programación [10].

Los lenguajes de programación funcional tienen cuatro elementos principales:

- Tipos de datos primitivos
- Constantes de cada tipo de dato

- Funciones entre tipos de dato
- Constructores

En el segundo elemento, la constante es un constructor con valor no paramétrico. También, el constructor es una función cuyas variables son funciones. Con estas características, cualquier lenguaje genera una categoría. Para componer esta categoría, necesitamos algunos cambios:

(1) Este lenguaje tiene que tener una función que no haga nada (id_A).
(2) Para llamar una constante con una función, el lenguaje necesita un tipo llamado 1. Este tipo se ve como el objeto terminal en la categoría del lenguaje.
(3) La composición del lenguaje tiene tupos de entrada y de salida. Por lo tanto, una composición es vista como un programa derivado.

Con estos cambios, para un lenguaje de programación funcional, L tiene la estructura de una categoría $C(L)$. $C(L)$ tiene las siguientes propiedades:

(1) Los objetos de $C(L)$ son los tipos de L.
(2) Las flechas de $C(L)$ son las operaciones de L.
(3) Las fuentes y los objetivos de las flechas son las entradas y salidas de de una operación en L.
(4) Composiciones de $C(L)$ son los constructores de L.
(5) Las flechas de identidad de $C(L)$ son las operaciones que no hacen nada en L.

De manera similar a esta estructura, la categoría de un lenguaje de programación funcional tiene functores y transformaciones naturales. Un ejemplo simple de un functor es List Set $\rightarrow$ Set que corresponde al constructor de tipos de lista en Haskell. La parte del objeto del functor mapea un conjunto A a un conjunto de listas sobre A, es decir secuencias de la forma $[x_1, ..., x_n]$ donde cada x_i es un elemento de A. La parte del morfismo del functor mapea una función f:A $\rightarrow$ B a la función normalmente escrita como el mapeo f en Haskell que manda una lista $[x_1, ..., x_n]$ a $[f(x_1), ..., f(x_n)]$. En la notación categórica, la función mapa f será escrita como List f: List A $\rightarrow$ List B.
Las transformaciones naturales del lenguaje de programación Haskell son funciones que satisfacen las condiciones de transformaciones naturales de la categoría. Por ejemplo, 'reverse' es una transformación natural de la categoría de Haskell (Hask). Esta función invierte la lista. para un functor:

$$List : C(L) \rightarrow C(L)$$

una transformación natural es:

$$\eta = reverse : Ob(C(L)) \rightarrow Mor(C(L))$$

Las transformaciones naturales representan funciones polimórficas. Las funciones polimórficas son mapeos entre constructores de tipos. Unos cuantos ejemplos de funciones polimórficas son:

$$append[A] : ListA \times ListA \rightarrow ListA$$
$$map[A, B] : [A \rightarrow B] \rightarrow [ListA \rightarrow ListB]$$
$$foldr[A, B] : [A \times B \rightarrow B] \times B \rightarrow [ListA \rightarrow B]$$

(Basado en [2])

Como resultado, Haskell nos da una categoría, un functor, y una transformación natural con la estructura del lenguaje [15]. Sea L un lenguaje de programación funcional con 3 tipos de data:

$$NAT\text{: Los naturales}$$
$$BOOLEAN = true, false$$
$$CHAR : ASCII$$

Entonces podemos construir la categoría $C(L)$. Los objetos de $C(L)$ son:

$$Ob(C(L)) = NAT, BOOLEAN, CHAR, 1$$

donde 1 es un singulete. Las flechas de $C(L)$ junto con:

$$Mor(C(L) \overset{s}{\underset{t}{\rightrightarrows}} Ob(C(L))$$

la función fuente y función objetivo son como sigue:

$$s(id_{NAT}) = t(id_{NAT}) = NAT$$
$$s(id_{CHAR}) = t(id_{CHAR}) = CHAR$$
$$s(id_{BOOLEAN}) = t(id_{BOOLEAN}) = BOOLEAN$$
$$s(id_1) = t(id_{BOOLEAN}) = BOOLEAN$$
$$s(id_1) = t(id_1) = 1$$
$$s(0) = 1 \text{ y } t(0) = NAT$$
$$s(c) = 1 \text{ y } t(c) = CHAR$$
$$s(false) = 1 \text{ y } t(false) = BOOLEAN$$
$$s(true) = 1 \text{ y } t(true) = BOOLEAN$$
$$s(n) = BOOLEAN \text{ y } t(n) = BOOLEAN$$
$$s(succ) = NAT \text{ y } t(succ) = NAT$$
$$s(ord) = CHAR \text{ y } t(ord) = NAT$$
$$s(chr) = NAT \text{ y } t(chr) = CHAR$$
$$s(id_1) \text{ y } t(id_1) = 1$$
$$s(x) = NAT \text{ y } t(x) = 1$$
$$s(y) = CHAR \text{ y } t(y) = 1$$
$$s(z) = BOOLEAN \text{ y } t(z) = 1$$

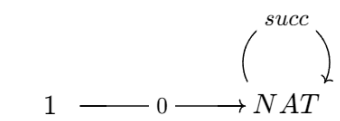
$$Mor(L) = (id_1, id_{NAT}, id_{CHAR}, id_{BOOLEAN}, 0, c, false, true, n, succ, ord, chr, x, y, z)$$

Es decir los morfismos de la categoría $C(L)$ son las identidades de los tipos correspondientes y las funciones sobre valores NAT, $CHAR$ o $BOOL$.

La constante 0 en NAT y la función succ son definidas por:

$$0\text{: } 1 \to NAT \text{ succ: } NAT \to NAT$$
$$x \to 0 \quad x \to x + 1$$

y todos los números naturales pueden ser generados con la composición de 0 y succ.



El ASCII ("American Standard Code for Information Interexchange"/ Código Americano Estándar para el intercambio de información), es un estándar que asigna letras, números, y otros caracteres en los 256 compartimientos disponibles en el código de 8-bits. chr y ord son funciones que convierten los caracteres a su valor ASCII y viceversa.

$$chr : NAT \to CHAR$$

toma un valor ASCII y retorna el caracter equivalente, y

$$ord : CHAR \to NAT$$

realiza la operación reversa convirtiendo un caracter en su valor numérico

$$c : 1 \to CHR$$

es una constante y todos los elementos del objeto CHR puede ser generado por la composición:

$$1 \xrightarrow{c} CHR \xrightarrow{ord} NAT \xrightarrow{chr} CHR \qquad (succ : NAT \to NAT)$$

Deberían haber dos constantes true y false:

$$\text{true: } 1 \to BOOLEAN \quad \text{false: } 1 \to BOOLEAN$$
$$x \to true \; x \to false$$

y una función n definida por:

$$\text{n: } BOOLEAN \to BOOLEAN$$
$$\text{true} \to \text{false}$$
$$\text{false} \to \text{true}$$

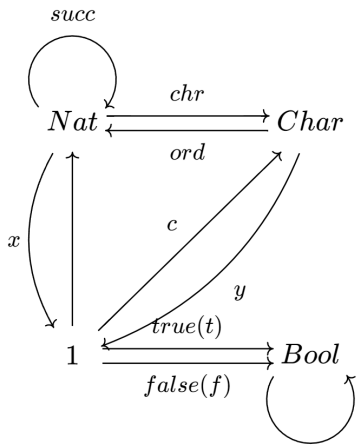
De ahora en adelante, todos los elementos del objeto BOOLEAN puede ser generado con la composición:

$$1 \xrightarrow{true} BOOLEAN \qquad (n : BOOLEAN \to BOOLEAN)$$

Entonces C(L) es una categoría con las siguientes ecuaciones siendo satisfechas:[1]

$$n \circ true = false$$
$$n \circ false = true$$
$$n \circ n = id_{BOOLEAN}$$
$$chr \circ ord = id_{CHAR}$$

El diagrama de la categoría C(L) se ve así:



3.7. **Ejemplos y aplicaciones.** Ya con nuestra base teórica construida, daremos algunos ejemplos de aplicación de Teoría de Categorías a C++[9] y finalmente a la programación funcional en Haskell [4].

3.8. **Functores en C++.** Los functores, dada la base funcional de Haskell, son expresados de manera fácil, pero pueden ser definidos en cualquier lenguaje que soporte programación genérica y funciones de alto orden. Consideremos el análogo en C++ de "Maybe", el tipo plantilla "optional". Aquí un bosquejo de la implementación (la implementación completa es más compleja dada las múltiples formas en que un argumento puede ser pasado, semánticas de copiado, y los problemas de manejos de recursos carecterísticos de C++ [9]).

```
template<class T>
class optional {
bool _isValid; // the tag
T _v;
public:
optional() : _isValid(false) {} // Nothing
optional(T x) : _isValid(true) , _v(x) {} // Just
bool isValid() const { return _isValid; }
T val() const { return _v; } };
```

Esta plantilla provee una parte de la definición de functor: el mapeo de tipos. Se mapea cualquier tipo T a un nuevo tipo opcional $< T >$. Se define esta acción en funciones:

```
template<class A, class B>
std::function<optional<B>(optional<A>)>
fmap(std::function<B(A)> f) {
return [f](optional<A> opt) {
if (!opt.isValid())
return optional<B>{};
else
return optional<B>{ f(opt.val()) };
};
}
```

Como veremos a continuación, por la naturaleza funcional de Haskell, la implementación de funciones como estas es mucho más fácil.

3.9. **Teoría de Categorías en Haskell.** Como lo mencionamos anteriormente, dada la naturaleza funcional en Haskell la implementación de conceptos de Teoría de Categorías es más directa. Comenzaremos definiendo cómo funciona la teoría de categorías en Haskell, posteriormente definiremos los functores y finalmente concluiremos con las mónadas y la aplicación de estas.

En Haskell el sistema subyacente es Hask. En Hask los objetos son los tipos de Haskell y los morfismos son las funciones de Haskell de tipo $A \to B$. En Haskell se ve fácilmente que se cumple, por lo anteriormente mencionado sobre la estructura de categoría de los lenguajes de programación funcional y además tenemos que:

- La existencia del morfismo identidad

```
id::  A -> A
```

- La composición de de morfismos

```
f . g = \x -> f (g x)
```

Para detalles y una exploración completa de Hask como categoría ver [5].

3.10. **Functores en Haskell.** Cuando hablamos de functores en Haskell, lo que se quiere decir realmente es un subconjunto de posibles functores, los functores de una categoría en sí misma, los endofunctores. En Hask, los endofunctores son representados como una clase de tipo con una sola función, fmap, que toma una función $a \to b$ y la lleva a una función $fa \to fb$, que realiza la misma computación con objetos envueltos. Frecuentemente usamos el operador fijo $ como un sinónimo de fmap, definido como

```
f $ x = fmap f x
```

3.11. **List es un functor.** Como ya vimos, podemos entender las funciones de Haskell como endofunctores en la categoría Hask, entonces haremos una exploración de List.

List es una construcción en Haskell similar a las listas en Python o los arreglos en C++, siendo estos arreglos de elementos del mismo tipo (Int,Bool,Char), su construcción completa y detalles se puede ver en [7].

Cuando examinamos estos en Haskell podemos ver que cumplen las condiciones para ser considerados functores: [4]

```
fmap id (x:xs)
=(id x) :fmap xs
=x : fmap xs
```

Lo cual inductivamente nos da:

```
fmap id (x:xs) = (x:xs)
```

cumpliéndose así la existencia de la identidad. Luego podemos ver que:

```
(fmap g) . (fmap h) (x :xs)
= fmap g (fmap h (x:xs))
= fmap g ((hx) : fmap h xs)
= (g . hx): fmap g (fmap h xs)
```

lo cual inductivamente nos da que está definida la composición, y por lo tanto "List" es un functor. Todas las instancias de functores en esta clase, cumplen estas leyes [7].

3.12. **Mónadas en Haskell.** La instancia de mónada en Haskell se define de la siguiente manera [7]:

```
Class Applicative m => Monad m where
    (>>=)       :: m a -> (a -> m b) -> m b
    (>>)        :: m a -> m b -> m b

    m >> k = m >>= \_ -> k

    return       :: a -> m a
    return       = pure
```

El primer uso de la mónada es para emular acciones secuenciales, de manera similar a como los punto y coma operan en los lenguajes similares a C. Esta concatenación de operaciones, envueltas en un contenedor de functores separado, nos da las herramientas para lidiar con algunos problemas de la programación funcional[4].

Trabajamos antes con la plantilla de "optional" en C++, ahora mostraremos su análogo ya mencionado en Haskell: "Maybe". Esta mónada se usa para el manejo de errores y otros tipos de funciones que no siempre tienen un valor que retornar [7].

```
instance Monad Maybe where
    return = pure
    (Just x) >>= f = f x
    Nothing >>= f = Nothing
```

Algunos ejemplos de su utilidad son:

```
--Prevee errores en el tiempo de compilado cuando se divide por cero
(/.) :: Integral a => a -> a -> Maybe a
x / . y = if y == 0 then Nothing else Just $ x `div` y

--Ejemplos
a = (3+2) /. 1
-- a = Just 5

b = (3+2) /. 0
-- b = Nothing

-- Obtener el elemento en la posición n-ésima de una lista.
-- Si no está, retorna "Nothing"
at :: [a]-> Int -> Maybe a
at(x : xs) 0 = Just x
at [] _= Nothing
at (_ : xs) n = at xs(n-1)
```

El sistema de entrada/salida de datos, está construido de manera puramente funcional, sobre la mónada I/O ("Input/Output") que nos permite obtener información del usuario, leer archivo entre otras funciones.

Por ejemplo, tenemos el siguiente programa [16] que:

(1) Le pregunta al usuario que inserte una cadena
(2) Lee la cadena
(3) Usa "map" para aplicar una función "shout" que pasa a mayúsculas todas las letras de la cadena
(4) Escribe el resultado de la cadena

```
module Main where

import Data.Char (toUpper)
import Control.Monad

main = putStrLn "Escriba su cadena: " >> fmap shout getLine >>= putStrLn

shout = map toUpper
```

Usando el compilador de GHCi podemos cargar nuestro programa para ver las definiciones de tipo, qué partes son funciones, qué partes acciones de tipo I/O u otros valores.

```
main :: IO ()
putStrLn :: String -> IO ()
"Escriba su cadena " :: [Char]
(>>) :: Mónada m => m a -> m b -> m b
fmap :: Functor m => (a -> b) -> m a -> m b
shout :: [Char] -> [Char]
getLine :: IO String
(>>=) :: Monad m => m a -> (a -> m b ) -> m b
```

Analizando el código tenemos que: `main` es `IO ()`. No es una función. Las funciones son de tipo `a -> b`. Todo nuestro programa es una acción de tipo IO. `putStrLn` es una función, pero resulta en una acción IO. El texto "Escriba su cadena: "es un `String` (que es un sinónimo de `[Char]`). Es usado como un argumento para `putStrLn` y es incorporado en la acción IO resultante. Así, `putStrLn` es una función, pero `putStrLn x` se evalúa como una acción IO. La parte `()` del tipo IO (llamada tipo unitario) indica que no hay nada disponible para ser pasado a funciones o acciones posteriores.

Esa última parte es clave. A veces decimos informalmente que una acción de IO "devuelve" algo; sin embargo, tomar eso demasiado literalmente lleva a confusión. Está claro lo que queremos decir cuando hablamos de funciones que devuelven resultados, pero las acciones de IO no son funciones. Pasemos a `getLine`, una acción de IO que sí proporciona un valor. `getLine` no es una función que devuelve un `String` porque `getLine` no es una función. Más bien, `getLine` es una acción de IO que, cuando se evalúa, materializa un `String`, que luego puede ser pasado a funciones posteriores mediante, por ejemplo, `fmap` y `(>>=)`.

Cuando usamos `getLine` para obtener un `String`, el valor es monádico porque está envuelto en un functor IO (que resulta ser una mónada). No podemos pasar el valor directamente a una función que toma valores simples (no monádicos, o no functoriales). `fmap` hace el trabajo de tomar una función no monádica mientras pasa y devuelve valores monádicos.

`(>>=)` hace el trabajo de pasar un valor monádico a una función que toma un valor no monádico y devuelve un valor monádico. Puede parecer ineficiente que `fmap` tome el resultado no monádico de su función dada y devuelva un valor monádico solo para que `(>>=)` luego pase el valor subyacente no monádico a la siguiente función. Sin embargo, es precisamente este tipo de encadenamiento lo que crea la secuenciación confiable que hace que las mónadas sean tan efectivas para integrar funciones puras con acciones de IO.

4. Conclusiones

En nuestro trabajo demostramos la relevancia de la Teoría de Categorías no solamente como tema académico en matemáticas pero también como una aplicación a las ciencias computacionales como a la programación en general. Vimos su aplicación a la teoría de lenguages funcionales en general, así como su aplicación a lenguajes no funcionales como C++. De igual forma vimos un ejemplo práctico como ser la aplicación de mónadas para el manejo de errores en Haskell.

Sin duda las posibilidades que abren tanto la Teoría de Categorías como la Programación Funcional, no son sólo provechosas para la comunidad matemática o de científicos computacionales, sino para la comunidad científica en general.

Para trabajos a futuro nos proponemos adentrarnos más tanto en la Teoría de Categorías, así como sus áreas aledañas como lo son la Teoría de Homotopía de Tipos, la Teoría de Lenguajes Formales, la Teoría de Topos y como estas junto a la Programación Funcional, las Ciencias Computacionales y la Lógica nos dan nuevas áreas de estudio multidisciplinario como ser la Teoría de la Demostración y la Lógica Categorial.

References

1. Charles Barr, *Category theory lecture notes for esslli*, Lecture Notes, 1999.
2. Michael Barr and Charles Wells, *Category theory for computing science*, Prentice Hall, New York, NY, USA, 1990.
3. Jean Gillibert and Christian Retoré, *Category theory, logic and formal linguistics: Some connections, old and new*, Journal of Applied Logic **12** (2014), no. 1, 1–22.
4. Samuel Grahn, *Monads in haskell and category theory*, 2019.
5. Makoto Hamana, *What is the category for haskell?*, 01 2007.
6. Saunders MacLane, *Categories for the working mathematician*, Graduate Texts in Mathematics, vol. 5, Springer-Verlag, New York, 1971, Graduate Texts in Mathematics, Vol. 5.
7. Simon Marlow, *Haskell 2010 language report*, 9 2019.
8. Jean-Pierre Marquis, *Category Theory*, The Stanford Encyclopedia of Philosophy (Edward N. Zalta and Uri Nodelman, eds.), Metaphysics Research Lab, Stanford University, Fall 2023 ed., 2023.
9. Bartosz Milewski, *Category theory for programmers*, Bartosz Milewski, 2019.
10. S. Poigné A. Rydeheard D. (eds.) Pitt, D. Abramsky, *Category theory and computer programming. tutorial and workshop*, Lecture Notes in Computer Science, vol. 240, Springer-Verlag, Berlin etc., 1986.

11. David Scott, *Using functional programming within an industrial product group: perspectives and perceptions: ACM SIGPLAN Notices: Vol 45, No 9 — dl.acm.org*, https://dl.acm.org/doi/abs/10.1145/1932681.1863557, [Accessed 08-07-2024].
12. ELİS SOYLU, *Relationships between category theory and functional programming with an application — journals.tubitak.gov.tr*, https://journals.tubitak.gov.tr/math/vol43/iss3/34/, [Accessed 08-07-2024].
13. Naotsugu Tsuchiya, Shigeru Taguchi, and Hayato Saigo, *Using category theory to assess the relationship between consciousness and integrated information theory*, Neurosci. Res. **107** (2016), 1–7 (en).
14. D. A. Turner, *Some history of functional programming languages*, Trends in Functional Programming (Berlin, Heidelberg) (Hans-Wolfgang Loidl and Ricardo Peña, eds.), Springer Berlin Heidelberg, 2013, pp. 1–20.
15. SP Vimal, *Functional programming*, BITS Press, Plain City, India, 2012.
16. Wikibooks, *Haskell — wikibooks, the free textbook project*, 2024, [Online; accessed 30-July-2024].

REGRESIÓN ROBUSTA

DAVID CRUZ

Resumen. Una buena base de datos permite estimar estadísticos de interés para una población de manera muy sencilla al utilizar métodos clásicos de estadística; por desgracia, no todas las bases de datos son buenas debido a un dato mal escrito, una anomalía en la población u otro factor que finalmente nos alejarán de una buena estimación. Para resolver estos problemas utilizamos la estadística robusta, capaz de realizar estimaciones precisas donde los datos que no están acorde a los demás no generan cambios bruscos en las estimaciones.

Abstract. A good database allows estimating statistics of interest for a population in a very simple way by using classical methods of statistics. Unfortunately, not all databases are good due to a misspelled data, an anomaly in the population or another factor that finally would move away from a good estimate. To solve these problems we use the robust statistics, capable of making precise estimates where the data that they are not in accordance with the others do not generate sudden changes in the estimates.

1. Introducción

Relacionar características sobre la vida es algo común para el ser humano: estudiar más para ganar más dinero; una empresa que invierte más en publicidad verá mejorías en sus ingresos; hacer calentamientos físicos reduce el riesgo de lesionarse... pero toda relación que se quiera realizar entre distintos factores deben ser estudiadas, pues no es posible simplemente decir "si estudio más ganaré más dinero", pues hay que comprobar que este hecho sea cierto o no.

Para ello recurrimos a la regresión lineal [1], un método en el cual se selecciona una variable de respuesta (dependiente) de la cual se quiere conocer su comportamiento, por ejemplo el dinero que gana una persona; y se selecciona una variable de predicción (independiente), por ejemplo el nivel de educación de la persona. Pero hay que ir un poco más allá, pues quizá la variable de respuesta no depende de un solo factor, de manera que se debe añadir variables de predicción en el modelo que se desea estudiar. Así es como aparece la regresión lineal múltiple, una extensión de lo anterior que utiliza un vector de variables para predecir a la variable de respuesta, por ejemplo si se desea estudiar el ingreso de una empresa (dependiente) se puede tomar como vector de predicción los factores gastos en publicidad, calidad de producto, precio del producto, etc.

Claro que para poder trabajar con estos modelos se debe partir de una serie de datos para realizar un análisis. Pero hay que recordar que los datos son producto de observaciones y/o debidas a acciones humanas por lo que están sujetas a errores

Date: 13 de agosto de 2024.
Key words and phrases. Dato atípico, estimación, inferencia, función distribución.

o desperfectos que afectarán las predicciones o análisis que se hagan. Para evitar que estos datos atípicos (que están fuera de lo usual) tengan gran peso en el estudio que se realiza, se recurre a la estadística robusta [2]. Con la estadística robusta se puede saber cuándo un dato está fuera de lo usual y qué hacer con estos para que se obtengan respuestas lo más acertadas posibles.

Cabe destacar que la regresión lineal es de suma importancia cuando se trata de realizar comparaciones entre datos, como influyen algunos de ellos en aquel que sea de interés, por ejemplo la calidad de sueño relacionada al insomnio durante la pandemia de Covid-19 [7].

La idea del trabajo es realizar comparaciones entre métodos clásicos y métodos robustos en datos con y sin datos atípicos (ya sean datos recolectados o simulaciones de datos) para mostrar las diferencias de estos métodos con distintos datos, tratando de indentificar cual es el de mejor acertación.

2. Antecedentes

En 1805, Legendre muestra por primera vez el método de mínimos cuadados. Años más tarde, Gauss profundiza en este método que resultaría en el estudio de variables que pueden ser comparadas [3].

Uno de los estudios de regresión lineal con mayor peso en el ámbito de la medicina, se culminó con la relación mortalidad y tabaquismo [4].

Los métodos robustos se remontan a finales del siglo XIX con Simon Newcomb, pero su mayor peso cae con los trabajos de Tukey, Huber y Hampel en las décadas $60's$ y $70's$.

Luego, George Box se dedica a recopilar estudios con estadística robusta y se dedica a esparcir técnicas y estrategias para implementar en los métodos clásicos que tuviesen problemas con datos atípicos [5].

Uno de los principales estudios utilizando estadística robusta descubrió que los agujeros de la capa de ozono se estaban formando años antes de ser descubiertos. Estos agujeros no habían sido detectados pues fueron considerados datos atípicos y correspondía eliminarlos del análisis. Al implementar métodos robustos (sin eliminar estos datos), se pudo concluir con mayor precisión cuándo comenzaron a formarse estos agujeros [9].

En 1999, Peña y Yohai describen un algoritmo determinista para obtener un punto de inicio a la regresión robusta, importante en algunas aplicaciones como riesgo financiero [2].

Finalmente, se debe comprender la diferencia entre los métodos que se aplican para realizar un buen estudio sobre regresión lineal. Es por ello que la comparación es inevitable para saber cuando un método es más efectivo que otro, que de hecho hay motivos para creer que mediante robustez se obtienen mejores resultados [6].

3. Regresión

3.1. Fundamentos Teóricos.

3.1.1. Estadística robusta. Utilizando [2] como línea base, se dará una explicación sobre estadística robusta.

Analizar una serie de datos requiere del uso de estadísticos que puedan representar los datos de la mejor forma posible. La estadística tradicional utiliza la media y la varianza (o su desviación estándar) como la representación más acertada para los

datos y, efectivamente, es el estadístico óptimo cuando se trata con una base de datos ideal.

Veamos qué sucede cuando en la base de datos se encuentra un dato que está fuera de lo normal (atípico) con el siguiente ejemplo: Los siguientes datos corresponden al contenido de cobre en harina integral (partes por millón) [8]

2.9	3.1	3.4	3.4	3.70	3.7
2.8	2.5	2.4	2.4	2.7	2.2
5.28	3.37	3.03	3.03	28.95	3.77
3.4	2.2	3.5	3.6	3.7	3.7

Entre estos datos se observa que la medición 28.95 está bastante alejada de las otras medidas, lo que da indicios de ser un valor atípico. Para ver su influencia en los estadísticos se calculan la media y desviación estándar con esta medición y sin ella

	Media	Desviación estaándar
Con el dato atípico	4.28	5.30
Sin el dato atípico	3.21	0.69

Claramente se puede observar que el dato influye de forma drástica en la desviación estándar y cambia en más de una unidad la media del conjunto de datos, por lo que se debe trabajar con este dato para que deje de tener tanta influencia en los estadísticos.

La solución más fácil de evitar la influencia de este tipo de datos es simplemente eliminarlos, pero ¿cuándo un dato es suficientemente atípico para ser eliminado? o ¿es este dato realmente significativo para el resto del conjunto?

Para evitar estas decisiones es mejor tratar con otro estadístico que se vea menos influenciado por datos atípicos, por ejemplo la mediana:
En el conjunto de datos completo la mediana es 3.38, mientras que sin el dato atípico la mediana es 3.37. Como el cambio entre una medida y otra no es tan grande, es conveniente utilizar este estadístico en lugar de la media para describir los datos.

Ahora, ¿qué sucede con la desviación estándar, que es el estadístico que más afectado se vio? Para evitar esta influencia se toma un estadístico de dispersión relacionado con la mediana (en lugar de la media) llamado MADN que se define como

$$\text{MADN}(\mathbf{x}) = \frac{\text{Med}\{|\mathbf{x} - \text{Med}(\mathbf{x})|\}}{0.6745}, \qquad \text{Med indica mediana}$$

De manera que el estadístico de disperción que se obtiene es $\text{MADN}(\mathbf{x}) = 0.53$ para el conjunto de datos y $\text{MADN}(\mathbf{x}) = 0.50$ quitando el dato atípico.

Con este ejemplo, se observa como un dato atípico cambia la forma en que se debe realizar el análisis de estos datos. ¿Será necesario cambiar la forma de analizar los datos cada vez se tenga un dato (o más) atípico(s)?
Media y desviación estándar. Supóngase que se tienen n datos $x_1, x_2, \ldots, x_n$ cuya media toma el valor de $\bar{x}_n$ y desviación estándar $s_n < \infty$.
Ahora, añádase un dato $x_{n+1} \in (-\infty, \infty)$, de manera que la nueva media será $\bar{x}_{n+1} \in (-\infty, \infty)$. Así que la nueva media puede tomar valores muy grandes o muy pequeños dependiendo del valor x_{n+1}. Además, la desviación estándar será afectada por este nuevo dato, de manera que la nueva desviación s_{n+1} sea un valor un poco menor a s_n o un valor mucho mayor a este.
De manera que no importa el conjunto de datos que se tome, un valor atípico

repercutirá en los valores de la media y desviación estándar, haciendo de estos estadísticos poco recomendados para análisis donde hay sospecha de atipicidad.
Mediana y MADN. Supóngase que se cuenta con n datos ordenados $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ de manera que la mediana será

$$\begin{cases} x_{(\lceil n/2 \rceil)} & \text{si } n \text{ es impar.} \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{si } n \text{ es par.} \end{cases}$$

donde $\lceil x \rceil$ denota la función techo (menor entero mayor a x). Como la MADN es una mediana, también es determinada según el orden de las diferencias entre cada dato con la mediana de ellos.
Ahora, añádase un dato $x_{n+1} \in (-\infty, \infty)$ al conjunto de datos. Si el nuevo dato se encuentra entre $x_{(1)}$ y $x_{(n)}$ entonces no es un dato atípico por lo que no es de mayor interés, así que véanse los otros dos casos

1. $x_{n+1} < x_{(1)}$.
   En este caso la mediana tomará el valor de
   $$\begin{cases} \dfrac{x_{(\lfloor n/2 \rfloor)} + x_{(\lceil n/2 \rceil)}}{2} & \text{si } n \text{ es impar.} \\ x_{(n/2)} & \text{si } n \text{ es par.} \end{cases}$$
   donde $\lfloor x \rfloor$ denota la función suelo (mayor entero menor que x).
2. $x_{(n)} < x_{n+1}$.
   En este caso la mediana tomará el valor de
   $$\begin{cases} \dfrac{x_{(\lfloor n+1/2 \rfloor)} + x_{(\lceil n+1/2 \rceil)}}{2} & \text{si } n \text{ es impar.} \\ x_{(n/2+1)} & \text{si } n \text{ es par.} \end{cases}$$

Por lo que la mediana no se verá muy afectada por los datos atípicos a menos que los datos estén muy dispersos en la parte central de estos. El análisis de MADN llevará a una conclusión parecida al tratarse de la mediana de un conjunto de datos al cual se le añade un elemento.

Dado que los estadísticos de media son más susceptibles que los estadísticos de mediana ante la aparición de datos atípicos, ¿por qué no usar estos estadísticos siempre para evitar una influencia innecesaria? Bueno, la respuesta es sencilla: la mayoría de datos no contienen datos atípicos y cuando esto sucede los mejores estimadores de son la media muestral y la desviación estándar.

3.1.2. Regresión lineal. De acuerdo a [1], se describe qué es una regresión lineal. Supóngase que se cuenta con un conjunto de datos con varias variables y se escoge una de ellas como variable de respuesta y las demás se consideran como variables predictoras. Con esto, se describe el modelo de regresión lineal de la siguiente manera

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

donde y_i representa la $i - ésima$ observación de la variable de respuesta; x_{ij}, la $i - ésima$ observación de la $j - ésima$ variable predictora; β_i, estadístico de peso para la $i - ésima$ variable predictora (β_0 es el intercepto o valor de base); y ε_i es la variable aleatoria para el error en la predicción.

Para que el modelo esté completo se debe asumir lo siguiente:

1. $E(\varepsilon_i) = 0$ para todo $i = 1, 2, \dots, n$.

2. $var\,(\varepsilon_i) = \sigma^2$ para todo $i = 1, 2, \ldots, n$.
3. $cov\,(\varepsilon_i, \varepsilon_j) = 0$ para todo $i \neq j$.

El modelo, escrito de forma matricial será el siguiente

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

que se abreviará como

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon.$$

La idea para realizar una regresión lineal gira en torno a estimar el vector β mediante mínimos cuadrados, encontrando el siguiente resultado

Teorema 3.1. *Sea* $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, *con* $\mathbf{X}$ *matriz de* $n \times (k+1)$ *con rango* $k+1 < n$. *Si* $E(\mathbf{y}) = \mathbf{X}\beta$, *entonces el estimador*

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

es insesgado para β.

Luego, para estimar la varianza σ^2 del modelo se utiliza el estimador

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(y_i - \mathbf{x_i'}\hat{\beta}\right)^2,$$

o con más simplicidad

$$s^2 = \frac{1}{n-k-1}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) = \frac{\mathbf{y'y} - \hat{\beta}'\mathbf{X'y}}{n-k-1}.$$

3.1.3. Regresión robusta. Para realizar la regresión lineal se utiliza el método de mínimos cuadrados, haciendo uso de la media de datos. Como ya se vio anteriormente, no es conveniente tomar como base estadísticos relacionados con la media por lo que se debe realizar un ajuste para los estimadores del modelo lineal. A continuación se detallan algunos métodos de ajuste.
Eliminación de datos atípicos. Para este método, se calculan los estimadores β con la regresión lineal mediante mínimos cuadrados.

Una vez obtenidos los estimadores, se procede a obtener las distancias de Cook para cada uno de los datos. Estas distancias miden la influencia que tiene cada dato en el modelo que se plantea con los estimadores β.

Estas distancias de Cook las obtenemos mediante [10]

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pECM}$$

donde

- $\hat{Y}_j$ es la predicción del modelo para la $j-ésima$ observación.
- $\hat{Y}_{j(i)}$ es la predicción del modelo estimado omitiendo la $i-ésima$ observación para la $j-ésima$ observación.
- p es el número de parámetros que se ajustan en el modelo.
- ECM es el error cuadrático medio del modelo de regresión.

Ahora que ya se tienen las distancias de Cook, se consideran influyentes (en este caso atípicos) los datos cuya distancia de Cook sea mayor que

$$\frac{4}{n-k-1}$$

donde n es el número de observaciones y k la cantidad de variables predictoras.

Una vez detectados los valores atípicos, se eliminan de la muestra y se vuelve a realizar la estimación del modelo de regresión lineal.

Modelo L_1. El problema de modelaje de la curva de mínima desviación absoluta es un problema de optimización para un conjunto de datos como sigue [11].

Definición 3.2. Sea A una matriz de observaciones de tamaño $n \times k$, $\mathbf{y}$ el vector de respuesta y ε_i el error de la $i-ésima$ observación. El problema desea encontrar un vector $c \in \mathcal{R}^k$ que minimice el valor de

$$||\mathbf{y} - Ac||_1, \qquad \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k} a_{i,j} c_j \right| \quad \text{ó} \quad \sum_{i=1}^{n} |\varepsilon_i|.$$

Estas tres expresiones son equivalentes, simplemente se escoge una con la cual trabajar de forma más sencilla.

El método más común para resolver este problema es el algoritmo de Barrodale y Roberts (BR) [12], que traduce el problema anterior al siguiente problema de programación lineal

$$\begin{array}{cc} \min_c & e'u + e'v \\ \text{sujeto a} & Ac^+ - Ac- + u - v = y, \\ & c^+, c^-, u, v \geq 0 \end{array}$$

donde e es un vector de 1s; u y v son la parte positiva y negativa, respectivamente, del vector ε; y c^+ c^- son la parte positiva y negativa, respectivamente, del vector c.

Una vez resuelto el problema de optimización, ya se cuenta con los estimadores deseados para el modelo de regresión.

Este método parte del uso o manipulación del valor de la mediana de los datos con los que se trabaja, de manera que no será necesario identificar los datos atípicos para quedar satisfecho con la estimación que realice.

Como se mencionó en la sección de estadística robusta, cuando hay datos atípicos es recomendable utilizar la mediana por sobre la media y es justo lo que hace el modelo con estimadores L_1, a partir de la mediana calcula iterativamente los valores buscados.

3.2. Comparación de métodos. Para comparar métodos, se utilizará el conjunto de datos de prestigio ocupacional de Duncan [13], que se describe a continuación

- Variable *type*: se refiere al tipo de ocupación (No será de interés).
- Variable *income*: Porcentaje de titulares con ingreso mayor a 3500 dólares en el año 1950 en Estados Unidos.
- Variable *education*: Porcentaje de titulares graduados de educación media al año 1950 en Estados Unidos.
- Variable *prestige*: porcentaje de respuesta a una encuesta que valora una ocupación con buen prestigio o incluso más.

- Se cuenta con 45 observaciones ocupacionales, es decir, 45 ocupaciones diferentes.

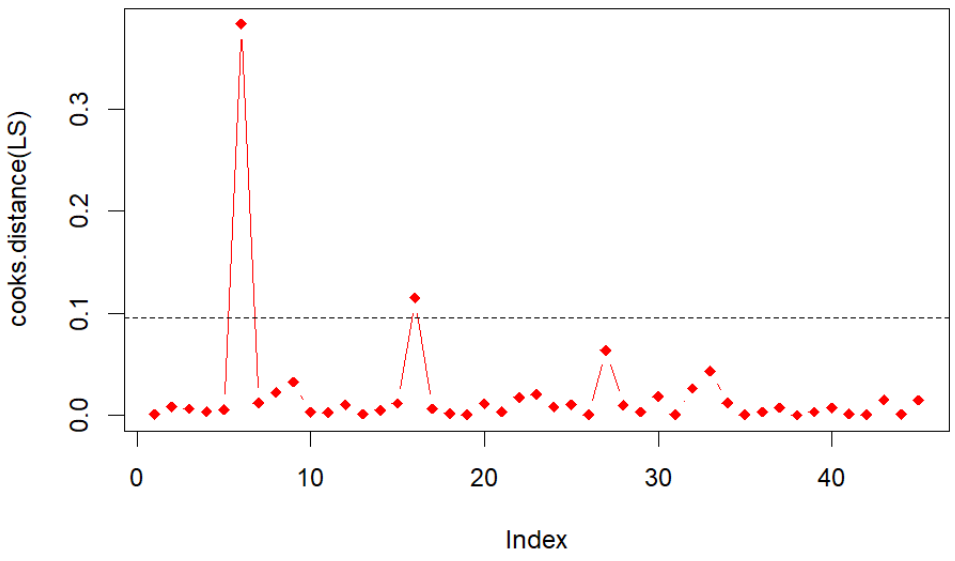
Para analizar distintos resultados se realizan las regresiones de estos datos que corresponden a income como la variable de respuesta y

1. prestige como variable predictiva.
2. education como variable predictiva.
3. prestige y education como variables predictivas.

3.2.1. Income, Prestige.
Eliminación de datos. Primero hay que realizar la regresión lineal sin considerar que haya valores atípicos, de manera que se obtienen los siguientes resultados

$$\begin{aligned}\hat{\beta}_0 &= 10.88\\ \hat{\beta}_1 &= 0.65\end{aligned}$$

Ahora, hay que calcular las distancias de Cook para cada observación, además del valor límite para los valores atípicos. En el siguiente gráfico se muestran los valores de las distancias de Cook junto a la recta que delimita cuáles pudiesen ser valores atípicos.



Como se ve en la gráfica, la medida 6 es claramente el dato que se eliminará, mientras que el 16 está cerca del límite por lo que se prefiere mantener.

Se prodece a realizar nuevamente la regresión lineal, pero ahora sin la observación *minister*, obteniendo los valores

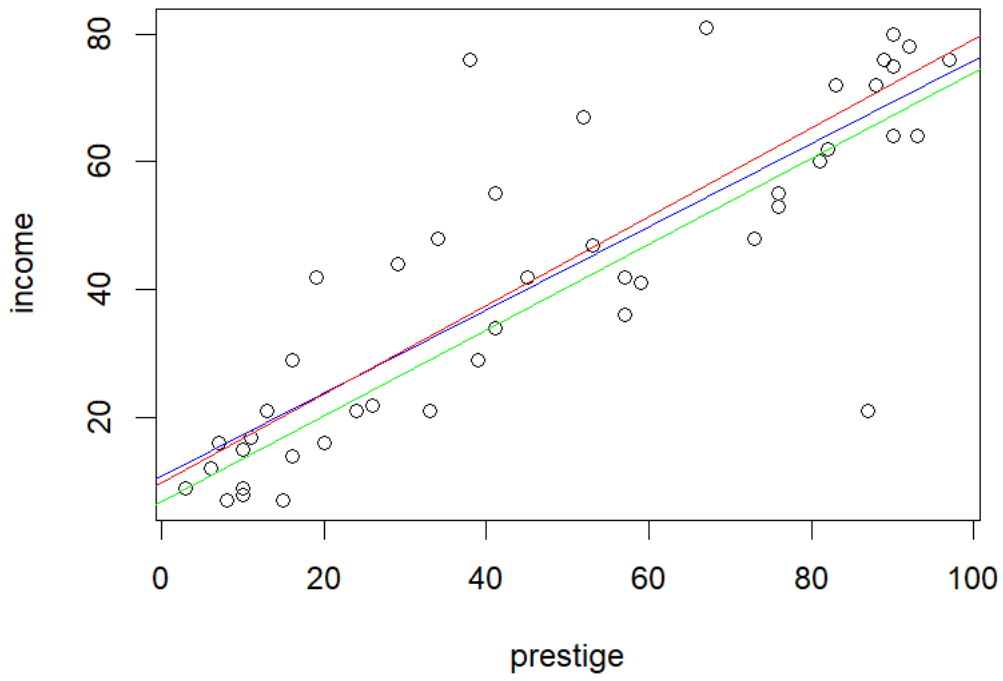
$$\begin{aligned}\hat{\beta}_0 &= 9.86\\ \hat{\beta}_1 &= 0.69\end{aligned}$$

Modelo L_1. Para este método se utiliza la herramienta en el software R llamada $l1fit$ que encuentra los estimadores utilizando el algoritmo de Barrodale y Roberts, así se encuentran los valores

$$\begin{aligned}\hat{\beta}_0 &= 6.99\\ \hat{\beta}_1 &= 0.67\end{aligned}$$

Comparación. El siguiente gráfico muestra los datos, además de los tres modelos calculados siendo

- Azul: Modelo de regresón lineal completo.
- Rojo: Modelo de regresión lineal sin atípicos.
- Verde: Modelo L_1.

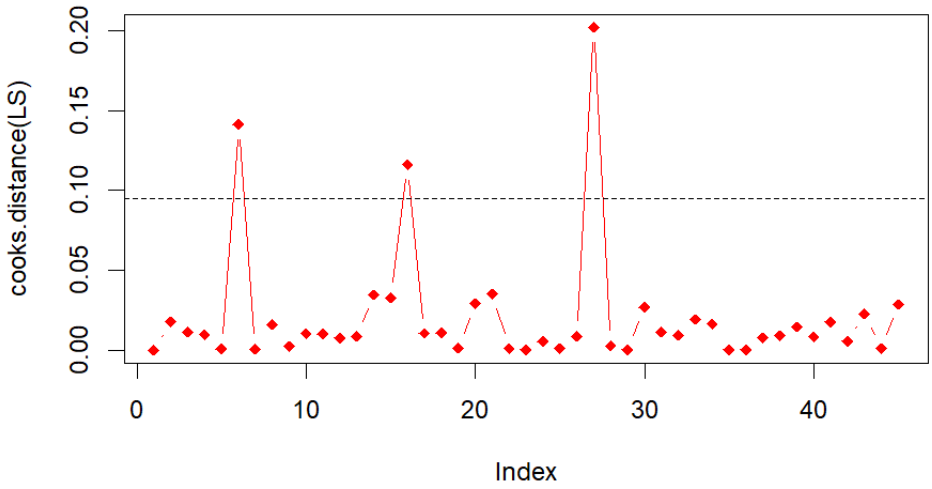


Aparentemente, el modelo L_1 se ajusta mejor para valores bajos de prestige, mientras que el modelo de regresión ajustado está mejor para valores altos de prestige.

3.2.2. Income, Education.
Eliminación de datos. Primero hay que realizar la regresión lineal sin considerar que haya valores atípicos, de manera que se obtienen los siguientes resultados

$$\begin{aligned}\hat{\beta}_0 &= 10.60\\ \hat{\beta}_1 &= 0.59\end{aligned}$$

Ahora, hay que calcular las distancias de Cook para cada observación, además del valor límite para los valores atípicos. En el siguiente gráfico se muestran los valores de las distancias de Cook junto a la recta que delimita cuáles pudieran ser valores atípicos.



Como se ve en la gráfica, las observaciones 6, 16 y 27 serán eliminadas.

Se prodece a realizar nuevamente la regresión lineal, pero ahora sin las observaciones *minister*, *conductor* y *RR.engineer* obteniendo los valores

$$\begin{aligned}\hat{\beta}_0 &= 4.31\\ \hat{\beta}_1 &= 0.69\end{aligned}$$

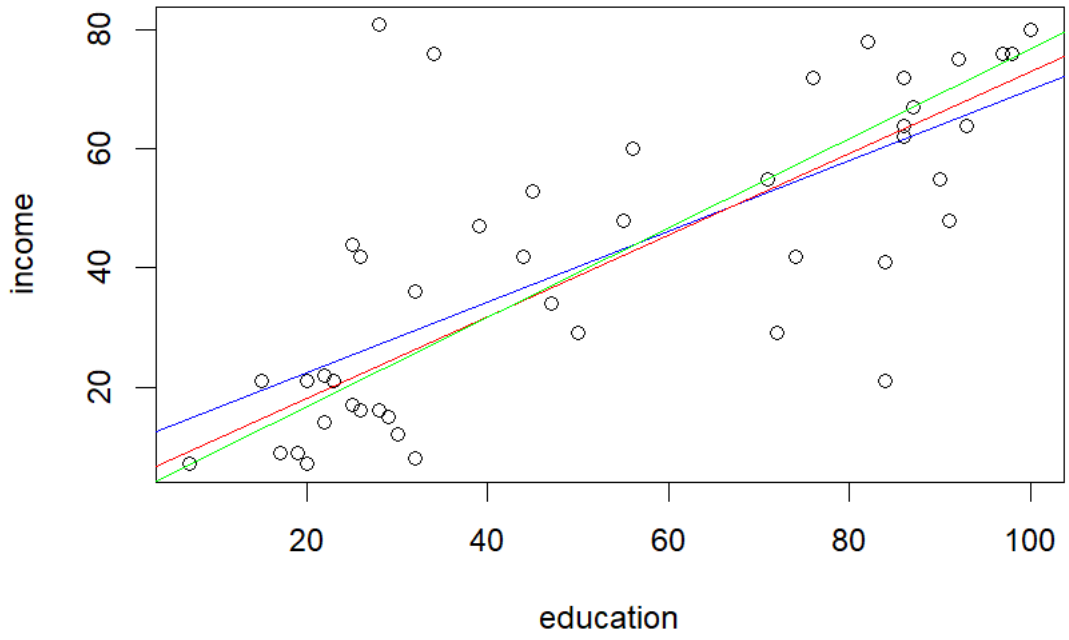
Modelo L_1. Para este método se utiliza la herramienta en el software R llamada $l1fit$ que encuentra los estimadores utilizando el algoritmo de Barrodale y Roberts, así se encuentran los valores

$$\begin{aligned}\hat{\beta}_0 &= 1.75\\ \hat{\beta}_1 &= 0.75\end{aligned}$$

Comparación. El siguiente gráfico muestra los datos, además de los tres modelos calculados siendo

- Azul: Modelo de regresón lineal completo.
- Rojo: Modelo de regresión lineal sin atípicos.
- Verde: Modelo L_1.

Cabe resaltar que a medida se ajusta el modelo para evitar la interferencia de datos atípicos, las rectas de regresión sun más bajas al inicio y con pendiente más elevada.



3.2.3. Income, Education y Prestige.
Eliminación de datos. Primero hay que realizar la regresión lineal sin considerar que haya valores atípicos, de manera que se obtienen los siguientes resultados

$$\begin{aligned}\hat{\beta}_0 &= 10.42\\ \hat{\beta}_1 &= 0.03\\ \hat{\beta}_2 &= 0.62\end{aligned}$$

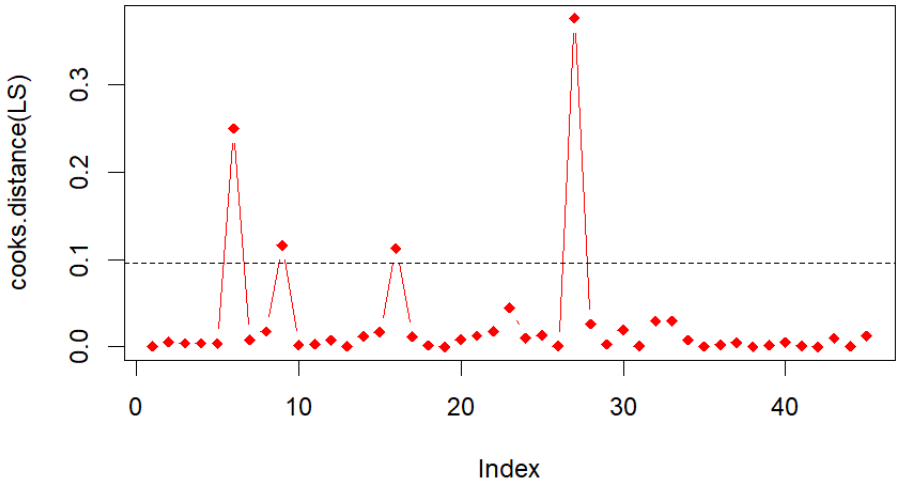
Ahora, hay que calcular las distancias de Cook para cada observación, además del valor límite para los valores atípicos. En el siguiente gráfico se muestran los valores de las distancias de Cook junto a la recta que delimita cuáles pudieran ser valores atípicos.

Como se ve en la gráfica, las observaciones 6 y 27 serán eliminadas, mientras que hay otras dos que están cercanas al límite que seguirán siendo parte del modelo.

Se prodece a realizar nuevamente la regresión lineal, pero ahora sin las observaciones *minister* y *RR.engineer* obteniendo los valores

$$\begin{aligned}\hat{\beta}_0 &= 7.72\\ \hat{\beta}_1 &= 0.15\\ \hat{\beta}_2 &= 0.56\end{aligned}$$

Modelo L_1. Para este método se utiliza la herramienta en el software R llamada $l1fit$ que encuentra los estimadores utilizando el algoritmo de Barrodale y Roberts,



así se encuentran los valores

$$\begin{aligned}\hat{\beta}_0 &= 3.93\\ \hat{\beta}_1 &= 0.19\\ \hat{\beta}_2 &= 0.55\end{aligned}$$

Comparación. Hay una clara diferencia entre el modelo de regresión completo y los que sugieren robustez. Entre ellos no hay mucha diferencia, siendo que el modelo L_1 le asigna un poco más de peso a la variable education.

4. Conclusiones

Se aprecian diferencias en los modelos al momento de realizar una regresión considerando los valores atípicos que al no tomarlos en cuenta. Realizar la validación de estos debería ser imprescindible para ejecutar el mejor análisis posible.

Se debe recordar que si al validar datos atípicos estos no se encuentran, lo mejor es proceder de la forma tradicional.

Referencias

1. Rencher, A. y Schaalje, G. (2008). *Linear Models in Statistics* (2da. edición). New Jersey: John Wiley & Sons, Inc.
2. Maronna, A., Martin, R., Yohai, V. y Salibián-Barrera, M. (2019). *Robust Statistics theory and methods (with R)* (2da. edición). New Jersey: John Wiley & Sons, Inc.
3. S. M. Stigler, "Gauss and the Invention of Least Squares," May 1981, doi: 10.1214/aos/1176345451.
4. R. Doll, R. Pető, K. Wheatley, R. Gray, and I. Sutherland, "Mortality in relation to smoking: 40 years' observations on male British doctors," Oct. 1994, doi: 10.1136/bmj.309.6959.901.
5. G. E. P. Box, "Robustness in the Strategy of Scientific Model Building," Jan. 1979, doi: 10.1016/b978-0-12-438150-6.50018-2.
6. Chun Yu, Weixin Yao, Xue Bai, "Robust Linear Regression: A Review and Comparison". Universidad de Cornell, arXiv:1404.6274

7. H. Jin et al., "Linear Regression Analysis of Sleep Quality in People with Insomnia in Wuhan City during the COVID-19 Pandemic" Apr. 2023, doi: 10.1155/2023/6746045.
8. Analytical Methods Committee, Robust statistics-how not to reject outliers. Part 1. Basic concepts"The Royal Society of Chemistry, Vol. 114, 1989, pp. 1693–1697
9. R. S. Kandel and M. Viollier, "Observation of the Earth's radiation budget from space," Apr. 2010, doi: 10.1016/j.crte.2010.01.005.
10. John Fox, (1991)*Regression Diagnostics: An Introduction*. Sage Publications.
11. Dongdong Lei, Iain Anderson and Maurice Cox, .A robust algorithm for least absolute deviations curve fitting".
12. Ronald Armstrong and James Godfrey, "Two linear programming algorithms for the linear discrete L_1 norm problem", enero 1979. Mathematics of computation, vol. 33, no 145, pp. 289–300
13. Duncan, O. D. (1961) A socioeconomic index for all occupations. In Reiss, A. J., Jr. (Ed.) Occupations and Social Status. Free Press [Table VI-1].

MODELADO DE POBLACIONES ESTRUCTURADAS POR EDAD CON LA ECUACIÓN DE VON FOERSTER-MCKENDRICK

MERARY ALEJANDRA GARCÍA SÁNCHEZ

Resumen. En el contexto del modelo Von Foerster-McKendrick, la ecuación de transporte es una herramienta fundamental para modelar la dinámica de las poblaciones. Esta ecuación describe cómo la distribución de edades en una población cambia a lo largo del tiempo. La formulación general de esta ecuación permite analizar el comportamiento de una población en función de diversas variables y condiciones iniciales. El modelo von Foerster-McKendrick es general porque su solución es independiente de las estructuras matemáticas específicas de las variables de entrada. Esto significa que el modelo se puede aplicar a diferentes contextos y sistemas sociales, siempre que se ajusten las tasas de mortalidad y natalidad adecuadas. En este trabajo vamos a investigar y analizar soluciones débiles de la ecuación de McKendrick-von Foerster, considerando tasas de nacimiento y muerte que varían con la edad de los individuos. El objetivo será explorar cómo estas variaciones afectan la dinámica poblacional a lo largo del tiempo, proporcionando ideas fundamentales para entender el impacto de las estructuras demográficas en modelos biológicos y epidemiológicos.

Abstract. In the context of the Von Foerster-McKendrick model, the transport equation is a fundamental tool for modeling population dynamics. This equation describes how the age distribution in a population changes over time. The general formulation of this equation allows for the analysis of population behavior based on various variables and initial conditions. The Von Foerster-McKendrick model is general because its solution is independent of the specific mathematical structures of the input variables. This means that the model can be applied to different contexts and social systems, as long as the appropriate birth and death rates are adjusted. In this work, we will investigate and analyze weak solutions of the McKendrick-von Foerster equation, considering birth and death rates that vary with the age of individuals. The objective will be to explore how these variations affect population dynamics over time, providing fundamental insights to understand the impact of demographic structures in biological and epidemiological models.

1. Introducción

El origen del modelo de transporte data del año 1941, cuando F. L. Hitchcock presentó un estudio titulado "La distribución de un producto desde diversos orígenes a numerosas localidades". Se cree que esta investigación fue la primera contribución para la resolución de los problemas de transporte. Hasta ahora, los fenómenos de transporte no se habían considerado como una asignatura con identidad propia. Este campo de estudio es muy básico y se relaciona con varias disciplinas clásicas

Date: Junio 10, 2024.

Key words and phrases. Ecuación de Transporte, Ecuación de Fluidos, Ecuación de McKendrick-von Foerster.

de la ciencia. Se considera que el estudio de los fenómenos de transporte es esencial para la ingeniería, al igual que lo son la termodinámica, la mecánica y el electromagnetismo [3].

En este trabajo lo que haremos es introducir el modelo de von Foerster-McKendrick para estudiar la variación de las poblaciones humanas en la estructura de edades. La ecuación esta dada por:

$$\frac{\partial}{\partial t}n(t,a) + \frac{\partial}{\partial a}n(t,a) = -\mu(a)n(t,a) \quad t > 0, a > 0.$$

Esta ecuación surge en el modelado de poblaciones donde es determinante la distribución por edad. Para describir cómo se distribuyen los miembros de una población por edades, considerando $n(t,a)$, nos indica el número de miembros con edad menor o igual a a en el instante de tiempo t, donde a y t toman valores positivos. Este modelo incluye ecuaciones diferenciales parciales lineales de primer orden, condiciones de contorno y condiciones iniciales. El modelo permite calcular la densidad de población (población por unidad de edad) en función del tiempo y la edad. Este modelo se puede utilizar para predecir las necesidades sociales a lo largo del tiempo para todas las edades, como las necesidades educativas o los recursos necesarios para ayudar a las personas mayores. En otros campos se han utilizado modelos similares, como los modelos Kermack-Mackendrick y Lotka-Mackendrick. Aunque la estructura de estos modelos es similar al modelo presentado en el artículo, sus objetivos difieren significativamente. Nuestro modelo es un modelo poblacional de von Foerster-McKendrick de población humana, es decir, un modelo determinista sin control en el que los nacimientos se calculan a través de la tasa de natalidad, no a través de la tasa de fertilidad, y se consideran los flujos migratorios. Por un lado, este modelo se caracteriza por tener en cuenta los flujos dinámicos que influyen en la dinámica de la población con la edad. Esto significa que la fertilidad, la mortalidad, la migración entre grupos de edad y la migración general de la población cambian con la edad. Por otro lado, este modelo calcula la migración de la población en diferentes edades (como primera aproximación), proporcional al tamaño de la población en el último siglo e inversamente proporcional a la duración del período de edad. [16].

Las ecuaciones en derivadas parciales de transporte tienen muchas aplicaciones en diversos campos de la ciencia y la ingeniería. Un ejemplo destacado se encuentra en las finanzas cuantitativas, donde se destaca la importancia de la Fórmula de Black-Scholes. Esta herramienta es fundamental en la economía moderna, utilizada para valorar diferentes activos financieros a lo largo del tiempo, como las acciones de empresas públicas [8].

2. Antecedentes

Las ecuaciones diferenciales en derivadas parciales son una rama del Análisis Matemático con múltiples aplicaciones en diversas áreas científicas, especialmente en el estudio de procesos que dependen de varias variables. Sofía Kovalevskaia hizo contribuciones fundamentales en este campo, y sus resultados son considerados los

más importantes y generales. Una característica distintiva de las ecuaciones diferenciales ordinarias es que se puede demostrar la existencia y unicidad de sus soluciones. En el caso para las ecuaciones diferenciales en derivadas parciales no se conocía una técnica unificadora que garantizara de forma general la existencia y unicidad de sus soluciones. No fue hasta finales del siglo XIX que Sofía Kovalevskaia, en su tesis doctoral, presentó un resultado fundamental en este sentido. Fue en el año de 1874, que Kovalevskaia recibió el Doctorado de Filosofía en Matemáticas de la Universidad de Gotinga por su trabajo titulado "Hacia una teoría de las ecuaciones diferenciales parciales". Con este logro, se convirtió en la primera mujer en obtener un doctorado en matemáticas. Kovalevskaia descubrió que algunas ecuaciones diferenciales no tienen una solución formal en series de potencias y determinó las condiciones necesarias para que cierto tipo de ecuaciones diferenciales parciales sean integrables. Dicha investigación se conoció como el Teorema de Cauchy-Kowalevsky. Kovalevskaia fue una matemática rusa que tuvo que enfrentar varios obstáculos para poder estudiar en la universidad. Para ello, tuvo que salir de Rusia, obtener permisos especiales para asistir a clases y solicitar tutorías privadas con distinguidos matemáticos [10].

Anderson Gray McKendrick, un destacado médico británico y matemático, fue precursor en el campo de la epidemiología matemática. Nacido en Edimburgo en 1876, McKendrick se graduó en medicina en la Universidad de Glasgow en 1900 y luego se unió al Servicio Médico de la India (IMS), donde comenzó su interés por la salud pública y la epidemiología durante una misión en Sierra Leona con Sir Ronald Ross. En 1905, fue nombrado en el Departamento de Investigación del Gobierno de India y trabajó en el Instituto Pasteur en Kausali, Punjab, donde estudió la rabia y comenzó a desarrollar su interés en los procesos epidemiológicos. En 1920, regresó a Gran Bretaña debido a problemas de salud y se estableció en Edimburgo.

McKendrick publicó un total de 58 artículos que abarcan temas médicos, estadísticos y demográficos, destacándose por su habilidad para aplicar métodos matemáticos a problemas epidemiológicos. Su trabajo sentó las bases para la epidemiología matemática moderna, influenciando a generaciones posteriores de investigadores en este campo crucial de la salud pública. Se retiró en 1941 y falleció en 1943 en Carrbridge, Inverness-Shire, dejando un legado duradero en la investigación epidemiológica y matemática [12].

Una ventaja de las ecuaciones diferenciales parciales es su capacidad para adaptarse fácilmente incluyendo detalles en el modelo, como la dependencia explícita del tiempo en los coeficientes y efectos no lineales. Una manera de modelar la evolución temporal de una población estructurada por edades es formular el proceso de evolución como una ecuación en derivadas parciales donde el tiempo y la edad son las variables independientes. La ecuación conocida como la ecuación de McKendrick-von Foerster, ha recibido atención por parte de matemáticos. Las condiciones iniciales y de contorno para la ecuación de McKendrick-von Foerster, impuestas por el modelo poblacional, difieren de las condiciones estándar en la teoría de ecuaciones diferenciales parciales para ecuaciones de evolución [7].

Los modelos de Von Foerster-McKendrick originados en la evolución, que es el concepto central en biología, resulta de una gran cantidad de procesos históricos únicos que, en su mayoría, no han dejado muchos registros directos. Además, a menudo se

carece de los registros necesarios para reconstruir algunas de las partes más interesantes de estos procesos históricos. Por otro lado, queda una dificultad importante: los cambios evolutivos significativos, como ocurrieron en el pasado, aparentemente tomaron una cantidad enorme de tiempo. Podemos asumir que los cambios evolutivos importantes resultan de la acumulación de cambios más pequeños, como los que estudian los biólogos de poblaciones, que describen poblaciones con estructura determinada por la edad, tamaño o nivel de maduración de individuos[1]. En [5, 6, 9] podemos encontrar información donde los modelos de Von Foerster-McKendrick describen lo antes mencionado.

La ecuación de McKendrick-von Foerster es una ecuación diferencial parcial lineal de primer orden utilizada en diversos campos de la biología matemática. Esta ecuación se aplica cuando la estructura por edades es una característica importante en el modelo matemático. Anderson Gray McKendrick la presentó por primera vez en 1926 como un límite determinista de modelos de redes aplicados a la epidemiología, y más tarde, en 1959, fue desarrollada de manera independiente por el profesor de biofísica Heinz von Foerster para describir los ciclos celulares [11]. En [2] podemos conocer más sobre la propuesta mencionada anteriormente de McKendrick en 1926.

3. Modelo Estructurado por edad

3.1. Generalidades. El modelo estructurado por edad, la ecuación de McKendrick-Von Foerster es:

$$\frac{\partial}{\partial t}n(t,a) + \frac{\partial}{\partial a}n(t,a) + \mu(a)n(t,a) = 0, \tag{3.1}$$

con la condición de frontera

$$n(t,0) = \int_0^{+\infty} b(a)n(t,a)da$$

y la condición inicial

$$n(0,a) = f(a).$$

Consideramos una población con $n(a,t)$ individuos de edad a en el momento t. La población total en el tiempo t es:

$$N(t) = \int_0^{+\infty} n(a,t)da \tag{3.2}$$

Suponiendo que la muerte ocurre con una tasa dependiente de la edad $\mu(a)$, tenemos: $\dfrac{dn(a,t)}{dt} = -\mu(a)n(a,t)$. Como el envejecimiento también depende del tiempo, $a \equiv a(t)$, y asumiendo que la edad y el tiempo se miden en la misma escala, $\dfrac{da}{dt} = 1$, la función $n(a,t)$ satisface la ecuación diferencial parcial lineal de primer orden:

$$\frac{\partial}{\partial t}n(t,a) + \frac{\partial}{\partial a}n(t,a) = -\mu(a)n(t,a) \tag{3.3}$$

La ecuación diferencial parcial nos indica que a medida que pasa el tiempo, las personas van envejeciendo. En el lado derecho de la ecuación, se muestra que la

única forma en que la población disminuye es a través de las muertes, y que la probabilidad de morir aumenta con la edad.

Para describir los nacimientos, se utiliza una función que distribuye la tasa de natalidad según diferentes clases de edad, denotada como $b(a)$. Los recién nacidos en el momento t (aquellos con edad $a = 0$) se determinan mediante la siguiente fórmula:

$$n(t,0) = \int_0^{+\infty} b(a)n(a,t)da. \tag{3.4}$$

La expresión 3.4 tiene la forma de una condición de contorno o de frontera.

Este método general, es útil incluso si los coeficientes en 3.1 varían o dependen de $n(t,a)$. En este caso que las curvas solución de $\frac{da}{dt} = 1$ y $\frac{ds}{dt} = 1$ son las líneas diagonales en el diagrama de Lexis. De hecho, la ecuación 3.1 se desarrolló observando cómo la derivada de $n(t,a)$ en la dirección de $\frac{da}{dt} = 1$ contribuye al equilibrio de la población.

3.2. Solución de la ecuación. En el contexto de las ecuaciones diferenciales parciales, la estabilidad significa la convergencia en el tiempo a algunas soluciones particulares (constantes). La existencia de soluciones de la ecuación 3.3 con las condiciones de contorno 3.4 ha sido demostrada por varios autores [14].

Para resolver una ecuación como la 3.3, que es una ecuación diferencial parcial lineal de primer orden, se utiliza un método estándar que aprovecha la existencia de curvas especiales en el plano $a - t$, llamadas características. A lo largo de estas características, la ecuación 3.3 se simplifica y se convierte en una ecuación diferencial ordinaria [7].

Ahora veamos que la solución de 3.1 se encuentra utilizando el método de características, en [7],[13] se proporciona un análisis detallado que incluye el cálculo de la solución paso a paso de la ecuación de McKendrick-von Foerster , ofreciendo una guía clara y comprensible del proceso. En esta sección, únicamente presentaremos la solución.

Al escribir 3.3 de la siguiente forma $\frac{\partial n(a,t)}{\partial t} = -\mu(a)n(a,t)$ las soluciones de 3.2 son también soluciones del sistema de ecuaciones diferenciales ordinarias,

$$\begin{cases} \frac{\partial n}{\partial t} = -\mu(a)n \\ \frac{\partial n}{\partial t} = 1 \end{cases} \tag{3.5}$$

Estas dos ecuaciones tienen soluciones generales, que se pueden expresar como:

$$\begin{cases} a - a_0 = t - t_0 \\ n(a,t) = n(a_0,t_0)e^{-\int_{t_0}^{t} \mu(s+a_0-t_0)ds} \end{cases} \tag{3.6}$$

Aquí, a_0 es la variable continua de edad en el instante $t = t_0$. La primera ecuación en 3.6 corresponde a la ecuación de las curvas características de la ecuación diferencial parcial 3.3. Al introducir la primera ecuación en la segunda, obtenemos la solución de la ecuación McKendrick-von Foerster cuando $t < a$.

$$n(a,t) = \phi(a-t)e^{-\int_{t_0}^{t}\mu(s+a_0)ds} = \phi(a-t)e^{-\int_{0}^{t}\mu(s+a-t)ds} \tag{3.7}$$

Aquí, $\phi(a-t) = \phi(a-t,0)$ representa la distribución inicial de densidad de la población en el tiempo $t_0 = 0$. Para $t < a$, la solución 3.7 es independiente de la condición de contorno 3.4. Para tener en cuenta la condición de contorno 3.4, damos la distribución inicial de la población en la forma de una función localmente integrable $\phi(a)$, definida para $a \geq 0$ [13].

3.3. Una clase de edad fértil. Ahora, supongamos que los nacimientos ocurren a una edad fija dada $a = \alpha$, y asumimos que la función de distribución de la tasa de natalidad por clase de edad es $b(a,t) = b_1\delta(a-\alpha)$, donde δ es la función delta de Dirac y b_1 es una constante no negativa. Para este caso particular, la reproducción ocurre únicamente en una clase de edad α y la condición de contorno 3.4 se convierte en

$$n(0,t) = b_1 n(\alpha,t). \tag{3.8}$$

Teorema 3.1. *Sea $n(a,0) = \phi(a) \in C^1([0,\infty))$ una condición inicial para la ecuación diferencial parcial de 3.3 con $a \geq 0$, y $t \geq 0$ y $\mu(a)$ una función no negativa y localmente integrable. Entonces en el interior de los conjuntos T_m con $m \geq 0$, la solución de la ecuacion de 3.3 con datos iniciales $\phi(a)$ y condición de frontera $n(0,t) = b_1 n(\alpha,t)$ es:*

1. *Si $(a,t) \in T_0$,*
$$n(a,t) = \phi(a-l)e^{-\int_0^t \mu(s+a_0)ds}$$
donde $a_0 = a - t$.

2. *Si $(a,t) \in T_m$ y $m \geq 1$,*
$$n(a,t) = b_1^m \phi(m\alpha + a - t)e^{-\int_0^{t-a-(m-1)\alpha}\mu(s+a_0 ds+(m-1)\int_0^\alpha \mu(s)ds+\int_0^\alpha \mu(s)ds}$$
donde $a_0 = m\alpha + a - t, m = \dfrac{(t-1)}{\alpha+1}, [x]$ representa la parte entera de x, $\alpha > 0$ y $b_1 > 0$.

Corolario 1. *Sea $n(a,0) = \phi(a) \in c^1([0,\infty))$, una condición inicial para la diferencia parcial de 3.3 con $a \leq 0$, $t \leq 0$ y $\mu \leq 0$. Entonces el interioir de los conjuntos T_M, con $m \leq 0$, la solución de 3.3 con datos iniciales $\phi(a)$ y condición de frontera $n(0,t) = b_1 n(\alpha,t)$ es:*

1. *Si, $(a,t) \in T_0$,*
$$n(a,t) = \phi(a-t)e^{-\mu t}$$

2. *Si, $(a,t) \in T_m$ y $m \leq 1$,*
$$n(a,t) = b_1^m \phi(m\alpha + a - t)e^{-\mu t}$$
donde $\dfrac{(t-a)}{\alpha+1}, [x]$ representa la parte entera de x, $\alpha > 0$ y $b_1 > 0$.

Definición 3.2. *Una función localmente integrable $n(a,t) \in L^1_{loc}(\mathbb{R}^2_+)$ es una solición débil en el sentido de las distribuciones de la ecuación diferencial parcial McKendrick, si*

$$\int\int_{\mathbb{R}^2_+} (\frac{\partial}{\partial t}\psi(a,t) + \frac{\partial}{\partial a}\psi(a,t) - \mu(a)\psi(a,t))n(a,t)dadt = 0$$

para cualquier $\psi(a,t) \in D(\mathbb{R}^2_+)$ [13].

Para ampliar la ecuación de McKendrick en función de los datos iniciales a todo el dominio de las variables independientes a y t, y teniendo en cuenta la condición de frontera 3.8, consideremos el caso cuando $t = a$. Tenemos que $n(a,t) = n(0,t-a)e^{-\int_0^a \mu(s)ds}$ y para $t = a$,

$$n(a,t) = n(0,0)e^{-\int_0^a \mu(s)ds}$$

Para calcular el número de recién nacidos en el tiempo $t = 0$, utilizamos la condición de contorno 3.8. Para que esta solución dependa de los datos iniciales, es necesario que tengamos lo siguiente,

$$n(a,t) = b_1(0)\phi(\alpha)e^{-\int_0^a \mu(s)ds} \quad \text{si} \quad t = a \tag{3.9}$$

donden $(\alpha, 0) = \phi(\alpha)$.

Es importante destacar que, para $t \leq a$, la solución $n(a,t)$ dada por la ecuación en 3.7, 3.9, generalmente presenta discontinuidades cuando se cruza transversalmente la línea $t = a$.

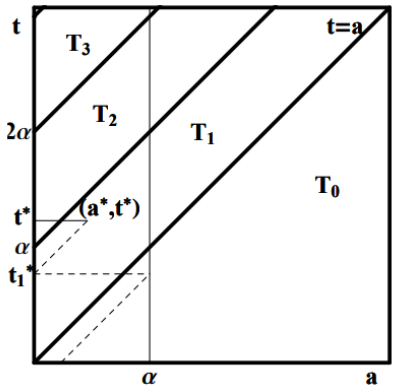


FIGURA 1. Curvas características $a - a_0 = t - t_0$ para la ecuación de McKendrick [15].

El siguiente teorema está relacionado con la función de fertilidad y el comportamiento a largo plazo de una función de población $n(a,t)$. Este teorema establece condiciones bajo las cuales $n(a,t)$ es tiempo periódico o converge a un comportamiento particular cuando el tiempo t tiende a infinito.

Teorema 3.3. *Supongamos que $\phi(a)$ es positiva y acotada en el intervalo $(0,\alpha]$ y la función de fertilidad b es una constante positiva. Entonces tenemos las siguientes condiciones:*

1. *Si,* $\ln b_1 = \int_0^\alpha \mu(s)ds$*, entonces, para todo* a *conocido* $t \geq a$,$n(a,t)$ *es tiempo periódico con período* $\tau = \alpha$

2. *Si,* $\ln b_1 = \int_0^\alpha \mu(s)ds$*, entonces, para todo* a *conocido* $t \geq a$,$n(a,t) \to \infty$ *con* $t \to \infty$

3. *Si,* $\ln b_1 = \int_0^\alpha \mu(s)ds$*, entonces, para todo* a *conocido* $t \geq a$,$n(a,t) \to 0$ *con* $t \to \infty$*. El comportamiento a largo plazo de* $n(a,t)$ *cuando* t *crece, para un valor fijo de* a*, depende de los datos iniciales* $\phi(a)$ *con* $a \in (0, \alpha]$.

Según el Teorema anterior, la estabilidad de las soluciones no nulas y su persistencia están condicionadas por la tasa de crecimiento r: si $r = 1$, las soluciones pueden ser estables o periódicas; si $r > 1$, se observa crecimiento exponencial; y si $r < 1$, la población tiende a extinguirse, a menos que $\phi(a)$ sea cero en el intervalo $(0, \alpha]$.

Teorema 3.4. *Sea* $n(a,0) = \phi(a)$ *(* $\in L^1_{loc}(\mathbb{R}_+)$ *) una condición inicial localmente integrable para la ecuación diferencial parcial de McKendrick (2.1), con* $a \geq 0$*,* $t \geq 0$*, y condición de frontera (3.1). Supongamos que* $\mu(a) \in L^1_{loc}(\mathbb{R}_+) \cap C^0(\mathbb{R}_+)$ *es una función no negativa y* $b_1(t) \in C^1(\mathbb{R}_+)$ *es positiva. Entonces, las soluciones débiles de la ecuación de McKendrick 3.3 son:*

1. *Si* $(a,t) \in \bar{T}_0$*,*

$$n(a,t) = \phi(a-t)\exp\left(-\int_0^t \mu(s+a_0)ds\right)$$

donde $a_0 = a - t$.

2. *Si* $(a,t) \in \bar{T}_m$ *y* $m \geq 1$*,*

$$n(a,t) = b_1(t-a)\cdots b_1(t-a-(m-1)\alpha)$$

$$\times\phi(m\alpha + a - t)\exp\left(-\int_0^{t-a-(m-1)\alpha} \mu(s+a_0)ds\right)$$

$$\times\exp\left(-(m-1)\int_0^\alpha \mu(s)ds - \int_0^a \mu(s)ds\right)$$

donde $a_0 = m\alpha + a - t$*,* $a > 0$*,* $m = \left\lfloor \frac{t-a}{\alpha} \right\rfloor + 1$*, y* $\lfloor x \rfloor$ *representa la parte entera de* x.

El siguiente teorema nos indica que la densidad de edad de la población y la población total tienen una modulación periódica a lo largo del tiempo [15].

Teorema 3.5. *Si* $\phi(a)$ *está acotada en el intervalo* $(0, \alpha]$ *y la función de fertilidad* b_1 *es una constante positiva, entonces, en las condiciones del teorema 3.4 para a fija y* $t \geq a$*,*

$$\frac{n(a,t)}{r^m}\chi(a,t)\phi(\alpha - \alpha\varepsilon(a,t))e^{-\int_\alpha^a \mu(s)ds}$$

es una función periódica del tiempo del período α*, y* $m = \dfrac{(t-a)}{\alpha+1}$.

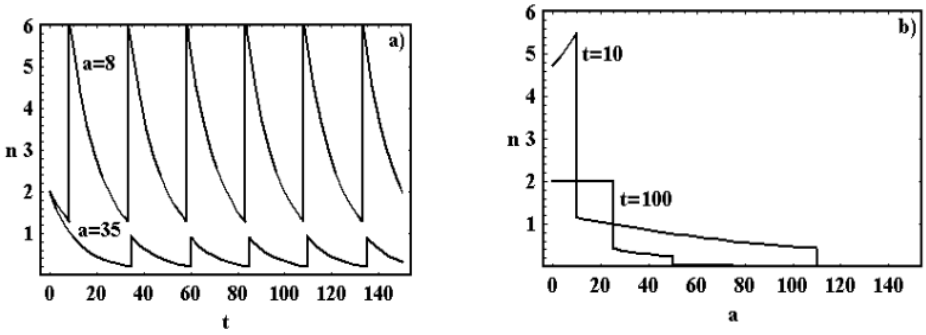


FIGURA 2. a) Evolución en el tiempo de la solución de la ecuación de McKendrick 3.3 para las edades $a = 8$ y $a = 35$, en una población con una clase de edad reproductiva $\alpha = 25$. b) Distribución de la densidad de individuos en función de la edad para $t = 10$ y $t = 100$. En ambos casos, la condición de estabilidad es $b_1 = e^{\mu\alpha}$ [15].

Para obtener una idea mas clara del teorema 3.5 mostramos la siguiente tabla a continuación que proporciona una visión detallada de la evolución de la población y la densidad poblacional en Honduras durante el período de 2016 a 2022. Los datos incluyen cifras anuales de la población total y la densidad de población correspondiente para cada año.

Año	Población	Densidad
2016	8,721,014	39
2017	8,866,351	39.8398158
2018	9,012,229	40.4952999
2019	9,158,345	41.1518535
2020	9,304,380	41.8080431
2021	9,450,711	42.4655628
2022	9,597,739	43.1262143

CUADRO 1. Datos de Población en Honduras por Instituto Nacional de Estadística INE- XVII Censo de población y VI de Vivencia [Elaboración porpia].

3.4. Un modelo de Tiempo Continuo. En esta sección, vamos a estudiar un modelo continuo que nos ayudará a entender el modelo de von Foerster-McKendrick. Este modelo se basa en una ecuación diferencial que nos permite hacer predicciones sobre cómo cambia la población humana con el tiempo.

Las variables que utilizaremos en este modelo son las siguientes:

- Tiempo (t): Representa el paso del tiempo.
- Población ($v(t)$): Es la cantidad de personas en un momento específico t.

- Flujo de nacimientos ($n(t)$): Mide el número de nacimientos por unidad de tiempo.
- Flujo de muertes ($d(t)$): Mide el número de muertes por unidad de tiempo.
- Flujo migratorio ($f(t)$): Es la diferencia entre el número de inmigrantes ($\tilde{f}(t)$) y el número de emigrantes ($fe(t)$).

En este modelo, la población y los flujos de nacimientos y muertes siempre son números no negativos ($\Re^+ \cup \{0\}$), mientras que el flujo migratorio ($f(t)$) puede ser cualquier número real ($\Re$).

La tasa de cambio de la población con respecto al tiempo, es decir, cómo varía la población con el paso del tiempo, se describe con la siguiente ecuación:

$$\frac{dv(t)}{dt} = n(t) - d(t) + f(t), \quad \text{con} \quad v(t_0) = v_0 \tag{3.10}$$

En 3.10 t_0 es el instante inicial y v_0 es la pobalción en ese instante. Esto significa que el cambio en la población a lo largo del tiempo depende de los nacimientos, las muertes y el saldo neto de la migración en ese período.

Sin embargo, si conocemos la tasa bruta de natalidad $a(t)$ y la tasa bruta de mortalidad $g(t)$, podemos calcular $n(t)$ y $d(t)$ de la siguiente manera:

$$n(t) = a(t) \cdot v(t) \quad \text{y} \quad d(t) = g(t) \cdot v(t).$$

Es importante notar que tanto $a(t)$ como $g(t)$ son variables que dependen del tiempo. Además, $f(t)$ también es una variable de entrada dependiente del tiempo, cuya forma se puede determinar utilizando el mismo procedimiento.

Así, la ecuación 3.10 se puede reescribir como:

$$\frac{dv(t)}{dt} = (a(t) - g(t)) \cdot v(t) + f(t), \quad v(t_0) = v_0$$

Este modelo proporciona una buena estimación para predecir la población. Una solución explícita de la ecuación 3.10, válida para $t \geq t_0$, es:

$$v(t) = \exp\left(\int_{t_0}^{t} (a(y) - g(y))dy\right) \cdot \left(v_0 + \int_{t_0}^{t} \exp\left(-\int_{t_0}^{y} (a(z) - g(z))dz\right) \cdot f(y) \cdot dy\right)$$

[16].

3.5. Caso de aplicación del modelo presentado.

3.5.1. Población. En este caso el sistema de población humana para validar el estudio del modelo de Von Foerster-McKendricks en este caso fue la ciudad de Valencia, dicha ciudad cuenta con aproximadamente 795,000 habitantes en la costa este de España (La siguiente información fue tomada de [16]).

Comencemos con la estructura matemática de las variables de entrada de edad, especialmente la densidad de población inicial, $u(x)$. El momento inicial elegido es el año $t_0 = 1991$. Tomando en cuenta que en $x = 0$, esta función debe ser igual al número de nacimientos en el momento inicial $t = t_0$. Dada la estructura matemática de las variables de entrada de edad, la representación de los datos de la muestra para el año 1991 se sugiere una función que sigue esta estructura:

$$a_1 \cdot \exp\left(-\frac{x}{b_1}\right) + a_2 \cdot \exp\left(-\frac{(x-c_2)^2}{2b_2}\right) + a_3 \cdot \exp\left(-\frac{(x-c_3)^2}{2b_3}\right) + a_4 \cdot \exp\left(-\frac{(x-c_4)^2}{2b_4}\right).$$

Los valores de c_2, c_3 y c_4 se han encontraron observando los picos de la función y variándolos, buscando la mejor función de ajuste. Ahora los datos utilizados para ajustar las composiciones familiares por edad de inmigrantes y emigrantes, $p_i(x)$ y $p_e(x)$, son el promedio de los datos correspondientes a los años 1998, 1999 y 2000 (observar tabla 1 en la sección de anexos de [16]). Aquí nuestras funciones $p_i(x)$ y $p_e(x)$ se obtienen asumiendo que tienen la misma estructura matemática que $u(x)$, dada por la ecuación mencionada al inicio de esta sección, observando los picos de los diagramas correspondientes y siguiendo el mismo procedimiento utilizado para $u(x)$. Los valores encontrados para los parámetros de la función se muestran en las siguientes tablas:

CUADRO 2. Valores de parámetros para $u(x)$ [16].

Parametro	valor	Parametro	valor
a_1	3290.78	a_3	3432.63
a_2	12331.5	a_4	8332.92
b_1	2.76422	b_3	63.2046
b_2	130.282	b_4	294.71
c_2	17.5	c_4	57.5

CUADRO 3. Valores de parámetros para $pi(x)$ [16].

Parametro	valor	Parametro	valor
a_1	0.0153467	a_3	0.00958857
a_2	0.0316831	a_4	0.00425549
b_1	9.56829	b_3	103.227
b_2	44.0264	b_4	32.4823
c_2	27.5	c_4	67.5

ahora podemos ver a continuación las representaciones gráficas de las funciones ajustadas $p_i(x)$ y $p_e(x)$, junto con los datos reales obeservando los picos de ambos diagramas.

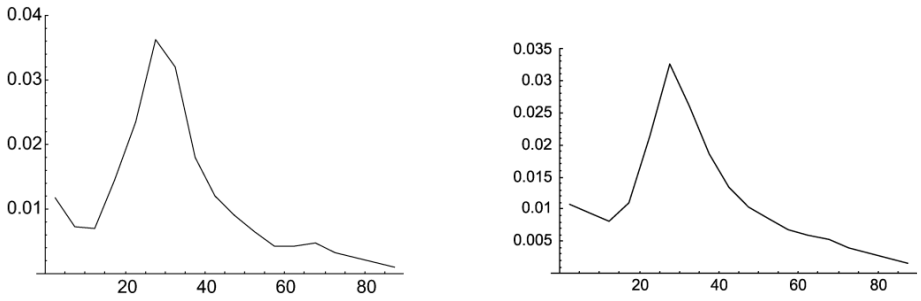


FIGURA 3. Valores promedio de los datos reales para la composición familiar edad-inmigrante en la ciudad de Valencia correspondientes a los años 1998, 1999 y 2000 [16].

La tasa de mortalidad por edad $b(x)$ que se encuentra en la base de datos es la de 1991 y 1996, respectivamente (observar tabla 1 en la sección de anexos de [16]). Veamos la siguiente figura donde se muestra los valores promedio de $b(x)$ en estadísticas actuariales.

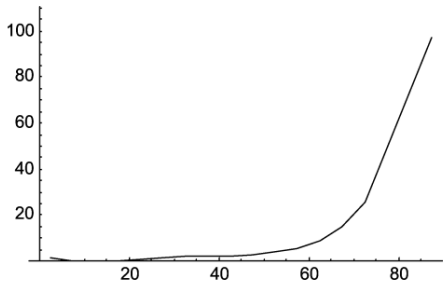


FIGURA 4. Valores promedio de los datos reales para la tasa de mortalidad por edad en la ciudad de Valencia correspondientes a los años 1991 y 1996 [16].

Se sugiere una función como:

$$a_1 \cdot \exp\left(-\frac{x}{b_1}\right) + a_2 \cdot \exp\left(-\frac{(x-c_2)^2}{2b_2}\right) + a_3 \cdot \exp\left(-\frac{(x-c_3)^2}{2b_3}\right). \tag{3.11}$$

De manera similar, los valores de c_2 y c_3 se han encontrado siguiendo el mismo procedimiento que para $u(x)$, $p_i(x)$ y $p_e(x)$.

Tomando en cuenta tres rangos diferentes para la variable de edad correspondientes a las tres funciones dadas en 3.11 con el fin de optimizar el proceso de ajuste. Los valores encontrados se muestran en la siguiente tabla, seguidamente con su representación gráfica de la función ajustada $b(x)$ junto con los datos reales.

CUADRO 4. Valores de parámetros para $pi(x)$ [16].

Parametro	valor	Parametro	valor
a_1	5.50524	b_3	10.915
a_2	1.24254	c_2	35.5
a_3	49.0905	c_3	80
b_1	1.97292	b_2	47.18585

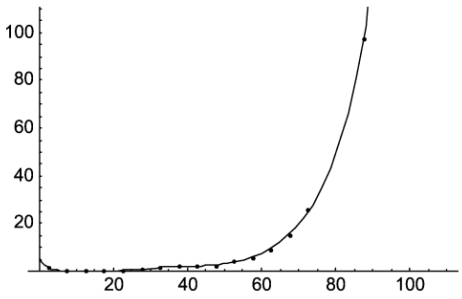


FIGURA 5. Función de ajuste (línea) y datos reales (puntos) para la edad-muerte [16].

Veamos ahora la estructura matemática de la fertilidad $a(t)$ y la mortalidad $g(t)$. Supongamos que ambos indicadores cambian con el tiempo bajo la influencia de factores socioeconómicos. El supuesto utilizado para ambas funciones es que son sumas de funciones logísticas. Los datos obtenidos en la base de datos para ambas tasas impositivas cubren los años 1985-2001 (revisar tabla 2 en la sección de anexos de [16]). Su representación gráfica asume una función logística en ambos casos:

$$\frac{\beta}{1+(\frac{\beta}{y_0}-1)exp(-\alpha(t-t_0)))}. \tag{3.12}$$

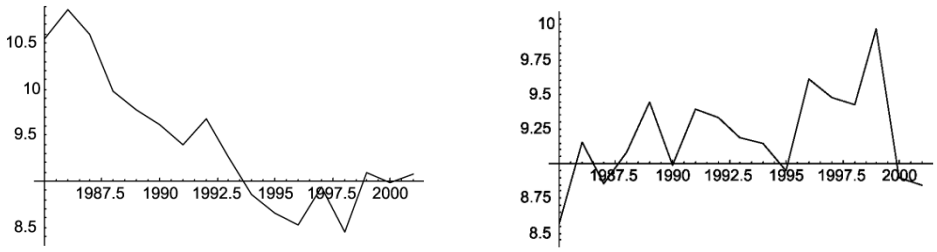


FIGURA 6. Valores de los datos reales de la tasa bruta de mortalidad de la ciudad de Valencia correspondientes al periodo 1985-2001 [16].

Si $b/y_0 < 1$, la función logística disminuye con el tiempo, como lo demuestran los datos de $a(t)$ (gráfico de la derecha), pero ha aumentado ligeramente en los últimos años. Y si $b/y_0 > 1$, entonces la función logística aumenta con el tiempo, como lo demuestran los datos correspondientes a $g(t)$ (gráfico de la izquierda). Los valores encontrados para ambos tasas se muestran en la siguiente tabla:

CUADRO 5. Valores de los parámetros para $a(t)$ y $g(t)$ [16].

	α	β	t_0	y_0
$a(t)$	0.129499	0.008474	1986	10.5688
$b(t)$	0.558679	0.009288	1985	8.36685

desde un punto de vista cualitativo o gráfico se puede deducir que las cifras presentadas en esta sección muestran que las curvas de densidad de población proyectadas para 1996 y 2001 tienen formas y tendencias similares a las curvas reales, es decir, las curvas tienen una estructura similar.

4. CONCLUSIONES

Hemos hecho un breve acercamiento relacionado a datos históricos, de quienes fueron las principales investigadores de esta ecuación y cual fue su aporte a la sociedad con esta misma (revisar la sección 2 de antecedentes). Hemos hablado desde las generalidades que existen en esta ecuación, hasta una solución de ella misma. También, hemos visto que es posible encontrar una solución para la ecuación de McKendrick en todo su dominio de las variables a y t bajo las condiciones iniciales y de frontera.

En conclusión el análisis de soluciones débiles para la ecuación de McKendrick-von Foerster con tasas variables ofrece una perspectiva más realista sobre la dinámica poblacional en comparación con modelos simplificados. Las variaciones en las tasas de natalidad y mortalidad reflejan mejor las condiciones biológicas y sociales reales, proporcionando información valiosa para la gestión de poblaciones en biología y epidemiología. Sin embargo, es importante considerar que los resultados pueden estar limitados por la precisión de las tasas utilizadas y los supuestos del modelo. La utilización de soluciones débiles para la ecuación de McKendrick-von Foerster demuestra una mayor flexibilidad en el modelado de la dinámica poblacional, especialmente cuando se consideran tasas de natalidad y mortalidad que varían con la edad. Este enfoque permite adaptar el modelo a diferentes contextos biológicos y sociales, proporcionando una herramienta más versátil para el análisis de poblaciones reales.

Para investigaciones futuras se puede tomar en consideración el estudio del caso de una edad limitada. Para poder hacer estos modelos de población, es fundamental considerar varios factores. Primero, debido a que las formas de vida no existen indefinidamente, es importante limitar la edad máxima de los humanos. Además, las tasas de natalidad y mortalidad podrían modelarse de forma no lineal para reflejar los cambios complejos causados por factores como la mortalidad y la disponibilidad de recursos.

Referencias

1. H. von Foerster: *Some remarks on changing populations, in: The Kinetics of Cellular Proliferation*, Grune and Stratton, New York, 1959, pp. 382-407.
2. M'Kendrick, A. G. (1925), *Applications of Mathematics to Medical Problems*
3. Cercignani, *The Boltzmann Equation and Its Applications*. Springer-Verlag, New York, 1988.
4. White, Frank (1991), *Viscous Fluid Flow. 3rd Edition*, McGraw-Hill Mechanical Engineering. ISBN-10: 0072402318.
5. A. Lasota, M. C. Mackey, M. Ważewska-Czyżewska: *Minimizing Therapeutically Induced Anemia*, J. Math. Biol. 13 (1981), 149-158.
6. M. E. Gurtin, R. C. McCamy: *Non-linear age-dependent Population Dynamics*, Arch. Rat. Mech. Anal. 54 (1974), 281.
7. B. L. KEYFIT, *The McKendrick Partial Differential Equation and Its Uses in Epidemiology and Population Study*, Department of Mathematics, University of Houston Houston, TX 77204-3476. U.S.A
8. Carlos Héctor Daniel Alliera, *Tesis de Licenciatura Estudio y Aplicaciones de Black Scholes*, Buenos Aires, Abril de 2007.
9. N. Kato: *A general model of size-dependent population dynamics with nonlinear growth rate*, J. Math. Anal. Appl. 297 (2004) 234-256.
10. Claudia Gisela Espinosa Guia, Cándido Eugenio Aguilar Aguilar, Raciel Vázquez Aguilar, *Sofía Kovalesvkaia su historia a través del género y las Matemáticas*, 02/08/2022.
11. Von Foerster equation (sin fecha) DBpedia. Disponible en: *https* : *//dbpedia.org/page/Von*$_{F}$*oerster*$_{e}$*quation* (Accedido: 24 June 2024).
12. *Anderson Gray Mckendrick*, 8 September 1876 - d. 30 May 1943.
13. R. Dilao and A. Lakmeche, *Diffusion in the Mckendrick-Von Foerster Equation*, Proeeedings ofDynamie Systems & Applieations 4 (2004) 647-653.
14. Morton E. Gurtin & Richard C. Maccamy, *Non-linear Age-dependent Population Dynamics*.
15. Rui Dilao and Abdelkader Lakmeche, *On The Weak Solutions Of The Mckendrick Equation*, Av. Rovisco Pais, 1049-001 Lisbon, Portugal.
16. Joan C. Micó, David Soler, Antonio Caselles, *Age-Structured Human Population Dynamics*, Departament de Matemática Aplicada , Universitat de València 16 Aug 2006.

Departamento de matemáticas, Universidad Nacional Autónoma de Honduras

Email address: magarcias@unah.hn

IMPLEMENTACIÓN DE MARTINGALAS AL MERCADO FINANCIERO

ALEJANDRO JOSÉ VÁSQUEZ

RESUMEN. En este artículo se presenta de forma detallada, los procesos estocásticos, las martingalas y los precios de acciones. El objetivo es presentar teoricámente las martingalas y construir un modelo matemático para los precios de acciones, donde se establece una relación entre la teoría de martingalas y los precios de acciones en el mercado finaciero. Además, se detallan los conceptos financieros, los de probabilidad y la teoría de martingalas. Finalmente se construye un modelo para los precios de acciones en el mercado finaciero aplicado a martingalas bajo ciertas condiciones a observar; como la ganacia o pérdida a futuro.

ABSTRACT. This article presents in detail stochastic processes, martingales and stock prices. The objective is to theoretically present martingales and build a mathematical model for stock prices. Where a relationship is established between the martingale theory and stock prices in the financial market. In addition, financial concepts, probability concepts and martingale theory are detailed. Finally, a model is built for share prices in the financial market applied to martingales under certain conditions to be observed; such as future profit or loss.

1. INTRODUCCIÓN

En estadística y teoría de la probabilidad un proceso estocástico o proceso aleatorio, es un concepto matemático relacionado con sucesiones de variables aleatorias que cambian obedeciendo al azar en función de otra variable; por ejemplo el tiempo [4]. Dentro de los procesos estocásticos se encuentran las martingalas, se popularizó en el siglo XVIII con fama de ser una estrategia ingenua y propia de mentes simples, puesto que aunque en apariencia es infalible, está, sin embargo, abocada a arruinar al jugador. Recibe el nombre de los habitantes de la localidad francesa de Martigues (martingales en francés), situada en las cercanías de Marsella, que por aquel entonces tenían fama de ser ingenuos y simplones. Tienen múltiples aplicaciones en finanzas, ciencias económicas, ingeniería, entre otras.

Por su parte en un mercado financiero una acción es un título emitido por una empresa o compañía, que le da al poseedor el derecho de ser propietario de una parte de la empresa; las ganancias se obtienen cuando el valor de las acciones aumentan y por el reparto de dividendos. Una acción corresponde a una operación financiera con riesgo pues su ganancia o valor futuro no se puede determinar con antelación; el valor futuro de una acción depende de muchos factores y variables que no necesariamente permanecen constante a través del tiempo [3].

Fecha: 12 de Agosto del 2024.

Palabras y frases clave. Martingalas, Modelos de Black-Shole y Binomial, Proceso Estocático.

El objetivo de este artículo es la presentación y descripción teórica de las martingalas y sus propiedades; de igual manera se construyen los principales modelos matemáticos que describen la evolución en el precio de una acción y de una opción sobre acciones, tanto para el caso discreto como para el caso continuo. Por último se establece la relación entre la teoría martingala y los modelos matemáticos que describen el precio de una acción, es decir se termina demostrando que dichos modlos bajo ciertas condiciones observadas constituyen una martingala y de ahí la relación entre martingalas y los modelos matemáticos que describen el precio de una acción o acciones.

También se presenta los principales conceptos de finanzas: operaciones con riesgos; operaciones sin riesgos, mercado de opciones, contrato de futuros y contrato forward; de igual modo lo concerniente a la teoría de la medida y probabilidad; concepto de σ-álgebra, funciones medibles, funciones simples y convergencia en funciones medibles, haciéndo especialmente énfasis en los tipos de convergencia para luego extenderlo a la convergencia de una variable aleatoria.

En cuanto a lo relacionado con los conceptos de probabilidad se presentan los elementos de función de distribución con algunos ejemplos y los teoremas de la ley de los grandes números y del límite central [4]. Se hace especial énfasis en los conceptos de valor esperado o esperanza de una variable aleatoria y sus propiedades, debido a su importancia para construir los conceptos de martingalas.

Se hace una descripción o desarrollo de la teoría de las martingalas, sus características y propiedades; teniendo en cuenta los conceptos de esperanza y valor esperado. Los elementos desarrollados son utilizados para mostrar la relación de la teoría de los procesos estocásticos y las martingalas con los modelos de precios de acciones [1, 3, 4].

Una relación de forma precisa al modelo de martingalas con respecto al mercado financiero, constituye la solución final al problema de implementación. Se desarrolla y se construyen los principales modelos de precios de acciones y de opciones sobre acciones; con un modelo binomial, el cual es apto para modelar el precio de una acción y de una opción sobre acciones en tiempo discreto, se muestra que este modelo tiene como función de densidad la función binomial.

El modelo de Black-Sholes, con el cual se describe el precio de una acción y de una opción sobre acciones en tiempo continuo, sólo permite el trabajo con opciones de tipo europeo, su desarrollo requiere el trabajo con ecuaciones diferenciales estocásticas y el uso de las integrales. Se demuestra que los modelos que describen la evolución en el precio de una acción constituyen una martingala, supermartingala o submartingala; en general se presentan las condiciones bajo las cuales la teoría de las martingalas son propicias para describir especialmente el precio de una acción en el mercado bursátil [2].

Esta es una línea de investigación en la orientación en estadística donde involugran la probabilidad, los procesos estocásticos y la economía vinculado con los temas prioritarios de la Univercidad Nacional Autónoma de Honduras (U.N.A.H.), como ser la globalización, productividad y competitividad.

2. Antecedentes

El concepto de martingala fue incorporado a la teoría de la probabilidad por Paul Levy (matemático frances, 1886-1971), y parte de su desarrollo inicial fue realizado por Joseph Doob (matemático americano, 1910 al 2004), cuya motivación fue demostrar la inexistencia de estrategias de juegos infalibles. Una martingala es un modelo matemático para una secuencia de apuestas justas que ha encontrado muchas aplicaciones en la probabilidad teórica y aplicada [5].
Actualmente encontramos muchos problemas de caminatas aleatorias que son resueltos a través de las martingalas, entre estos tenemos la dinámica de juego de la ruleta americana, la predicción en los resultados electorales, optimización del interés de retorno en una inversión y múltiples aplicaciones en la economía y por otra parte, el cálculo estocástico moderno involucra integración estocástica con respecto a martingalas, las cuales incluyen el movimiento browniano y el proceso de Poisson compensado.
Una de las áreas donde hoy se aplica la teoría general de la integración estocástica es la economía y la aplicacion de la fórmula de la opción de Black-Scholes, una de las ecuaciones más famosas de las finanzas, se desarrolla en 1970 pero se publica en 1973 [5].

Fischer Black (1938–1995) y Myron S. Scholes (1941, Premio Nobel de Economía 1997), comenzaron a trabajar conjuntamente en el tema en 1969. Fischer Black recuerda que envió el artículo al "Journal of Political Economyz rápidamente lo recibió de vuelta con una carta de rechazo. Decía que era demasiado especializado para ellos, y que sería mejor en el "Journal of Finance". Entonces lo envió a "The Review of Economics and Statistics", y también volvió con una carta de rechazo. Decía que sólo podían publicar una cantidad muy pequeña de los ensayos que recibían, ninguno de estos Jounals quieso pasar el ensayo a un revisor por que creía que era por su dirección postal que no era académica; despues de resivir las carta reescribieron el ensayo para poner énfasis en los aspectos econónmicos subyacentes (la acción ordinaria) en la derivación de su fórmula [6].
Dos academicos de la Universidad de Chicago, Merton Miller y Eugene Fama se enteraron del rechazo de su ensayo y sugirieron al "Journal of Political Economy"que el ensayo deveria tener una segunada consideración por sus buenos resultados. En agosto de 1971 el Journal aceptó el ensayo, de forma condicinal por si el revisor les pedia hacer cambios al ensayo, el borrador final fue emitido en mayo de 1972, tuvo el título de "The pricing of options and corporate liabilities", apació en la edición de mayo a junio de 1973.

Black y Scholes comezaron, a penzar como aplicar su fórmula para cuantificar los valores de los bonos con riesgo y las acciones. El argumento de que la posición compensada de opciones y acciones con transacciones continuas es literalmente sin riesgos fue sugerida por Merton Miller, Comentando que la verción final del ensayo de la fórmula que dedujeron estaba en su forma más general en comparción con la fórmula que habían hecho anteriormente, suponiendo que el riesgo total de la acci'on puede eliminarse por diversificación, lo mismo pasa con el riesgo de la opción [6].
Todo el desarrollo e interacciones que tubo la fórmula del modelo de Black-Scholes fueron mencionado en el artículo de Merton Miller "The theory of rational option

pricing", publicado por Bell Joarnal of Economics en 1973, el cual fue publicado posteriormente después del ensayo de Black y Scholes.
En 1990, en el libro "Continuous Time Finance", Merton Miller comenta que los modelos en tiempo continuo reafirman el analisis de los modelos financieros, no hay mejor ejemplo que la contribuci'on del modelo de Black-Scholes que desde el día de su publicación, renovó el campo de la valuaci'on de opciones y la deuda de una empresa; asi esta área abre nuevas puertas, el estudio aplicado y empírico en las finanzas [6].

Esta teoría de valuación de opciones creada por Fisher Black y Myron Scholes donde escribieron sobre la valoración de opciones compradoras y utilizando su fórmula crearon una cartera que instante a instate no tenía riesgo. Esta cartera consistía en compra de una acción y la venta de la opción. Esta posición se revisa continuamente de acuerdo con los cambios del valor de la acción, de modo que la rentabilidad de esta cartera es segura. Para evitar oportunidades de arbitraje el valor de esta cartera en cualquier momento tiene que ser su valor final descontado a la tasa sin riesgo. Una ecuación en derivadas parciales se obtine de la dinámica que gobierna esta cartera cuya solución con las condiciones de contorno habituales es la fórmula de Black-Scholes.
Merton en 1977 encontró un modo general de valoración de activos derivados. Un activo derivado es un activo cuyo valor en una futura fecha especificada está unicamente determinado por el valor de otro activo, el activo principal. Una opción compradora es un activo derivado cuyo activo principal es la acción y por tanto, la valoración de opciones compradoras es una aplicación particular de este método de valoración de Merton que es más general [6].

Las hipótesis del modelo de valoración de activos de Merton son la existencia de un activo sin riesgo cuya tasa de retorno r, donde es constante y conocida sobre el tiempo. Un activo para el cual se conoce el proceso estocástico que genera su valor sobre el tiempo y que puede ser representado por un proceso de difusión y un segundo activo, el activo derivado. Merton da el valor de este segundo activo como solución de una ecuación de derivadas parciales con condiciones de contorno dependiendo de los términos del activo derivado. También especifica la" trading strategy", a un tiempo continuo que produce una cartera compuesta del activo sin riesgo y del primer activo, cuyo valor en el primer instante es el valor del activo derivado. Este modelo permite que el activo principal y el activo derivado paguen a sus poseedores una corriente continua de dividendos.
Un título diferente de la valoración de activos derivados fue sugerido por Cox y Ross en 1976 y desarrollado por Harrison y Kreps en 1979. Ellos dan condiciones sobre el precio de activos para que sean martingalas en una adecuada filtración y de un adecuado espacio de probabilidad. El valor actual de esos activos es su valor esperado futuro en ese espacio de probabilidad [6].

3. modelo de precios de acciones aplicado a martingala

3.1. Modelo Binomial. Toda la referencia del modelo binomial aplicado a martingala fue tomado de [7].

La importancia del modelo binomial es que se puede utilizar principalmente para los conceptos de arbitraje, valuación y su relación con la probabilidad neutra al riesgo; es un modelo discreto, que con un número suficiente de pasos da una buena aproximación al modelo de tiempo continuo donde podemos aplicarle la teoría y conceptos de esperanza condicional y martimgala.
Para determinar las características del modelo binomial y cómo este es apropiado para describir el precio de la acción ordinaria (subyacente) y valuar una opción sobre acciones utilizando períodos o pasos de nodos múltibles.

Definición 3.1. (Distribución binomial). Sea X una variable aletoria, se dice que X tiene una distribución binomial si y sólo si su distribución de probabilidad está dada por:

$$B(n,p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3.1}$$

donde $n \in N$; $\quad x = 0, 1, 2, ..., n$; $\quad 0 < p < 1$.

En una distribución binomial se cumple que:

$$E(X) = np; \quad Var(X) = np(1-p). \tag{3.2}$$

Definición 3.2. (Oportunidad Arbitraje). Una oportunidad de arbitraje es la posibilidad de recibir ganancias sin usar el capital propio; en una oportunidad de arbitraje no hay la posibilidad o riesgos de obtener pérdidas.

Definición 3.3. Sea $S(t)$ el capital de un inversionista que invierte en la bolsa de valores, con $0 < t \leq T$; se dice que existe una oportunidad de arbitraje para el inversionista, si empezando con el capital $S(0) = 0$, el capital $S(T)$ cumple que:

$$S(T) \geq 0 \ \ y \ \ P(\{S(T) \geq 0\}) \geq 0. \tag{3.3}$$

Esta ecuación nos dice que el inversor no tiene nunca deudas y siempre tiene posibilidad de ganar, sabemos que en la práctica cuando se invierte en acciones, siempre hay posibilidad de ganar o de perder; por lo que se impondrán algunas restricciones al modelo binomial, de forma tal que no exista oportunidad de arbitraje. por lo que se conoce como mercado de acciones aplicado al modelo binomial.
Un mercado de acciones en un período es cuando el supuesto es $t = 0$; para un inversionista existen como oportunidades de inversión compra o venta de acciones a un precio actual
$S_1(0) = S_0 > 0$ y el precio a futuro como se muesta en (3,4) y (3,5):

$$S(T) = \begin{cases} C(1+r)^t + HS_0U \text{ probabilidad p} \\ \\ C(1+r)^t + HS_0D \text{ probabilidad q} \end{cases} \tag{3.4}$$

$$S(T) = \begin{cases} Ce^{rT} + HS_0U \text{ probabilidad p} \\ \\ Ce^{rT} + HS_0D \text{ probabilidad q.} \end{cases} \tag{3.5}$$

donde $q = 1 - p$ y $0 < p < 1$. C representa el capital invertido en $t = 0$ y H el número de acciones negociadas en $t = 0$.

De acuerdo a (3,4) en $t = 0$ un inversionista podrá invertir su dinero comprando o solicitando acciones y dinero prestado e invertirlo. En caso de necesitar más acciones de las que le permite el capital inicial C, podrá prestar dinero; en caso que invierta menos dinero del capital inicial C, invertirá el resto en algunas operación sin riesgos, por ejemplo en depósito a término fijo.

Si

$$U > D \geq e^{rT},$$

entonces se tomaría como estrategia tomar el préstamo, se compran acciones, en el período de maduración (T) se paga el préstamo y se recogen los beneficios de las acciones.

Si $-C = HS_0 > 0$, entonces obtenemos los casos de interés discreto y continúo:

$$S(T) = \begin{cases} -gS_0C(1+r)^t + HS_0U \text{ probabilidad p} \\ \\ -gS_0C(1+r)^t + HS_0D \text{ probabilidad q} \end{cases} \tag{3.6}$$

$$S(T) = \begin{cases} -gS_0Ce^{rT} + HS_0U \text{ probabilidad p} \\ \\ -gS_0Ce^{rT} + HS_0D \text{ probabilidad q} \end{cases} \tag{3.7}$$

En ambos casos se cumple $S(T) \geq 0$, y por consiguiente se puede ganar beneficio de arbitraje. Para evitar oportunidad de arbitraje se impone la restricción:

$$D < e^{rT} < U \;\; o \;\; D < 1 + r < U,$$

para los casos de interés continuo y discreto respectivamente.
En este modelo binomial se está definiendo el número de movimiento del precio de las acciones donde se distribuye según una binomial de parámetros 1 y p, es decir $B(1, p)$ y se le conoce como modelo binomial a un paso o de un período.

Con este modelo se va a describir el comportamiento del precio de las acciones a través del tiempo; supongamos que S_0 es el precio de una acción en el momento $t = 0$, es evidente que transcúrrio un período, el precio de la acción S_1 puede ser mayor que S_0 o menor que S_0.

Ejemplo: Sea un árbol binomial que modela el precio de un activo (una acción) en un paso

$$S_0 = 2000; \;\; s_1(x_1) = \frac{3}{2}S_0 = 3000; \;\; S_1(x_2) = \frac{3}{4}S_0 = 1500,$$

es decir que dentro de un período la acción puede tomar un valor de 3000 o 1500.
Ahora se va a considerar una combinación que consiste en la adquisición de un número de acciones ordinarias o subyacentes al mismo tiempo que se emite una opción de compra sobre ellas, de modo que la cartera proporcione el mismo flujo de caja tanto si las acciones ordinarias suben como si bajan; teniendo asi una cartera libre de riesgo.

Sea H el número de acciones ordinarias que se adquirió, el valor de una acción ordinatria dentro de un período es de 3000 y el valor de una opción es de 1000, el valor de portafolio o flujo de caja viene dado por la expresión:

$$3000H - 1000.$$

En el caso de que el valor de la acción ordinaria sea de 1500 y el de la opción 0 se observa que la opción no se realiza y esta dado por:

$$1500H - 0.$$

Ahora igualando los dos flujos de caja, y como el portafolio es libre de riesgos, se obtiene que $H = \frac{2}{3}$, es decir que el portafolio formado por $\frac{2}{3}$ de una acción ordinaria y la venta de una opción de compra sobre ella no tiene ningún riesgo (siempre tendrá el mismo valor) y el rendimiento que se obtendrá de ella a lo largo del tiempo será un rendimiento sin riesgo (r), definido por:

$$\frac{\text{flujo caja}}{\text{sobre inversión}} = 1 + r. \tag{3.8}$$

Observación: r representa la tasa de interés dada en el mercado de dinero, es decir el porcentaje de dinero que se debe pagar por cada unidad monetaria tomada en préstamo y $1+r$ el valor de cada unidad monetaria invertida en el sigiente período. Donde S_0 es el precio de la acción en el momento $t = 0$. Sea U factor de ascenso y D factor de descenso, c el precio de la opción de compra, c_u y c_d precios de la opción de compra en caso que la acción ordinaria haya ascendido o bajado respectivamente. El portafolio o flujo de caja esperado está dada por:

$$H = \frac{c_u - c_d}{S_0(U - D)}. \tag{3.9}$$

Es decir que H es la razón de cambio entre el precio de las opciones y el cambio entre el precio de las acciones. Sabemos que S_t representa el valor de la acción para un periodo de tiempo determinado t. Ahora obteniendo una expresión para el valor de la opción de compra c, operando através de la relación de rentabilidad libre de riesgo que esta dada por:

$$c = \frac{c_u p + c_d(1 - p)}{(1 + r)}, \tag{3.10}$$

$$c = \frac{c_u p + c_d(1 - p)}{e^r}. \tag{3.11}$$

Se puede concluir que el valor actual de una opción de compra consiste en calcular la media ponderada de los flujos compra, tanto si el precio de la acción ordinaria asciende o si desciende. Ahora bien para aplicar el modelo binomial en dos períodos usamos el mismo ejemplo anterior, donde U es el factor de crecimiento y D factor de descenso. El valor de la acción ordinaria a podido llegar a un valor máximo de 4500, un valor mínimo de 1125 y un valor intermedio de 2250. Por otra parte el valor de la opción de compra sería de 2500, 250 y 0.

Podemos calcular el precio de acción ordinaria (S) con el modelo binomial en

dos períodos donde se obtienen tres posibilidades de precios: S_0U^2, precio alcista; S_0UD, precio medio y S_0D^2, precio descentente es decir:

$$S_2 = \begin{cases} S_0U^2 & \text{probabilidad} \quad p^2 \\ S_0UD & \text{probabilidad} \quad p(1-p) \\ S_0D^2 & \text{probabilidad} \quad (1-p)^2 \end{cases} \tag{3.12}$$

En general para el modelo binomial multiPle podemos utilizar la siguiente proposición.

Proposición: Sea S_t el precio de acción ordinaria en un período de tiempo determinado t, X el número de pasos hacia arriba en la evolución del precio de la acciÓn ordinaria y t el numero de períodos entonces

$$S_t = S_0U^XD^{t-X}, \tag{3.13}$$

donde la variable aleatoria X se distribuye binomial $X \sim B(t,p)$, p representa la probabilidad con la que aparece U Y t el número de períodos.

Para el caso general el precio de acción de compra se puede obtener por medio del triángulo de Pascal y la conbinatoria; esto es, el valor actual de una opción de compra que es el valor actual de los flujos de cajas esperado a lo largo de un árbol binomial y para un número de períodos t que esta dada por:

$$c = \frac{1}{(1+r)^n}\sum_{x=0}^{n}\{\binom{n}{x}p^x(1-p)^{n-x}max\{S_0U^xD^{n-x} - S_0, 0\}\}, \tag{3.14}$$

$$c = \frac{1}{e^{rn}}\sum_{x=0}^{n}\{\binom{n}{x}p^x(1-p)^{n-x}max\{S_0U^xD^{n-x} - S_0, 0\}\}. \tag{3.15}$$

Concluimos que (3,14) y (3,15) cumplen los siguientes supuestos:

1. Los factores U Y D son constantes en todos los períodos o pasos; así mismo las varianzas de los rendimientos.
2. La distribución de los precios de las acciones es una binomial multiplicativa.
3. Los tipos de interés sin riesgos son constantes.
4. No existe coste de transacción, se puede establecer una cobertura sin riesgo para cada período entre la opción y la acción ordinaria (subyacente).

Como hemos visto la aplicación del modelo binomial al precio de acciones ahora podemos adaptarlo a una martingala como alternativa, para describir la evolución del precio de una acción en tiempo discreto; es decir se va establecer que dicho modelo ya sea un paso, dos pasos, o pasos múltiples es una martingalas.

Proposición: Sea S_t el precio de una acción en el momento t, el valor esperado de la acción es

$$E(S_t) = SUp + SD(1-p). \tag{3.16}$$

Proposición: El valor esperado de una acción para el modelo binomial de periodos múltibles es

$$E(S_t) = S_{t-1}(SUp + SD(1-p)). \tag{3.17}$$

Proposición: Sea S_t el precio de una acción, el modelo matemático $\{S_t, F_t\}$ repesenta una martingalas.

Vemos que una martingala se cumple que $E(S_t) < \infty$ es integrable, sea adaptado a la filtración F_t y $E(S_{t+1}|F_t) = S_t$ casi seguro.
Sabemos que $E(S_t) < \infty$ (finita) por el modelo binomial decimos que es integrable. Supongamos que x es el elemento alcista en el precio de la acción y y es el elemento bajista en el precio de la misma acción, de modo que $S_1(x) = S_0D$ y $S_1(y) = S_0D$, y asi por el modelo binomial múltiple tendriamos los conjuntos F_t que son los σ-álgebra para $t = 0, 1, 2, 3, ...$, y σ-álgebra tiene la propiedad que

$$F_0 \subseteq F_1 \subseteq F_2 \subseteq F_3 \subseteq ...$$

entonces es monotona por tanto es adactada a la firtaración F_t, ahora consideremos el precio de la acción S_t con $t \geq 0$ y $S_{t+1}(x) = S_0D$, $S_{t+1}(y) = S_0D$ y asi sucesivamente para $t = 0, 1, 2, 3, ..., n$ vamos a deternimar $E(S_{t+1}|F_t)$.

$$\begin{aligned}\int_\Omega E(S_{t+1}|F_t)\ dp &= E(S_{t+1}|F_t) \\ &= S_0(U_P + D(1-P))^t \\ &= S_{t-1}(U_P + D(1-P)) \\ &= S_t(U_P + D(1-P)).\end{aligned}$$

Se cumple que $E(S_{t+1}|F_t) = S_t(U_P + D(1-P))$ por tanto se concluye que el modelo binomial que decribe el comportamiento del precio de una acción es una martingalas.
Observación: S_{t+1} es F_t-medible, donde F_t es la σ-álgebra que contiene toda la información correspondiente a los valores o precios anteriores de la acción es decir, $\{S_0, S_1, S_2, ..., S_{t-1}\} \subseteq F_t$. Por otra parte S_{t+1} será una martingala si

$$(U_P + D(1-P)) = 1$$

una submartingala si

$$(U_P + D(1-P)) \geq 1$$

o una supermartingala si

$$(U_P + D(1-P)) \leq 1.$$

3.2. Modelo de Black-Scholes. Toda la referencia del modelo de Black-Scholes aplicado a las martingala fue tomado de [8].

Para formular el modelo de mercado comenzaremos definiendo el conjunto temporal T que es la fecha terminal de toda actividad económica. donde los activos se pueden negociar en cualquier tiempo t entre 0 y T $(0 \leq t \leq T)$, tomamos un espacio de probabilidad (Ω, F, P) y sea $\omega \in \Omega$ se puede interpretar como una completa descipción de un posible estado del mundo. F es la σ-álgebra de sucesos distinguible en el momento T y P la probabilidad del espacio (Ω, F). Es necesario en el modelo que todos los inversores estén de acuerdo en que los sucesos F tienen una probabilidad de ocurrir pero no es necesario que estén de acuerdo en la asignación de las probabilidades.

Los inversores están informados de la estructuta de la familia de sub-σ-álgebra de $F = \{F_t : 0 \leq t \leq T\}$ que es la completa información a lo largo el tiempo.
Supondremos que F es creciente ($F_t \subset F_s$ *para* $t \leq s$) es decir, que la información

revelada no se olvida. $F_T = F$ es decir, que en el momento T nosotros conocemos cual es el vedadero estado del mundo. En el momento 0 sólo conocemos los sucesos que tienen probabilidad 0 ó 1 de ocurrir en otras palabras F es una filtración.

Los activos disponibles en el mercado financiero de Black-Scholes son:

1. Un activo principal que por conveniencia supondremos que no paga dividendos antes del momento T. Podemos pensar en una acción pero podría ser cualquier otro activo.
2. Un activo sin riesgo con tasa de rentabilidad conocida y constante r.

Supondremos que los precios S_t del activo principal a lo largo del tiempo siguen una distribución logarítmica normal es decir:

$$S_t = exp((\mu - \frac{1}{2}\sigma^2)t + \sigma\omega(t))$$

donde ω(t) es un movimiento Browniano standard.

El valor del activo sin riesgo en el momento t lo representamos como $S_0(t)$. Este activo da derecho a un dólar en el momento T, por lo que

$$S_0(t) = exp(-r(T-t)).$$

La información F_t disponible en el momento t es la que revelan los precios del activo principal hasta el momento t, no se conoce los precios despues del momento t, por tanto F_t es la sub-σ-álgrebra generada por los S_s para $0 \leq s \leq t$.
También supondremos que estamos en un mercado sin fricciones, es decir que no hay ni impuesto ni costes de transacciones. El tomar prestado y la venta a crédito está permitido sin resticciones y los tipos de interés para tomar prestado y para prestar son los mismos. La negociación tiene lugar continuamente (estamos en un mercado continuo).
Una martingala sobre un espacio de medida (Ω, F, Q) con respecto a la filtración F_t, es un proceso estocástico X_t cuyo valor esperado en una fecha futura es su valor actual

$$E_Q(X_t|F_t) = X_t \; para \; t \leq s.$$

Los precios descontados de un activo no son los valores actuales de ese activo, sino que son los precios que tendrían en una economía en la que los tipos de interés fueran cero.

Un activo derivado se dice que es negociable si existe una estrategia autofinanciada que se duplica en cualquier momento el valor del activo derivado. Si un activo derivado es negociable su valor tiene que ser el precio de esta estrategia duplicante autofinanciada. Si no hay oportunidades de arbitraje, todas las estrategias duplicantes tienen que tener el mismo precio.

Proposición: Si nuestro sistema de precios (S_0, S_1) no permite oportunidad de arbitraje, existe una probabilidad Q sobre (Ω, F), es equivalente a P, para el cual el precio descontado del primer activo es

$$S_1^0(t) = \frac{S_1(t)}{S_0(t)}$$

es una martingala. Si la probabilidad es Q bajo el cual S_1^0 es una martingala, es única, entonces todo activo derivado tiene un único precio compatible con (S_0, S_1) que es el valor neto del futuro o valor esperado bajo Q.

Como nuestro propósito es cómo se valora un activo derivado X cuyo valor en el momento $T \times T$ es una función continua g del valor S_T en el momento T de nuesto activo: $X(T) = g(S_T)$, donde su formula de valoración se puede definir de la siguiente manera:

Dado

$$S_0(t) = exp(-r(T-t))$$

$$S_1(t) = exp((\mu - \frac{1}{2}\sigma^2)t + \sigma\omega(t))$$

los precios desconatdos de nuestros activos son

$$S_0^0 = 1$$

$$S_1^0(t) = \frac{S_1(t)}{S_0(t)} exp(rT) exp((\mu - r - \frac{1}{2}\sigma^2)t + \sigma\omega(t))$$

como Q es una medida de probablidad sobre (Ω, F) equivalente a P y si suponemos una condición necesaria y por teorema de Girsanov, ω representa un movimiento Browniano standard bajo P definido como:

$$S_1^0(t) = exp(rT) exp(\sigma\omega^0(t) - \frac{1}{2}\sigma^2 t)$$

que es una martingala con respecto a Q.

Ahora para calcular el valor de X en cualquier momento t, sabemos que

$$X^0(t) = \frac{X(t)}{S_0(t)} = E^0(\frac{X(s)}{S_0(s)}|F_t) \quad para \ \ t \leq s \tag{3.18}$$

Como $\omega^0(t)$ es un movimiento Browniano standard con respecto a Q, $\omega^0(T) - \omega^0(t)$ está Q-normalmente distribuida con media 0 y varianza $T - t$ y por tanto

$$X(t) = exp(-r(T-t))$$
$$\int_{-\infty}^{+\infty} g(exp(r(T-t))S_1(t)e^{\sigma y - \frac{1}{2}\sigma^2(T-t)}) \frac{1}{\sqrt{2\pi(T-t)}} e^{-\frac{y^2}{2(T-t)}} dy$$

Que da el valor del activo derivado en el momento t.

Utilizando esta consecuencia podemos encontrar la formula de Black-Scholes, para el valor de una opción compradora sobre un activo principal, aplicado a $g(x) =$

$max(x-k,0)$ donde k es el precio de la opción y por tanto

$$X(t) = exp(-r(T-t))$$

$$\int_{-\infty}^{+\infty} max(exp(r(T-t))S_1(t)e^{\sigma y - \frac{1}{2}\sigma^2(T-t)} - k)$$

$$\frac{1}{\sqrt{2\pi(T-t)}} e^{-\frac{y^2}{2(T-t)}} dy$$

$$= S_1(t) \int_{-d_1}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{\mu^2}{2}} du - kexp(-r(T-t)) \int_{-d_2}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{\mu^2}{2}}$$

$$= S_1(t)N(d1) - kexp(-r(T-t))N(d_2).$$

Se cumple y N es es la función de distribución normal standard, en este caso la estrategia duplicante es mantener en cada instante $tN(d_1)$ unidades del activo S_1 y el resto del dinero $X(t) - N(d_1)S_1$ invertirlo en el activo sin riesgo.

Vemos que este modelo de Black-Scholes nos permite estimar la volatilidad del mercado de un activo subyacente en función del precio y el tiempo, es también una estrategia de autofinanciación cuyo pago final es igual al pago de un valor derivado siempre y cuando el inversor pueda comprar y vender activos sin costes por las transacciones.

4. Conclusiones

Los principales aportes de este trabajo con los modelos matemáticos describe la valuación que es el cálculo numérico que se realiza con el fin de asignar un valor monetario de un determinado bien, propiedad o inversión. Que fue realizado para el análisis de una acción tanto para el caso discreto como para el caso continuo y haciendo énfasis en como se implementa la teoría de las martingalas.
Las martingalas constituyen una herramienta importante para ver la tendencia en el precio de las acciones y en general para precios con ciertas condiciones con las que pueden modelar, esto nos dice que tienen una referencia y su precio actual depende del precio del pasado.
El modelo de Black-Scholes tiene sus limitantes debido a que se basa en algunas suposiciones sobre el mercado y tampoco toma en cuenta las comisiones y coste de transacción, que en la realidad si es un gasto extra, además solo se permite manejar las opciones de tipo europeo.

La función binomial para el caso discreto, la normal logarítmica para el caso continuo constituyen funciones de distribución aptas para modelar el precio acción y de una opción sobre una acción. La importancia del modelo binomial radica en que se puede trabajar todas las propiedades de las martingalas y para conceptos importantes de la teoría financiera como es el arbitraje y mercado eficientes que es una base para el modelo continuo, para una cantidad suficientes de periodos se puede tener una buena aproximación.

Estos metodos mencionados son utilizados por organizaciones o empresas para cubrirse ante un eventual cambio en las variables de mercado, debido a que permite pactar precios futuros sobre los productos y así disminuir posibles pérdidas futuras por cambios inesperados en los precios.
Asimismo estos modelos establecen que la predicción de precios, les permiten a las organizaciones o empresas disminuir la incertidumbre de la evolución futura de precios, lo que les ayuda a que puedan proyectar su presupuesto y el desarrollo de sus operaciones de manera correcta.

Referencias

1. Arbeláez J. y Cárcamo. *Un curso rápido de cálculo estocástico para aplicaciones a modelos económicos.*
2. Black, F. and Schole, *The Princing of the Options and Corporate Liabilities. J. Political Economy*, Volumen 81 de 1973.
3. Balbás A, *las Matemáticas de la Economía Financiera*, España. Volumen 102 del 2008.
4. Luis Rincón , *Curso intermedio de probabilidad*, Facultad de Ciencias, UNAM, 2007.
5. Edwin R. Cuero *Solución de algunos Problemas de Caminata Aleatorias através de las Martingalas*, Facultad de Ciencias, Universidad Tecnológica de Pereira, 2013.
6. Ricardo A. Fornero, *Cronología Fotográfica de las Finansas*, volumen 4 de 1970 a 1980.
7. Luis E. Riascos *Aplicaión de las Martingalas al estudio del modelo de precio de acciones*, Facultad de Ciencias Básicas, Universidad Tecnológica de Pereira, 2012
8. Miguel A. Ariño *Valoración de activos por el metodo de las martingalas* , Investigaciones Económicas vol. XVI (1992)