UNAH
UNIVERSIDAD NACIONAL
AUTÓNOMA DE HONDURAS

Maestría en
Matemática

BOLETÍN DIVULGATIVO

ENERO 2026

Volumen I

TEMÁTICA:

- REGRESIÓN CUANTÍLICA
- PRE-TRAINING
- REDES NEURONALES CONVOLUCIONALES
- RANDOM FOREST
- VALORES EXTREMOS
- MODELOS VAR INTEGRADO
- ESTADÍSTICA ROBUSTA
- INFERENCIA CAUSAL
- MODELADO ESPACIAL

Presentación

Este documento fue desarrollado por la Coordinación de Investigación y Vinculación de la Maestría en Matemática de la UNAH, presenta artículos divulgativos y de investigación desarrollados por estudiantes del Seminario de Investigación de Estadística Matemática de la quinta promoción del programa, cursos desarrollado durante el tercer período académico del año 2025.

Se abarca una temática bastante amplia: Regresión cuantílica, Pre-training, Redes neuronales convolucionales, Random forest, Teoria de valores extremos, Modelos VAR, Estadística robusta, Inferencia causal y Modelado Espacial; 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 avances en sus trabajos de tesis que incluso incluyen experimentación.

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 el programa de maestría en Matemáticas y por parte de la Coordinación de Investigación y Vinculación, se está desarrollando en los estudiantes un espíritu investigador.

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, productividad, infraestructura y desarrollo territorial. Esto evidencia que la Coordinación de Investigación y Vinculación de la Maestría en Matemática está sumamente interesada en colaborar con las prioridades investigativas de la universidad y mantiene un compromiso con vincularse con la sociedad.

Enero del año 2026, Ciudad Universitaria

Tegucigalpa, M.D.C., Honduras

Edificio F1, Segundo Piso, Ciudad Universitaria
Tegucigalpa, M.D.C. Honduras.
https://mm.unah.edu.hn/
maestria.matematica@unah.edu.hn
Tel. 2216-3000 Ext. 100647

Contenido

REGRESIÓN CUANTÍLICA: UNA BREVE REVISIÓN BIBLIOGRÁFICA DE SU EVOLUCIÓN Y MÉTODOS

ERLIN VASQUEZ

Resumen. La Regresión Cuantílica es una extensión de la regresión lineal que permite relajar los supuestos de normalidad y homocedasticidad. Fue propuesta por Koenker & Bassett, y el objetivo es estimar los cuantiles condicionales de la variable de respuesta dadas las covariables. Se explica como este método extiende la regresión lineal al permitir el estudio de cualquier cuantil de la distribución condicional. El siguiente trabajo detalla la fundamentación matemática de los cuantiles, la formulación del problema de regresión cuantílica y la inclusión de técnicas de regularización para controlar la complejidad del modelo y seleccionar variables relevantes, como el LASSO. Ademas, expone principales estrategias computacionales como el ADMM y métodos de punto interior para resolver estos problemas en alta dimension, acompañando la revisión con simulaciones que ilustran los beneficios metodológicos frente a la regresión OLS convencional. Se destaca que la regresión cuantílica penalizada constituye un marco robusto y flexible para el análisis estadístico, permitiendo caracterizar con mayor precisión la distribución condicional completa de la variable de respuesta. Su robustez ante valores atípicos (*outliers*), la hace un método útil en modelado de riesgos financieros, y es aplicable a economía, ciencias sociales, medicina donde la heterogeneidad y los efectos no lineales son comunes.

Abstract. Quantile Regression is an extension of linear regression that relaxes the assumptions of normality and homoscedasticity. It was proposed by Koenker & Bassett, and its objective is to estimate the conditional quantiles of the response variable given the covariates. It is explained how this method extends linear regression by allowing the study of any quantile of the conditional distribution. The following work details the mathematical foundations of quantiles, the formulation of the quantile regression problem, and the inclusion of regularization techniques to control model complexity and select relevant variables, such as LASSO. In addition, it presents the main computational strategies such as ADMM and interior-point methods to solve these problems in high dimensions, accompanied by simulations that illustrate the methodological benefits compared to conventional OLS regression. It is emphasized that penalized quantile regression constitutes a robust and flexible framework for statistical analysis, allowing for a more precise characterization of the full conditional distribution of the response variable. Its robustness to outliers makes it a useful method in financial risk modeling, and it is applicable to economics, social sciences, and medicine, where heterogeneity and nonlinear effects are common.

Date: Diciembre 2025.
Key words and phrases. Regresión cuantílica, heterogeneidad, outliers, LASSO.

1. Introducción

El estudio de la regresión cuantílica introducida formalmente por Koenker y Bassett ([11]), representa un avance importante en la estadística aplicada y la econometría, al permitir la estimación de las relaciones condicionales subyacentes no solo en la media de la variable de respuesta, sino en cualquier cuantil de la distribución. La regresión cuantil puede verse como una extensión de regresión lineal ordinaria (OLS), que asume efectos homogéneos de las covariables y se centra en la media condicional de la variable de respuesta dadas las covariables, la regresión cuantílica a diferencia de OLS captura variaciones en la intensidad de estos efectos a lo largo de la distribución, lo cual tiene relevancia en escenarios donde los datos presentan heterogeneidad, como por ejemplo en estudios de distribución de ingresos, impactos de políticas educativas o análisis de riesgos en salud ([12]).

Este método ofrece un aporte valioso en contextos o situaciones donde las distribuciones son asimétricas o presentan colas pesadas, situaciones que suelen presentarse comúnmente en datos reales del área social o económica. En este contexto, por ejemplo, en el análisis de salarios, la regresión cuantílica puede dar indicios de como el efecto de la educación varia entre trabajadores de bajos ingresos (cuantiles inferiores) y altos ingresos (cuantiles superiores) ofreciendo perspectivas mas reales o matizadas que la OLS, que podría subestimar o sobrestimar efectos en los extremos ([4]).

El objetivo principal de este trabajo es brindar una revisión comprehensiva de la regresión cuantílica, desde su heurística en sus bases teóricas. Secundariamente, se busca resaltar su importancia metodológica para investigadores en estadística, promoviendo su uso y robustez. La relevancia de esta temática radica en la capacidad para detectar e informar políticas publicas mas equitativas, al poder identificar efectos importantes que afectan desproporcionadamente a aquellos sectores mas vulnerables. En las siguientes secciones, se proporciona una justificación sobre el uso y aplicación de este método en el contexto de Honduras, se revisan los antecedentes históricos mas relevantes y destacados, se detalla el marco teórico, por ultimo se concluye con algunas implicaciones futuras en este escenario.

2. Justificación

El estudio y exploración de la regresión cuantílica contribuye directamente a abordar desafíos estructurales en Honduras, como ser la desigualdad en la distribución de ingresos y la pobreza multifacetica, que impactan gran parte de la población según el Instituto Nacional de Estadística (INE) en 2024 ([10]). La regresión cuantílica permite separar los efectos de variables como remesas en diferentes sectores socioeconómicos, el nivel educativo, el acceso a servicios básicos que pueden beneficiar en el diseño de intervenciones focalizadas que fomenten la inclusión social y el crecimiento sostenible.

De acuerdo a las prioridades de investigación establecidas por la Universidad Nacional Autónoma de Honduras (UNAH), este trabajo se alinea con el eje de *Desarrollo Económico y Social*, de manera mas especifica en el tema prioritario de *Pobreza e inequidad* ([29]), donde se le da prioridad al análisis de desigualdad sobre los grupos sociales mas vulnerables. Pero ademas, en el marco de la Maestría en Matemática con Orientación en Estadística Matemática de la UNAH, este estudio se introduce en la línea de investigación *Econometría y Actuaria* con el enfoque de

manejo, procesamiento y presentación de la información, pero también la predicción de tendencias de un proceso, promoviendo así herramientas avanzadas para el análisis predictivo y la toma de decisiones.

Según el estudio realizado por Díaz ([5]) donde se usa como variable de respuesta a la pobreza laboral definida como la cantidad de personas que no tienen acceso a la "canasta básica" y algunas variables explicativas consideradas en estudio son la inflación, crecimiento económico (PIB), etc; donde para estimar el impacto de estas variables sobre la distribución de la pobreza se emplea modelos de regresión cuantílica; esto hace resaltar la importancia de desentrañar este método tanto en forma teórica como en contextos aplicados.

3. Antecedentes

El desarrollo formal de la regresión cuantílica fue liderado por Roger Koenker y Gilbert Bassett, quienes en su artículo de 1978 definieron el método como la solución a un problema de minimización de la pérdida asimétrica, extendiendo así el principio de mínimos cuadrados ordinarios a cuantiles arbitrarios ([11]). Koenker, ha sido el principal impulsor, publicando el libro de referencia *Quantile Regression* en 2005, donde se expone algoritmos computacionales, la teoría inferencial y aplicaciones ([12]).

Aunque la regresión cuantil se formalizó en 1978, tiene sus raíces en el siglo XIX, con contribuciones iniciales de matemáticos como Pierre-Simon Laplace, quien en 1818 propuso estimadores basados en cuantiles para resumir distribuciones ([13]). Edgeworth extendió estas ideas en la década de 1880, introduciendo conceptos de profundidad estadística que anticipan la robustez moderna ([6]). Mas tarde, en el siglo XX, Ragnar Frish, pionero de la econometría exploro formas robustas de regresión en los años de 1920, aunque no llego a formalizar los cuantiles condicionales.

Después del tratamiento formal de la regresión cuantílica por Koenker y Bassett en las décadas siguientes se produjeron avances significativos. En los años 1990, se extendió a datos censurados y de supervivencia, con trabajos como el de Ying ([34]) sobre regresión mediana censurada. En los 2000, Yu & Moyeed ([35]) introdujeron enfoques bayesianos utilizando la distribución Laplace asimétrica (ALD), facilitando inferencia en modelos complejos ([35]). Extensiones a datos de panel incluyeron efectos fijos por Lamarche ([17]), y modelos factoriales para alta dimensionalidad por Koenker ([13]).

Se han propuesto extensiones ingeniosas ah este método, Meinshausen propone una extensión de *Ramdon Forest* para estimar cuantiles condicionales ([20]), Takeuchi propone un método no paramétrico para estimar cuantiles condicionales, usando técnicas de *kernel methods/ maquinas de soporte (SVM/RKHS)* que son mas flexibles que modelos lineales ([25]), Yichao & Yufeng desarrollan métodos de selección de variables dentro del marco de la regresión cuantil, focalizando en penalizaciones tipo LASSO adaptativo y SCAND aplicadas a la regresión cuantil ([32]).

Aportes recientes en regresión cuantílica como ser, métodos para estimar cuantiles extremos condicionales combinando *teoría de valores extremos* y *gradient boosting* ([30]), Steven & Padilla proponen un método en un marco no paramétrico que mezcla la función de pérdida de cuantiles con una penalización de LASSO aplicado sobre un grafo de vecinos cercanos (K-nearest neighbors, KNN) ([33]). Li & Megiddo proporcionan un método que permite estimar simultáneamente los coeficientes de regresión para varios cuantiles, suavizandolos como funciones suavizadas de los

cuantiles mediante spline ([18]). Cuando los datos tienen muchas variables (mas que observaciones) y queremos hacer regresión cuantil, los métodos clásicos (QR penalizado con L_1, "Quantile LASSO") subren dificultades, Tan & Wang & Zhou proponen una combinación de *Convolution smoothing* para suavizar la función de pérdida de cuantiles y una regularización cóncava plegada que reduce el sesgo de una penalización L_1 ([26]).

La evolución de la regresión cuantílica ha pasado de un enfoque complementario a la media a una herramienta estándar en econometría, con más de 20,000 citas al trabajo fundacional de Koenker y Bassett. Recientes revisiones, como la de ([31]), enfatizan sus aplicaciones en experimentos estocásticos y modelos paramétricos. En ciencias del desarrollo, ([22]) demostró su utilidad para analizar efectos diferenciales en logros educativos, revelando variaciones no capturadas por la OLS.

Estos aportes han expandido la regresión cuantílica a datos correlacionados, censurados y de alta dimensión, consolidándola como método robusto para heterogeneidad.

4. Marco Teórico

En el estudio de la Regresión Cuantílica, comprender los cuantiles, sus propiedades estadísticas y la estructura de problemas de optimización que permiten su calculo es esencial para analizar la distribución condicional de la variable de respuesta mas allá de la media. La regresión cuantílica surge como una extensión del enfoque por mínimos cuadrados la reeemplazar el error cuadrático por una función de perdida asimétrica que permite capturar relaciones heterogéneas a lo largo de cada cuantil de la distribución condicional.

La regresión cuantil caracteriza el comportamiento de la variable de respuesta, lo cual resulta indispensable en situaciones donde el efecto de las covariables presentan colas pesadas, asimetría, heterogeneidad o valores atípicos. Como se ha mencionado antes, el conocimiento de los cuantiles resulta fundamental para comprender este enfoque. A continuación, se presenta una introducción detallada sobre este concepto.

4.1. Cuantiles. Sea X una variable caracterizada por su función de distribución acumulada $F_X(x)$, continua por la derecha y definida como:

$$F_X(x) = P(X \leq x).$$

Para $\tau \in (0, 1)$ el τ-ésimo cuantil de X es,

$$Q_X(\tau) = F_X^{-1}(\tau) = \text{ínf}\{x \in \mathbb{R} : F_X(x) \geq \tau\}.$$

La función $Q_X(\tau)$ es continua por la izquierda y, en este marco, la mediana $F_X^{-1}(\frac{1}{2})$, juega un rol central. Cuando X es una variable aleatoria continua, el cuantil es único y la igualdad se satisface estrictamente. Ademas, los cuantiles se pueden ver como solución al problema de optimización,

$$Q_X(\tau) \in \arg\min_{q \in \mathbb{R}} \mathbb{E}[\rho_\tau(X - q)]$$

donde $\rho_\tau(\mu) = \mu(\tau - I(\mu < 0))$ es la función de pérdida cuantílica.

Dado que F_X es monótona no decreciente, cualquier elemento de $\{x : F_X(x) = \tau\}$ minimiza la pérdida esperada. Cuando la solución es única, $\hat{x} = F_X^{-1}(\tau)$; de lo contrario, tenemos un intervalo de cuantiles τ-ésimos del cual podemos elegir el

elemento más pequeño, para adherirnos a la convención de que la función cuantil empírica sea continua por la izquierda ([12]).

Los cuantiles suelen agruparse para dividir la distribución en partes iguales, tales como:

1. Cuartiles, que segmentan a la distribución en cuatro partes correspondientes a los cuantiles 0.25, 0.5 y 0.75.
2. Deciles, que la dividen en diez partes, asociados a los cuantiles 0.1, 0.2, $\cdots$, 0.8, 0.9.
3. Percentiles, que la particionan en cien partes.

Dada una muestra aleatoria $X_1, X_2, \cdots, X_n$ de F_X es posible ordenarla de forma ascendente y expresarla como $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ donde $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, y $X_{(i)}$ es la i-ésima estadística de orden. Al estimar F_X mediante la función de distribución empírica F_n, se tiene

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x),$$

o equivalente,

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < X_{(1)} \\ \frac{i}{n} & \text{si} \quad X_{(i)} \leq x < X_{(i+1)}, \ \ i = 1, \cdots, n-1 \\ 1 & \text{si} \quad x \geq X_{(n)} \end{cases}$$

da lugar a los cuantiles muestrales ([7]).

Para $\tau \in (0,1)$ se define el τ-cuantil muestral de X como el cuantil τ de la función de distribución empírica F_n y se denota por $Q_n(\tau)$ y esta dada por:

$$Q_n(\tau) = \begin{cases} X_{(1)} & \text{si} \quad 0 < \tau \leq \frac{1}{n} \\ X_{(2)} & \text{si} \quad \frac{1}{n} < \tau \leq \frac{2}{n} \\ \vdots & \\ X_{(n)} & \text{si} \quad \frac{n-1}{n} < \tau \leq 1 \end{cases}$$

Estas expresiones ilustran la relación subyacente que se presenta entre los cuantiles muestrales y las estadísticas de orden ([7]).

4.2. Regresión cuantílica. Consideremos un conjunto de covariables o matriz de diseño $X \in \mathbb{R}^{n \times p}$ y una variable de respuesta $Y \in \mathbb{R}^{n \times 1}$. El modelo lineal multivariado esta dado por:

$$Y = X\beta + \varepsilon \tag{4.1}$$

donde $\beta \in \mathbb{R}^{p \times 1}$ es el vector de parámetros, $\varepsilon \in \mathbb{R}^{n \times 1}$ es una perturbación aleatoria que recoge todos aquellos factores distintos de las variables X_i influyendo en Y_i. En regresión lineal multivariada se busca estimar la media de la variable de respuesta Y condicionada a que $X = x$ es decir,

$$E(Y|X = x) = x^\top \beta.$$

El procedimiento mas utilizado para estimar β es el de mínimos cuadrados ordinarios (OLS) que involucra la minimización de la suma de las desviaciones al cuadrado, es decir,

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - X_i\beta)^2$$

para estimar β basta con derivar e igualar después a 0, obteniéndose de forma cerrada el estimador para β:

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} XY$$

pero el método OLS requiere hipótesis previas sobre la aleatoriedad de la relación (4.1) expresadas en términos $\varepsilon_i \sim N(0, \sigma^2)$.

Los objetivos que se persiguen en regresión cuantílica son los mismos que en OLS, es decir, describir las relaciones entre las variables. De forma análoga al modelo de mínimos cuadrados ordinarios, en el que $E(Y|X = x) = x^\top\beta$, y por lo tanto $E(\varepsilon|X = x) = 0$, aquí $Q_Y(\tau|X = x) = x^\top\beta_\tau$ lo que implica que $Q_Y(\varepsilon|X = x) = 0$, siendo el único supuesto que se hace sobre los errores aleatorios. La regresión cuantílica busca estimar el τ-ésimo cuantil esperado de la variable de respuesta condicionado a las observaciones. Es decir, dada la muestra de tamaño n, $\{X_i, Y_i\}$, $i = 1, \ldots, n$, del modelo lineal de cuantiles

$$Q_Y(\tau|X = x) = x^\top\beta_\tau,$$

el estimador del coeficiente del τ-ésimo cuantil de regresión de Koenker y Bassett (1978) ([11]) es

$$\hat{\beta}_\tau = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^\top\beta) \tag{4.2}$$

donde $\rho_\tau(\mu) = \mu(\tau - I(\mu < 0))$ es la función de pérdida cuantílica. El parámetro β_τ describe el cambio en el cuantil condicional de Y ante variaciones en X. La variación de β_τ en función de τ permite detecta heterogeneidad en la relación respuesta–covariables. Este enfoque (4.2) permite modelar cuantiles condicionales distintos de la media, ofreciendo robustez ante *outliers*, heterocedasticidad y distribuciones no normales.

El problema planteado anterior, presenta el inconveniente de que $\rho_\tau(\mu)$ la función de perdida cuantílica no es diferenciable, lo que hace necesario convertir el problema (4.2) a un problema de programación lineal bajo algunas transformaciones, introduciendo $2n$ variables artificiales, o «de holgura», $\{u_i, v_i : i = 1, \ldots, n\}$ para representar las partes positiva y negativa del vector de residuos,

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}} \left\{\tau\mathbf{1}_n^\top u + (1-\tau)\mathbf{1}_n^\top v \mid \mathbf{1}_n X\beta_\tau + u - v = y\right\}, \tag{4.3}$$

donde $\mathbf{1}_n$ denota un vector de n unos. Claramente, en (4.3) estamos minimizando una función lineal en un conjunto de restricciones poliédrico, que consiste en la intersección del hiperplano de dimensión $2n + 1$ determinado por las restricciones de igualdad lineal y el conjunto $\mathbb{R}^p \times \mathbb{R}_+^{2n}$. Dicho problema puede ser resuelto mediante diversos algoritmos que trataremos mas adelante. Muchas características de la solución son inmediatamente evidentes a partir de este simple hecho. Por ejemplo, mín$\{u_i, v_i\}$ debe ser cero para todo i, ya que, de lo contrario, la función objetivo puede reducirse sin violar la restricción al disminuir dicho par hacia cero.

Esto se conoce comúnmente como complementariedad en la terminología de la programación lineal. De hecho, por esta misma razón, podemos restringir la atención a «soluciones básicas» de la forma $\xi = Y_i$ para alguna observación i. Observe que la función objetivo es convexa y lineal por tramos, con puntos de inflexión en los valores observados Y_i ([12]).

4.3. Regularización en Regresión cuantílica. En la practica contemporánea, la regresión cuantílica a menudo incorpora penalizaciones para controlar la complejidad del modelo y mitigar el sobreajuste. El problema general penalizado se formula como:

$$\hat{\beta}_\tau = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^\top \beta) + \lambda P(\beta) \tag{4.4}$$

donde $\rho_\tau(.)$ representa la perdida cuantílica en el cuantil τ, y $P(\beta)$ denota el termino de penalización, comúnmente basado en normas L_1 similar a Lasso o L_2 similar a Ridge ([19]).

Históricamente, la regularización surge de la necesidad de equilibrar fidelidad al dato y suavidad al ajuste. Hoy en día, se busca minimizar:

$$\min_f \{L(f) + \lambda P(f)\} = L(f) + \lambda P(f) \to \min_f ! \tag{4.5}$$

donde $L(f)$ cuantifica la infidelidad o falta de ajuste, o incluso mejor la perdida al ajuste f y $P(f)$ impone una penalización afectada por una parámetro de regularización $\lambda > 0$. Una línea de desarrollo independiente, ajena inicialmente de probabilidad, y alineada con la "combinación de observaciones" según Stigler ([19]), remite a Hadamard ([8]),quien señalo de que no todos los problemas están realmente bien planteados. Tikhonov ([27, 28]) propuso la regularización como familia de técnicas para estabilizar soluciones, cuya forma mas exitosa coincida con la expresión anterior 4.5.

Tibshirani ([24]) inicia con la penalización restringida, motivado por el *nonnegative garrote* de Breiman ([2]). Ambas perspectivas – *Tikhonov y Phillips* – están estrechamente vinculadas vía multiplicadores de Lagrange, bajo convexidad de L y P ([19]).

4.3.1. Ajuste del parámetro. Whittaker y Robinson ([36]) indicaban que el grado de sacrificio de fidelidad por suavidad varia según el problema, recomendando probar valores de λ y seleccionar el mas satisfactorio. Actualmente, predominan métodos automáticos para la selección de λ.

Un enfoque común es la validación cruzada, como sugiere Hastie ([9]), preferiblemente con poco grupos (k-fold). Valores típicos como $k = 2$ o $k = 10$ suelen proporcionar resultados fiables, aunque la selección aleatoria puede inducir volatilidad ([19]).

La validación cruzada *leave-one-out* (n pliegues) es computacionalmente costosa para penalizaciones no cuadrática debido a su no linealidad. Para L_1, se prefiere la noción de grados de libertad, equivalentes al numero de ajuste exactos en cero como verificaron en Koenker ([16]) y posteriores estudios ([19]).

4.4. Métodos computacionales. La regresión cuantílica clásica se formula como un problema de optimización lineal que se puede resolver mediante método simplex (punto exterior) o métodos de barrera/interior; la elección dependen de n, p y la estructura de dispersión. Para n y p moderados, el método de punto exterior puede ser competitivo; para gran escala, el método de punto interior es preferible por su complejidad amortizada. El método ADMM es fundamental para trabajar con la regresión cuantílica regularizada. Esta sección ofrece un recorrido por técnicas computacionales clave en regresión cuantílica.

4.4.1. Metodos de punto exterior. El algoritmo de Barrodale y Roberts ([1]) exploto la dualidad con variables acotadas en la regresión mediana. El problema primal de regresión mediana es:

$$\text{mín}\{1_n^\top u + 1_n^\top v \mid y - Xb = u - v;\ (u, v) \geq 0\}$$

de dimensión $(2n + p)$. El dual tiene resulta más simple:

$$\max_a \{y^\top a \mid X^\top a = \tfrac{1}{2} X^\top 1_n;\ a \in [0, 1]^n\}.$$

Implementa una estrategia dual de tipo Edgeworth: dada una solución básica

$$b^{(h)} = (X^{(h)})^{-1} y^{(h)},$$

se identifica la dirección de descenso mas pronunciada.

La extensión a cuantiles $\tau \neq 0.5$ es directa: nn el problema primal, solo reemplazamos los 1_n por pesos asimétricos apropiados; en el dual, simplemente cambiamos el $\frac{1}{2}$ por $1 - \tau$. Variaciones en τ generan trayectorias de soluciones; Portnoy ([23]) demostró que el numero esperado de soluciones distintas es $O(n \log n)$. Técnicas paramétricas similares aplican a problemas penalizados tipo Lasso. Aunque los métodos simplex facilitan el trazado de trayectorias, el numero de soluciones puede volverse prohibitivo, requiriendo aproximaciones ([15]).

4.4.2. Métodos de punto interior. A diferencia de los métodos de punto exterior, que transitan vértices del conjunto factible, los de punto interior parten del centro hacia un vértice. El método de barrera logarítmica de Frisch para programación lineal canónica:

$$\text{mín}\{c^\top x \mid Ax = b;\ x \geq 0\}$$

reemplaza desigualdades por:

$$\text{mín}\left\{ c^\top x - \mu \sum_{j=1}^{p} \log x_j \mid Ax = b \right\}.$$

Relajando $\mu \to 0$, se converge a un vértice. Explotando primal y dual:

$$\max_y \{b^\top y \mid A^\top y + z = c;\ z \geq 0\},$$

la optimalidad implica que $c - \mu X^{-1} e = A^\top y$, por lo que podemos establecer $z = \mu X^{-1} e$ para satisfacer la restricción dual y obtener el sistema

$$
\begin{aligned}
Ax &= b,\\
A^\top y + z &= c,\\
Xz &= \mu e,\\
x &\geq 0,\\
z &\geq 0.
\end{aligned}
$$

La trayectoria paramétrica $(x(\mu), y(\mu), z(\mu))$ describe la *trayectoria central* desde el centro del conjunto de restricciones hasta una solución en el borde del conjunto de restricciones que satisface la condición clásica de holgura complementaria, $Xz = 0$, cuando $\mu = 0$. Cuando la dimensión paramétrica del modelo es grande, las implementaciones de punto interior puede ser bastante lentas, pero en la mayoría de las aplicaciones no paramétricas, como las que abarca el modelo aditivo penalizado por variación total descrito en Koenker ([14]) e implementado en **rqss** de **quantreg**, la matriz de diseño es extremadamente dispersa. En estos caso, la factorización de Cholesky viabiliza problemas con miles de parámetros ([15]).

4.4.3. Método de dirección alterna de multiplicadores (ADMM). Es común en aplicaciones estadísticas encontrarse con problemas de optimización con componentes convexos aditivamente separables. El algoritmo resuelve problemas de la forma

$$
\begin{aligned}
&\text{minimizar } f(x) + g(z)\\
&\text{sujeto a } Ax + Bz = c
\end{aligned}
$$

con f, g convexas; variables $x \in \mathbb{R}^n$ y $z \in \mathbb{R}^m$, donde $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$, y $c \in \mathbb{R}^p$. Un ejemplo familiar sería f como la log-verosimilitud (negativa) y g una penalización paramétrica tipo lasso.

El Lagrangiano aumentado es:

$$
L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2.
$$

El ADMM consta de las iteraciones

$$
x^{k+1} := \arg\min_x L_\rho(x, z^k, y^k) \tag{4.6}
$$

$$
z^{k+1} := \arg\min_z L_\rho(x^{k+1}, z, y^k) \tag{4.7}
$$

$$
y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c), \tag{4.8}
$$

donde $\rho > 0$. Este algoritmo del *método de dirección alterna de multiplicadores (ADMM)* tiene amplia aplicabilidad y se ha demostrado que converge bajo condiciones suaves ([15], [3]).

4.4.4. Forma Escalada. Combinando terminos lineales y cuadraticos en el lagrangiano aumentado, y escalando la variable dual $u = \frac{1}{\rho}y$, el ADMM puede escribirse en una forma ligeramente diferente, que suele ser más conveniente. Definiendo el residuo $r = Ax + Bz - c$, tenemos

$$y^T r + \frac{\rho}{2}\|r\|_2^2 = \frac{\rho}{2}\left\|r + \frac{1}{\rho}y\right\|_2^2 - \frac{1}{2\rho}\|y\|_2^2 = \frac{\rho}{2}\|r + u\|_2^2 - \frac{\rho}{2}\|u\|_2^2,$$

Usando la variable dual escalada, podemos expresar el ADMM como

$$x^{k+1} := \arg\min_x \left\{ f(x) + \frac{\rho}{2}\left\|Ax + Bz^k - c + u^k\right\|_2^2 \right\} \tag{4.9}$$

$$z^{k+1} := \arg\min_z \left\{ g(z) + \frac{\rho}{2}\left\|Ax^{k+1} + Bz - c + u^k\right\|_2^2 \right\} \tag{4.10}$$

$$u^{k+1} := u^k + Ax^{k+1} + Bz^{k+1} - c. \tag{4.11}$$

Definiendo el residuo en la iteración k como $r^k = Ax^k + Bz^k - c$, vemos que

$$u^k = u^0 + \sum_{j=1}^{k} r^j,$$

la suma acumulada de los residuos ([3]).

4.4.5. Parámetro de Penalización Variable. Para acelerar convergencia, se emplea ρ_k variable (posiblemente diferentes para cada iteración), con el objetivo de mejorar la convergencia en la práctica y hacer que el rendimiento dependa menos de la elección inicial del parámetro de penalización. Aunque puede ser difícil probar la convergencia del ADMM cuando ρ varía en cada iteración, la teoría para ρ fijo aún se aplica si se asume que ρ se fija después de un número finito de iteraciones.

Un esquema simple que a menudo funciona bien es

$$\rho^{k+1} := \begin{cases} \tau^{\text{incr}}\rho^k & \text{si } \|r^k\|_2 > \mu\|s^k\|_2, \\ \rho^k/\tau^{\text{decr}} & \text{si } \|s^k\|_2 > \mu\|r^k\|_2, \\ \rho^k & \text{en caso contrario,} \end{cases} \tag{4.12}$$

donde $\mu > 1$, $\tau^{\text{incr}} > 1$ y $\tau^{\text{decr}} > 1$ son parámetros. Elecciones típicas podrían ser $\mu = 10$ y $\tau^{\text{incr}} = \tau^{\text{decr}} = 2$. La idea detrás de esta actualización del parámetro de penalización es intentar mantener las normas de los residuos primal y dual dentro de un factor μ entre sí a medida que ambos convergen a cero ([3]).

Las ecuaciones de actualización del ADMM sugieren que valores grandes de ρ imponen una gran penalización a las violaciones de la factibilidad primal y, por lo tanto, tienden a producir residuos primales pequeños. Por el contrario, la definición de s^{k+1} sugiere que valores pequeños de ρ tienden a reducir el residuo dual, pero a costa de disminuir la penalización sobre la factibilidad primal, lo que puede resultar en un residuo primal más grande. El esquema de ajuste (3.13) aumenta ρ por τ^{incr} cuando el residuo primal parece grande en comparación con el residuo dual, y reduce ρ por τ^{decr} cuando el residuo primal parece demasiado pequeño en relación con el residuo dual. Este esquema también puede refinarse considerando las magnitudes relativas de ϵ^{pri} y ϵ^{dual} ([3]).

Cuando se usa un parámetro de penalización variable en la forma escalada del ADMM, la variable dual escalada $u^k = (1/\rho)y^k$ también debe reescalarse después de actualizar ρ; por ejemplo, si ρ se reduce a la mitad, u^k debe duplicarse antes de continuar ([3]).

Este método es particularmente relevante en regresión cuantílica penalizada cuando las funciones son convexas.

4.5. Simulaciones. Se muestran a continuación simulaciones considerando tres escenarios:

1. Primer escenario (1): Se consideran 500 observaciones tomadas de una variable predictora distribuida uniformemente en el intervalo de (0,100). La variable de respuesta se genera mediante la expresión:

$$Y = 2 + 0.5X + \varepsilon$$

donde ε se selecciono de una de una distribución normal con media 0 pero varianza dada por $1 + 0.3X$. En este escenario se esta considerando la heterocedasticidad, se aplica la regresión cuantílica en los cuantiles Q_1, Q_2, Q_3 y se hace una comparativa visual con respecto a la regresión lineal que predice la media condicional, que pasa por el centro del diagrama de dispersión por lo cual no captura que la variabilidad de los datos cambia con X. Ademas nótese que la regresión cuantílica es mas informativa que la regresión lineal cuando los datos presentan heterocedasticidad y la distribución de los errores no es simétrica.

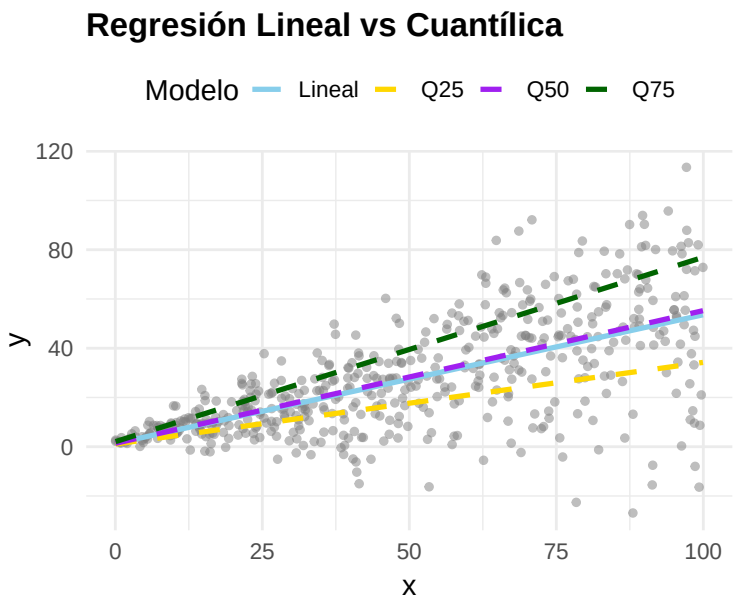


FIGURA 1. Comportamiento de la regresión cuantílica $\tau = 0.25, 0.5, 0.75$ vs Lineal.

2. Segundo escenario (2): Se eligió una variable predictora X distribuida normal estándar de la cual se selecciono una muestra de 100 observaciones y se formulo el siguiente modelo:

$$Y = 2 + 3X + \varepsilon$$

donde ε se tomo de una distribución t-Student con dos grados de libertad para ilustrar el caso de colas pesadas, también se consideran *outliers*. En (2)

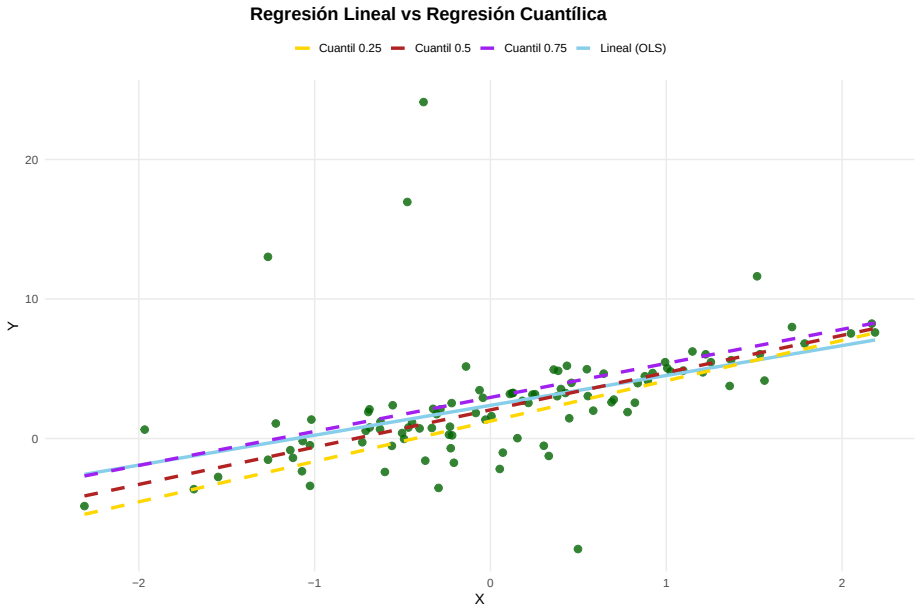


FIGURA 2. Comportamiento de la regresión cuantílica $\tau = 0.25, 0.5, 0.75$ vs Lineal considerando colas pesadas y valores extremos.

se puede observar como la media condicional es afectada por los valores extremos, y como los cuantiles como ser la mediana no se ven afectados por esos valores. Esto ilustra el potencial de la regresión cuantílica sobre escenarios donde solo se modela la media condicional de la variable de respuesta.

3. Tercer escenario: En el contexto multivariado se generaron cuatro variables predictoras de una distribución normal estándar considerando nuevamente errores distribuidos t-Student con tres grados de libertad. Por cada variable se generaron 500 observaciones, y la variable de respuesta se diseño bajo el siguiente modelo

$$Y = 2 + 3X_1 - 1.5X_2 + 2X_3 + 0.5X_4 + \varepsilon$$

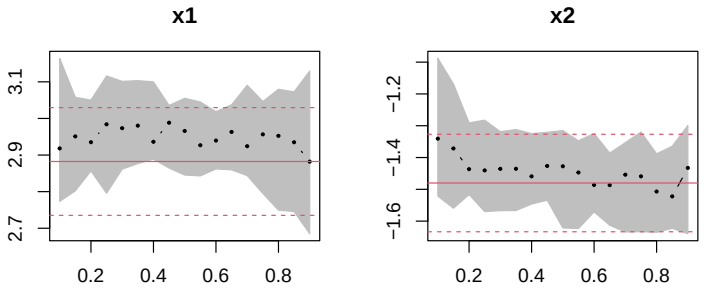
En la tabla (1) se presentan los coeficientes obtenidos mediante regresión lineal multivariada, regresión cuantílica y regresión cuantílica LASSO penalizada en el cual el parámetro de regularización se obtuvo mediante validación cruzada usando por defecto k-folds igual a 10.

Nótese, que la variable X_4 tiene coeficiente cero, esto indica que la regresión cuantílica penalizada por LASSO permite hacer selección de variables.

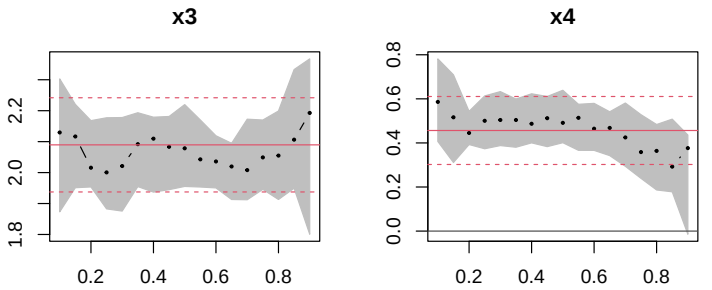
Predictor	OLS	RQ	RQ LASSO
(Intercepto)	2.13946	1.995	2.0338
x1	2.88242	2.96584	2.5791
x2	−1.48015	−1.42750	−1.0682
x3	2.08958	2.07843	1.6679
x4	0.45645	0.49118	0.0000

RQ LASSO usa $\lambda = 0.1233$ seleccionado por validación cruzada. Coeficientes exactamente cero indican que la variable fue eliminada por la penalización LASSO.

CUADRO 1. Comparación de coeficientes estimados para la mediana ($\tau = 0.5$)



(A) Cambio X_1 (B) Cambio en X_2



(C) Cambio en X_3 (D) Cambio en X_4

FIGURA 3. Cambio de los coeficientes sin penalizar según los cuantiles

Ademas, en la figura (3) se muestra como van cambiando los coeficientes de las variables a lo largo de los cuantiles. Los cambios que se presentan los coeficientes representan el efecto de las colas pesadas y los *outliers*.

5. CONCLUSIONES

La regresión cuantílica constituye un marco flexible y robusto para analizar la distribución condicional completa de una variable de respuesta. Su formulación

convexa, sus extensiones no paramétricas y los métodos computacionales modernos como ADMM y punto interior permiten abordar aplicaciones de alta dimensión y alto volumen de datos. La integración de regularización mediante penalizaciones L_1 y L_2 equilibra ajuste y complejidad en contextos de alta dimensión, articulando la equivalencia entre formulaciones con restricción de perdida y con multiplicadores de lagrange, y habilitando selección de hiperparámetros con validación cruzada y nociones de grados de libertad. En conjunto, la teoría de cuantiles, la regularización convexa y los avances algorítmicos conforman un ecosistema metodológico maduro y versátil que se extiende hacia análisis mas ricos accionables a lo largo de la distribución condicional, desde el centro hasta las colas.

Referencias

1. Barrodale, I., & Roberts, F. (1974). Solution of an overdetermined system of equations in the L_1 norm. *Communications of the ACM*, 17, 319–320.
2. Breiman, L. (1995). Better subset selection using the nonnegative garrote. *Technometrics*, 37, 373–384.
3. Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Foundations and Trends in Machine Learning*, 3(1), 1–122.
4. Buchinsky, M. (1994). *Changes in the US Wage Structure 1963–1987: Application of Quantile Regression. Econometrica*, 62(2), 405–458.
5. Díaz Carreño, M. Á. (2023). *Pobreza laboral e inflación en México 2006–2022. Apertura Económica*, 38(97). Recibido: 23 de junio de 2022; aceptado: 14 de septiembre de 2022; publicado: 20 de enero de 2023.
6. Edgeworth, F. Y. (1888). *On a New Method of Reducing Observations Relating to Several Quantities. Philosophical Magazine*, 26, 360–376.
7. Gibbons, J.D., & Chakraborti, S. (2010). *Nonparametric Statistical Inference (5th ed.). Chapman and Hall/CRC.*
8. Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 49–52.
9. Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
10. Instituto Nacional de Estadística (INE). (2024). *Encuesta Permanente de Hogares y Bienes Móviles (EHPM) 2024.* Tegucigalpa, Honduras.
11. Koenker, R., & Bassett, G. (1978). *Regression Quantiles. Econometrica*, 46(1), 33–50.
12. Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.
13. Koenker, R. (2017). *Quantile Regression 40 Years On. Annual Review of Economics*, 9, 155–177.
14. Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25, 239–262.
15. Koenker, R. (s.f.). *Computational methods for quantile regression.* University of Illinois at Urbana-Champaign.
16. Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81, 673–680.
17. Lamarche, C. (2010). *Robust Penalized Quantile Regression Estimator for Panel Data Models. Journal of Econometrics*, 157(2), 246–264.
18. Li, T.-H., & Megiddo, N. (2025). *Spline Quantile Regression.* In *Proceedings of the Second ACM-SIAM Symposium on Discrete Algorithms*, 225–233.
19. Mizera, I. (s.f.). *Quantile regression: Penalized.* University of Alberta, Edmonton, Canada.
20. Meinshausen, N. (2006). *Quantile Regression Forests. Journal of Machine Learning Research*, 7, 983–999.
21. Phillips, D. L. (1962). A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM*, 9, 84–97.
22. Petscher, Y., & Logan, J. A. R. (2014). *Quantile Regression in the Study of Developmental Sciences. Child Development*, 85(3), 861–881.

23. Portnoy, S. (1989). Asymptotic behavior of the number of regression quantile breakpoints. *SIAM Journal on Scientific and Statistical Computing*, 12, 867–883.
24. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267–288.
25. Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. (2006). *Nonparametric Quantile Regression. Journal of Machine Learning Research*, 7, 1231–1264.
26. Tan, K. M., Wang, L., & Zhou, W.-X. (2021). *High-Dimensional Quantile Regression: Convolution Smoothing and Concave Regularization. Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1), 205–233.
27. Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39, 195–198.
28. Tikhonov, A. N. (1963). On the solution of incorrectly posed problems and the regularization method. *Doklady Akademii Nauk SSSR*, 151, 501–504.
29. Universidad Nacional Autónoma de Honduras (UNAH). (2012). *Prioridades de Investigación UNAH*. Dirección de Investigación Científica, Tegucigalpa.
30. Velthoen, J., Dombry, C., Cai, J.-J., & Engelke, S. (2021). *Gradient Boosting for Extreme Quantile Regression. Advances in Neural Information Processing Systems*, 34, 24942–24955.
31. Waldmann, E. (2017). *Quantile Regression: A Short Story on Quantiles and Percentiles. arXiv preprint arXiv:1707.00907*.
32. Wu, Y., & Liu, Y. (2009). *Variable Selection in Quantile Regression. Statistica Sinica*, 19(2), 801–817.
33. Ye, S. S., Padilla, O. H. M., Balakrishnan, S., & Scott, C. (2020). *Non-Parametric Quantile Regression via the K-NN Fused Lasso. Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:10703–10713.
34. Ying, Z., Jung, S. H., & Wei, L. J. (1995). *Survival Analysis with Median Regression Models. Journal of the American Statistical Association*, 90(429), 178–184.
35. Yu, K., & Moyeed, R. A. (2001). *Bayesian Quantile Regression. Statistics & Probability Letters*, 54(4), 437–447.
36. Whittaker, E. T., & Robinson, G. (1924). *The calculus of observations*. Blackie.

Maestría en Matemática, Universidad Nacional Autónoma de Honduras.
Email address: evasquez@unah.edu.hn

APLICACIÓN DE PRE-TRAINING EN SERIES DE TIEMPO CLIMÁTICAS EN HONDURAS

NATHALYE NICOL DERAS DURON

Resumen. Esta investigación explora la integración del pre-training como paradigma del aprendizaje estadístico en series de tiempo climáticas de Honduras. Motivada por los avances teóricos recientes en el aprendizaje estadístico con pre-training. Esta investigación tiene como objetivo evaluar si los modelos preentrenados pueden mejorar la estimación y predicción en variables como la temperatura y la precipitación. Este estudio busca conectar los desarrollos teóricos del pre-training con los desafíos prácticos del análisis de datos climáticos.

Abstract. This research explores the integration of pre-training as a paradigm of statistical learning in climatic time series from Honduras. Motivated by recent theoretical advances in statistical learning with pre-training. This study aims to evaluate whether pre-trained models can improve estimation and prediction for variables such as temperature and precipitation. The study seeks to bridge the theoretical developments of pre-training with the practical challenges of climate data analysis

1. Introducción

El pre-entrenamiento es un paradigma poderoso en el aprendizaje automático para transferir información entre modelos. Por ejemplo, supongamos que se tiene un conjunto de datos de tamaño moderado con imágenes de gatos y perros y se planea ajustar una red neuronal profunda para clasificarlos. Con el pre-entrenamiento, se comienza con una red neuronal entrenada en un corpus grande de imágenes no solo de gatos y perros, sino de cientos de clases. Se fijan todos los pesos de la red excepto las capas superiores y luego se realiza un ajuste fino usando nuestro conjunto de datos. Esto suele producir un rendimiento dramáticamente mejor que entrenar únicamente con nuestro propio conjunto de datos [2].

En el aprendizaje estadístico contemporáneo, el pre-training ha emergido como un enfoque fundamental para mejorar la eficiencia de los modelos predictivos, o ahorrar al equipo desarrollador tiempo y dinero. El pre-training consiste en entrenar un modelo de manera previa en una tarea o conjunto de datos relacionados, de modo que las representaciones o parámetros aprendidos se utilicen como punto de partida para una tarea específica posterior (fine-tuning) [5].

En este sentido, el pre-training no garantiza por sí mismo un mejor desempeño; su éxito radica en la calidad de los datos iniciales, la representatividad de las condiciones climáticas empleadas durante el preentrenamiento y la adecuada calibración del modelo al contexto local. En el caso de Honduras el uso de pre-training

Fecha: Octubre 2025.
Palabras y frases clave. Aprendizaje estadístico, pre-training, aprendizaje supervisado.

se propone como una alternativa metodológica prometedora, siempre que los datos globales utilizados reflejen patrones compatibles con la dinámica climática regional.

Esta estrategia, ampliamente estudiada en contextos de machine learning y transfer learning, ha sido demostrado que consigue recuperar el soporte verdadero [2].

Desde una perspectiva estadística, el pre-training puede entenderse como una forma de incorporar información previa en el proceso de estimación. En modelos lineales penalizados, por ejemplo, el trabajo de Tibshirani y colaboradores [2] formaliza el impacto del pre-training sobre el modelo LASSO, mostrando que la inclusión de representaciones previas reduce la varianza del estimador y mejora la precisión predictiva en alta dimensión. De forma análoga, [3] extienden estos resultados al contexto del aprendizaje estadístico heterogéneo, demostrando que el pre-training puede mejorar la inferencia y la predicción en entornos con variabilidad estructural entre unidades o dominios.

2. Justificación

El contexto climático de Honduras ofrece una oportunidad idónea para aplicar y evaluar los efectos del pre-training. Las series de temperatura y precipitación disponibles presentan patrones temporales y espaciales complejos, así como posibles sesgos derivados de limitaciones en la naturaleza de los datos. En estos escenarios, los métodos tradicionales entrenados desde cero pueden sufrir de sobreajuste o inestabilidad, especialmente cuando los tamaños muestrales son pequeños.

Implementar modelos que integren pre-training permitiría transferir conocimiento aprendido a partir de datos globales o regionales a las condiciones locales hondureñas, mejorando la robustez de las estimaciones y la calidad de las predicciones climáticas. Además, comparar cuantitativamente los modelos con y sin pre-training proporcionará evidencia empírica sobre los beneficios reales de esta estrategia en contextos de análisis estadístico aplicado a la climatología.

Por tanto, esta investigación no solo contribuirá a la comprensión teórica del pre-training en modelos estadísticos, sino también a su aplicación práctica en el análisis y modelado de series de tiempo ambientales, un área de creciente relevancia para la planificación y adaptación climática, así como en la mitigación del riesgo debido a las condiciones climáticas potencialmente inestables en Honduras.

Este estudio se enmarca en las líneas de investigación de la UNAH, particularmente en el eje de Investigación de Ambiente, Biodiversidad y Desarrollo. La aplicación del pre-training en series climáticas integra métodos contemporáneos de aprendizaje estadístico con necesidades nacionales en monitoreo ambiental, aportando herramientas matemáticas relevantes para comprender y anticipar variaciones climáticas locales.

3. Antecedentes

El estudio del pre-training aplicado a datos climáticos ha evolucionado de manera acelerada en los últimos años, impulsado por la necesidad de mejorar la estabilidad y precisión de los modelos predictivos en contextos complejos. Diversos trabajos

recientes han explorado el uso de representaciones preentrenadas, tanto en modelos lineales penalizados como en arquitecturas profundas, destacando su capacidad para reducir el sobreajuste y aprovechar información proveniente de dominios amplios o heterogéneos.

Un primer trabajo es [6] donde los autores realizan un estudio en profundidad sobre métodos de preentrenamiento para evaluar sus impactos en la predicción meteorológica global, prestando especial atención al control del sobreajuste y al análisis de la relación entre la dificultad de la tarea y el rendimiento. Además, se presenta un modelo llamado *Baguan*, que está basado en transformadores, y que utiliza un paradigma de preentrenamiento y fine-tuning con un autoencoder enmascarado.

En este estudio, se utilizó el conjunto de datos ERA5 como los valores reales para el entrenamiento de modelos e inferencia. Este conjunto de datos incluye una amplia gama de variables, como temperatura, humedad, precipitación y presión media al nivel del mar, entre otras.

Se emplean dos métricas cuantitativas específicas para evaluar la precisión de la predicción: el Error Cuadrático Medio Ponderado por Latitud (RMSE, por sus siglas en inglés) y el Coeficiente de Correlación de Anomalías Ponderado por Latitud (ACC, por sus siglas en inglés). Para la optimización, se utilizó la función de pérdida de Error Cuadrático Medio (MSE) durante el preentrenamiento y la función de pérdida de Error Absoluto Medio (MAE) durante el ajuste fino.

Finalmente, en lo que a resultados respecta, Baguan demuestra un rendimiento superior, superando a IFS y Pangu-Weather en una variedad de experimentos, además de sobresalir en diversas tareas posteriores, incluyendo predicción de subestacional a estacional (S2S) y predicción regional, demostrando su versatilidad y aplicabilidad en diferentes escalas temporales y espaciales en la predicción meteorológica.

Un segundo trabajo, presentado por M. Schuessler, E. Sverdrup y R. Tibshirani [3], propone estrategias de preentrenamiento que aprovechan un fenómeno presente en aplicaciones del mundo real: los factores que son pronósticos del resultado suelen ser también predictivos de la heterogeneidad del efecto del tratamiento. Los objetivos planteados en este estudio son:

- Proponer una estrategia de preentrenamiento que va más allá de tratar la función del resultado promedio como un mero parámetro irrelevante en el marco del R-learner, aprovechando el soporte compartido entre los factores pronósticos y predictivos para la estimación del efecto promedio condicional del tratamiento (CATE).
- Establecer un enfoque con tres metas principales: primero, aumentar la precisión de la estimación del CATE explotando las sinergias entre tareas de predicción aparentemente independientes; segundo, mejorar la recuperación del soporte de los modificadores del efecto o efectos de interacción; y tercero, obtener mayor información sobre la supuesta existencia de un soporte compartido entre factores pronósticos y predictivos al estimar el CATE.
- Demostrar la viabilidad de este enfoque mediante el desarrollo de un conjunto de marcos de estimación que utilizan el R-learner basado en lasso (R-lasso)

y modelos no paramétricos, mostrando además cómo este enfoque puede extenderse a entornos no lineales mediante expansiones en funciones base y bosques aleatorios.

En cuanto a los resultados, se concluyó que el empleo de esta estrategia de preentrenamiento en el R-learner produce tasas de error más bajas, mayor capacidad para detectar heterogeneidad y menores tasas de descubrimientos falsos, lo cual es particularmente relevante en campos como el descubrimiento de biomarcadores. No obstante, se identificaron algunas limitaciones: este enfoque no ofrece beneficios de rendimiento en escenarios con poca o ninguna coincidencia entre factores predictivos y pronósticos. Otra limitación es la dependencia de la función de pérdida R (R-loss) para la elección adecuada de α y otros hiperparámetros; si el error de estimación de los parámetros irrelevantes es elevado, la R-loss se vuelve menos confiable para estos hiperparámetros.

Un tercer estudio es el presentado en [2] donde se desarrolla un marco para el lasso en el que un modelo se ajusta a un conjunto de datos grande y luego se afina utilizando un conjunto de datos más pequeño, este tiene una amplia variedad de aplicaciones, incluyendo modelos estratificados, respuestas multinomiales, modelos de múltiples respuestas, estimación del efecto promedio condicional del tratamiento e incluso gradient boosting, los cuales se evalúan durante el estudio. El algoritmo utilizado para este objetivo, es el algoritmo 1.

Un resultado de particular utilidad para este estudio, es que en respuestas ordenadas en el tiempo y encadenamiento de resultados, se probó de forma empírica que para todos los puntos temporales, el preentrenamiento casi iguala o supera la alternativa de ajustar modelos por separado.

Algoritmo 1 Lasso Pre-Entrenado con grupos de entrada fijos

Entrada: Conjunto de entrenamiento, número de grupos K, parámetro $\alpha \in [0, 1]$
Salida: Modelos ajustados para cada grupo con `cv.glmnet`

1: Ajustar un único modelo lasso "global" al conjunto de entrenamiento, por ejemplo usando `cv.glmnet` en R.
2: A partir de este modelo, elegir el vector de pesos $\hat{\beta}_0$ a lo largo del camino de λ, usando por ejemplo `lambda.min`, el valor que minimiza el error de validación cruzada.
3: Fijar $\alpha \in [0, 1]$. Definir los valores de `offset` y `penalty.factor` como sigue:
   - Definir $\texttt{offset} = (1 - \alpha) \cdot (X_k \hat{\beta}_0 + \hat{\mu}_0)$.
   - Sea S el soporte de $\hat{\beta}_0$. Definir el factor de penalización `pf` como:

$$\texttt{pf}_j = I(j \in S) + \frac{1}{\alpha} \cdot I(j \notin S).$$

4: Para cada clase $k = 1, \ldots, K$, ajustar un modelo individual usando `cv.glmnet` con los parámetros `offset` y `penalty.factor`.
5: Usar estos modelos para realizar predicciones dentro de cada grupo.

Definiciones de términos técnicos. A continuación se presentan definiciones breves de algunos términos utilizados a lo largo del documento, con el fin de mantener claridad y consistencia conceptual:

- **ERA5:** Conjunto de reanálisis climático desarrollado por el *European Centre for Medium-Range Weather Forecasts* (ECMWF), que integra observaciones atmosféricas globales con modelos numéricos, proporcionando series históricas de alta resolución espacial y temporal.
- **Baguan:** Modelo de predicción meteorológica basado en la arquitectura *transformer*, preentrenado mediante un autoencoder enmascarado y posteriormente ajustado (*fine-tuning*) para tareas climáticas específicas. Se ha destacado por su desempeño superior en predicción de variables atmosféricas.
- **Pre-training:** Etapa en la que un modelo se entrena inicialmente sobre un conjunto amplio o distinto de datos, con el propósito de aprender representaciones generales que luego serán refinadas en la tarea específica de interés.
- **Fine-tuning:** Fase de ajuste final del modelo preentrenado, en la cual los parámetros aprendidos previamente se adaptan a las características particulares del conjunto de datos objetivo.

4. Marco Teórico

4.1. Pre-Training y Aprendizaje por Transferencia. El pre-entrenamiento y el aprendizaje por transferencia son técnicas fundamentales en el aprendizaje automático, y representan estrategias para aprovechar el conocimiento de tareas relacionadas con el fin de mejorar el rendimiento y la eficiencia en una nueva tarea objetivo. Ambos pueden describirse mediante formulaciones matemáticas claras, comúnmente referenciadas en revisiones de literatura.

Para comprender correctamente el aprendizaje por transferencia y su respectiva definición, es necesario plantear unas definiciones previas.

4.1.1. Dominio. Un dominio $\mathcal{D}$ consiste de dos componentes. Un espacio de características χ y una istribución marginal de probabilidad $P(X)$, donde $X = (x_1, x_2, \ldots, x_n) \in \chi$. En general, si dos dominios son diferentes, podrían tener distinto espacio de características, o distribuciones de probabilidad marginales.

4.1.2. Tarea. Dado un dominio específico, $\mathcal{D} = (\chi, P(X))$, una *tarea* consiste en dos componentes. Un espacio de etiquetas $\mathcal{Y}$, y una función predictiva $f(\cdot)$ (denotada por $\mathcal{T} = (\mathcal{Y}, f(\cdot))$ que no es observada pero puede ser aprendida por los datos de entrenamiento, que consiste en pares (x_i, y_i), donde $x_i \in X$ y $y_i \in \mathcal{Y}$. La función $f(\cdot)$ puede ser usada para predecir la etiqueta correspondiente $f(x)$ para algún x. Con esto, podemos definir el aprendizaje por transferencia.

4.1.3. Aprendizaje por Transferencia. Dado algún dominio $\mathcal{D}_s$ y una tarea de aprendizaje $\mathcal{T}_s$, un dominio objetivo $\mathcal{D}_T$ y una tarea de aprendizaje $\mathcal{T}_T$, el aprendizaje por transferencia busca mejorar el aprendizaje de la función predictiva objetivo $f_t(\cdot)$ en $\mathcal{D}_T$ usando el aprendizaje de $\mathcal{D}_s$ y $\mathcal{T}_s$, donde $\mathcal{D}_s \neq \mathcal{D}_T$ o $\mathcal{T}_s \neq \mathcal{T}_T$.
Cabe aclarar en la definición de aprendizaje por transferencia que la condición $\mathcal{D}_s \neq \mathcal{D}_T$ implica que $\chi_s \neq \chi_T$ o $P_s(X) \neq P_T(X)$ [11].

El pre-training, o pre-entrenamiento, en aprendizaje automático, es una etapa de entrenamiento que entrena un modelo de propósito general (a veces llamado foundation model) utilizando datos de acceso público. El pre-entrenamiento suele

ir seguido de un ajuste fino (fine-tuning) para dotar al modelo de información específica para una tarea determinada [12].

Sea:

- $\mathcal{D}_s = \{(x_s^{(i)}, y_s^{(i)})\}$ el conjunto de datos fuente, extraído del dominio fuente $\mathcal{S}$ con distribución $P_s(X, Y)$.
- θ los parámetros del modelo f_θ.

El objetivo del preentrenamiento generalmente consiste en minimizar la pérdida L sobre los datos fuente:

$$\min_{\theta} \ \mathbb{E}_{(x_s, y_s)\sim P_s}\big[\, L(f_\theta(x_s),\, y_s)\,\big]$$

Este paso ayuda a que f_θ aprenda representaciones transferibles [11].

4.2. Pre-Training en Contextos Heterogéneos. En el trabajo de Schuessler, Sverdrup y Tibshirani (2025) amplía la comprensión del pre-training al introducirlo dentro del marco del aprendizaje estadístico heterogéneo, donde las relaciones entre variables difieren entre subpoblaciones o dominios. Los autores demuestran que, en muchos problemas empíricos, los factores que son pronósticos del resultado suelen ser también predictivos de la heterogeneidad del efecto del tratamiento. Aprovechando esta coincidencia, proponen una estrategia de preentrenamiento basada en el R-learner con penalización tipo LASSO (R-lasso), que mejora la precisión en la estimación del efecto condicional promedio del tratamiento (CATE). El valor teórico de este planteamiento radica en que el pre-training deja de ser solo una herramienta de predicción para convertirse en un instrumento de mejor inferencia estadística, al reducir la varianza en la estimación de los modificadores de efecto y fortalecer la recuperación del soporte compartido entre tareas. Esto implica que el conocimiento adquirido durante el preentrenamiento no solo acelera la convergencia del modelo, sino que también mejora la calidad inferencial del proceso, extendiendo su aplicabilidad a contextos causales y de inferencia estructurada [3].

4.3. Pre-Training en Cambio Climático. Un ejemplo es VITA (Variational Pretraining of Transformers for Climate Applications), que utiliza datos meteorológicos detallados durante el pre-entrenamiento para aprender patrones climáticos complejos y su relación con resultados agrícolas como los rendimientos de maíz y soya. Este enfoque mejora significativamente la precisión de las predicciones, especialmente para eventos climáticos extremos que se han vuelto más frecuentes debido al cambio climático. El pre-entrenamiento de VITA le permite generalizar bien a lo largo del tiempo y en diferentes geografías, capturando dinámicas universales entre clima y agricultura sin depender en gran medida de datos auxiliares como información del suelo. Esto demuestra cómo el pre-entrenamiento con datos históricos del clima puede mejorar la resiliencia y la precisión en la predicción de rendimientos agrícolas frente a los impactos del cambio climático [7].

Otra aplicación del pre-entrenamiento en ciencia climática es la restricción de proyecciones climáticas a largo plazo. Las redes neuronales profundas, pre-entrenadas con extensas simulaciones de modelos climáticos y observaciones históricas, pueden capturar mejor relaciones complejas como los cambios entre el CO_2 atmosférico y la temperatura. Esto mejora la confiabilidad y precisión de las proyecciones climáticas

futuras, reduce la incertidumbre en las estimaciones del aumento de temperatura y ayuda a evaluar cuándo podrían superarse umbrales críticos, como el límite de calentamiento global de 1.5 ∘ C [8].

También se están desarrollando modelos de lenguaje pre-entrenados como ClimateBERT, diseñados para manejar y analizar mejor textos y literatura científica relacionada con el clima, mejorando tareas como la clasificación de textos y el análisis semántico en investigación climática. En resumen, el pre-entrenamiento en la ciencia del clima ayuda a desarrollar modelos más precisos, generalizables y robustos frente a condiciones climáticas complejas y cambiantes. Esto beneficia las predicciones agrícolas, las proyecciones climáticas, las evaluaciones de impactos locales y también el procesamiento de información climática en forma de textos, apoyando así los esfuerzos de mitigación y adaptación en el contexto del cambio climático [10].

En resumen, el pre-entrenamiento en la ciencia del clima ayuda a desarrollar modelos más precisos, generalizables y robustos frente a condiciones climáticas complejas y cambiantes. Esto beneficia las predicciones agrícolas, las proyecciones climáticas, las evaluaciones de impactos locales y también el procesamiento de información climática en forma de textos, apoyando así los esfuerzos de mitigación y adaptación en el contexto del cambio climático.

5. Metodología y Resultados Obtenidos

Para la evaluación empírica se consideraron cuatro modelos generadores de datos distintos, cada uno representando un conjunto particular de supuestos de distribución. Adicionalmente, se aplicó un esquema de pre-entrenamiento en uno de estos modelos con el fin de analizar hasta qué punto el conocimiento adquirido bajo ese escenario específico podía transferirse a los otros tres generadores El entrenamiento y pre-entrenamiento se lleva a cabo usando LASSO.

La idea base de los modelos que fueron puestos a prueba para efectos de estos experimentos, es como sigue:
Sea un modelo lineal:

$$Y_g = X_g\beta_g + \epsilon_g$$

Donde:

- X_g es una matriz,
- β_g es un vector de coeficientes,
- $\epsilon_g \sim N(0, \sigma_g^2 I)$

Si todos los grupos comparten sus características $\beta_g = \beta_0$, el pre-entrenamiento no agrega ningún beneficio al ajuste individual.

5.1. MODELO I. El modelo I es como sigue:

$$Y_g = X_g\beta_g + \varepsilon_g, \quad \beta_g = \beta_0 + \delta_g, \quad \delta_g \sim \mathcal{N}(0, \tau^2 I_p)$$

La motivación detrás de este, es que todos los grupos comparten los mismos predictores y estructura general del modelo, pero cada uno tiene variaciones contextuales en los coeficientes, con respecto a la intución de esta propuesta es que el pre-entrenamiento puede estimar la componente compartida β_0 eficientemente y

luego realizar ajuste fino en las desviaciones locales δ_g. Aunque la estructura del Modelo I,

$$\beta_g = \beta_0 + \delta_g, \qquad \delta_g \sim N(0, \tau^2 I_p),$$

puede recordar a la formulación de un modelo jerárquico o de efectos aleatorios, en este estudio no se interpreta como tal.
El objetivo del modelo es únicamente generar variaciones controladas entre grupos a través del término δ_g, sin especificar una estructura multinivel completa ni realizar inferencia sobre componentes de varianza, como suele hacerse en los modelos jerárquicos formales. Por tanto, el Modelo I comparte una forma matemática similar, pero no se considera un modelo jerárquico en sentido estricto dentro del enfoque adoptado.

5.2. MODELO II. El modelo II es como sigue:

$$Y_g = X_g\beta_0 + \varepsilon_g, \quad \varepsilon_g \sim \mathcal{N}(0, \sigma_g^2 I_n), \quad \sigma_g^2 \in \{1, 1{,}5, 2, 3\}$$

La motivación detrás de este modelo es que los grupos comparten la misma estructura media pero difieren en su nivel de ruido. Esto pretende capturar heterocedasticidad a través de subpoblaciones, con respecto a la intuición, se pretende probar si el pre-entrenamiento estabiliza las estimaciones en grupos con varianzas más grandes.

5.3. MODELO III.

$$\beta_1 = [1, 0{,}8, 0{,}5], \quad \beta_2 = [1, -0{,}8, 0{,}5]$$

$$Y_g = X_g\beta_g + \varepsilon_g$$

La motivación tras este modelo es que los grupos siguen la misma estructura de regresión pero difieren en la dirección de uno de los efectos, pretendiendo representar similaridad parcial entre poblaciones.

5.4. MODELO IV.

$$Y_4 \;=\; X_{4,1}\,\beta_{4,1} \;+\; X_{4,2}\,\beta_{4,2} \;+\; \varepsilon_4.$$

El modelo 4, utiliza únicamente las primeras dos columnas predictoras de X para explicar Y_4.

5.5. Resultados obtenidos. Fijando $n = 100, p = 3$ obtenemos los resultados que siguen:

Grupo	n	MSE Pre-Entrenamiento	MSE_Entrenamiento
2	100	2.335063	2.274642
3	100	3.653632	3.636336
4	100	8.478320	8.601884

CUADRO 1. Resumen de MSE para los distintos grupos.

En los resultados obtenidos en la tabla 1 se evidencia que bajo las condiciones evaluadas (número de muestra y características fijo, el pre-entrenamiento no ofrece mejoras significativas en los grupos 2 y 3, lo cual se refleja en los valores de MSE muy similares.

Esto indica que, para estos grupos, los coeficientes preentrenados no aportan información adicional útil, ya sea porque los modelos objetivo difieren sustancialmente del modelo preentrenado o porque los datos disponibles ya son suficientes para una estimación precisa.

En contraste, el Grupo 4 muestra una reducción ligeramente mayor, pero significativa en el MSE al usar LASSO con preentrenamiento, lo que sugiere que el preentrenamiento es beneficioso cuando el grupo objetivo tiene una señal más débil, mayor ruido o menos predictores informativos.

Además, los resultados sugieren que el impacto del pre-entrenamiento depende fuertemente del grado de similitud entre el modelo fuente y el modelo objetivo. En los grupos 2 y 3, donde la estructura del modelo verdadero coincide con la del entrenamiento base, pero la señal es suficientemente fuerte o los datos son informativos, el pre-entrenamiento no aporta mejoras sustanciales. Esto coincide con la teoría previa, que afirma que el beneficio del pre-training disminuye cuando los modelos locales ya pueden estimarse con baja varianza. En contraste, el Grupo 4 presenta una estructura distinta, utilizando únicamente dos de los predictores para generar la respuesta. En este caso, el pre-entrenamiento actúa como un mecanismo regularizador, ayudando al modelo a estabilizar los coeficientes en presencia de una señal más débil y mayor incertidumbre. Este patrón refuerza la idea de que el pre-training es más útil en escenarios con heterogeneidad estructural o cuando los datos por grupo poseen menos información útil. En conjunto, estos hallazgos son coherentes con los resultados de la literatura reciente, donde el pre-entrenamiento tiende a mejorar el desempeño cuando existe algún componente global compartido entre dominios, pero su beneficio disminuye cuando los modelos específicos son suficientemente robustos o cuando el soporte entre tareas difiere marcadamente.

6. Conclusiones

El aprendizaje por transferencia se ha consolidado como un paradigma del aprendizaje automático con un alto potencial para mejorar el rendimiento de los modelos bajo las condiciones adecuadas. Dentro de este marco, el preentrenamiento ha ganado especial relevancia en los últimos años, pues permite aprovechar modelos de gran escala entrenados con vastas cantidades de información y transferir ese conocimiento a tareas más específicas. Esto se traduce en reducciones importantes de tiempo, recursos computacionales y costos para quienes implementan estos métodos.

No obstante, a pesar de sus numerosas ventajas, el preentrenamiento no garantiza mejoras en todos los casos. Existen escenarios en los que un modelo preentrenado no supera de manera significativa a un modelo entrenado desde cero, ya sea por diferencias sustanciales entre el dominio original y el dominio objetivo, o por la disponibilidad suficiente de datos específicos para la tarea final. Tal comportamiento se observó también en la sección de resultados de este trabajo, donde el preentrenamiento no produjo mejoras consistentes en todos los grupos evaluados.

Como trabajo futuro, se plantea evaluar y validar la eficacia del preentrenamiento en datos propios de una región particularmente vulnerable, como Honduras, con el

fin de generar conocimiento que contribuya a la mitigación del riesgo asociado a la inestabilidad climática.

Referencias

1. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer, New York, 2013.
2. E. Craig, M. Pilanci, T. Le Menestrel, B. Narasimhan, M. A. Rivas, S.-E. Gullaksen, R. Dehghannasiri, J. Salzman, J. Taylor, and R. Tibshirani, *Pretraining and the Lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), qkaf050, 2025. Disponible en: `https://doi.org/10.1093/jrsssb/qkaf050`.
3. M. Schuessler, E. Sverdrup, and R. Tibshirani, *Statistical Learning for Heterogeneous Treatment Effects: Pretraining, Prognosis, and Prediction*, arXiv preprint arXiv:2505.00310, 2025. Disponible en: `https://arxiv.org/abs/2505.00310`.
4. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, 2009.
5. C. Stryker, *Pretrained Models*, IBM Think, Staff Editor, AI Models, 2025. Disponible en: `https://www.ibm.com/think/topics/pretrained-model`.
6. P. Niu, Z. Ma, T. Zhou, W. Chen, L. Shen, R. Jin, and L. Sun, *Utilizing Strategic Pre-training to Reduce Overfitting: Baguan – A Pre-trained Weather Forecasting Model*, Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25), Association for Computing Machinery, New York, 2025, pp. 2186–2197. Disponible en: `https://doi.org/10.1145/3711896.3737178`.
7. A. Hasan, M. Roozbehani, and M. Dahleh, *VITA: Variational Pretraining of Transformers for Climate-Robust Crop Yield Forecasting*, arXiv preprint arXiv:2508.03589, 2025. Disponible en: `https://arxiv.org/abs/2508.03589`.
8. F. Immorlano, V. Eyring, T. le Monnier de Gouville, G. Accarino, D. Elia, S. Mandt, G. Aloisio, & P. Gentine, *Transferring climate change physical knowledge*, Proc. Natl. Acad. Sci. U.S.A. 122 (15) e2413503122, Disponible en: `https://doi.org/10.1073/pnas.2413503122`, (2025).
9. J. González-Abad, M. Iturbide, A. Hernanz, and J. M. Gutiérrez, *Pre-training for Deep Statistical Climate Downscaling: A case study within the Spanish National Adaptation Plan (PNACC)*, EGUsphere, vol. 2025, pp. 1–25, 2025. Disponible en: `https://egusphere.copernicus.org/preprints/2025/egusphere-2025-3754/`. DOI: `https://doi.org/10.5194/egusphere-2025-3754`.
10. N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold, *ClimateBert: A Pretrained Language Model for Climate-Related Text*, CoRR, vol. abs/2110.12010, 2021. Disponible en: `https://arxiv.org/abs/2110.12010`.
11. S. J. Pan and Q. Yang, *A Survey on Transfer Learning*, IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010. Disponible en: `https://doi.org/10.1109/TKDE.2009.191`.
12. NIST, *pre-training*, CSRC Glossary, 2025. Available at: `https://csrc.nist.gov/glossary/term/pre_training`.

Maestría en Matemática, Universidad Nacional Autónoma de Honduras.
Dirección actual: Dirección actual
Dirección de correo electrónico: `nathalye.deras@unah.hn`

ESTIMACIÓN DE VELOCIDAD Y DENSIDAD VEHICULAR MEDIANTE REDES NEURONALES CONVOLUCIONALES PARA EL AJUSTE DE MODELOS DE REGRESIÓN

RUTH EUNICE MORENO MELARA

Dedicado a mi familia

Resumen. En esta investigación se plantea un estudio orientado a la estimación de velocidad vehicular mediante técnicas de visión por computadora y aprendizaje automático. Los videos capturados en carreteras serán procesados mediante un modelo de Redes Neuronales Convolucionales (CNN) entrenado con una base de datos elaborada por los propios autores, a partir de imágenes y secuencias de video etiquetadas manualmente. Este modelo permitirá la detección de vehículos, sobre la cual se desarrollarán algoritmos de seguimiento y estimación de velocidad. Con los datos obtenidos se calculará la densidad vehicular en intervalos de tiempo definidos, aplicando métodos de muestreo que garanticen la representatividad de la información. Finalmente, se implementará una regresión lineal entre velocidad y densidad, cuyos coeficientes servirán como parámetros de entrada en un modelo de congestión vehicular formulado mediante ecuaciones diferenciales parciales. El objetivo de la investigación es generar información precisa y validada que contribuya al desarrollo de modelos avanzados para el análisis y predicción del tráfico vehicular.

Abstract. This research presents a study focused on vehicle speed estimation using computer vision and machine learning techniques. The videos captured on roadways will be processed through a Convolutional Neural Network (CNN) model trained with a self-developed database, built from manually labeled images and video sequences. This model will enable vehicle detection, upon which tracking and speed estimation algorithms will be developed. Using the data obtained, vehicle density will be calculated over defined time intervals, applying sampling methods that ensure representativeness and reduce potential bias. Finally, a linear regression between speed and density will be implemented, whose coefficients will serve as input parameters for a traffic congestion model formulated through partial differential equations. The objective of this research is to generate accurate and validated information that contributes to the development of advanced models for traffic analysis and prediction.

1. Introducción

En esta investigación se aborda el desafío de estimar la velocidad vehicular y analizar su relación con la densidad de tráfico mediante técnicas de visión por computadora y aprendizaje automático. El estudio combina el uso de Redes Neuronales Convolucionales (CNNs) para la detección de vehículos, desarrollo de algoritmos

Fecha: Octubre 2025.
Palabras y frases clave. Redes Neuronales, Detección, estimación, regresión.

de seguimiento de objetos y la estimación de la velocidad respectiva de cada objeto. A partir de esta herramienta, se pretende realizar la estimación de coeficientes de regresión, con el propósito de generar modelos cuantitativos que describan el comportamiento del tránsito en entornos reales.

Los videos capturados en carreteras se procesan mediante un modelo de CNN entrenado con una base de datos de elaboración propia, construida a partir de imágenes y secuencias de video etiquetadas manualmente. Este modelo se encarga de la detección de vehículos en cada cuadro del video (frames), sobre la cual se desarrolla un algoritmo de seguimiento que permite identificar, etiquetar y dar continuidad a cada vehículo detectado a lo largo de la secuencia. A partir de la información generada por este seguimiento, se implementa un algoritmo adicional para la estimación de la velocidad, con el fin de calcular el desplazamiento de cada vehículo en función del tiempo y la posición. Posteriormente, con los resultados de estos dos procesos, se diseña un tercer algoritmo para el cálculo de la densidad vehicular, permitiendo obtener las dos variables fundamentales, velocidad y densidad, que serán empleadas en la regresión lineal destinada a modelar su relación.

El estudio también contempla la validación de los datos recolectados y la evaluación de los métodos de muestreo más apropiados, con el objetivo de garantizar la representatividad y reducir posibles sesgos en las estimaciones.

En conjunto, esta investigación pretende aportar herramientas metodológicas y analíticas que contribuyan al avance del conocimiento en el campo de la visión por computadora aplicada al transporte, favoreciendo la comprensión y modelización de la dinámica vehicular en contextos urbanos y carreteros.

2. Justificación

La creciente congestión vehicular en los principales corredores urbanos del país representa un problema de gran impacto económico y social, generando pérdidas de tiempo, aumento en el consumo de combustible y mayores niveles de contaminación ambiental. En este contexto, la estimación precisa de la velocidad y la densidad vehicular constituye una herramienta para el diseño de políticas públicas orientadas a la optimización del tránsito, la planificación de infraestructura vial y la mejora de la movilidad urbana.

Esta investigación propone un enfoque basado en redes neuronales convolucionales (CNN) y técnicas de aprendizaje automático para la detección, seguimiento y estimación de velocidad de vehículos a partir de secuencias de video, complementado con un análisis de regresión lineal entre la velocidad y la densidad vehicular. La aplicación de estos métodos permite generar información que puede ser utilizada en modelos de predicción y simulación del tráfico, contribuyendo a la formulación de estrategias que promuevan una gestión vial más eficiente y sostenible.

El desarrollo de esta temática se alinea con los temas prioritarios del Eje 1 de la UNAH, “Desarrollo Económico y Social”, específicamente en el apartado de “Infraestructura y desarrollo territorial”, ya que sus resultados pueden apoyar la planificación y modernización del sistema vial nacional.

El enfoque interdisciplinario de este trabajo combina el rigor matemático con herramientas de inteligencia artificial, lo que fortalece la capacidad de análisis y

predicción de fenómenos complejos relacionados con la movilidad urbana, aportando así al desarrollo científico y tecnológico del país.

3. Antecedentes

El desarrollo de sistemas automáticos para la estimación de velocidad vehicular constituye una línea de investigación dentro de la visión por computadora y los sistemas inteligentes de transporte. La necesidad de mejorar la seguridad vial, optimizar el flujo de tráfico y reducir la congestión ha impulsado la implementación de tecnologías capaces de analizar de manera automática las secuencias de video obtenidas por cámaras de vigilancia. Estos sistemas permiten estimar la velocidad, la densidad y la clasificación de vehículos en tiempo real, lo cual representa un insumo esencial para la planificación urbana, la detección de infracciones, entre otras. [7].

Los primeros enfoques para determinar la velocidad vehicular a partir de secuencias de video se basaron principalmente en el análisis de flujo óptico, técnica que estima el movimiento de los píxeles entre fotogramas consecutivos para calcular la dirección y magnitud del desplazamiento. Ruimin Ke et al. [2] propusieron un método que combina el flujo óptico con el algoritmo de agrupamiento *K-Means*, aplicado a videos aéreos capturados por vehículos no tripulados. Este enfoque permite calcular la velocidad promedio de los vehículos en escala de imagen y posteriormente convertirla a unidades reales, alcanzando un error relativo de aproximadamente el 12 %.

Con el objetivo de mejorar la precisión de las mediciones, se introdujeron modelos de calibración de cámara que consideran la altura de instalación y el ángulo de inclinación. Karim et al. [3] demostraron que el uso de parámetros geométricos permite transformar las coordenadas del plano de imagen al plano del mundo real, reduciendo los errores asociados a la perspectiva. Sin embargo, la calibración manual de cada cámara representa una limitación significativa para la escalabilidad de estos sistemas en entornos urbanos complejos.

Makwana y Goel [4] introdujeron un modelo que integra detección, clasificación y seguimiento de vehículos mediante la conversión de coordenadas del centro geométrico del objeto desde el sistema de imagen al sistema del mundo real. Su propuesta estableció la base para los algoritmos posteriores que incorporan técnicas de seguimiento, como el filtro de Kalman [5] y el algoritmo húngaro [6], para mejorar la continuidad de las trayectorias en múltiples fotogramas.

Con el auge del aprendizaje profundo, los modelos de Redes Neuronales Convolucionales (CNN) han revolucionado la detección y el seguimiento de objetos en video. Estos avances facilitaron la identificación automática de vehículos, su clasificación en múltiples categorías y el cálculo de sus trayectorias mediante sistemas de seguimiento en tiempo real. En este contexto, Grents et al. [1] desarrollaron un sistema que combina un detector Faster R-CNN de dos etapas con el algoritmo de seguimiento SORT (*Simple Online and Real-Time Tracking*), logrando determinar la velocidad de los vehículos con un error porcentual absoluto promedio inferior al 22 %. El modelo fue entrenado con más de 52,000 objetos extraídos de videos urbanos y mostró un desempeño robusto frente a condiciones variables de iluminación y densidad vehicular.

Las CNN, al permitir la extracción jerárquica de características espaciales, han superado ampliamente las limitaciones de los métodos tradicionales de flujo óptico y de calibración geométrica. La combinación con filtros estadísticos como el de Kalman posibilita la predicción de trayectorias bajo ruido o interrupciones momentáneas, mientras que algoritmos de optimización como el húngaro permiten una asignación eficiente de detecciones entre cuadros consecutivos. No obstante, las principales dificultades actuales se centran en la dependencia de grandes volúmenes de datos etiquetados, la sensibilidad a la resolución de video y el procesamiento en tiempo real en entornos urbanos congestionados.

En síntesis, la literatura muestra una evolución progresiva desde los modelos basados en flujo óptico hasta los enfoques híbridos que integran aprendizaje profundo y técnicas de seguimiento probabilístico. La tendencia actual se orienta hacia sistemas que no solo estimen la velocidad vehicular, sino que también incorporen la densidad del tráfico, la detección de patrones anómalos y la predicción de congestión. Este panorama refleja la relevancia científica y tecnológica del tema, y justifica la continuidad de investigaciones orientadas a mejorar la precisión y eficiencia de los modelos de estimación de velocidad vehicular mediante CNN y métodos estadísticos.

4. Construcción de la Base de Datos

La calidad del conjunto de datos incide directamente en el rendimiento de los modelos de visión por computador. En esta sección se detalla el procedimiento para la adquisición, preprocesamiento, anotación y organización del dataset.

4.1. Adquisición de datos. En primer lugar, se debe realizar la adquisición del material visual, entendida como la recopilación de secuencias de video o imágenes estáticas bajo condiciones controladas o naturales. Este proceso implica definir el entorno de captura, la resolución objetivo, la tasa de fotogramas y el posicionamiento de las cámaras, con el fin de minimizar sesgos asociados a iluminación, oclusiones o variabilidad excesiva del fondo.

Sea $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$ el conjunto de videos obtenidos con una cámara de resolución 1920×1080 píxeles y frecuencia de muestreo FPS $= 30$

Cada video V_k es una secuencia temporal de fotogramas

$$V_k = \{I_{k,1}, I_{k,2}, \ldots, I_{k,T_k}\},$$

donde cada imagen

$$I_{k,t} \in \mathbb{R}^{H\times W\times 3}, \quad H = 1080,\ W = 1920.$$

Con $\mathbb{R}^{H\times W\times 3} = \{A \mid A_{i,j,c} \in \mathbb{R},\ 1 \le i \le H,\ 1 \le j \le W,\ 1 \le c \le 3\}$.

Es decir, cada imagen es un tensor (ver sección 7.1.1) cuya primera dimensión corresponde a la altura (H), la segunda al ancho (W), y la tercera a los tres canales de color (RGB).

Los frames extraídos se recopilan en el conjunto

$$\mathcal{I} = \bigcup_{k=1}^{N} \{I_{k,t} : 1 \leq t \leq T_k\}.$$

En este contexto, un frame se define como una imagen estática obtenida a partir de una secuencia de video. Cada video k está compuesto por T_k imágenes ordenadas temporalmente, denotadas como $I_{k,1}$, $I_{k,2}$, ..., I_{k,T_k}. Por tanto, el conjunto $\mathcal{I}$ representa la colección total de imágenes individuales derivadas de todos los videos considerados.

Se debe realizar una selección manual de imágenes relevantes dentro del conjunto $\mathcal{I}$, de modo que únicamente aquellas muestras que aportan información útil sean retenidas para el entrenamiento del modelo. Aunque $\mathcal{I}$ contiene todos los frames extraídos, no todos ellos presentan condiciones adecuadas para la tarea de visión por computador que se pretende abordar. Esta depuración inicial facilita las etapas posteriores de anotación, ya que elimina muestras que podrían inducir ambigüedad o inconsistencias en el etiquetado.

Una vez identificado el subconjunto de imágenes relevantes $|\mathcal{I}|$, el siguiente paso consiste en aplicar un proceso de preprocesamiento destinado a normalizar y estandarizar el material visual.

5. Preprocesamiento de imágenes

Para cada imagen original $I \in \mathcal{I}$ se aplicaron transformaciones definidas como una función $\Phi : \mathbb{R}^{H \times W \times 3} \longrightarrow \mathbb{R}^{h \times w \times 3}$ donde $(h, w) = (416, 416)$ es la resolución utilizada por el modelo CNN que se han analizado, de esta forma se realiza una normalización y redimensionamiento a cada imagen mediante $I' = \Phi(I) = \text{Resize}(I, 416, 416)$.

El redimensionamiento garantiza la homogeneidad espacial del conjunto de imágenes, pero no aborda la presencia de información irrelevante en la escena. Para focalizar el procesamiento en los objetos de interés, se incorpora una etapa adicional: la segmentación de fondo.

5.1. Segmentación de fondo. Tras normalizar la resolución y estructura tensorial de cada imagen, es necesario aplicar operaciones que permitan reducir la presencia de información irrelevante dentro de la escena. Muchas imágenes contienen amplias regiones de fondo que no aportan contenido significativo para la tarea de detección y seguimiento. Para eliminar este efecto y centrar el procesamiento en los objetos de interés, se incorpora una etapa de segmentación de fondo, descrita a continuación.

Sea $I' \in \mathbb{R}^{h \times w \times 3}$ una imagen preprocesada. Definimos una función de segmentación

$$S : \mathbb{R}^{h \times w \times 3} \longrightarrow \{0, 1\}^{h \times w},$$

la cual asigna a cada píxel de la imagen un valor binario. El resultado es una máscara $M = S(I')$ donde

$$M(x,y) = \begin{cases} 1, & \text{si el píxel pertenece a la región de interés,} \\ 0, & \text{si el píxel pertenece al fondo.} \end{cases}$$

La imagen segmentada (región de interés) se obtiene aplicando la máscara

$$I_{\text{ROI}}(x,y) = I'(x,y) \cdot M(x,y),$$

es decir, se conservan únicamente los píxeles donde $M = 1$ y se eliminan (se vuelven cero) los píxeles donde $M = 0$.

6. Datos a utilizar: Anotación y clases definidas

Dado que se utilizará un modelo de CNN en un esquema de aprendizaje supervisado, es indispensable contar con un conjunto de datos previamente anotado. Esto implica especificar, para cada imagen, tanto la ubicación de los objetos de interés como la clase a la que pertenecen. Dichas anotaciones constituyen la información necesaria para que la red neuronal pueda aprender a detectar y clasificar correctamente los vehículos presentes en las imágenes.

Se debe definir $\mathcal{C}$ el conjunto de clases consideradas, como por ejemplo

$$\mathcal{C} = \{\texttt{car}, \texttt{big\_car}, \texttt{motorcycle}, \texttt{small\_bus}\}.$$

Para cada imagen I_{ROI} se define un conjunto de anotaciones

$$\mathcal{B}(I_{\text{ROI}}) = \{(b_i, c_i)\}_{i=1}^{m_I},$$

donde:

- $b_i \in \mathbb{R}^4$ representa un *bounding box* en formato
$$b_i = (x_i, y_i, w_i, h_i),$$
siendo (x_i, y_i) el centro del recuadro y (w_i, h_i) su ancho y alto.
- $c_i \in \mathcal{C}$ es la clase anotada.
- m_I es el número total de objetos anotados en esa imagen.

Cada imagen anotada satisface $m_I \geq 1$, es decir, contiene al menos un objeto etiquetado.

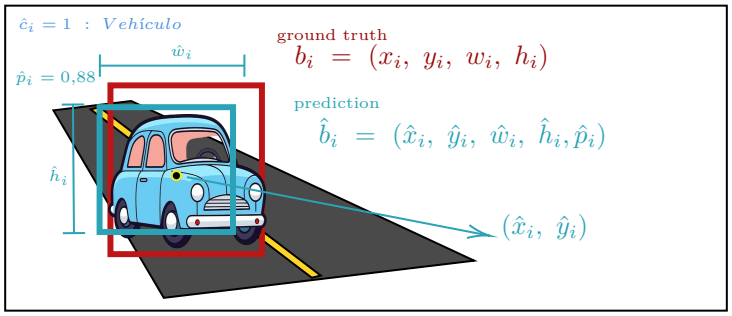


Figura 1. Ejemplo gráfico de un objeto anotado y su correspondiente predicción. El recuadro rojo indica la anotación de referencia, mientras que el recuadro amarillo muestra el bounding box estimado por la red

La Figura 6 ilustra la relación entre las anotaciones manuales utilizadas durante el entrenamiento y las predicciones generadas por la CNN. El recuadro rojo representa el ground truth $b_i = (x_i, y_i, w_i, h_i)$, mientras que el recuadro celeste corresponde a la predicción del modelo $\hat{b}_i = (\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i, \hat{p}_i)$, que incluye tanto las coordenadas estimadas como la probabilidad asociada a la presencia del objeto.

Cabe señalar que la anotación se realiza sobre el mismo tipo de imagen que será utilizada para entrenar a la CNN. Si el modelo recibe la imagen original I, las etiquetas deben definirse sobre I; si recibe únicamente la región segmentada I_{ROI}, entonces la anotación debe hacerse sobre I_{ROI}. De este modo se garantiza coherencia geométrica entre las anotaciones y los datos empleados en el entrenamiento.

Estos elementos constituyen la base semántica sobre la cual el modelo aprende a distinguir y localizar los objetos de interés. La forma en que la CNN procesa estas anotaciones, transforma la información visual y aprende representacionesse detallará en la sección 7.1.

6.1. Estructura y partición del dataset. Una vez completado el proceso de etiquetado, se debe definir el dataset total, usualmente como

$$\mathcal{D} = \{(I_{\text{ROI}}, \mathcal{B}(I_{\text{ROI}}))\}.$$

En un escenario típico de aprendizaje supervisado, este conjunto se particiona en tres subconjuntos disjuntos destinados a funciones específicas dentro del esquema de entrenamiento.

$$\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}} = \mathcal{D}, \qquad \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{val}} = \varnothing, \quad \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \varnothing,$$

con proporciones

$$|\mathcal{D}_{\text{train}}| = 0{,}70\,|\mathcal{D}|, \qquad |\mathcal{D}_{\text{val}}| = 0{,}20\,|\mathcal{D}|, \qquad |\mathcal{D}_{\text{test}}| = 0{,}10\,|\mathcal{D}|.$$

Donde

- $\mathcal{D}_{\text{train}}$: datos utilizados para entrenar el modelo,
- $\mathcal{D}_{\text{val}}$: datos usados para ajustar hiperparámetros y evitar sobreajuste,
- $\mathcal{D}_{\text{test}}$: datos reservados para evaluar el desempeño final.

Esta partición debe asegurar independencia entre subconjuntos y representatividad de las clases para obtener una evaluación confiable del modelo. Estas cifras pueden variar según la disponibilidad de datos y la complejidad del problema. La finalidad de esta estructura es proporcionar un marco experimental reproducible y consistente, sobre el cual se puedan comparar distintos modelos o configuraciones.

7. Arquitectura General del Sistema

El sistema opera sobre una secuencia de imágenes extraída de un video, la cual puede representarse como $\{I_t\}_{t=1}^{T}$, donde cada elemento I_t corresponde al fotograma capturado en el instante t. En la práctica, un video puede considerarse como un conjunto de imágenes ordenadas temporalmente, y su captura está determinada por la tasa de muestreo de la cámara.

Sea f_s la tasa de muestreo expresada en fotogramas por segundo (FPS). Este parámetro indica cuántas imágenes son registradas en un segundo de grabación. De

esta forma, el intervalo temporal entre dos fotogramas consecutivos viene dado por $\Delta t = 1/f_s$. Por ejemplo, si el video se registra a 30 FPS, entonces cada imagen se obtiene cada $\Delta t \approx 0{,}033$ segundos.

El objetivo consiste en transformar esta secuencia en un conjunto de trayectorias vehiculares $\gamma^{(k)}(t)$ y, posteriormente, en estimaciones de velocidad $v^{(k)}(t)$.

La trayectoria de cada vehículo se representa mediante la función $\gamma^{(k)}(t)$, la cual indica la posición del vehículo k en cada instante t del video. De forma simple, se define como

$$\gamma^{(k)}(t) = (X_t^{(k)},\, Y_t^{(k)}),$$

donde $(X_t^{(k)}, Y_t^{(k)})$ corresponde a la ubicación del vehículo en el plano real en el tiempo t. A partir de esta trayectoria, la velocidad del vehículo se describe mediante la función $v^{(k)}(t)$, definida como la razón entre el desplazamiento entre dos fotogramas consecutivos y el tiempo transcurrido entre ellos:

$$v^{(k)}(t) = \frac{\|\gamma^{(k)}(t + \Delta t) - \gamma^{(k)}(t)\|}{\Delta t}.$$

De esta manera, $\gamma^{(k)}(t)$ indica en donde se encuentra el vehículo en cada instante, mientras que $v^{(k)}(t)$ indica la velocidad entre un frame y el siguiente.

Para lograrlo, el sistema se divide en tres componentes:

1. **Detección**: identificación de vehículos en cada imagen mediante una CNN.
2. **Seguimiento**: asociación entre detecciones de fotogramas consecutivos y estimación de la posición del vehículo en el tiempo.
3. **Estimación cinemática**: cálculo de desplazamientos reales y de velocidades.

De esta manera, cada módulo transforma la información del anterior, permitiendo obtener descripciones coherentes del movimiento vehicular.

7.1. Modelo de Detección Basado en CNN. El modelo que se planea utilizar en esta investigación se enmarca dentro del campo de la *Inteligencia Artificial* (IA), entendida como el conjunto de métodos que permiten que un sistema computacional realice tareas que, tradicionalmente, requieren de capacidades humanas tales como percepción, toma de decisiones o clasificación. Dentro de este campo, el *Machine Learning* (ML) constituye la rama que se enfoca en el diseño de algoritmos capaces de aprender patrones a partir de datos.

A su vez, el *Deep Learning* (DL) es una subcategoría de ML basada en modelos compuestos por múltiples capas no lineales, capaces de aproximar funciones de alta complejidad. Cuando estos modelos se entrenan utilizando datos etiquetados, el enfoque se denomina *aprendizaje supervisado*.

En particular, se empleará una *Red Neuronal Convolucional* (CNN), un arquitectura de DL especialmente diseñada para el procesamiento de imágenes, debido a su capacidad para extraer automáticamente características espaciales y jerárquicas

relevantes para la detección de objetos.

Según [12], así como en los organismos biológicos necesitan estímulos externos para el aprendizaje, en las redes neuronales artificiales el estímulo externo lo proporcionan los datos de entrenamiento que contienen ejemplos de pares entrada-salida de la función que se va a aprender. Por ejemplo, la formación los datos pueden contener representaciones en píxeles de imágenes (entrada) y sus etiquetas anotadas (por ejemplo, auto, motocicleta) como salida.

Estos pares de datos de entrenamiento se introducen en la red neuronal mediante el uso de representaciones de entrada para hacer predicciones sobre las etiquetas de salida. Los datos de entrenamiento proporcionan retroalimentación sobre la exactitud de los pesos en la red neuronal dependiendo de qué tan bien coincida la salida predicha (por ejemplo, la probabilidad de auto) para una entrada particular con la etiqueta de salida anotada en los datos de entrenamiento.

Supongamos que tenemos un conjunto de datos de entrenamiento

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(n)}, y^{(n)})\}$$

donde $x^{(i)}$ es la representación de entrada y $y^{(i)}$ es la etiqueta de salida correspondiente para el ejemplo i.

Una red neuronal toma la representación de entrada $x^{(i)}$ y produce una salida predicha $\hat{y}^{(i)}$. Esto se puede expresar como una función $f(x^{(i)}; \theta)$, donde θ son los parámetros (pesos) de la red neuronal [12].

La exactitud de la predicción se puede cuantificar utilizando una función de pérdida $L(y^{(i)}, \hat{y}^{(i)})$, que mide la discrepancia entre la salida predicha y la etiqueta de salida verdadera. La retroalimentación sobre la exactitud de los pesos en la red neuronal se obtiene minimizando esta función de pérdida sobre el conjunto de datos de entrenamiento.

Por lo tanto, el objetivo es encontrar los parámetros θ que minimicen la función de pérdida promedio sobre todos los ejemplos de entrenamiento:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L(y^{(i)}, \hat{y}^{(i)}) = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L(y^{(i)}, f(x^{(i)}; \theta))$$

Donde $L(y^{(i)}, \hat{y}^{(i)})$ es una función de pérdida específica, como la entropía cruzada en el caso de la clasificación, y $\frac{1}{n} \sum_{i=1}^{n} L(y^{(i)}, \hat{y}^{(i)})$ es la pérdida promedio sobre todos los ejemplos de entrenamiento [12].

7.1.1. Arquitectura general de las CNN. La arquitectura general de las CNN's, se componen de tres tipos de capas: las **capas convolucionales**, las **capas de agrupamiento** (pooling) y las **capas totalmente conectadas** (fully-connected).

Capas de convolución: El nombre "red neuronal convolucional" indica que la red utiliza una operación matemática llamada **convolución**. La convolución es un

tipo especializado de operación lineal. Las redes convolucionales *son simplemente redes neuronales que utilizan la convolución en lugar de la multiplicación de matrices general en al menos una de sus capas* [13].

Sea x y w dos funciones continuas, el producto especial denotado por $x * w$ se define mediante la integral

$$s(t) = (x * w)(t) = \int_{-\infty}^{\infty} x(\tau) w(t - \tau)\, d\tau \tag{7.1}$$

se llama **convolución** de x y w.

En la terminología de redes convolucionales, el primer argumento de la Eq. 7.1 de la convolución a menudo se denomina **input** y el segundo argumento (w) como **kernel** [12].

Normalmente, según [12], cuando trabajamos con datos en una computadora, el tiempo será *discretizado*. Si ahora asumimos que x y w están definidos solo en el número entero t, podemos definir la convolución discreta como:

$$S(i, j) = (K * I)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \tag{7.2}$$

En aplicaciones de aprendizaje automático, la entrada suele ser un arreglo multidimensional de datos y el kernel suele ser un arreglo multidimensional de parámetros que son adaptados por el algoritmo de aprendizaje. Nos referiremos a estos arreglos multidimensionales como **tensores** [12].

Debido a que cada elemento de la entrada y el kernel deben ser almacenados explícitamente por separado, usualmente asumimos que estas funciones son cero en todas partes excepto en el conjunto finito de puntos para los cuales almacenamos los valores. Esto significa que en la práctica podemos implementar la sumatoria infinita como una sumatoria sobre un número finito de elementos del arreglo.

Capa Pooling: La capa *pooling* realiza una operación de submuestreo sobre los mapas de características con el fin de reducir su resolución espacial y aumentar la robustez del modelo frente a pequeñas traslaciones o ruido. En el caso de *max pooling*, la salida se define como

$$Y_{i,j} = \text{máx}\{X_{p,q} : (p, q) \in \text{ventana}(i, j)\},$$

donde cada ventana corresponde a una región local (por ejemplo, 2×2) del mapa de activaciones X. Esta operación conserva las características de mayor relevancia mientras reduce la dimensionalidad [12] .

Capa Fully Connected: La capa *fully connected* (o densamente conectada) integra la información extraída por las capas convolucionales para producir la decisión final del modelo. Implementa una transformación afín del tipo

$$y = Wx + b,$$

donde x es el vector de características aplanado, W es la matriz de pesos y b es el vector de sesgos. Esta capa se utiliza típicamente en las etapas finales de la red

para tareas de clasificación o regresión [12].

De este modo, cada imagen I_t es procesada por un modelo CNN de detección $\mathcal{F}_\theta$ que devuelve un conjunto de cajas delimitadoras

$$\mathcal{B}_t = \{b_{t,i}\}_{i=1}^{N_t},$$

donde cada detección $b_{t,i}$ contiene:

$$b_{t,i} = (x_{t,i},\, y_{t,i},\, w_{t,i},\, h_{t,i},\, c_{t,i},\, p_{t,i}),$$

siendo:

- $(x_{t,i}, y_{t,i})$: centro de la caja,
- $w_{t,i}, h_{t,i}$: ancho y alto en píxeles,
- $c_{t,i}$: clase del objeto (vehículo),
- $p_{t,i}$: probabilidad o confianza asignada por la red.

8. Algoritmos de Seguimiento de Objetos

Una vez obtenidas las detecciones cuadro a cuadro mediante el modelo de detección $\mathcal{F}_\theta$, es necesario establecer correspondencias temporales entre ellas para construir trayectorias coherentes. Este proceso se conoce como *seguimiento de objetos* (object tracking). En este trabajo se analizan dos enfoques distintos: un método puramente geométrico basado en las salidas de la CNN y un método probabilístico basado en el Filtro de Kalman combinado con el Algoritmo Húngaro.

8.1. Seguimiento Geométrico Basado en Detecciones. Cada objeto detectado en un cuadro t está delimitado por una caja delimitadora o *bounding box*, la cual se describe mediante sus coordenadas espaciales. El seguimiento entre cuadros consecutivos se realiza utilizando el criterio denominado *Índice de Unión sobre la Intersección* (Intersection over Union, IoU). De acuerdo con [11], para dos cajas delimitadoras b_{t-1} y $b_{t,i}$ en los cuadros $t-1$ y t, respectivamente, el IoU se define como

$$\text{IoU}(b_{t-1}, b_{t,i}) = \frac{\text{Área}\left(b_{t-1} \cap b_{t,i}\right)}{\text{Área}\left(b_{t-1} \cup b_{t,i}\right)}.$$

Este índice cuantifica el grado de superposición entre dos cajas: un valor alto indica que ambas delimitan esencialmente la misma región en la imagen.

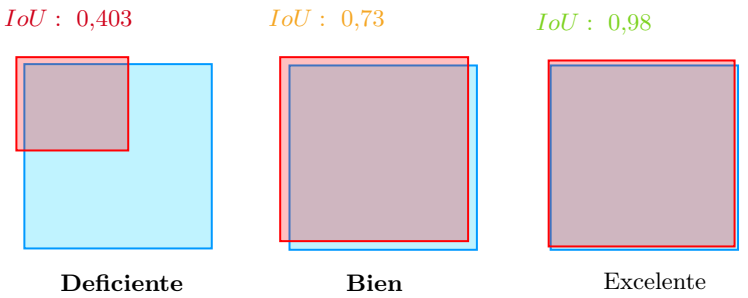


Figura 2. IoU obtenido entre dos bounding boxes, con distintos niveles de superposición

Para cada nuevo cuadro, se calcula el IoU entre la caja asociada al objeto seguido en el cuadro anterior y cada una de las detecciones presentes en el cuadro actual. A continuación, se selecciona la caja que maximiza la superposición:

$$b_t^* = \arg\max_{b_{t,i}} \text{IoU}(b_{t-1}, b_{t,i}).$$

Finalmente, se establece un umbral de decisión $\tau = 0{,}8$ por ejemplo, de tal forma que si

$$\text{IoU}(b_{t-1}, b_t^*) \geq \tau,$$

entonces la detección b_t^* se considera la continuación del mismo objeto en el cuadro t. En caso contrario, se descarta la asociación por no existir suficiente evidencia geométrica de continuidad [11].

Este método presenta la ventaja de ser simple y computacionalmente eficiente, basándose únicamente en la información geométrica contenida en los *bounding boxes*. Sin embargo, su desempeño se ve afectado por oclusiones, cambios bruscos en la forma de las cajas, o inconsistencias en las detecciones de la red neuronal.

8.2. Seguimiento Basado en Filtro de Kalman. Además del método geométrico basado en IoU, es posible emplear un enfoque más robusto que combine un modelo dinámico explícito con un algoritmo de asignación óptima. Este enfoque es el utilizado en sistemas modernos de seguimiento como SORT y DeepSORT, los cuales integran un Filtro de Kalman con el Algoritmo Húngaro.

En este caso, el movimiento de cada objeto se modela mediante un estado latente x_t que incluye su posición y velocidad en el plano de la imagen. El Filtro de Kalman permite predecir la evolución del estado entre cuadros consecutivos mediante el modelo lineal

$$x_t = F_{t-1}x_{t-1} + w_{t-1} \tag{8.1}$$

$$y_t = H_t x_t + v_t \tag{8.2}$$

donde $\mathbf{x}_k$ y $\mathbf{y}_k$ son los vectores de estado y de medición en el instante k. Las matrices $\mathbf{F}_k$ y $\mathbf{H}_k$ representan, respectivamente, la matriz de transición del sistema y la matriz de observación [9].

Los términos $\mathbf{w}_k$ y $\mathbf{v}_k$ corresponden al ruido del proceso y al ruido de medición. Se asume que ambos son independientes, de media cero, ruido blanco Gaussiano, con matrices de covarianza $\mathbf{Q}_k$ y $\mathbf{R}_k$, respectivamente $\mathbf{w}_k \sim \mathcal{N}(0,\, \mathbf{Q}_k)$ y $\mathbf{v}_k \sim \mathcal{N}(0,\, \mathbf{R}_k)$ [9]

Cuando el detector proporciona una medición z_t, típicamente asociada a las coordenadas del centro del *bounding box*, el filtro actualiza su estimación utilizando

$$z_t = Hx_t + v_t,$$

donde H es la matriz de observación y v_t es el ruido de medición con covarianza R.

Luego de obtener las predicciones del Filtro de Kalman para cada objeto rastreo activo, es necesario asociarlas con las detecciones provenientes del modelo $\mathcal{F}_\theta$. Para ello se construye una matriz de costos C, cuyas entradas representan la discrepancia entre la predicción de cada objeto y cada detección nueva. Con esta matriz se resuelve un problema de asignación óptima utilizando el Algoritmo Húngaro, obteniendo así la correspondencia entre objetos y detecciones en el cuadro actual. [10]

Esta combinación de predicción dinámica y asignación óptima permite manejar oclusiones breves, detecciones perdidas y mediciones ruidosas, proporcionando un seguimiento mucho más estable que el método basado únicamente en IoU. En particular, la incorporación de un modelo de movimiento evita saltos abruptos en la trayectoria y permite mantener la identidad del objeto aun cuando su *bounding box* varíe significativamente entre cuadros consecutivos.

9. Estimación de datos a escala real

Para poder trasladar las coordenadas obtenidas en la imagen hacia un sistema de coordenadas métricas coherente con la carretera, es necesario describir de manera precisa la transformación geométrica que relaciona ambos planos. Este proceso se basa en una transformación proyectiva que modela cómo un plano tridimensional, al ser observado desde una cámara, se representa como una superficie bidimensional en la imagen. Dicho mapeo se describe mediante una matriz de perspectiva que captura la deformación generada por la proyección central de la cámara.

Según [1] ,la transformación proyectiva puede expresarse mediante el siguiente sistema, que relaciona las coordenadas del punto en la imagen (x, y) con las coordenadas en escala real (u, v) :

$$\begin{pmatrix} w'u \\ w'v \\ w' \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}.$$

El vector homogéneo $(w'u,\ w'v,\ w')$ incorpora la escala necesaria para preservar la información de la perspectiva. Para recuperar las coordenadas físicas es suficiente con normalizar

$$u = \frac{ax + by + c}{gx + hy + 1}, \qquad v = \frac{dx + ey + f}{gx + hy + 1}.$$

Los parámetros (a, b, c, d, e, f, g, h) contienen la información geométrica de la proyección, tales como rotación, escala, traslación y deformación proyectiva. Su determinación requiere comparar puntos de la imagen con sus equivalentes en la escala real.

9.1. Obtención de los parámetros proyectivos. Para estimar estos parámetros es necesario contar con cuatro puntos visibles en la carretera, cuyas coordenadas se conocen tanto en píxeles (x_i, y_i) como en la escala real (u_i, v_i). Cada correspondencia genera un sistema de dos ecuaciones [1]:

$$\begin{cases} ax_i + by_i + c - gx_iu_i - hy_iu_i = u_i, \\ dx_i + ey_i + f - gx_iv_i - hy_iv_i = v_i. \end{cases}$$

Al utilizar cuatro puntos, se obtienen ocho ecuaciones lineales independientes que permiten resolver el vector de parámetros

$$\alpha = (a, b, c, d, e, f, g, h)^T.$$

Este sistema puede resolverse mediante diversos métodos numéricos: eliminación gaussiana, factorización LU, descomposición QR o inversión matricial, siempre que la matriz asociada sea no singular. La singularidad puede producirse si los puntos elegidos están mal distribuidos sobre la carretera, demasiado cercanos entre sí o alineados de manera que no permitan reconstruir una perspectiva única [1].

El resultado de este proceso es una transformación proyectiva completamente definida que permite convertir coordenadas de la imagen en posiciones métricas directamente sobre la superficie de la carretera.

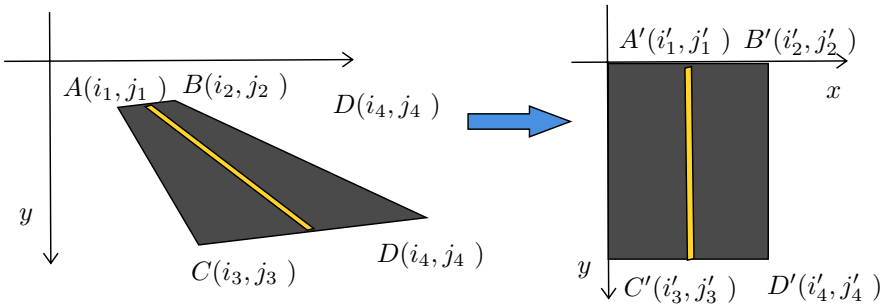


FIGURA 3. Visualización de la región de interés dependiendo de la perspectiva. La imagen a la izquierda corresponde a los puntos en pixeles y a la derecha sus equivalentes en escala real (metros).

9.2. Conversión de píxeles a coordenadas reales. De acuerdo con [1], una vez estimada la matriz proyectiva, cualquier punto de la imagen puede convertirse a su ubicación real aplicando

$$\begin{pmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{w} \end{pmatrix} = H \begin{pmatrix} x \\ y \\ 1 \end{pmatrix},$$

y normalizando

$$u = \frac{\tilde{u}}{\tilde{w}}, \qquad v = \frac{\tilde{v}}{\tilde{w}}.$$

De esta forma, las coordenadas registradas por el sistema de visión en unidades de píxeles se transforman en posiciones reales expresadas en metros, corregidas de los efectos de la perspectiva de la cámara. Cada punto de seguimiento del vehículo puede representarse ahora en un sistema de referencia físico directamente asociado a la carretera.

Todo esto se realiza para que la trayectoria del vehículo en la imagen se convierta en una curva sobre el plano real que pueda ser interpretable. Al trabajar en coordenadas homogéneas, la transformación proyectiva conserva la estructura geométrica de la escena, y la normalización garantiza que cada punto corresponda a una ubicación verdadera sobre la carretera.

10. Estimación de coeficientes de Regresión sobre la densidad vehicular

Una vez obtenida la velocidad de los vehículos a partir de la detección, seguimiento y proyección al plano real, el siguiente paso consiste en estimar la densidad vehicular observable en la escena. En este contexto, la densidad se interpreta como el número de vehículos presentes en un tramo específico de la carretera durante un intervalo de tiempo. A diferencia de la velocidad, que se asocia al movimiento individual de cada vehículo, la densidad es una magnitud colectiva que describe el grado de ocupación de la vía.

La relación entre velocidad y densidad vehicular ha sido estudiada extensamente como base para el análisis macroscópico del tráfico. Kerner [14] propone un esquema de identificación empírica del diagrama fundamental del tráfico y demuestra, a partir de datos observacionales, que dicha relación no es universal, lo que justifica la necesidad de estimaciones contextuales como las que se plantean en este trabajo.

Para cuantificar la densidad vehicular se adopta la formulación empírica clásica empleada en estudios macroscópicos de tráfico, en la cual la densidad se define como el número total de vehículos que ocupan un tramo de carretera de longitud determinada. Sea $L > 0$ la longitud fija de un tramo de carretera observado, y sea $\mathcal{I} = [t, t+\Delta t]$ un intervalo de tiempo de observación. Denotamos por $N(t)$ el número de vehículos contenidos en el segmento espacial de longitud L en el instante t. La densidad vehicular instantánea, entendida como una función escalar del tiempo, se define como

$$\rho(t) = \frac{N(t)}{L}, \quad t \in \mathcal{I}.$$

Esta expresión corresponde a la definición empírica adoptada en modelos macroscópicos de flujo vehicular, como los estudiados por Kerner [14]. En dicho marco, $\rho(t)$ representa una variable de estado cuya evolución describe el grado de ocupación de la vía en función del tiempo. Su estimación, junto con la velocidad media $v(t)$, permite caracterizar el estado dinámico del sistema vehicular en el tramo considerado.

Una vez teniendo las estimaciones de velocidad y densidad posibilita el estudio de su relación funcional, que en el contexto del análisis de tráfico se modela frecuentemente mediante una regresión entre $v(t)$ y $\rho(t)$. Esta relación, denotada típicamente por $v = f(\rho)$, nos permitirá realizar estimación de los coeficientes de un modelo de regresión.

11. Conclusiones

Este proyecto establece una estructura metodológica para abordar la estimación de velocidad vehicular y su relación funcional con la densidad del tráfico, integrando técnicas de visión por computadora con herramientas estadísticas. El enfoque contempla el uso de redes neuronales convolucionales (CNN) como mecanismo principal para la detección automática de vehículos a partir de secuencias de video, generando identificaciones cuadro a cuadro en el plano imagen.

Sobre estas detecciones se plantea la implementación de algoritmos de seguimiento, tanto geométricos como probabilísticos, con el objetivo de construir trayectorias temporales coherentes para cada vehículo. El seguimiento se concibe como un proceso de asociación entre detecciones sucesivas, y su formulación incluye técnicas como el Índice de Unión sobre Intersección (IoU), el Filtro de Kalman y el Algoritmo Húngaro. Estos métodos permiten resolver la correspondencia temporal bajo condiciones de oclusión, variabilidad geométrica o ruido en las detecciones.

Una vez definidas las trayectorias en el plano imagen, se proyectarán sobre un sistema de coordenadas métricas mediante una transformación homográfica calibrada con puntos de control en la escena. A partir de estas trayectorias físicas, se plantea calcular la velocidad vehicular mediante derivación numérica del desplazamiento, ajustada a escala real.

De forma complementaria, se define la densidad vehicular como una función escalar en el tiempo, calculada a partir del número de vehículos por unidad de longitud en una región espacial fija. Esta magnitud, junto con la velocidad, permite caracterizar el estado macroscópico del sistema vehicular.

La disponibilidad conjunta de las variables $v(t)$ y $\rho(t)$ permite formular una relación funcional que será modelada mediante regresión, con el objetivo de estimar coeficientes que describan empíricamente el comportamiento del flujo vehicular. Estos coeficientes podrán ser utilizados como insumos en modelos basados en ecuaciones diferenciales parciales.

En conjunto, la estructura planteada proporciona un marco metodológico para la generación de datos observacionales y su integración en modelos de dinámica vehicular.

Referencias

1. Grents, A., Varkentin, V., & Goryaev, N. (2020). Determining vehicle speed based on video using convolutional neural network. *Transportation Research Procedia, 50*, 192–200.
2. Ke, R., Kim, S., Li, Z., & Wang, Y. (2015). Motion-vector clustering for traffic speed detection from UAV video. *IEEE International Smart Cities Conference*, 1–5.
3. Karim, M. R., Dehghani, A., & Ram, A. G. (2010). Vehicle speed detection in video image sequences using CVS method. *International Journal of the Physical Sciences, 5*(17), 2555–2563.
4. Makwana, B., & Goel, P. (2013). Moving vehicle detection and speed measurement in video sequence. *International Journal of Engineering Research & Technology, 2*(10), 3534–3537.
5. Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45.
6. Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*(1–2), 83–97.

7. Shepelev, V., Aliukov, S., Nikolskaya, K., & Shabiev, S. (2020). The capacity of the road network: data collection and statistical analysis of traffic characteristics. *Energies, 13*(7), 1765.
8. V. Garg, S. Kumar, and S. Chandra, "A Mathematical Model for Video-Based Traffic Parameter Estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3833–3842, 2019, doi: 10.1109/TITS.2019.2896847.
9. S. S. Teoh and T. Bräunl, *A Reliability Point and Kalman Filter-based Vehicle Tracking Technique*.
10. A. Bewley, Z. Ge, L. Ott, F. T. Ramos, and B. Upcroft, *Simple Online and Realtime Tracking*, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016.
11. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You Only Look Once: Unified, Real-Time Object Detection*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2016.
12. C. C. Aggarwal, *Neural Networks and Deep Learning*, Springer, 2018.
13. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
14. B. S. Kerner, "Fundamental diagram of traffic flow: New identification scheme and further evidence from empirical data," *Physical Review E*, vol. 68, no. 3, p. 036130, 2003. doi: 10.1103/PhysRevE.68.036130.

Maestría en Matemática, Universidad Nacional Autónoma de Honduras.
Dirección actual: Tegucigalpa, Honduras
Dirección de correo electrónico: `ruth.moreno@unah.hn`

UNA REVISIÓN BIBLIOGRÁFICA SOBRE RANDOM FOREST (BOSQUES ALEATORIOS)

ALLAN MAURICIO CÓRDOVA MARTÍNEZ

Resumen. En este artículo se presenta una revisión bibliográfica sobre Random Forest, un algoritmo de aprendizaje automático para clasificación y regresión. Random Forest combina múltiples árboles de decisión para mejorar precisión, estabilidad y generalización, entrenando cada árbol con una muestra aleatoria y un subconjunto de variables, lo que reduce la correlación entre ellos y previene el sobreajuste. Las predicciones se obtienen por votación en clasificación o por promedio en regresión, haciéndolo robusto frente a ruido y datos incompletos. Además, permite identificar variables relevantes, consolidándose como una técnica versátil y confiable en contextos complejos. Sus aplicaciones incluyen medicina y bioinformática, finanzas, marketing, ingeniería, medio ambiente y ciencias sociales.

Abstract. This article presents a bibliographic review of Random Forest, a machine learning algorithm for classification and regression. Random Forest combines multiple decision trees to improve accuracy, stability, and generalization, training each tree with a random sample of the data and a subset of variables, which reduces correlation among trees and prevents overfitting. Predictions are obtained by voting in classification or by averaging in regression, making the method robust to noise and incomplete data. Additionally, it allows the identification of relevant variables, establishing it as a versatile and reliable technique in complex contexts. Its applications include medicine and bioinformatics, finance, marketing, engineering, environmental sciences, and social sciences.

1. Introducción

En la actualidad, el volumen y la complejidad de los datos generados en los ámbitos social, económico y ambiental exigen herramientas analíticas capaces de procesar información masiva y extraer patrones relevantes de manera eficiente y confiable. En este contexto, los **bosques aleatorios** (*Random Forests*) [4] se consolidan como una de las metodologías más robustas del aprendizaje estadístico moderno, debido a su capacidad para combinar múltiples árboles de decisión y generar modelos de alta precisión, estabilidad y generalización.

El algoritmo *Random Forest*, propuesto por Leo Breiman en 2001 [4], pertenece a la familia de métodos de aprendizaje en conjunto (*ensemble learning*), los cuales buscan mejorar el desempeño de los modelos individuales mediante la agregación de múltiples clasificadores o regresores. A diferencia de los árboles de decisión tradicionales, los bosques aleatorios introducen aleatoriedad tanto en la selección de los datos como de las variables, reduciendo la correlación entre los árboles y, en consecuencia, la varianza del modelo final. Esta característica los hace especialmente

Fecha: Diciembre 2025.

Palabras y frases clave. Random Forest, aprendizaje automático, clasificación y regresión, árboles de decisión.

útiles en escenarios donde existen relaciones no lineales, alta dimensionalidad o presencia de ruido en los datos.

El presente trabajo tiene como **objetivo principal** analizar los fundamentos teóricos y metodológicos del algoritmo *Random Forest*, destacando su relevancia en el análisis estadístico y su potencial aplicación en problemáticas nacionales. En particular, se busca describir su funcionamiento, las medidas de pureza utilizadas en la construcción de árboles, los criterios de importancia de variables y otras definiciones muy importantes obtenidas a partir de revisiones bibliográficas.

Desde una perspectiva aplicada, el estudio de los bosques aleatorios adquiere gran importancia en el contexto hondureño, ya que su implementación en investigaciones sociales, económicas y ambientales permite abordar con mayor rigor científico desafíos prioritarios del país, tales como la medición de la pobreza, la planificación territorial, la seguridad alimentaria y el análisis educativo. Asimismo, el fortalecimiento de las capacidades nacionales en ciencia de datos e inteligencia artificial contribuye al desarrollo de competencias técnicas avanzadas y al impulso de la investigación científica en la Universidad Nacional Autónoma de Honduras (UNAH).

2. Justificación

Random Forest (Bosques Aleatorios) [4] es un algoritmo de conjunto que combina múltiples árboles de decisión para generar predicciones más precisas y robustas. Su capacidad para manejar grandes volúmenes de datos, identificar patrones complejos y realizar predicciones confiables lo convierte en una herramienta invaluable para el análisis de fenómenos sociales, económicos y ambientales que afectan al país.

En el escenario hondureño, Bosques Aleatorios puede afrontar diversos retos nacionales fundamentales:

1. Análisis de Pobreza y Desigualdad: Mediante el procesamiento de datos socioeconómicos, el algoritmo puede identificar factores determinantes de la pobreza, predecir áreas de riesgo social y evaluar el impacto de políticas públicas, contribuyendo a la focalización eficaz de programas sociales.

2. Planificación Territorial y Desarrollo: La técnica permite examinar patrones de expansión urbana, mejorar la distribución de la infraestructura y prever necesidades de servicios básicos en diferentes regiones del territorio nacional

3. Seguridad Alimentaria: A través del análisis de variables climáticas, socioeconómicas y productivas, Bosques Aleatorios puede predecir zonas o áreas de vulnerabilidad alimentaria y optimizar estrategias de seguridad nutricional.

4. Educación y Desarrollo Humano: El algoritmo puede identificar factores de deserción escolar, predecir rendimiento académico y optimizar la asignación de recursos educativos, contribuyendo al fortalecimiento del capital humano nacional.

La implementación de Bosques Aleatorios en investigaciones nacionales fortalece las capacidades científicas del país en el área de ciencia de datos e inteligencia artificial, campos emergentes de gran importancia a nivel global. Esta técnica representa una oportunidad para que Honduras desarrolle competencias técnicas avanzadas, genere conocimiento aplicado y contribuya a la producción científica regional en metodologías cuantitativas.

Por las razones ya mencionadas, la investigación en Bosques Aleatorios se sitúa dentro del **Eje de Investigación : Población y Condiciones de Vida**, específicamente en el **tema prioritario: b) Cultura, ciencia y educación** de las líneas de investigación de la Universidad Nacional Autónoma de Honduras (UNAH) , ya que representa una contribución al desarrollo científico-tecnológico nacional que impulsa las capacidades investigativas del país y crea herramientas aplicables para la solución de diversas problemáticas nacionales primordiales. Tambien Random Forest (Bosques Aleatorios) pertenece a la familia de modelos de aprendizaje supervisado, más específicamente dentro de los modelos de ensamble (ensemble methods), y aún más concretamente, de los métodos de bagging (bootstrap aggregating).Se encuentra en la línea de investigación de **Modelación Matemática (Maestría en Matemática con Orientación en Ingeniería Matématica de la UNAH)**.

3. Antecedentes

El tema de los **Random Forest** (bosques aleatorios) ha evolucionado a partir de los árboles de decisión y los métodos de agregación estadística.Un resumen histórico sobre su desarrollo, los principales autores involucrados y los aportes más recientes a la teoría se muestran a continuación:

3.1. Orígenes y desarrollo histórico. Los primeros trabajos que dieron origen a los bosques aleatorios se encuentran en los estudios sobre árboles de decisión, particularmente el método **CART (Classification And Regression Trees)** propuesto por Breiman, Friedman, Olshen y Stone (1984). Este enfoque estableció las bases para construir modelos interpretables, aunque con alta varianza y sensibilidad a los datos.

Posteriormente, **Leo Breiman** propuso en 1996 el método *Bagging (Bootstrap Aggregating)* [1], que consistía en generar múltiples conjuntos de entrenamiento mediante remuestreo con reemplazo, entrenar varios modelos y combinar sus predicciones mediante promedio o voto mayoritario, reduciendo así la varianza de los árboles individuales.

De forma paralela, **Tin Kam Ho** introdujo en 1995 el *Random Subspace Method* [2], que sugiere entrenar clasificadores en subespacios aleatorios de las características disponibles, reduciendo la correlación entre los árboles. Más adelante, **Amit y Geman** (1997) extendieron la idea de aleatorizar tanto la selección de variables como los puntos de corte en los nodos [3].

3.2. Formalización de Random Forests. En 2001, **Leo Breiman** consolidó todas estas ideas en su influyente artículo *Random Forests* [4], en el que formalizó un método de *ensemble learning* que combina árboles de decisión entrenados sobre muestras bootstrap y subconjuntos aleatorios de variables. Además, introdujo la estimación del error de generalización mediante observaciones fuera de la bolsa (*out-of-bag, OOB*) y la medición de la importancia de las variables mediante permutación.

Breiman también estableció fundamentos teóricos que relacionan el error de generalización con la fuerza promedio de los clasificadores individuales y la correlación media entre ellos, mostrando que un equilibrio entre ambos factores produce modelos más robustos y precisos [4].

3.3. Desarrollos teóricos recientes. Durante las dos últimas décadas, se han producido importantes avances en el análisis teórico y en las variantes del método. **Biau, Devroye y Lugosi** (2008) demostraron resultados de consistencia para bosques aleatorios y variantes simplificadas, estableciendo una base teórica sólida para su comportamiento asintótico [5].

Recientemente, se han propuesto múltiples extensiones y mejoras del enfoque original de Breiman:

- **Chen, Wang y Lei** (2024) presentaron el *Data-driven Multinomial Random Forest*, una variante con consistencia fuerte y formulación multinomial más estable [6].
- **Dorador** (2024) propuso estrategias de *Forest Pruning* para eliminar árboles redundantes sin pérdida de precisión [7].
- **Ignatenko, Surkov y Koltcov** (2024) desarrollaron *Random Forests* con criterios de información basados en entropías paramétricas, mejorando la calidad de las divisiones [8].
- **Ren, Zhu, Bai y Zhang** (2024) introdujeron el modelo *Intuitionistic Fuzzy Random Forest*, que combina conjuntos difusos con aprendizaje de bosques [9].
- **Konstantinov, Utkin, Lukashin y Muliukha** (2023) propusieron los *Neural Attention Forests*, que integran mecanismos de atención derivados de redes neuronales [10].

Estos trabajos reflejan la tendencia actual hacia modelos combinados más eficientes, capaces de manejar incertidumbre y datos de alta dimensionalidad, manteniendo la esencia del enfoque propuesto por Breiman en 2001.

Cuadro 1. Evolución histórica y desafíos de Random Forest

Etapa / Autor	**Contribución principal**	**Desafíos actuales**
CART (1984) [11]	Base de los árboles de decisión	Alta varianza
Bagging (Breiman, 1996) [1]	Reducción de varianza	Falta de diversidad entre modelos
Random Subspace (Ho, 1995)	Selección aleatoria de variables	Sensibilidad a parámetros
Random Forest (Breiman, 2001) [4]	Combinación bagging + aleatoriedad	Interpretabilidad limitada
Avances recientes (2023–2024)	Consistencia fuerte, integración con redes neuronales y lógica difusa	Escalabilidad y eficiencia

3.4. Síntesis comparativa. En resumen, los bosques aleatorios constituyen una de las técnicas más exitosas y versátiles del aprendizaje estadístico moderno. Su robustez, precisión y facilidad de uso han llevado a su aplicación en prácticamente todas las áreas de la ciencia de datos, y continúan siendo objeto de investigación activa en teoría estadística y optimización computacional.

4. Conceptos Preliminares

Leo Breiman en su artículo *Random Forest* (2001) [4], compara resultados obtenidos con modelos ya antes establecidos como ser:

- **Adaboost (Adaptive Boosting)**: Propuesto por Freud y Shapire en 1996, es un algoritmo determinista que ajusta iterativamente los pesos del conjunto de entrenamiento. En cada iteración, las observaciones mal clasificadas reciben un mayor peso, de modo que los clasificadores posteriores se enfoquen en los errores cometidos por los anteriores. La predicción final es una combinación ponderada de todos los clasificadores entrenados, lo que reduce el sesgo y mejora la precisión.

- El método **Bagging (Bootstrap Aggregating)** es el antecesor de Random Forest y fue propuesto por Breiman en 1996 [1]. Es un método general de reducción de la varianza que se basa en el remuestreo *bootstrap* (con reemplazo) junto con un modelo de regresión o clasificación. Su procedimiento se puede resumir en los siguientes pasos:
  1. Se generan múltiples subconjuntos de entrenamiento mediante remuestreo con reemplazo del conjunto original.
  2. Se entrena un modelo (por ejemplo, un árbol de decisión) en cada subconjunto.
  3. Para predecir una nueva observación, se promedian (en regresión) o se votan (en clasificación) las predicciones de los distintos modelos.

  Este proceso produce un estimador más estable, especialmente útil para modelos con alta varianza como los árboles de decisión. El Bagging es una técnica fundamental en el aprendizaje en conjunto (*ensemble learning*), y sentó las bases teóricas sobre las que más tarde se construiría el método Random Forest.

- Los **árboles de decisión** son modelos jerárquicos de aprendizaje automático utilizados tanto para tareas de clasificación como de regresión. Su estructura divide recursivamente el espacio de los predictores en regiones homogéneas respecto a la variable respuesta. Algunas definiciones basicas son:
  1. **Nodo raíz:** punto inicial que contiene el conjunto completo de datos.
  2. **Nodos internos:** condiciones o preguntas basadas en características de los datos.
  3. **Ramas:** conexiones entre nodos que representan las respuestas a dichas condiciones.
  4. **Hojas:** representan las decisiones o predicciones finales.

  Una de las principales ventajas de los árboles de decisión es su interpretabilidad, ya que pueden visualizarse fácilmente y las reglas de decisión se expresan de forma lógica y transparente. No obstante, un solo árbol tiende a tener alta varianza: pequeños cambios en los datos pueden alterar significativamente la estructura del árbol.

En síntesis, las ideas de Bagging, selección aleatoria de variables y el método CART se unifican en el algoritmo de Random Forest, pero con una gran mejora en robustez y capacidad de generalización.

5. Estimación mediante Bosques Aleatorios

El término "bosques aleatorios" (*random forests*) tiene diferentes interpretaciones según el contexto. Algunos investigadores lo utilizan como un término general para cualquier método que combine múltiples árboles de decisión con componentes aleatorios, sin importar la técnica específica de construcción. Otros autores lo reservan exclusivamente para el algoritmo desarrollado por Breiman en 2001 [4]. En este documento, adoptaremos principalmente esta segunda perspectiva.

Los bosques aleatorios son lo suficientemente flexibles para resolver dos tipos de problemas: clasificación supervisada (asignar categorías) y regresión (predecir valores numéricos). Para facilitar la comprensión inicial, se enfocara en problemas de regresión y posteriormente se revisara brevemente el caso de clasificación. El objetivo es presentar el algoritmo de manera clara y matemáticamente rigurosa.

El contexto general es el de la regresión no paramétrica. Observamos una variable de entrada aleatoria $X \in \mathcal{X} \subset \mathbb{R}^p$ y queremos predecir una respuesta numérica aleatoria $Y \in \mathbb{R}$ estimando la función de regresión:

$$m(x) = \mathbb{E}[Y \mid X = x].$$

Para lograr esto, disponemos de un conjunto de entrenamiento:

$$\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$$

compuesto por variables aleatorias independientes con la misma distribución que el par prototipo (X, Y). La meta u objetivo es utilizar estos datos para construir un estimador $m_n : \mathcal{X} \to \mathbb{R}$ que aproxime la función m.

Decimos que el estimador m_n es *consistente en error cuadrático medio* si:

$$\mathbb{E}[(m_n(X) - m(X))^2] \to 0 \quad \text{cuando} \quad n \to \infty,$$

donde la esperanza considera tanto la aleatoriedad de X como la de la muestra $\mathcal{D}_n$.

Un bosque aleatorio es esencialmente una colección de M árboles de regresión, cada uno construido con cierta aleatoriedad. Para el árbol número j en esta colección, el valor predicho en un punto x se denota:

$$m_n(x; \Theta_j, \mathcal{D}_n),$$

donde $\Theta_1, \ldots, \Theta_M$ son variables aleatorias independientes e idénticamente distribuidas.

La variable aleatoria Θ es independiente del conjunto de entrenamiento $\mathcal{D}_n$ y se utiliza para introducir dos tipos de aleatoriedad: primero, para remuestrear (seleccionar aleatoriamente un subconjunto de) los datos antes de construir cada árbol; segundo, para seleccionar qué variables considerar al hacer cada división en el árbol.

Matemáticamente, cada árbol individual predice mediante:

$$m_n(x; \Theta_j, \mathcal{D}_n) = \sum_{i \in \mathcal{D}_n^*(\Theta_j)} \frac{\mathbf{1}_{\{X_i \in A_n(x; \Theta_j, \mathcal{D}_n)\}} Y_i}{N_n(x; \Theta_j, \mathcal{D}_n)},$$

donde:

- $\mathcal{D}_n^*(\Theta_j)$ representa el subconjunto de datos seleccionados para construir este árbol (la muestra remuestreada),
- $A_n(x; \Theta_j, \mathcal{D}_n)$ es la región o celda terminal del árbol que contiene al punto x, y
- $N_n(x; \Theta_j, \mathcal{D}_n)$ es la cantidad de puntos (de los seleccionados) que caen dentro de esa celda.

En términos simples, cada árbol particiona el espacio de características en regiones (como dividir un mapa en zonas), y para predecir en un punto nuevo x, encuentra en qué región cae y promedia los valores Y de los puntos de entrenamiento en esa región.

La predicción final del bosque con M árboles se obtiene promediando:

$$m_{M,n}(x;\Theta_1,\ldots,\Theta_M;D_n) = \frac{1}{M}\sum_{j=1}^{M} m_n(x;\Theta_j;D_n). \tag{1}$$

En la implementación estándar de **R** (paquete `randomForest`) [60], el valor predeterminado es `ntree = 500`, es decir, se construyen 500 árboles. Como podemos elegir M tan grande como queramos (limitado solo por recursos computacionales), tiene sentido desde el punto de vista teórico considerar el límite cuando M tiende a infinito:

$$m_{\infty,n}(x;D_n) = \mathbb{E}_\Theta[m_n(x;\Theta;D_n)].$$

Aquí, $\mathbb{E}_\Theta$ denota el valor esperado respecto a la aleatoriedad de Θ, dado el conjunto de datos D_n. La ley de los grandes números justifica esta operación, estableciendo que:

$$\lim_{M\to\infty} m_{M,n}(x;\Theta_1,\ldots,\Theta_M;D_n) = m_{\infty,n}(x;D_n)$$

casi seguramente, dado D_n.

Para simplificar la notación se escribirá simplemente $m_{\infty,n}(x)$ en lugar de $m_{\infty,n}(x;D_n)$.

6. Descripción del Algoritmo

El procedimiento para construir un bosque aleatorio con M árboles es el siguiente:

Paso 1: Remuestreo previo. Antes de construir cada árbol, se seleccionan aleatoriamente a_n observaciones del conjunto original, ya sea con o sin reemplazo. Únicamente estas a_n observaciones se utilizarán para construir y hacer predicciones con ese árbol particular.

Paso 2: Construcción del árbol mediante divisiones aleatorias. En cada nodo del árbol, el algoritmo no considera todas las p variables disponibles, sino que selecciona aleatoriamente un subconjunto de `mtry` variables. Entre estas variables seleccionadas, elige la que produce la mejor división según el criterio CART .

Paso 3: Criterio de parada. La construcción de cada árbol continúa hasta que cada nodo terminal (hoja) contiene menos de `nodesize` observaciones.

Paso 4: Predicción individual. Para un nuevo punto x, cada árbol predice promediando los valores Y_i de las observaciones cuyos X_i caen en la misma celda terminal que x.

Paso 5: Agregación. La predicción final se obtiene promediando las predicciones de todos los árboles.

El Algoritmo 1 describe formalmente este proceso.

Algoritmo 1. Construcción y predicción de un bosque aleatorio.

1. **Entrada:** Datos $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, número de árboles M, parámetros a_n, `mtry`, `nodesize`.
2. **Para cada árbol** $j = 1,\ldots,M$:
   a) Seleccionar aleatoriamente a_n observaciones de $\mathcal{D}_n$ (con o sin reemplazo).

b) Construir un árbol de regresión:
   - En cada nodo, seleccionar aleatoriamente `mtry` variables de las p disponibles.
   - Encontrar la mejor división entre estas `mtry` variables usando el criterio CART.
   - Continuar dividiendo hasta que cada nodo terminal tenga menos de `nodesize` observaciones.

c) Almacenar el árbol resultante $m_n(x; \Theta_j, \mathcal{D}_n)$.

3. **Salida:** Predicción final mediante promedio: $m_{M,n}(x; \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(x; \Theta_j, \mathcal{D}_n)$.

Aunque pueda parecer complejo inicialmente, el algoritmo se basa en ideas simples. Los tres parámetros clave son:

1. $a_n \in \{1, \ldots, n\}$: Tamaño de la submuestra para cada árbol;
2. `mtry` $\in \{1, \ldots, p\}$: Número de variables candidatas consideradas en cada división;
3. `nodesize` $\in \{1, \ldots, a_n\}$: Tamaño mínimo de nodo (criterio de parada).

En el modo de regresión de `randomForest` en R, los valores predeterminados son: `mtry` $= \lceil p/3 \rceil$, $a_n = n$ (usar todos los datos con reemplazo, es decir, bootstrap), y `nodesize` $= 5$.

Forma extendida del algoritmo:

Algoritmo 1 Valor predicho del bosque aleatorio de Breiman en x.

Entrada: Conjunto de entrenamiento $\mathcal{D}_n$, número de árboles $M > 0$, $a_n \in \{1, \ldots, n\}$, mtry $\in \{1, \ldots, p\}$, nodesize $\in \{1, \ldots, a_n\}$ y $x \in \mathcal{X}$.
Salida: Predicción del bosque aleatorio en x.

para $j = 1, \ldots, M$ **hacer**
  Seleccionar a_n puntos, con (o sin) reemplazo, uniformemente en $\mathcal{D}_n$. En los siguientes pasos, solo se usan estas a_n observaciones.
  Sea $\mathcal{P} = (\mathcal{X})$ la lista que contiene la celda asociada con la raíz del árbol.
  Sea $\mathcal{P}_{\text{final}} = \emptyset$ una lista vacía.
  mientras $\mathcal{P} \neq \emptyset$ **hacer**
    Sea A el primer elemento de $\mathcal{P}$.
    si A contiene menos de `nodesize` puntos o si todos los $X_i \in A$ son iguales **entonces**
      Eliminar la celda A de la lista $\mathcal{P}$.
      $\mathcal{P}_{\text{final}} \leftarrow$ Concatenar$(\mathcal{P}_{\text{final}}, A)$.
    si no
      Seleccionar uniformemente, sin reemplazo, un subconjunto $\mathcal{M}_{\text{try}} \subset \{1, \ldots, p\}$ de cardinalidad `mtry`.
      Seleccionar la mejor división en A optimizando el criterio de división CART a lo largo de las coordenadas en $\mathcal{M}_{\text{try}}$ *(ver texto para detalles)*.
      Dividir la celda A según la mejor división. Llamar A_L y A_R a las dos celdas resultantes.
      Eliminar la celda A de la lista $\mathcal{P}$.
      $\mathcal{P} \leftarrow$ Concatenar$(\mathcal{P}, A_L, A_R)$.
    fin si
  fin mientras
  Calcular el valor predicho $m_n(x; \Theta_j, \mathcal{D}_n)$ en x igual al promedio de los Y_i que caen en la celda de x en la partición $\mathcal{P}_{\text{final}}$.
fin para
Calcular la estimación del bosque aleatorio $m_{M,n}(x; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n)$ en el punto x de acuerdo con (2)

El criterio de división CART. [11]

Por simplicidad, consideremos primero un árbol sin remuestreo que usa todos los datos $\mathcal{D}_n$.

Sea A una celda (región) cualquiera, y sea $N_n(A)$ el número de puntos que caen en A. Una división potencial en A se define por un par (j, z), donde:

- j es el índice de una variable (coordenada) en $\{1, \ldots, p\}$
- z es el valor umbral de corte en esa coordenada

Denotemos por $\mathcal{C}_A$ el conjunto de todas las divisiones posibles en A, y sea $X_i = (X_i^{(1)}, \ldots, X_i^{(p)})$.

Para cualquier división candidata $(j, z) \in \mathcal{C}_A$, el criterio CART de regresión mide la reducción en varianza lograda por la división:

(6.1)

$$L_{\text{reg},n}(j,z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2 \, \mathbf{1}_{\{X_i \in A\}} - \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \, \mathbf{1}_{\{X_i^{(j)} < z\}} - \bar{Y}_{A_R} \, \mathbf{1}_{\{X_i^{(j)} \geq z\}} \right)^2 \mathbf{1}_{\{X_i \in A\}},$$

donde:

$$A_L = \{x \in A : x^{(j)} < z\}, \qquad A_R = \{x \in A : x^{(j)} \geq z\},$$

y $\bar{Y}_A$, $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$ representan el promedio de los valores Y_i en las regiones A, A_L y A_R respectivamente. Por convención, si ningún punto cae en una región, su promedio se define como 0.

La mejor división $(j_n^\star, z_n^\star)$ para la celda A se encuentra maximizando:

$$(j_n^\star, z_n^\star) \in \arg \max_{j \in \texttt{Mtry}, (j,z) \in \mathcal{C}_A} L_{\text{reg},n}(j,z).$$

En caso de empate, se elige el punto medio entre dos observaciones consecutivas. Este procedimiento se adapta naturalmente al caso con remuestreo, optimizando sobre las a_n observaciones seleccionadas en lugar de todo $\mathcal{D}_n$.

En resumen, el algoritmo en cada nodo: (1) selecciona aleatoriamente `mtry` coordenadas, (2) evalúa el criterio (2) para todas las divisiones posibles en esas direcciones, y (3) selecciona la mejor. Este criterio, introducido por Breiman et al. (1984) [27] en el algoritmo CART, mide esencialmente cuánto disminuye la varianza al realizar una división.

Existen tres diferencias fundamentales entre CART [11] tradicional y los bosques aleatorios de Breiman:

1. **Selección aleatoria de variables:** En bosques aleatorios, solo se consideran `mtry` variables seleccionadas aleatoriamente en cada división, no todas las p variables.
2. **Sin poda:** Los árboles individuales crecen completamente hasta que cada nodo terminal contiene como máximo `nodesize` observaciones (o todos los puntos son idénticos).
3. **Remuestreo:** Cada árbol se construye con una submuestra de a_n observaciones. Cuando $a_n = n$ con reemplazo, tenemos el modo *bootstrap*; cuando $a_n < n$, tenemos *submuestreo*.

6.1. Clasificación Supervisada. Para mayor claridad, se enfocará en clasificación binaria, aunque los bosques aleatorios pueden manejar naturalmente problemas

multiclase. En este contexto, la variable de respuesta Y toma valores en $\{0,1\}$ y el objetivo es predecir Y dado X.

Un *clasificador* m_n es una función medible que intenta estimar la etiqueta Y a partir de X y los datos $\mathcal{D}_n$. Se considera *consistente* si su probabilidad de error converge a la del clasificador óptimo de Bayes:

$$L(m_n) = \mathbb{P}[m_n(X) \neq Y] \xrightarrow[n\to\infty]{} L^\star,$$

donde $L^\star$ es el error del clasificador de Bayes:

$$m^\star(x) = \begin{cases} 1, & \text{si } \mathbb{P}[Y=1 \mid X=x] > \mathbb{P}[Y=0 \mid X=x], \\ 0, & \text{en otro caso.} \end{cases}$$

En clasificación, el bosque aleatorio realiza un **voto por mayoría** entre los árboles:

$$m_{M,n}(x;\Theta_1,\ldots,\Theta_M,\mathcal{D}_n) = \begin{cases} 1, & \text{si } \dfrac{1}{M}\displaystyle\sum_{j=1}^{M} m_n(x;\Theta_j,\mathcal{D}_n) > \dfrac{1}{2}, \\ 0, & \text{en otro caso.} \end{cases}$$

Cada árbol individual clasifica usando voto mayoritario en su celda terminal:

$$m_n(x;\Theta_j,\mathcal{D}_n) = \begin{cases} 1, & \text{si hay más puntos de clase 1 que de clase 0 en la celda de } x, \\ 0, & \text{en otro caso.} \end{cases}$$

Más formalmente, para una región A:

$$m_n(x;\Theta_j,\mathcal{D}_n) = \begin{cases} 1, & \text{si } \displaystyle\sum_{i\in\mathcal{D}_n^*(\Theta)} \mathbf{1}_{\{X_i\in A,\, Y_i=1\}} > \sum_{i\in\mathcal{D}_n^*(\Theta)} \mathbf{1}_{\{X_i\in A,\, Y_i=0\}}, \quad x\in A, \\ 0, & \text{en otro caso.} \end{cases}$$

El criterio CART para clasificación se modifica para medir la pureza de nodos. Para una celda A, sean $p_{0,n}(A)$ y $p_{1,n}(A)$ las proporciones empíricas de las clases 0 y 1. El criterio de división es:

(6.2)

$$L_{\text{class},n}(j,z) = p_{0,n}(A)p_{1,n}(A) - \frac{N_n(A_L)}{N_n(A)}\,p_{0,n}(A_L)p_{1,n}(A_L) - \frac{N_n(A_R)}{N_n(A)}\,p_{0,n}(A_R)p_{1,n}(A_R),$$

Este criterio se basa en el **índice de impureza de Gini** $2p_{0,n}(A)p_{1,n}(A)$, que mide qué tan mezcladas están las clases en un nodo. Un nodo puro (todos de la misma clase) tiene Gini $= 0$, mientras que un nodo con 50 % tiene el valor máximo.

Para clasificación binaria ($Y \in \{0,1\}$), optimizar $L_{\text{class},n}$ es equivalente a optimizar $L_{\text{reg},n}$, por lo que los árboles resultantes son idénticos. La diferencia está en la predicción: clasificación usa voto mayoritario, regresión usa promedio.

Los valores recomendados para clasificación son: `nodesize` $= 1$ y `mtry` $= \sqrt{p}$.

6.2. Ajuste de Parámetros. La literatura sobre el ajuste óptimo de M, `mtry`, `nodesize` y a_n es limitada, con contribuciones notables de Díaz-Uriarte y de Andrés (2006) [41], Bernard et al. (2008) [20] y Genuer et al. (2010) [44]. El ajuste de parámetros puede ser computacionalmente costoso, especialmente con grandes conjuntos de datos.

Número de árboles (M): La varianza del bosque disminuye al aumentar M, y más árboles generalmente mejoran las predicciones. Importante: aumentar M *no causa sobreajuste*. Siguiendo a Breiman (2001) [4]:

$$\lim_{n\to\infty} E\left[(m_{M,n}(X;\theta_1,\ldots,\theta_M) - m(X))^2\right] = E\left[(m_{\infty,n}(X) - m(X))^2\right]$$

El costo computacional crece linealmente con M, por lo que se busca un equilibrio entre precisión y tiempo de cómputo. Díaz-Uriarte y de Andrés (2006) [41] argumentan que M es irrelevante (siempre que sea suficientemente grande) en problemas de microarrays. Genuer et al. (2010) [44] ofrecen una discusión exhaustiva sobre este parámetro.

Tamaño mínimo de nodo (`nodesize`): Los valores predeterminados (1 para clasificación, 5 para regresión) son generalmente buenos, aunque carecen de respaldo teórico riguroso. Kruppa et al. (2013) [54] discuten un algoritmo para ajustar este parámetro en clasificación.

Número de variables por división (`mtry`): Díaz-Uriarte y de Andrés (2006) [41] encuentran que `mtry` tiene poco impacto, aunque valores muy grandes pueden reducir el rendimiento. Genuer et al. (2010) [44] sugieren que el valor predeterminado a menudo es óptimo o demasiado pequeño, por lo que un enfoque conservador es usar `mtry` tan grande como sea computacionalmente factible.

Una ventaja importante: los parámetros se pueden ajustar usando la estimación **out-of-bag** (OOB), sin necesitar un conjunto de validación separado. Como cada árbol se construye con una muestra bootstrap, aproximadamente un tercio de las observaciones quedan fuera y pueden usarse como conjunto de prueba interno. El error OOB, calculado sobre estas observaciones excluidas, permite ajustar parámetros de manera eficiente.

7. Modelos Simplificados y Promedios Locales

7.1. Modelos Simplificados. A pesar de su uso extensivo, existe una brecha entre la teoría y la práctica de los bosques aleatorios. La complejidad del algoritmo completo dificulta el análisis matemático riguroso de sus propiedades fundamentales.

Como observó Denil et al. (2014) [39], esto ha creado una división en la literatura: los trabajos empíricos proponen extensiones elaboradas sin garantías teóricas claras, mientras que los trabajos teóricos se enfocan en versiones simplificadas donde el análisis es más manejable.

Un marco básico para el análisis teórico involucra **bosques aleatorios puros**, donde los árboles se construyen independientemente de los datos de entrenamiento D_n. El ejemplo más estudiado es el **bosque centrado**, que opera así:

1. Sin remuestreo (se usan todos los datos);
2. En cada nodo, se selecciona uniformemente una coordenada de $\{1,\ldots,p\}$;
3. Se divide en el centro de la celda a lo largo de esa coordenada.

Este proceso se repite recursivamente k veces, produciendo un árbol binario completo con 2^k hojas. El parámetro k controla el tamaño de las celdas finales y actúa como parámetro de suavizado: debe ser lo suficientemente grande para capturar variaciones locales, pero no tanto que impida el promediado efectivo.

Los **bosques uniformes** son una variante donde los cortes se realizan en posiciones aleatorias uniformes en lugar del centro, con análisis matemático similar.

Breiman (2004) [31], Biau et al. (2008) [23] y Scornet (2015a) [71] demostraron que los bosques centrados son consistentes (para clasificación y regresión) siempre que $k \to \infty$ y $\frac{n}{2^k} \to \infty$ simultáneamente. La demostración se basa en resultados generales de consistencia para árboles aleatorios [40].

Si X es uniforme en $[0,1]^p$, hay en promedio $\frac{n}{2^k}$ puntos por nodo terminal. La elección $k \approx \log n$ (árboles completamente crecidos) no satisface $\frac{n}{2^k} \to \infty$, revelando una limitación del análisis. Además, como no hay bagging, la consistencia proviene del árbol individual, no del ensamble.

Para tasas de convergencia, Breiman (2004) [31] y Biau (2012) [22] consideran variables $X^{(j)}$ independientes con función de regresión $m(x)$ que depende solo de un subconjunto S (Strong) de variables. Si la probabilidad de dividir según la variable j tiende a $1/|S|$ y m es Lipschitz, entonces:

$$\mathbb{E}\big[m_{\infty,n}(X) - m(X)\big]^2 = O\Big(n^{-0{,}75/(|S|\log 2 + 0{,}75)}\Big).$$

Esto muestra que la tasa depende solo de $|S|$ (variables relevantes), no de p (dimensión total), demostrando adaptación a esparsidad. Esta tasa es más rápida que la tasa estándar $n^{-2/(p+2)}$ cuando $|S| \leq \lfloor 0{,}54\, p \rfloor$.

Genuer (2012) [45] estudia **bosques puramente uniformes** (PURF) en una dimensión, demostrando consistencia y, bajo suposiciones Lipschitz, la tasa:

$$\mathbb{E}\big[m_{\infty,n}(X) - m(X)\big]^2 = O\big(n^{-2/3}\big),$$

que es minimax óptima para funciones Lipschitz [75, 76].

Biau (2012) [22] muestra que los bosques centrados reducen el error de estimación (a tasa lenta $1/\log n$) incluso con árboles completamente crecidos ($k \approx \log n$), un beneficio del promediado. Arlot y Genuer (2014) [16] demuestran que ciertos bosques también mejoran la tasa de error de aproximación comparado con árboles individuales.

7.2. Bosques, Vecinos y Kernels. Consideremos variables i.i.d. $X_1, \ldots, X_n$. En geometría aleatoria, X_i es un **vecino más cercano por capas** (LNN) de x si el hiperrectángulo definido por x y X_i no contiene ningún otro punto de datos. El número de LNN de x suele ser mayor que uno y depende de la configuración de los puntos.

Sorprendentemente, los bosques sin remuestreo están íntimamente relacionados con los LNN. Si cada hoja contiene exactamente un punto y no hay remuestreo, entonces la predicción del bosque en x es un promedio ponderado de los Y_i cuyos X_i son LNN de x:

$$m_{\infty,n}(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \tag{3}$$

donde $W_{ni}(x) = 0$ si X_i no es LNN de x y $\sum_{i=1}^{n} W_{ni}(x) = 1$.

Esta conexión fue señalada por Lin y Jeon (2006) [61], quienes demostraron que si X es uniforme en $[0,1]^p$ y el crecimiento es independiente de $Y_1, \ldots, Y_n$, entonces:

$$\mathbb{E}\big[m_{\infty,n}(X) - m(X)\big]^2 = O\left(\frac{1}{n_{\text{máx}} (\log n)^{p-1}}\right),$$

donde $n_{\text{máx}}$ es el número máximo de puntos en celdas terminales.

Desafortunadamente, los pesos exactos $(W_{n1}, \ldots, W_{nn})$ para el bosque de Breiman son desconocidos, y una teoría general de bosques en el marco LNN aún no existe.

Sin embargo, la ecuación (3) permite analizar bosques mediante **promediado local**. Para un bosque finito sin remuestreo:

$$m_{M,n}(x; \Theta_1, \ldots, \Theta_M) = \sum_{i=1}^{n} W_{ni}(x) Y_i,$$

donde:

$$W_{ni}(x) = \frac{1}{M} \sum_{j=1}^{M} \frac{\mathbf{1}_{\{X_i \in A_n(x;\Theta_j)\}}}{N_n(x; \Theta_j)}.$$

Los pesos son no negativos y suman uno. Las observaciones en celdas densamente pobladas contribuyen menos que aquellas en celdas menos pobladas, una característica importante cuando los bosques se construyen independientemente de los datos.

Al hacer $M \to \infty$, la estimación puede escribirse (aproximadamente) como:

$$m_{\infty,n}(x) \approx \frac{\sum_{i=1}^{n} Y_i K_n(X_i, x)}{\sum_{j=1}^{n} K_n(X_j, x)}, \tag{4}$$

donde:

$$K_n(x, z) = \mathbb{P}_\Theta[z \in A_n(x; \Theta)].$$

La función $K_n(\cdot, \cdot)$ se llama **kernel del bosque** y caracteriza la forma de las celdas. $K_n(x, z)$ es la probabilidad de que x y z caigan en la misma celda en un árbol aleatorio, sirviendo como medida de proximidad. Cada bosque tiene su propio kernel, pero el asociado a divisiones CART depende fuertemente de los datos y es difícil de analizar.

Nótese que K_n no necesariamente pertenece a la familia de kernels Nadaraya-Watson [66, 80], que tienen la forma homogénea:

$$K_h(x, z) = \frac{1}{h} K\left(\frac{x - z}{h}\right)$$

Por ejemplo, Scornet (2015b) [72] demostró que para un bosque centrado en $[0, 1]^p$ con parámetro k:

$$K_{n,k}(x, z) = \sum_{k_1, \ldots, k_p} \frac{k!}{k_1! \cdots k_p!} \left(\frac{1}{p}\right)^k \prod_{j=1}^{p} \mathbf{1}_{2^{k_j} x_j = 2^{k_j} z_j}$$

La conexión entre bosques y estimación por kernel fue mencionada por Breiman (2000a) [29] y desarrollada por Geurts et al. (2006) [46]. Arlot y Genuer (2014) [16] muestran que ciertos bosques simplificados pueden escribirse como estimadores kernel y proporcionan sus tasas de convergencia.

Davies y Ghahramani (2014) [35] incorporan kernels basados en bosques en algoritmos de procesos gaussianos, demostrando empíricamente que superan a kernels lineales y de base radial. Los kernels de bosques también pueden usarse como entrada para métodos kernel estándar como Análisis de Componentes Principales con Kernels y Máquinas de Vectores de Soporte.

8. Fundamentos teóricos y variantes de los bosques aleatorios de Breiman

Esta sección aborda el algoritmo original de Breiman (2001) [4]. Como la construcción depende de toda la muestra D_n, un análisis matemático completo es difícil. Para avanzar, se han investigado los mecanismos individuales por separado.

8.1. El Mecanismo de Remuestreo. El algoritmo de Breiman selecciona n veces de entre n puntos *con reemplazo* para cada árbol. Este procedimiento, que se remonta a Efron (1982) [42], se denomina **bootstrap**. Generar muchas muestras bootstrap y promediar los predictores se conoce como **bagging** (bootstrap-aggregating), propuesto por Breiman (1996) [28] para mejorar aprendices débiles o inestables.

Aunque el bootstrap es simple conceptualmente, su teoría es compleja. La distribución de la muestra bootstrap D_n^* difiere de la original D_n. Por ejemplo, si X tiene densidad y muestreamos con reemplazo, con probabilidad positiva al menos una observación se selecciona múltiples veces, creando puntos idénticos en D_n^*. Por tanto, D_n^* no puede ser absolutamente continua.

El papel del bootstrap en bosques aleatorios permanece poco comprendido. La mayoría de análisis lo reemplazan por **submuestreo**, donde cada árbol se construye con $a_n < n$ ejemplos elegidos *sin reemplazo* [78, 73]. Frecuentemente se asume que $a_n/n \to 0$, excluyendo el régimen bootstrap.

El análisis de **bosques medianos** [71] proporciona intuición sobre el submuestreo. Un bosque mediano es similar al centrado, pero:

- El corte se realiza en la mediana empírica (no el centro)
- La construcción continúa hasta que cada celda contiene exactamente una observación

Aunque cada árbol individual es generalmente inconsistente (el número de casos en hojas no crece con n), Scornet (2015a) [71] demuestra que si $a_n/n \to 0$, el bosque mediano es consistente.

La condición $a_n/n \to 0$ garantiza que:

- Cada observación (X_i, Y_i) se usa en el árbol j con probabilidad pequeña cuando n crece
- El punto x queda desconectado de (X_i, Y_i) en una gran proporción de árboles

Si esto no ocurriera, el valor predicho en x estaría excesivamente influenciado por pares individuales (X_i, Y_i), haciendo el ensamble inconsistente. El error de estimación es pequeño cuando la probabilidad máxima de conexión entre x y todas las observaciones es pequeña. Así, $a_n/n \to 0$ es una forma conveniente de controlar estas probabilidades, asegurando que las particiones sean suficientemente diversas.

Biau y Devroye (2010) [25] aplicaron bagging al vecino más cercano (1-NN). El estimador 1-NN estándar:

$$r_n(x) = Y_{(1)}(x),$$

donde $Y_{(1)}(x)$ corresponde al $X_{(1)}(x)$ más cercano a x, no es generalmente consistente.

Mediante **subbagging**, se transforma en consistente con submuestras suficientemente pequeñas. El predictor elemental r_{a_n} es la regla 1-NN aplicada a una submuestra aleatoria de tamaño a_n. El estimador submuestreado es:

$$r_n^*(x) = \mathbb{E}^*[r_{a_n}(x)],$$

donde $\mathbb{E}^*$ es la esperanza respecto al remuestreo, dado D_n.

Biau y Devroye (2010) [25] demostraron que r_n^* es universalmente consistente en media cuadrática (sin condiciones sobre la distribución de (X, Y)) siempre que $a_n \to \infty$ y $a_n/n \to 0$. La demostración se basa en que r_n^* es un estimador de promediado local con pesos: $W_{ni}(x) = \Pr[X_i$ es el vecino más cercano de x en una selección aleatoria de tamaño $a_n]$.

Biau et al. (2010) [24] demuestran además que r_n^* alcanza la tasa óptima de convergencia sobre clases Lipschitz, independientemente de si el remuestreo es con o sin reemplazo.

8.2. Divisiones de Decisión. El proceso de división por coordenadas es difícil de comprender porque utiliza tanto X_i como Y_i para decidir las divisiones.

Basándose en Bühlmann y Yu (2002) [32], Banerjee y McKeague (2007) [18] establecen una ley límite para la ubicación de divisiones en un modelo de regresión:

$$Y = m(X) + \varepsilon,$$

donde X es real y ε es ruido gaussiano independiente.

Supongamos que la distribución de (X, Y) es conocida, y sea d^* la división óptima que maximiza el criterio teórico CART. Los estimadores de regresión en los hijos son:

$$\beta_{\ell,n}^* = \mathbb{E}[Y \mid X \le d^*], \quad \beta_{r,n}^* = \mathbb{E}[Y \mid X > d^*].$$

Cuando la distribución es desconocida, se estiman empíricamente:

$$(\hat{\beta}_{\ell,n}, \hat{\beta}_{r,n}, \hat{d}_n) \in \arg\min_{\beta_\ell,\, \beta_r,\, d} \sum_{i=1}^{n} \left(Y_i - \beta_\ell \mathbf{1}_{\{X_i \le d\}} - \beta_r \mathbf{1}_{\{X_i > d\}}\right)^2.$$

Bajo condiciones de regularidad (densidad f de X y m continuamente diferenciables), Banerjee y McKeague (2007) [18] demuestran:

$$n^{1/3} \begin{pmatrix} \hat{\beta}_{\ell,n} - \beta_\ell^* \\ \hat{\beta}_{r,n} - \beta_r^* \\ \hat{d}_n - d^* \end{pmatrix} \xrightarrow{D} \begin{pmatrix} c_1 \\ c_2 \\ 1 \end{pmatrix} \arg\max_t \left(aW(t) - bt^2\right), \tag{5}$$

donde W es un movimiento browniano estándar bilateral y a, b son constantes positivas. Esta distribución límite permite construir intervalos de confianza para las divisiones CART.

Ishwaran (2013) [51] analiza la **Preferencia por Cortes Extremos** (End-Cut Preference, ECP) del criterio CART: las divisiones sobre variables no informativas tienden a ubicarse cerca de los bordes de la celda. Esta es una propiedad deseable porque:

- Con aleatorización, existe probabilidad positiva de que ninguna variable preseleccionada sea informativa
- Si el corte se realiza en el centro, el tamaño muestral se reduce drásticamente (factor de dos)
- Un corte cerca del borde maximiza el tamaño muestral del nodo, permitiendo recuperación en niveles posteriores

Ishwaran (2013) [51] argumenta que ECP puede ser beneficiosa incluso para variables informativas si la región correspondiente contiene poca señal.

Scornet et al. (2015) [73] demuestran que los bosques aleatorios, asintóticamente, realizan divisiones con alta probabilidad a lo largo de las variables informativas. Denotando por $j_{n,1}(X), \ldots, j_{n,k}(X)$ las primeras k direcciones de corte para la

celda de X, bajo condiciones de regularidad y una modificación donde todas las direcciones se preseleccionan:

$$\text{Con probabilidad } 1-\xi, \text{ para } n \text{ grande y todo } 1 \leq q \leq k: \quad j_{n,q}(X) \in \{1, \ldots, |S|\}.$$

Esto explica por qué los bosques se adaptan a esparsidad: seleccionan cortes principalmente sobre variables informativas, proyectando efectivamente los datos sobre el subespacio generado por esas variables.

Variantes del algoritmo:

- **Extra-Trees** [46]: Selecciona aleatoriamente puntos de corte y elige el que maximiza CART. Rendimiento similar con mayor eficiencia computacional.
- **PERT** Árboles de ajuste perfecto con divisiones aleatorias. Aunque los árboles individuales sobreajustan, el ensamble es consistente porque los clasificadores están casi no correlacionados.
- **Divisiones oblicuas** [4, 77]: Divisiones a lo largo de combinaciones lineales de características. Menze et al. (2011) [65] notan que las divisiones ortogonales producen superficies de decisión en forma de cajas, óptimas para algunos datos pero subóptimas para datos colineales.

Selección de variables ponderada: La selección uniforme inevitablemente incluye variables irrelevantes. Varias modificaciones proponen ponderación basada en datos:

- Kyrillidis y Zouzias (2014) [57]: Selección no uniforme de características en árboles de clasificación.
- **Enriched Random Forests** [15]: Muestreo ponderado favoreciendo características informativas.
- **Reinforcement Learning Trees** [85]: En cada nodo, construyen un bosque para determinar la variable con mayor mejora futura, no efecto marginal inmediato.

Corrección de sesgos: Las divisiones CART están sesgadas hacia covariables con muchas divisiones posibles [27, 74] o valores faltantes [52]. Hothorn et al. (2006) [48] proponen un procedimiento de dos pasos: (1) seleccionar la variable, (2) seleccionar la posición del corte.

Regularización:

- **Regularized Random Forest (RRF)** [36]: Penaliza la selección de una nueva característica cuando su ganancia es similar a características usadas previamente.
- **Guided RRF (GRRF)** [37]: Usa puntuaciones de importancia de un bosque ordinario para guiar la selección en RRF.
- Penalización convexa tipo Garrote [64]: Selecciona grupos funcionales de nodos para estimaciones más parsimoniosas.

Konukoglu y Ganz (2014) [53] abordan el control de tasa de falsos positivos, presentando una forma fundamentada de determinar umbrales para selección de características relevantes sin carga computacional adicional.

8.3. Consistencia, Normalidad Asintótica y Otros Resultados. Se ha demostrado matemáticamente muy poco sobre el procedimiento original de Breiman. Un resultado fundamental [4] muestra que el error del bosque es pequeño cuando el poder predictivo de cada árbol es bueno y la correlación entre errores de árboles es baja:

$$E_{X,Y}\left[Y - m_{\infty,n}(X)\right]^2 \leq \bar{\rho}\, E_{\Theta,X,Y}\left[Y - m_n(X;\Theta)\right]^2,$$

donde:

$$\bar{\rho} = \frac{E_{\Theta,\Theta'}[\rho(\Theta,\Theta')g(\Theta)g(\Theta')]}{E_{\Theta}[g(\Theta)]^2},$$

con Θ y Θ' independientes e idénticamente distribuidos,

$$\rho(\Theta,\Theta') = \mathrm{Corr}_{X,Y}(Y - m_n(X;\Theta),\, Y - m_n(X;\Theta')),$$

y

$$g(\Theta) = \sqrt{E_{X,Y}[Y - m_n(X;\Theta)]^2}.$$

Friedman et al. (2009) [43] descomponen la varianza del bosque como producto entre correlación de árboles y varianza de un árbol:

$$\mathrm{Var}[m_{\infty,n}(x)] = \rho(x)\,\sigma(x),$$

donde $\rho(x) = \mathrm{Corr}[m_n(x;\Theta), m_n(x;\Theta')]$ y $\sigma(x) = \mathrm{Var}[m_n(x;\Theta)]$.

Scornet (2015a) [71] establece una conexión entre bosques finitos e infinitos:

$$0 \leq E[m_{M,n}(X;\Theta_1,\ldots,\Theta_M) - m(X)]^2 - E[m_{\infty,n}(X) - m(X)]^2 \leq \frac{8}{M}\left(\|m\|_\infty^2 + \sigma^2(1 + 4\log n)\right).$$

Esta desigualdad proporciona una solución para elegir M: permite que el error del bosque finito se aproxime arbitrariamente al del infinito.

Normalidad asintótica: Reemplazando bootstrap por submuestreo y simplificando las divisiones, se han demostrado resultados de normalidad.

Wager (2014) [78] demuestra normalidad asintótica bajo las suposiciones:

1. Los cortes se distribuyen sobre todas las p direcciones y no separan una fracción pequeña de datos.
2. Se usan dos conjuntos de datos distintos: uno para construir el árbol y otro para estimar valores en hojas.

Además, el *jackknife infinitesimal* estima consistentemente la varianza del bosque [79].

Mentch y Hooker (2015) demuestran un resultado similar para bosques finitos, basándose en que la predicción no varía significativamente al modificar ligeramente una etiqueta. Si $a_n = o(\sqrt{n})$ y $\lim_{n\to\infty} n/M_n = 0$, entonces para x fijo:

$$\frac{\sqrt{n}\,(m_{M,n}(x;\Theta_1,\ldots,\Theta_M) - \mathbb{E}[m_{\infty,n}(x)])}{\sqrt{a_n^2\,\zeta_{1,a_n}}} \xrightarrow{D} \mathcal{N}(0,1)$$

donde $\mathcal{N}(0,1)$ es la distribución normal estándar y:

$$\zeta_{1,a_n} = \mathrm{Cov}\big(m_n(X_1,X_2,\ldots,X_{a_n};\Theta),\, m_n(X_1,X_2',\ldots,X_{a_n}';\Theta')\big).$$

Tanto Mentch y Hooker (2015) [**?**] como Wager et al. (2014) [79] proporcionan estimadores para la varianza ζ_{1,a_n}.

Scornet et al. (2015) [73] demostraron consistencia para modelos aditivos en la versión podada del bosque de Breiman. Desafortunadamente, la consistencia sin poda depende de una conjetura sobre CART que es difícil de verificar.

Resultado negativo [23]: Considere un ejemplo donde k es fijo, `mtry` $= 1$, y cada árbol minimiza la verdadera probabilidad de error. Sea X uniforme en $[0,1]^2 \cup [1,2]^2 \cup [2,3]^2$ con Y función de X ($L^* = 0$):

- $[0,1] \times [0,1]$: franjas verticales alternando $m(x) \in \{0,1\}$
- $[2,3] \times [2,3]$: franjas horizontales alternando
- $[1,2] \times [1,2]$: tablero de ajedrez 2×2

Ningún árbol cortará correctamente el rectángulo central, independientemente de las direcciones y profundidad. La probabilidad de error es al menos $1/6$. Esto ilustra que la consistencia de bosques construidos codiciosamente es delicada. Sin embargo, con el algoritmo original de Breiman (exactamente un punto por celda), se obtiene una regla consistente.

Nótese que m no es Lipschitz, una suposición de suavidad en la que se basan muchos resultados.

9. Importancia de Variables

9.1. Medidas de Importancia. Los bosques aleatorios ofrecen dos medidas para clasificar la importancia de variables:

1. Mean Decrease Impurity (MDI): Basada en la disminución total de impureza al dividir por cada variable, promediada sobre todos los árboles. Para la variable $X^{(j)}$:

$$\hat{\mathrm{MDI}}(X^{(j)}) = \frac{1}{M} \sum_{\ell=1}^{M} \sum_{t \in T_\ell, j = j^*_{n,t}} p_{n,t}\, L_{\mathrm{reg},n}\big(j^*_{n,t}, z^*_{n,t}\big),$$

donde:

- $p_{n,t}$ es la fracción de observaciones en el nodo t
- $\{T_\ell\}$ es la colección de árboles
- $(j^*_{n,t}, z^*_{n,t})$ es la división óptima en t

MDI calcula la disminución ponderada de impureza para divisiones según $X^{(j)}$ y promedia sobre árboles. Para clasificación, se reemplaza $L_{\mathrm{reg},n}$ por $L_{\mathrm{class},n}$.

2. Mean Decrease Accuracy (MDA): Basada en permutación de valores y estimación out-of-bag. Para $X^{(j)}$, se permutan aleatoriamente sus valores en observaciones OOB y se mide el incremento en error de predicción.

Sea $D_{\ell;n}$ el conjunto OOB del árbol ℓ y $D^j_{\ell;n}$ el mismo conjunto con valores de $X^{(j)}$ permutados. Entonces:

$$\mathrm{MDA}(X^{(j)}) = \frac{1}{M} \sum_{\ell=1}^{M} \left(R_n\big(m_n(\cdot;\Theta_\ell), D^j_{\ell;n}\big) - R_n\big(m_n(\cdot;\Theta_\ell), D_{\ell;n}\big) \right), \tag{6}$$

donde para $D = D_{\ell;n}$ o $D = D^j_{\ell;n}$:

$$R_n\big(m_n(\cdot;\Theta_\ell), D\big) = \frac{1}{|D|} \sum_{i:(X_i,Y_i)\in D} \big(Y_i - m_n(X_i;\Theta_\ell)\big)^2. \tag{7}$$

La versión poblacional de MDA es:

$$\mathrm{MDA}^*(X^{(j)}) = \mathbb{E}\Big[(Y - m_n(X'_j;\Theta))^2\Big] - \mathbb{E}\Big[(Y - m_n(X;\Theta))^2\Big],$$

donde $X'_j = (X^{(1)}, \ldots, X'^{(j)}, \ldots, X^{(p)})$ con $X'^{(j)}$ copia independiente de $X^{(j)}$.

Para clasificación, MDA satisface (6) y (7) con $Y_i \in \{0, 1\}$, representando R_n la proporción de puntos correctamente clasificados.

10. Algunas Extensiones

10.1. Bosques Ponderados. En el bosque de Breiman, la predicción final es el promedio simple de árboles. Una mejora natural es incorporar pesos a nivel de árbol para enfatizar los más precisos [81]. Bernard et al. (2012) [21] proponen guiar la construcción mediante remuestreo y aleatorización para que cada árbol complemente los existentes. El **Dynamic Random Forest (DRF)** resultante muestra mejora significativa en 20 conjuntos de datos reales.

10.2. Bosques en Línea. El bosque original es **offline**: recibe todos los datos inicialmente. Los algoritmos **online** no requieren el conjunto completo de una vez, apropiados para escenarios de streaming donde los datos se generan continuamente.

Extensiones online incluyen Saffari et al. (2009) [69], Denil et al. (2013) [38], y Lakshminarayanan et al. (2014) [58]. Los **Mondrian forests** [58] se construyen online con desempeño competitivo y mayor velocidad.

Una dificultad importante es decidir cuándo hay suficientes datos para dividir una celda. Yi et al. (2012) [84] proponen **Information Forests**, que difieren la clasificación hasta que una medida de confianza sea suficientemente alta, dividiendo datos para maximizar esta medida. Biau y Devroye (2013) [26] proporcionan teoría relacionada con estos árboles codiciosos.

10.3. Bosques de Supervivencia. El análisis de supervivencia estudia el tiempo hasta que ocurren eventos, frecuentemente con datos censurados a la derecha. Enfoques paramétricos como hazards proporcionales fallan en modelar efectos no lineales.

Ishwaran et al. (2008) [49] extendieron bosques al contexto de supervivencia con **Random Survival Forests (RSF)**, probando consistencia para variables categóricas. Yang et al. (2010) [83] demostraron que incorporar funciones kernel en RSF (algoritmo KIRSF) mejora resultados en muchas situaciones. Ishwaran et al. (2011) [50] revisan el uso de profundidad mínima para medir calidad predictiva de variables.

10.4. Bosques de Ranking. Clémençon et al. (2013) [34] extendieron bosques para problemas de ranking con **Ranking Forests**, basados en ranking trees [33]. El enfoque se basa en puntuación no paramétrica y optimización de la curva ROC mediante el criterio AUC.

10.5. Bosques de Clustering. Yan et al. (2013) [82] presentan **Cluster Forests (CF)** para clasificación no supervisada. CF explora aleatoriamente un data cloud de alta dimensión para obtener buenos agrupamientos locales, luego los agrega mediante spectral clustering. La búsqueda está guiada por una medida de calidad de clúster, y CF mejora progresivamente cada agrupamiento de manera similar al crecimiento de árboles en bosques.

10.6. Bosques de Cuantiles. Meinshausen (2006) [63] muestra que los bosques proporcionan información sobre la distribución condicional completa de la respuesta, permitiendo estimación de cuantiles.

11. Ejemplo simulado

11.1. Descripción de los datos. Para ilustrar el funcionamiento de un modelo de regresión basado en *Random Forest*, se generó un conjunto de datos simulado con $n = 1000$ observaciones y cinco variables explicativas:

$x_1 \sim \mathcal{U}(0, 10), \quad x_2 \sim \mathcal{U}(-3, 3), \quad x_3 \sim \mathcal{N}(5, 2^2), \quad x_4 \sim \mathcal{N}(0, 1^2), x_5 \sim \mathcal{U}(-5, 5).$

La variable respuesta y se generó a partir de la siguiente relación no lineal con componente aleatoria:

$$y = 5 + 2\sin(x_1) - 0 \cdot 5x_2^2 + 0 \cdot 8x_3 - 0 \cdot 3x_4 + \varepsilon,$$

donde el término de error se distribuye como

$$\varepsilon \sim \mathcal{N}(0, 1).$$

Además en el experimento se tomaron $N = 1000$ arboles y $mtry = 2$ para entrenamiento del modelo.

El objetivo del ejercicio es ajustar un modelo de *Random Forest* para estimar y a partir de las covariables $(x_1, x_2, x_3, x_4, x_5)$ y evaluar su capacidad predictiva en un conjunto de prueba.

11.2. Resultados. La tabla 2 presenta los resultados obtenidos tras entrenar el modelo *Random Forest* en el conjunto de prueba, mostrando la comparación entre los valores observados y las predicciones generadas. Esta tabla permite evaluar de manera clara la precisión del modelo en las observaciones no utilizadas durante el entrenamiento.

Cuadro 2. Comparación entre valores reales y predichos del modelo Random Forest para algunos puntos de prueba

Observación	**Valor Real**	**Predicción RF**
1	8.641	7.106
5	3.163	5.186
6	11.592	11.880
21	8.246	9.454
23	9.914	9.894
24	10.523	9.527
25	4.394	6.455
29	7.679	9.210
33	9.387	7.906
35	12.720	11.384
37	8.373	8.818
42	8.388	7.859
44	7.252	6.975
54	8.303	7.882
59	8.606	9.723
62	7.717	7.787
63	9.398	8.057
65	7.518	6.169
66	5.456	5.304
69	8.367	6.543

En la tabla 3 el modelo *Random Forest* presenta un error cuadrático medio (**MSE**) de $\mathbf{1 \cdot 622}$ y un error cuadrático medio de la raíz (**RMSE**) de $\mathbf{1 \cdot 274}$ en el conjunto de prueba. Dado que la variable respuesta y toma valores entre $-1 \cdot 19$ y $15 \cdot 39$, el error promedio equivale aproximadamente al $\mathbf{7 \cdot 7}\,\%$ del rango total de y. Este resultado sugiere que el modelo logra un ajuste adecuado y una capacidad predictiva razonablemente buena, manteniendo errores moderados en relación con la variabilidad de los datos observados.

CUADRO 3. Desempeño del modelo *Random Forest* en el conjunto de prueba

Métrica	**Símbolo**	**Valor**
Error Cuadrático Medio	MSE	1.622
Raíz del Error Cuadrático Medio	RMSE	1.274

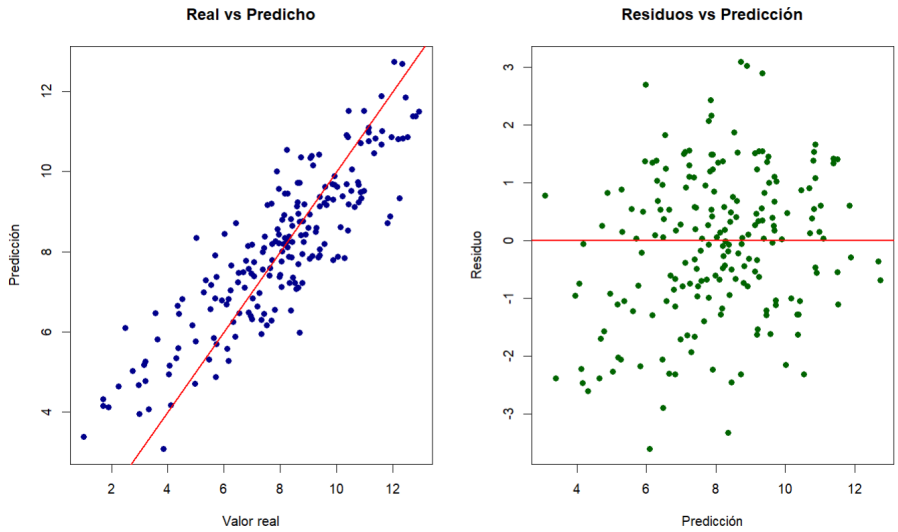
El modelo *Random Forest* obtuvo un **Error Out-of-Bag (OOB)** de

$$\text{OOB} = 1 \cdot 716$$

lo que representa una estimación interna del error de predicción promedio. Este valor indica que, en promedio, el modelo comete un error cuadrático medio aproximado de $1 \cdot 716$ al predecir observaciones no utilizadas durante el entrenamiento de cada árbol.

La Figura 1 presenta una comparación detallada del desempeño del modelo *Random Forest*, mostrando tanto la relación entre los valores observados y las predicciones generadas como la distribución de los residuos asociados. Esta representación permite evaluar visualmente la precisión y el ajuste del modelo.

FIGURA 1. Comparación de desempeño del modelo Random Forest: a la izquierda, (*Real vs Predicho*); a la derecha, (*Residuos vs Predicción*).



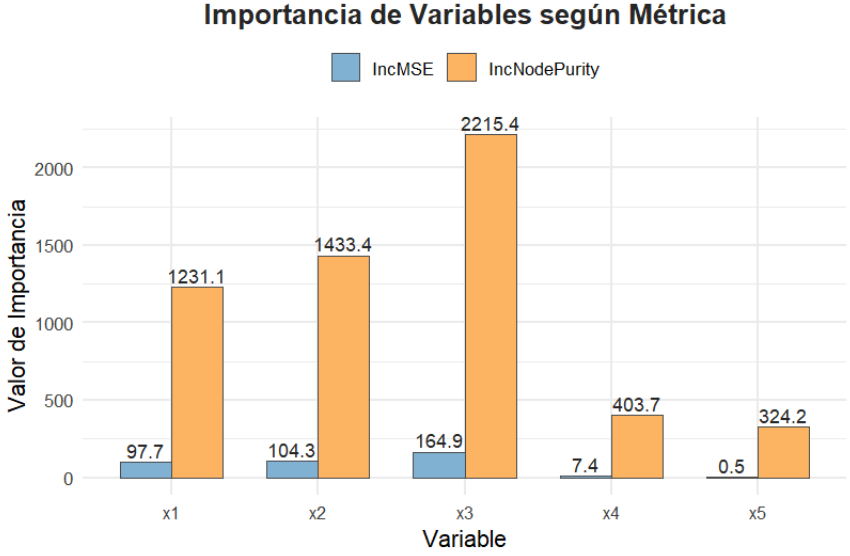
La Tabla 4 muestra la importancia relativa de cada variable en el modelo *Random Forest*. Se observa que x_3, x_2 y x_1 son las variables más influyentes tanto en términos de %IncMSE como de IncNodePurity, lo que indica que ejercen un mayor impacto sobre las predicciones del modelo. Por el contrario, x_4 y, especialmente, x_5 presentan valores bajos, sugiriendo que aportan poca información para la predicción de la variable respuesta y. Esta información resulta útil para identificar las covariables que dominan la estructura del modelo.

CUADRO 4. Importancia de las variables en el modelo *Random Forest* para regresión segun la métrica

Variable	**%IncMSE**	**IncNodePurity**
x_1	97.709	1231.129
x_2	104.299	1433.419
x_3	164.855	2215.393
x_4	7.381	403.692
x_5	0.537	324.187

La Figura 2 ofrece una representación visual refinada de la relevancia de las variables en el modelo, mediante un gráfico de barras comparativo entre métricas de importancia. Esta visualización proporciona una comprensión más intuitiva de las variables que ejercen una mayor influencia en el desempeño predictivo del modelo.

FIGURA 2.



12. Conclusiones

En este trabajo se ha demostrado que los *Random Forests* constituyen una de las metodologías más robustas y versátiles de aprendizaje estadístico moderno, gracias a su capacidad para manejar datos de alta complejidad, identificar patrones relevantes y realizar predicciones precisas y estables. Se destaca que la aleatorización en la selección de datos y variables permite reducir la correlación entre árboles, disminuyendo la varianza del modelo final y haciéndolo especialmente útil en escenarios con relaciones complejas y presencia de ruido.

Desde una perspectiva teórica, la solidez del método fue establecida por Leo Breiman, siendo posteriormente reforzada por diversos autores quienes han garantizado la consistencia y adaptabilidad de los bosques ante distintos tipos de datos. En el ámbito aplicado, los *Random Forests* resultan especialmente apropiados para abordar problemas sociales, económicos y ambientales de Honduras, como la medición de pobreza, la planificación territorial y el análisis educativo.

Metodológicamente, se ha subrayado la importancia del algoritmo y sus criterios de construcción, así como los métodos de evaluación de la importancia de las variables, incluyendo las medidas *Mean Decrease Impurity* (MDI) y *Mean Decrease Accuracy* (MDA). Asimismo, la flexibilidad para realizar tareas de regresión y clasificación, junto con la posibilidad de ajustar parámetros utilizando la estimación *out-of-bag*, representa una fortaleza adicional.

En resumen, la contribución principal de este trabajo es resaltar la integración de robustez teórica, eficiencia práctica y versatilidad aplicada por parte de los *Random Forests*, consolidándolos como una herramienta primordial para el análisis, la predicción y la toma de decisiones informadas en contextos multidisciplinarios. Además, su implementación y estudio favorecerían el fortalecimiento de competencias nacionales en ciencia de datos e inteligencia artificial, aportando significativamente al desarrollo científico de la región.

Referencias

[1] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[2] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1995.

[3] Y. Amit and D. Geman, "Randomization of decision trees," *Machine Learning*, vol. 29, no. 2–3, pp. 223–244, 1997.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] G. Biau, L. Devroye, and G. Lugosi, "Consistency of random forests and other averaging classifiers," *Journal of Machine Learning Research*, vol. 9, pp. 2015–2033, 2008.

[6] J. Chen, X. Wang, and F. Lei, "Data-driven multinomial random forest: a new random forest variant with strong consistency," *Journal of Big Data*, 2024.

[7] A. Dorador, "Theoretical and empirical advances in forest pruning," *ArXiv preprint arXiv:2401.05535*, 2024.

[8] V. Ignatenko, A. Surkov, and S. Koltcov, "Random forests with parametric entropy-based information gains for classification and regression problems," *PeerJ Computer Science*, 2024.

[9] Y. Ren, X. Zhu, K. Bai, and R. Zhang, "A new random forest ensemble of intuitionistic fuzzy decision trees (IFRF)," *ArXiv preprint arXiv:2403.07363*, 2024.

[10] A. V. Konstantinov, L. V. Utkin, A. A. Lukashin, and V. A. Muliukha, "Neural attention forests: transformer-based forest improvement," *ArXiv preprint arXiv:2304.05980*, 2023. 2023.

[11] Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees," Chapman and Hall/CRC, 1984.

[12] Breiman, L. (1996). *Bias, variance, and arcing classifiers*. Technical Report, University of California, Berkeley.

[13] M. R. Segal, "Machine learning benchmarks and random forest regression," *Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco*, 2004.
[14] G. Biau and E. Scornet, "A random forest guided tour," *Sorbonne Universités, UPMC Univ Paris 06, Institut Universitaire de France*, 2016.
[15] D. Amaratunga, J. Cabrera, and Y.-S. Lee, "Enriched random forests," *Bioinformatics*, vol. 24, no. 18, pp. 2010–2014, 2008.
[16] S. Arlot and R. Genuer, "Analysis of purely random forests bias," *arXiv preprint arXiv:1407.3939*, 2014.
[17] Z.-D. Bai, L. Devroye, H.-K. Hwang, and T.-H. Tsai, "Maxima in hypercubes," *Random Structures & Algorithms*, vol. 27, no. 3, pp. 290–309, 2005.
[18] M. Banerjee and I. W. McKeague, "Confidence sets for split points in decision trees," *The Annals of Statistics*, vol. 35, no. 2, pp. 543–574, 2007.
[19] O. Barndorff-Nielsen and M. Sobel, "On the distribution of the number of admissible points in a vector random sample," *Theory of Probability and Its Applications*, vol. 11, no. 2, pp. 249–269, 1966.
[20] S. Bernard, L. Heutte, and S. Adam, "Forest-RK: A new random forest induction method," *in Proceedings of the 4th International Conference on Intelligent Computing*, pp. 430–437, 2008.
[21] S. Bernard, L. Heutte, and S. Adam, "Dynamic random forests," *Pattern Recognition Letters*, vol. 33, no. 12, pp. 1580–1586, 2012.
[22] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, pp. 1063–1095, 2012.
[23] G. Biau, L. Devroye, and G. Lugosi, "Consistency of random forests and other averaging classifiers," *Journal of Machine Learning Research*, vol. 9, pp. 2015–2033, 2008.
[24] G. Biau, F. Cérou, and A. Guyader, "On the rate of convergence of the bagged nearest neighbor estimate," *Journal of Machine Learning Research*, vol. 11, pp. 687–712, 2010.
[25] G. Biau and L. Devroye, "On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification," *Journal of Multivariate Analysis*, vol. 101, no. 10, pp. 2499–2518, 2010.
[26] G. Biau and L. Devroye, "On the risk of estimates for block decreasing densities," *Journal of Multivariate Analysis*, vol. 119, pp. 176–189, 2013.
[27] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees," *Wadsworth International Group*, Belmont, CA, 1984.
[28] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
[29] L. Breiman, "Some infinity theory for predictor ensembles," *Technical Report 579, Statistics Department, University of California at Berkeley*, 2000.
[30] L. Breiman and A. Cutler, "Random forests," *http://www.stat.berkeley.edu/ breiman/RandomForests/*, 2003.
[31] L. Breiman, "Consistency for a simple model of random forests," *Technical Report 670, Statistics Department, University of California at Berkeley*, 2004.
[32] P. Bühlmann and B. Yu, "Analyzing bagging," *The Annals of Statistics*, vol. 30, no. 4, pp. 927–961, 2002.
[33] S. Clémençon and N. Vayatis, "Tree-based ranking methods," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4316–4336, 2009.
[34] S. Clémençon, M. Depecker, and N. Vayatis, "Ranking forests," *Journal of Machine Learning Research*, vol. 14, pp. 39–73, 2013.
[35] A. Davies and Z. Ghahramani, "The random forest kernel and other kernels for big data from random partitions," *arXiv preprint arXiv:1402.4293*, 2014.
[36] H. Deng and G. Runger, "Feature selection via regularized trees," *in Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2012.
[37] H. Deng and G. Runger, "Gene selection with guided regularized random forest," *Pattern Recognition*, vol. 46, no. 12, pp. 3483–3489, 2013.
[38] M. Denil, D. Matheson, and N. de Freitas, "Narrowing the gap: Random forests in theory and in practice," *in Proceedings of the 30th International Conference on Machine Learning*, pp. 665–673, 2013.
[39] M. Denil, D. Matheson, and N. de Freitas, "Consistency of online random forests," *in Proceedings of the 31st International Conference on Machine Learning*, pp. 1256–1264, 2014.
.

[40] L. Devroye, L. Györfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition," *Springer-Verlag*, New York, 1996.
[41] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, pp. 1–13, 2006.
[42] B. Efron, "The jackknife, the bootstrap, and other resampling plans," *SIAM*, Philadelphia, 1982.
[43] J. H. Friedman, B. E. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2009.
[44] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.
[45] R. Genuer, "Variance reduction in purely random forests," *Journal of Nonparametric Statistics*, vol. 24, no. 3, pp. 543–562, 2012.
[46] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
[47] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, "A Distribution-Free Theory of Nonparametric Regression," *Springer-Verlag*, New York, 2002.
[48] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, 2006.
[49] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.
[50] H. Ishwaran, U. B. Kogalur, X. Chen, and A. J. Minn, "Random survival forests for high-dimensional data," *Statistical Analysis and Data Mining*, vol. 4, no. 1, pp. 115–132, 2011.
[51] H. Ishwaran, "The effect of splitting on random forests," *Machine Learning*, vol. 99, no. 1, pp. 75–118, 2013.
[52] H. Kim and W.-Y. Loh, "Classification trees with unbiased multiway splits," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 589–604, 2001.
[53] E. Konukoglu and M. Ganz, "Approximate false positive rate control in selection frequency for random forest," *arXiv preprint arXiv:1410.2838*, 2014.
[54] J. Kruppa, Y. Liu, G. Biau, M. Kohler, I. R. König, J. D. Malley, and A. Ziegler, "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory," *Biometrical Journal*, vol. 56, no. 4, pp. 534–563, 2013.
[55] J. Kruppa, Y. Liu, H.-C. Diener, T. Holste, C. Weimar, I. R. König, and A. Ziegler, "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications," *Biometrical Journal*, vol. 56, no. 4, pp. 564–583, 2014.
[56] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, "Consumer credit risk: Individual probability estimates using machine learning," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8336–8346, 2014.
[57] A. Kyrillidis and A. Zouzias, "Non-uniform feature sampling for decision tree ensembles," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4548–4552, 2014.
[58] B. Lakshminarayanan, D. M. Roy, and Y. W. Teh, "Mondrian forests: Efficient online random forests," *in Advances in Neural Information Processing Systems*, pp. 3140–3148, 2014.
[59] P. Latinne, O. Debeir, and C. Decaestecker, "Limiting the number of trees in random forests," *in Multiple Classifier Systems*, Springer, pp. 178–187, 2001.
[60] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[61] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 578–590, 2006.
[62] J. D. Malley, J. Kruppa, A. Dasgupta, K. G. Malley, and A. Ziegler, "Probability machines: Consistent probability estimation using nonparametric learning machines," *Methods of Information in Medicine*, vol. 51, no. 1, pp. 74–81, 2012.
[63] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, pp. 983–999, 2006.
[64] N. Meinshausen, "Forest Garrote," *Electronic Journal of Statistics*, vol. 3, pp. 1288–1304, 2009.

[65] B. H. Menze, B. M. Kelm, D. N. Splitthoff, U. Koethe, and F. A. Hamprecht, "On oblique random forests," *in Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 453–469, 2011.
[66] E. A. Nadaraya, "On estimating regression," *Theory of Probability and Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.
[67] D. N. Politis, J. P. Romano, and M. Wolf, "Subsampling," *Springer-Verlag*, New York, 1999.
[68] S. S. Qian, K. H. Reckhow, J. Zhai, and G. McMahon, "Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach," *Water Resources Research*, vol. 41, no. 7, 2005.
[69] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, "On-line random forests," *in IEEE 12th International Conference on Computer Vision Workshops*, pp. 1393–1400, 2009.
[70] D. F. Schwarz, I. R. König, and A. Ziegler, "On safari to Random Jungle: A fast implementation of Random Forests for high-dimensional data," *Bioinformatics*, vol. 26, no. 14, pp. 1752–1758, 2010.
[71] E. Scornet, "Learning with random forests," *PhD thesis, Université Pierre et Marie Curie-Paris VI*, 2015.
[72] E. Scornet, "On the asymptotics of random forests," *Journal of Multivariate Analysis*, vol. 146, pp. 72–83, 2016.
[73] E. Scornet, G. Biau, and J.-P. Vert, "Consistency of random forests," *The Annals of Statistics*, vol. 43, no. 4, pp. 1716–1741, 2015.
[74] M. R. Segal, "Regression trees for censored data," *Biometrics*, vol. 44, no. 1, pp. 35–47, 1988.
[75] C. J. Stone, "Optimal rates of convergence for nonparametric estimators," *The Annals of Statistics*, vol. 8, no. 6, pp. 1348–1360, 1980.
[76] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *The Annals of Statistics*, vol. 10, no. 4, pp. 1040–1053, 1982.
[77] A. Truong, "Fast growing and interpretable oblique trees via logistic regression models," *PhD thesis, University of Oxford*, 2009.
[78] S. Wager, "Asymptotic theory for random forests," *arXiv preprint arXiv:1405.0352*, 2014.
[79] S. Wager, T. Hastie, and B. Efron, "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1625–1651, 2014.
[80] G. S. Watson, "Smooth regression analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 26, no. 4, pp. 359–372, 1964.
[81] S. J. Winham, R. R. Freimuth, and J. M. Biernacka, "A weighted random forests approach to improve predictive performance," *Statistical Analysis and Data Mining*, vol. 6, no. 6, pp. 496–505, 2013.
[82] D. Yan, A. Huang, and R. Jordan, "Cluster forests," *Computational Statistics & Data Analysis*, vol. 66, pp. 178–192, 2013.
[83] H. Yang, M. Rantalainen, and J. K. Nicholson, "Toward improved identification of true positives in ADME genes: Kernel-based regularized least squares with random forests for QSAR studies of imbalanced high-throughput data," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 1737–1750, 2010.
[84] C. Yi, Y. Tian, R. Peng, and M. Bennamoun, "Information forests," *in British Machine Vision Conference (BMVC)*, 2012.
[85] R. Zhu, D. Zeng, and M. R. Kosorok, "Reinforcement learning trees," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1770–1784, 2015.

Maestría en Matemática, Universidad Nacional Autónoma de Honduras.
Dirección actual: Departamento de Estadística Matemática, Facultad de Ciencias, Universidad Nacional Autónoma de Honduras
Dirección de correo electrónico: allan.cordova@unah.edu.hn

EVALUACIÓN DEL RIESGO CLIMÁTICO AGRICOLA EN HONDURAS MEDIANTE UN MODELO ACTUARIAL ECONOMÉTRICO BASADO EN FUNCIONES DE VALORES EXTREMOS

AXEL JOSAPHET CRUZ LOPEZ

Resumen. La producción agrícola en Honduras enfrenta una alta vulnerabilidad a eventos climáticos extremos, exacerbados por el cambio climático. Esta investigación desarrolla y evalúa un modelo actuarial econométrico híbrido para la cuantificación del riesgo climático agrícola en el país. El modelo propuesto fusiona tres enfoques: el análisis econométrico de series temporales (ARIMA-X y VAR) para capturar la dinámica y co-dependencia entre variables macro-climáticas (precipitación, temperatura) y los rendimientos de producción;la Teoría de Valores Extremos (EVT), utilizando distribuciones GEV y GPD, para modelar específicamente la frecuencia y severidad de los eventos catastróficos (sequías, inundaciones) que residen en las colas de la distribución; y la ciencia actuarial para el cálculo de primas de riesgo. Mediante simulaciones de Monte Carlo, se integra la dinámica base con los shocks extremos para generar una distribución agregada de pérdidas agrícolas. Esta distribución permite el cálculo de la Pérdida Esperada (EL) y el Tail Value at Risk (TVaR), fundamentando una metodología robusta y cuantitativa para la tarificación de seguros agrícolas adaptada a las condiciones de riesgo extremo en Honduras.

Palaras claves : Riesgo Climático, Teoría de Valores Extremos, Modelos Actuariales, Series Temporales, Honduras.

Abstract. Agricultural production in Honduras faces high vulnerability to extreme weather events, exacerbated by climate change. This research develops and evaluates a hybrid econometric-actuarial model for quantifying agricultural climate risk in the country. The proposed model combines three approaches: econometric time series analysis (ARIMA-X and VAR) to capture the dynamics and co-dependence between macro-climatic variables (precipitation, temperature) and production yields; Extreme Value Theory (EVT), using GEV and GPD distributions, to specifically model the frequency and severity of catastrophic events (droughts, floods) that lie in the tails of the distribution;and actuarial science for the calculation of risk premiums. Through Monte Carlo simulations, the base dynamics are integrated with extreme shocks to generate an aggregated distribution of agricultural losses. This distribution allows for the calculation of Expected Loss (EL) and Tail Value at Risk (TVaR), providing a robust and quantitative methodology for pricing agricultural insurance adapted to the conditions of extreme risk in Honduras.

Key words : Climate Risk, Extreme Value Theory, Actuarial Models, Time Series, Honduras

Fecha: Noviembre 2025.
Palabras y frases clave. Riesgo climatico, Econométria.

1. Introducción

La economía de Honduras y la seguridad alimentaria de su población dependen intrínsecamente del desempeño del sector agrícola. Este sector, sin embargo, es uno de los más vulnerables a la variabilidad climática y, de forma creciente, a la intensificación de eventos extremos como sequías prolongadas, huracanes e inundaciones torrenciales, fenómenos exacerbados por el cambio climático. La evaluación precisa de este riesgo es fundamental, pero presenta un desafío metodológico significativo. Los modelos econométricos tradicionales, como los de series temporales ARIMA o VAR, son eficientes para capturar la dinámica promedio y las interacciones entre variables como la precipitación, la temperatura y la producción (Zulfiqar et al., 2024). No obstante, estos modelos fallan sistemáticamente al subestimar la probabilidad y el impacto de los eventos catastróficos, ya que sus supuestos a menudo de normalidad no pueden modelar adecuadamente las colas pesadas de las distribuciones de pérdidas.

Por otro lado, la Teoría de Valores Extremos (EVT) ofrece un marco estadístico robusto, fundamentado matemáticamente, diseñado específicamente para modelar el comportamiento de estos eventos raros y severos (Coles, 2001). Modelos como la Distribución Generalizada de Valores Extremos (GEV) o la Distribución Generalizada de Pareto (GPD) permiten una caracterización precisa de la cola de la distribución, es decir, del riesgo de pérdidas extremas en los cultivos (Van Tassell, 2024). Sin embargo, la EVT por sí sola no captura la dinámica temporal subyacente ni las co-dependencias econométricas del sistema climático-agrícola. La literatura reciente busca cerrar esta brecha, reconociendo que ni los modelos econométricos por sí solos, ni los modelos de EVT de forma aislada, son suficientes para una evaluación integral del riesgo.

Esta investigación propone y desarrolla un modelo actuarial econométrico híbrido para la evaluación del riesgo climático agrícola en Honduras. El modelo fusiona estas disciplinas para superar sus limitaciones individuales.

El objetivo general de este trabajo es cuantificar el riesgo de pérdida agrícola debido a factores climáticos extremos en Honduras, desarrollando una metodología para el cálculo de primas de seguro actuarialmente justas. Para alcanzar esto, se plantean los siguientes objetivos específicos:

- Modelar la dinámica y co-dependencia de línea base entre las series temporales de producción agrícola, precipitación y temperatura mediante modelos econométricos ARIMA-X y/o VAR.
- Ajustar modelos de Teoría de Valores Extremos GEV y GPD a los residuos extremos de los modelos climáticos o directamente a los eventos extremos ejemplo. sequías para caracterizar la frecuencia y severidad del riesgo catastrófico.
- Integrar ambos componentes (econométrico y EVT) a través de simulaciones de Monte Carlo para generar una distribución de pérdida agregada anual para el sector agrícola.
- Calcular métricas de riesgo actuarial, como la Pérdida Esperada (EL) y el Tail Value at Risk (TVaR), para establecer una base técnica para la prima pura de riesgo y los requerimientos de capital.

La importancia de este articulo radica en su aplicación práctica. Al fusionar estas técnicas, el modelo no solo describe el riesgo, sino que lo cuantifica en términos monetarios y probabilísticos. Esto proporciona una herramienta cuantitativa esencial para el diseño de seguros agrícolas paramétricos o de índice, la estructuración de fondos de contingencia y la toma de decisiones de política pública para la adaptación al cambio climático. Siguiendo enfoques similares aplicados en otros contextos (Ly et al., 2024), este trabajo ofrece una metodología robusta para fortalecer la resiliencia financiera de los agricultores hondureños ante un futuro climático incierto y extremo.

2. Jutificación

La presente tesis, centrada en la Evaluación del Riesgo Climático Agrícola en Honduras mediante un Modelo Actuarial Econométrico, se justifica por la urgente necesidad de dotar al país de herramientas cuantitativas y técnicamente avanzadas para gestionar las crecientes amenazas que el cambio climático impone a su estabilidad económica, sostenibilidad productiva y seguridad alimentaria.

1. Relevancia y Justificación Nacional (Resolución de Problemas de País):
   La justificación de este trabajo radica en la respuesta directa y cuantificable a problemas nacionales críticos, utilizando datos oficiales y especializados
   - Riesgo Catastrófico y Sostenibilidad Financiera: La agricultura, pilar económico hondureño, está expuesta a la volatilidad climática. El uso de la Teoría de Valores Extremos (EVT) (Coles, 2001) sobre datos de COPECO e IHCIT permite modelar la probabilidad de eventos catastróficos (sequías e inundaciones) que impactan la producción. Los modelos tradicionales subestiman este riesgo, un fallo que ha costado al país pérdidas significativas.
   - Fundamento para el Seguro Agrícola Nacional: La adopción de seguros agrícolas es mínima debido a la falta de metodologías transparentes para la tarificación. Esta investigación utiliza el PIB Agrícola y datos de precios del BCH y SEFIN para monetizar el impacto de las pérdidas. Al fusionar la EVT con el enfoque actuarial (Klugman et al., 2019; Ly et al., 2024), el estudio establece el cálculo de primas puras basadas en la Pérdida Esperada y el TVaR, proporcionando la base técnica indispensable para que SENASA y otras instituciones financieras puedan desarrollar e implementar productos de transferencia de riesgo sostenibles.
   - Integración de Datos y Política Pública: El trabajo exige la integración rigurosa de estadísticas agroclimáticas SENASA, IHCIT, meteorológicas (COPECO) y económicas BCH, SEFIN en modelos ARIMA-X/VAR (Enders, 2014). Esta integración demuestra la viabilidad de utilizar la información nacional dispersa para la toma de decisiones basada en evidencia, transitando de una gestión de crisis reactiva a una gestión de riesgo predictiva y cuantificada.
2. Alineación con las Líneas de Investigación Prioritarias de la UNAH
   Este trabajo se alinea de manera fundamental y directa con las prioridades de investigación de la Universidad Nacional Autónoma de Honduras (UNAH):

Línea de Investigación Prioritaria UNAH	Justificación de la Alineación
Cambio Climático, Ambiente y Gestión de Riesgos	Se centra en la evaluación y gestión cuantitativa del riesgo climático, utilizando datos meteorológicos IHCIT, COPECO para calibrar modelos de Teoría de Valores Extremos (EVT) que miden la severidad del impacto ambiental sobre el sector productivo nacional.
Desarrollo Económico, Pobreza, Desigualdad y Desarrollo Humano	El estudio aborda la fragilidad económica del sector agrícola. Al proveer la metodología para el seguro, contribuye a la resiliencia financiera y la estabilidad de ingresos de los agricultores, actuando como una herramienta contra la pobreza rural y para la planificación económica BCH, SEFIN.
Ciencia y Tecnología	La tesis es una aplicación de modelación avanzada, fusionando disciplinas matemáticas (EVT), estadísticas (Simulación Monte Carlo) y econométricas (ARIMA-X, VAR) para resolver un problema nacional, contribuyendo a la innovación metodológica y la generación de conocimiento científico-cuantitativo.

3. Línea de Investigación de la Maestría
   El trabajo sigue y fusiona de forma sinérgica las siguientes líneas de investigación de la Maestría en Estadística:
   a) Econometría y actuaría : Esta es la línea central. El estudio es la aplicación de modelos estadísticos y matemáticos para la evaluación de riesgos y la tarificación de seguros(actuaría) en un contexto económico PIB, precios, producción.
   b) Teoría de los valores extremos : El modelo depende fundamentalmente de la EVT para calcular la probabilidad de eventos o valores más extremos que los observados previamente sequías, inundaciones, lo cual es su principal aporte metodológico.
   c) Series de tiempo : El análisis de las series históricas de producción y variables climáticas (Enders, 2014) es esencial para la etapa de diagnóstico y modelado de la **dinámica temporal** base del fenómeno.

3. Antecedentes

1. El Pilar de la Econometría de Series Temporales y el Clima
   El análisis de series temporales se formalizó en la década de 1970 con la metodología Box-Jenkins. Esta proporcionó un marco sistemático para la identificación, estimación y verificación de modelos univariados ARIMA (Autorregresivo Integrado de Media Móvil), fundamentales para la modelización de variables que exhiben dependencia temporal, estacionalidad o tendencia (Enders, 2014).

   - Primeros Aportes (1980 - 1990): Inicialmente, los modelos de series temporales se aplicaron al análisis macroeconómico. Sin embargo, su extensión a variables climáticas y agrícolas fue evidente al buscar la relación entre variables económicas y factores exógenos. La incorporación de la variable climática como un factor exógeno dio origen al modelo ARIMA con variables exógenas (ARIMA-X), permitiendo capturar cómo la precipitación o la temperatura afectan la producción (Zulfiqar et al., 2024).
   - Desarrollo Multivariado (1990 - 2000): La comprensión de que las variables económicas y climáticas se influyen mutuamente llevó al desarrollo de modelos multivariados, principalmente el Vector Autorregresivo (VAR), popularizado por Christopher A. Sims. Estos modelos son esenciales para analizar la causalidad y la dinámica de corto y largo plazo entre variables interconectadas ejemplo, la precipitación en la producción y el PIB agrícola.
   - Posteriormente, los trabajos de Robert F. Engle y Clive W. J. Granger sobre la Cointegración permitieron modelar la relación de equilibrio a largo plazo entre series, incluso si estas son no estacionarias. Este concepto es vital en la agricultura, ya que la producción y los precios, aunque volátiles, pueden mantener una relación estable a largo plazo.
   - Aportes Recientes: Los trabajos más recientes buscan refinar la relación entre el clima y los resultados económicos. Estudios como el de Sarker y Sarker (2024) usan modelos econométricos para vincular la volatilidad del clima con las dinámicas de las exportaciones agrícolas, estableciendo la metodología para monetizar el impacto de las variables climáticas.

2. El Pilar de la Teoría de Valores Extremos (EVT)
   La Teoría de Valores Extremos (EVT) es la rama de la estadística que se enfoca en el comportamiento probabilístico de los valores atípicos, es decir, de los máximos o mínimos de una secuencia de datos. Sus fundamentos se remontan a principios del siglo XX, pero su formalización clave ocurrió en la segunda mitad.

   - Ronald Fisher y Leonard Tippett (1928): Publicaron el teorema que establece que la distribución de los máximos normalizados de una gran muestra debe converger a una de las tres formas asintóticas (Gumbel, Fréchet, Weibull).
   - Boris V. Gnedenko (1943): Demostró formalmente el Teorema de Fisher-Tippett-Gnedenko, el pilar de la distribución GEV (Generalized Extreme Value), utilizada para modelar los máximos por bloques (Coles, 2001).
   - Laurens de Haan y A. L. M. Rootzén (Década de 1970): Sus trabajos formalizaron el enfoque de Picos Sobre el Umbral (POT), demostrando

que la distribución de los excesos por encima de un umbral alto converge a la Distribución Generalizada de Pareto (GPD). Este enfoque es preferido en la práctica actuarial y financiera por su uso eficiente de los datos extremos.

- Aportes Recientes a la Teoría: El desarrollo reciente de la EVT se centra en hacer que los modelos sean no estacionarios (Coles, 2001). Dado que el riesgo climático está cambiando, el modelado moderno requiere que los parámetros de las distribuciones GEV/GPD como la localización o la escala sean funciones del tiempo o de covariables como tendencias de temperatura. Esto es crucial para proyectar el riesgo en un escenario de cambio climático y no solo describirlo históricamente.

3. El Pilar Actuarial y la Fusión con el Riesgo Extremo La ciencia actuarial, históricamente centrada en seguros de vida y pensiones, se expandió a la modelación de pérdidas no vida Propiedad y Daños a partir de la década de 1980.

    - Desarrolladores Clave y Modelos de Pérdida: Los trabajos de Stuart A. Klugman, Harry H. Panjer y Gordon E. Willmot formalizaron los modelos de pérdida agregada, que combinan distribuciones de frecuencia (cuántos eventos ocurren) y severidad (cuál es la magnitud de cada pérdida) (Klugman et al., 2019). Esta es la base para las Simulaciones de Monte Carlo utilizadas para generar la distribución de pérdida total.
    - Fusión EVT-Actuarial (Siglo XXI): La crisis financiera de 2008 y la creciente amenaza del riesgo catastrófico natural impulsaron la adopción de la EVT en la tarificación de seguros y la gestión de capital (Solvencia II). La EVT se convirtió en la herramienta estándar para modelar la cola de la distribución de pérdidas (severidad) antes de ser agregada mediante Monte Carlo. Esto permite calcular métricas actuariales robustas como el Value at Risk (VaR) y el Tail Value at Risk (TVaR), esenciales para la fijación de la prima pura y el recargo por riesgo (Dickson, 2016; Osepa & Mailafia, 2024).
    - Integración Actuarial-Econométrica: El aporte más reciente y relevante para esta tesis es la integración de los tres pilares. Ly, Riam, y Hizam (2024) ejemplifican este enfoque al utilizar modelos de cointegración (econometría) para vincular los rendimientos de cultivos con factores climáticos extremos (EVT), aplicando el resultado para diseñar un sistema de tarificación de primas.
    - Modelado del Rendimiento como Riesgo: El trabajo de Van Tassell (2024) valida el uso de la EVT para modelar la cola izquierda de la distribución de rendimientos agrícolas es decir, las grandes pérdidas o fallas de cosecha, proveyendo el insumo directo para el cálculo actuarial del riesgo de seguro.

4. Integración Actuarial-Econométrica y Riesgo Agrícola

    - El aporte más reciente es la integración de los tres pilares en aplicaciones sectoriales. El trabajo de Van Tassell (2024) valida la aplicación de la EVT para modelar la cola izquierda de las distribuciones de rendimientos

de cultivos las grandes pérdida, proporcionando el insumo directo para el cálculo de pérdidas.

- Finalmente, la investigación de Ly, Riam, y Hizam (2024) ejemplifica la integración final: utilizan la cointegración econometría y la cuantificación de extremos (EVT) para alimentar directamente un sistema actuarial de tarificación de primas, que es el objetivo último de esta tesis.

Cuerpo del articulo

Marco teorico

Modelado de serie temporales en agricultura. El modelado de serie temporales agricolas consiste en analizar la evolución temporal de variable agroclimatica como rendimiento,producción precipitación o temperatura con el fin de capturar sus tendencia, estacionalidades y perturbaciones aleatorias.Este enfoque permite identificar relaciones dinámicas entre los factores climáticos y los resultados productivos, esenciales para pronósticos y evaluación de riesgo en agricultura.[3]
Propiedades fundamentales de las series temporales

Definición Estacionariedad: Una serie Y_t es estacionaria si su media y varainza son constantes a lo largo del tiempo y la covarianza depende solo de la distancia temporal (h), no del tiempo absoluto.

$$E[Y_t] = \mu, \quad Var(Y_t) = \sigma^2, \quad Cov(Y_t, Y_{t-h}) = \gamma(h) \tag{3.1}$$

La estacionariedad garantiza que los parámetros estimados sean estables y que el proceso sea predecible.[8]

Definición Autocorrelación y dependencia temporal: Las observaciones sucesivas de Y_t puede estar correlacionadas, capturando persistencia climática o agricola como por ejemplo el rendimiento que afecta las condiciones previas de humedad o temperatura.[8]
La función de autocorrelación (FAC) se define como :

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \tag{3.2}$$

Definición Estacionalidad : En agricultura, es común observar patrones estacionales asociados a ciclos de cosecha o precipitaciones. Se puede eliminar mediante diferenciación estacional :

$$Y_t^{'} = Y_t - Y_{t-s} \tag{3.3}$$

donde s representa la periodicidad como por ejemplo puede ser perido de 12 meses o 4 trimestre.[3]

Modelo ARIMA(p,d,q)
El modelo ARIMA(AutoRegressive Integrated Moving Average) representa la relación entre el valor actual de una serie y sus valores pasados,junto con errores pasados.

$$Y_t = \phi_1 Y_{t-1} + \phi Y_{t-2} + ... + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{3.4}$$

donde :

- p : Es el orden autorregresivo (AR).
- d: Es el número de diferencias aplicadas para lograr estacionariedad.

- q : Es el orden del promedio móvil (MA).
- ε_t: Es el error o ruido blanco con media vero y varainza constante.

La version integrada se aplica cuando la serie presenta tendencia, transformando Y_t en diferencia de orden d

$$\bigtriangledown^d Y_t = (1-B)^d Y_t \tag{3.5}$$

[8], [3]

Modelo ARIMA-X(ARIMAX)

En contextos agricolas, las fluctuaciones en el rendimiento no dependen únicamente de la dinámica temporal interna, sino tambien de factores exógenos climáticos.
El modelo ARIMA-X o ARIMAX amplia la formulación clásica incluyendo variables externas X_t, tales como precipitación acumulada, temperatura media o indices climáticos :

$$Y_t = \alpha + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{r} B_k X_{t-k} + \varepsilon_t \tag{3.6}$$

donde

- Y_t : Rendimiento o producción agricola.
- X_t : Variale exógena climática.
- B_k sensibilidad del rendimiento ante el factor climático.

Este modelo permite evaluar cómo los choques climáticos afectan la producción y cuantifican su elasticidad frente a la precipitación o temperatura.[3],[8]

Análisis de cointegración agricola

Cuando las series no son estacionarias en nivel, pero una combinación lineal de ella si lo es,se dice que están cointegradas.Esto implica una relación de equilibrio de largo plazo entre la variables agroclimáticas.
El modelo de cointegración es :

$$Y_t = \alpha + \beta X_t + \varepsilon_t \tag{3.7}$$

donde ε_t es estacionario, aunque Y_t y X_t no lo sean individualmente.
Este analisis de cointegración se usa para modelar relaciones entre rendimiento agricola y eventos climáticos extremos o indices globales como ENSO, temperatura oceánica etc, evitando regresiones espurias.[1]

Teoria de valores Extremos (EVT) en riesgo climaticos. Lateoria de valores extremos EVT, por sus siglas en ingles: Extreme Value Theory proporciona el marco estadistico para modelar eventos raros o extremos, es decir, aquellos que se sitúan en las colas de una distribución
En el contexto climático y agricola, la EVT permite estimar la probabilidad y magnitud de fenómenos pocos frecuentes pero de gran impacto, como sequias prolongadas, lluvias torreciales o temperaturas extremas, que afectan el rendimiento y la estabilidad económica del sector agricola.

Según Cole(2001)[7], la EVT se fundamente en el estudio del comportamiento asintótico de los maximos (o minimos) de una secuencia de variales aleatorias, extendiendo los principios de la probabbilidad clásica hacia las colas de la distribución.De esta manera, en lugar de analizar el comportamiento promedio, la EVT se

centra en los riesgos extremos, es decir, en los eventos que se sobrepasan un cierto umbral critico.[7]

Modelo de máximos por bloques(Block Maxima Approach)
El enfoque de máximos por bloques consiste en dividir una serie temporal en bloques de igual longitud por ejemplo años o estaciones y luego toma el valor máximo (o minimo) de cada bloque.
Si los datos son independintes e idénticamente distribuidos, el teorema de Fisher Tippett Gnedenko establece que

Teorema de Fisher Tippett Gnedenko Para un tamaño de bloques suficientemente grande, la distribución de los máximos se aproximan a una de las tres formas conocidas como distribución de valores extremos generalizada (GEV)

$$G(x) = exp\{-[1+\xi(\frac{x-\mu}{\sigma})]^{\frac{-1}{\xi}}\} \quad donde \quad 1+\xi\frac{x-\mu}{\sigma} > 0 \tag{3.8}$$

donde

- μ : Es el parametro de localización.
- $\sigma > 0$: Es el parámetro de escala.
- ξ: Es el parámetro de forma que determina el tipo de cola.

Consideremos la interpretacion del parámetro de forma (ξ):

- $\xi = 0$: Es de tipo Gumbel, cola exponencial(eventos moderadamente extremos).
- $\xi > 0$: Es de tipo Fréchet, cola pesada(eventos muy extremos , como lluvias torrenciales).
- $\xi < 0$: Es de tipo Weibull, cola finita(limite superior natural, útil para temperatura máxima.

[7]

Modelo de excedencias sobre Umbral(Peaks Over Threshold, POT)
El segundo enfoque, propuesto por Pickands(1975) y formalizado en cole(2001)[7], consiste en modelar directamente las excedencias sore un umbral alto u.
Si $Y = X - u$ representa el exceso sobre u, entonces para valores sufucientente grandes de u, la distribución condicional de Y sigue aproximadamente una distribución Pareto generalizada (GPD) :

$$H(y) = 1-(1+\xi\frac{y}{\beta})^{\frac{-1}{\xi}}, \quad y > 0, \quad 1+\xi\frac{y}{\beta} > 0 \tag{3.9}$$

donde

- $\beta > 0$: Parámetro de escala.
- ξ : parámetros de forma, compartido con la GEV.

Este modelo es especialmente útil para estimar probabilidades de eventos extremos raros, incluso cuando no existen observaciones directas de tales eventos en el historial. [7], [2]

Tenemos algunas propiedades fundamentales de la EVT :

1. Invarianza a tranformaciones lineales : Si X sigue un GEV o GPD , cualquier tranformación lineal a+bX con $b > o$ también pertenece a la mima familia.

2. Estabilidad de la forma : El parámetro ξ define la clase de cola(Gumbel, Frécheto Weibull) y se mantiene constante bajo escalas temporales razonables.
3. Interpretación de riesgo : La EVT permite calcular métricas de riesgo como el nivel de retorno y el periodo de retorno :

$$x_T = \mu + \frac{\sigma}{\xi}[(-ln(1-1/T))^{-\xi} - 1] \tag{3.10}$$

donde x_T es el evento esperado una vez cada T periodos por ejemplo una sequia centeraia.

[7]
La EVT ha sido ampliamente utilizada para estimar la probabilidad de lluvias extremas que superan la capacidad de drenaje agricola, tambien determian el riesgo de perdida por seguias prolongadas y evalua el impacto potencial de eventos ENSO(El Niño/La Niña) sobre el rendimiento agricola.
Según Van Tassell(2024), al combinar la EVT con información agroclimática, se obtiene una estimación más precisa del riesgo de pérdida extrema, lo cual es fundamental para diseñar seguros indexados climáticos y determinar primas actuariales justas. [2]

Modelos actuariales y de riesgo. Los modelos actuariales de riesgo constituyen la base matematica de la valoración de perdida, estimación de reserva y cálculo de primas en seguros.
En el contexto agricola , estos modelos permiten cuantificar la frecuencia e intensidad de eventos climáticos adversos(como sequias o lluvias excesivas) y estimar la pérdida esperada total de los productores.

De acuerdo con Klugman, Panjer y Willmot(2019)[11], el riesgo se representa como una suma aleatoria de pérdidas individuales, donde cada evento climático genera una pérdida X_i, y el número total de eventos N sigue una distribución discreta como por ejemplo Poisson o una Binomial.Asi, el modelo de pérdida agregada se define como :

$$S = \sum_{i=1}^{N} X_i \tag{3.11}$$

donde

- S : Es la pérdida total en un periodo, por ejemplo una temporada agricola.
- N : Es el número de evento extremos.
- X_i : Es la pérdida individual causada por el evento i.

[9]

Componentes del modelo de riesgo

1. Frecuencia de evento (N) :El numero de evento extremos se modela mediante una distribución discreta de conteo. Las mas utilizadas son :
   - Distribución de Poisson

$$P(N = n) = \frac{\lambda^n e^{-\lambda}}{n!} \qquad n = 0, 1, 2... \tag{3.12}$$

donde $\lambda = E[N]$ es la frecuencia esperada de eventos como por ejemplos lluvias intensas por años.

- Distribución Binomial : Si el número de observaciones es finito

$$P(N = n) = \binom{m}{n} p^n (1-p)^{m-n} \tag{3.13}$$

con m el número máximo de ensayos(años o parcelas) y p la probabilidad de un evento extremo.

2. Severidad de pérdida (X_i) :Cada pérdida individual X_i se modela con una distribución continua no negativa. Las mas comunes son
   - Distribución Lognormal : Es útil para pérdidas moderadas y variables climáticas multiplicativas.
   - Distribución Gamma o Weibull :aplicables a daños acumulativos.
   - Distribución Pareto o Generalized Pareto (GPD) : ideal para pérdidas extremas, en conexión con la EVT.

$$f(x;\xi,\beta) = \frac{1}{\beta}(1+\xi\frac{x}{\beta})^{\frac{-1}{\xi-1}} \qquad x > 0 \tag{3.14}$$

donde β es el parametro de escala y ξ es el parámetro de forma(cola pesada si $\xi > 0$)

[9], [7]

Momentos y métricas de riesgo

El riesgo total S combina la aleatoriedad de N(frecuencia) y X_i(severidad).Bajo independencia entre ambos, los momentos del total de pérdida son :

$$E[S] = E[N]E[X] \tag{3.15}$$

$$Var(s) = E[N]Var(X) + Var(N)(E[X])^2 \tag{3.16}$$

Estas expresiones permiten estimar el valor esperado de la perdida total y su variabilidad, base para el cálculo de reservas y primas.[9]

Distribución de pérdida agregadas

Cuando no existe una forma analitica simple para S, se utiliza métodos de simulación Monte Carlos o aproximaciones numéricas (Panjer recursion) para obtener la distribución de perdida totales. [13]

Simulación Monte Carlos:

- Genera $N \sim$ Poisson(λ).
- Simular N pérdidas X_i según su distribución.
- Calcular $S = \sum X_i$.
- Repetir hasta obtener la distribución empirica de S.

[9]

Probabilidad de ruina y control de solvencia

Según Dickson(2016)[10], el analisis actuarial no solo evalúa pérdidas esperadas, sino tambien la probabilidad de ruina del asegurador, es decir, la probabilidad de que las pérdidas acumuladas excedan el capital inicial u

$$\psi(u) = P(ruina|capital\ inicial\ = u) \tag{3.17}$$

En el modelo clásico de riesgo de Cramér-Lundberg[12], con primas c cobradas a tasa constante, la reserva del asegurado al tiempo t se expresa como :

$$U(t) = u + ct - S(t) \tag{3.18}$$

donde u es el capital inicial, c el ingreso por primas y S(t) es la pérdida acumulada hasta t.
La condición de equilibrio para evitar la ruina es

$$c > E[S]/t \tag{3.19}$$

[10]

Integración metodológica.

1. Los residuos de ARIMA-X representan las desviaciones no explicadas por los factores normales, es decir, los eventos anómalos o extremos.
2. Dichos residuos se analizan mediante EVT, obteniendo párametros de cola (ξ, β) que describe la severidad de las pérdidas.
3. Los parámetros de frecuencias (λ) y severidad (GPD) se integran en el modelo actuarial de riesgo agregado, permitiedo cuantificar la pérdida total esperada y la prima justa que compensa al asegurador.

Metodológia

Datos y procesamiento. Esta subsección repasa la fuente de datos y las series utilizadas ,PIB, precios de BCH,SEFIN; precipitación, temperatura y producción de SENASA,IHCIT,COPECO, la necesidad de homogeneidad de series intertemporales como menciona Enders, 2014,[8]) pero aquí usamos las otras series, y el procesamiento de la serie cronológica de eventos meteorológicos extremos para capturar las variables climáticas de interés.
De Sarker et al. (2024)[4] el artículo discute cómo los eventos meteorológicos extremos influyen en la dinámica de las exportaciones agrícolas y en las expectativas económicas intertemporales.

Definición de eventos meteorológicos extremos : las sequías y las inundaciones son los eventos meteorológicos extremos, ya que son grandes desviaciones de la norma e influyen en la producción agrícola de manera positiva o negativa, influyendo así en las exportaciones agrícolas. Estos eventos son choques que influyen en la economía para estar desbalanceada.

Propiedades de los eventos meteorológicos extremos: los eventos meteorológicos extremos están altamente sesgados con colas pesadas que ilustran el gran impacto son altamente arriesgados ejemplo cosechas pobres. Estos eventos meteorológicos extremos son no estacionarios, ya que están influenciados por el cambio climático, de ahí la necesidad de ajustar las series a la homogeneidad de series intertemporales.

Honduras necesita examinar los datos de temperatura de precipitaciones de SENASA,IHCIT,COPECO para homogeneizar tendencias y aislar mejor los extremos, ya que las tendencias pueden ser no climáticas ,variaciones estacionales, errores de medición, etc. Esto es útil para analizar los datos económicos, utilizando los datos

del PIB de BCH,SEFIN.

el artículo describe el uso de análisis de regresión para determinar la relación entre eventos climáticos extremos y exportaciones:

$$ln(EX_t) = \alpha + \beta_1 ln(EX_{t-1}) + \beta_2 EXT_t + \epsilon_t \tag{3.20}$$

donde EX_t es el volumen de exportaciones agricola en el periodo t, EXT_t es un indicador de evento extremo por ejemplo, desviación de precipitación, y ϵ_t es el error.Esta ecuación ayuda a procesar series temporales para detectar impactos en producción agrícola hondureña.

La Academia Americana de Actuarios (2024): El Índice de Riesgo Climático de los Actuarios (ACRI)[6] evalúa riesgos climáticos globales específicos, incluyendo la agricultura.

Definición ACRI : El ACRI es un índice compuesto que evalúa la magnitud del riesgo climático por ejemplo, sequías, inundaciones en varios sectores, incluyendo la agricultura, utilizando datos históricos y proyecciones.

Es aditivo y escalable, lo que permite homogeneizar series temporales de múltiples fuentes por ejemplo, precipitación de COPECO con precios de SEFIN. Incluye componentes sobre la frecuencia e intensidad de los extremos.
Para Honduras, ACRI puede ser utilizado para corroborar la homogeneidad de la serie temporal, asegurando que los datos de producción agrícola reflejen el verdadero riesgo climático, es decir, cambios en la precipitación que impulsan la cosecha.

El ACRI se calcula como :

$$ACRI = w_1.FREQ + w_2.INT + w_3.EXP \tag{3.21}$$

donde FREQ es la frecuencia de eventos extremos,INT su intensidad,EXP la exposición económica por ejemplo, basada en PIB agrícola ,w_i son pesos. Esto apoya el procesamiento de datos hondureños para identificar umbrales de riesgo.

Modelo econométrico base. Aquí se describe la selección y ajuste de modelos como ARIMA-X o VAR para series de producción y clima, capturando dinámicas base y tendencias relacionado con Enders, 2014,[8] .
El artriculo de Osepa et al.(2024)[5] combina EVT con machine learning para pronosticar riesgos de inversión, aplicable a modelos econométricos base.

Definición : Un modelo de pronóstico de riesgo de inversión integra EVT para extremos con técnicas econométricas por ejemplo, VAR para capturar tendencias no lineales en series temporales.

Una propiedad de esto es que es hibrido robusto a no estacionariedad, y usa machine learning para ajustar parámetros dinámicos, mejorando la predicción de shocks climáticos en producción agrícola.

En Honduras, se puede aplicar para ajustar un modelo VAR a series de producción (SENASA) y clima (COPECO), incorporando tendencias de cambio climático. Esto complementa ARIMA-X al incluir variables exógenas extremas.

El modelo combina EVT con regresión como se muestra:

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \gamma X_t + \delta EVT_t + \epsilon_t \tag{3.22}$$

donde Y_t es la producción agricola, X_t son variables climáticas, EVT_t es el compomente extremo por ejemplo de GPD , y ϵ_t es el error. esto captura dinámicas base con tendencias. Sarker et al(2024)[4] Vincula eventos extremos con dinámicas económicas, útil para modelos base.

Definición : Las expectativas climáticas se refieren a proyecciones de eventos extremos que afectan modelos econométricos, como VAR para producción agrícola.

Una propiedad muy importante es que son prospectivas, incorporando incertidumbre, y permiten ajustar modelos para tendencias estacionales o de largo plazo.

Para series hondureñas, esto ayuda a seleccionar ARIMA-X al incluir precipitación como variable exógena, capturando tendencias de sequía.

Similar a la anterior, pero enfocada en exportaciones:

$$\Delta EX_t = \alpha + \beta_1 \Delta EX_{t-1} + \beta_2 CLIM_t + \epsilon_t \tag{3.23}$$

donde CLM_t, representa expectativas climaticas por ejemplo temperaturas extremas, aplicable a producción.

Modelado de evento extremos(EVT). En esta sección se define umbrales, para sequias y exceso, ajusta GPD(POT) y modelos no estacionario[7] Se usa EVT para modelar extremos en riesgos de inversion[5]

Definición : EVT modelala distribución de valores extremos por ejemplo, minimo de precipitación para seguias usando GPD para excesos sobre umbrales.

Esto nos lleva a las siguientes propiedades
Propiedades

- Es adecuado para colas pesadas, con parametros como forma ξ y escala σ.
- Permite modelos no estacionarios al incluir covariables por ejemplo como tiempo para cambio climatico.

Para Honduras, define umbrales de sequias por ejemplo precipitación < 50 mm/mes y ajusta GPD a exceso, incorporando tendencias climaticas.

para GPD(POT)

$$P(X > x | X > u) = (1 + \xi \frac{x-u}{\sigma})^{\frac{-1}{\xi}} \tag{3.24}$$

donde u es el umbral, ξ la forma y σ la escala.

para el caso no estacionario

$$\sigma(t) = \sigma_o + \beta \tag{3.25}$$

donde t es el tiempo , capturando el cambio climático.

De American Academy of Actuaries(2024)[6] el ACRI incluye componentes EVT para extremos.

Definición : EVT en ACRI modela el riesgo extremos como distribuciones de pérdidas agricolas.
Es probabilistico, con foco en retornos de nivel por ejemplo Var climatico,esto se aplica a umbrales hondureños ajustado GPD para precipitación o temperaturas.

Integración actuarial y simulación. Aqui se integra EVT con el modelo econometrico para shocks en producción, simulación Monte Carlos y métricas de riesgo [9] Tratar riesgos de seguros y ruina , aplicable a integración actuarial[10] no proporciona la siguiente definición.

Definición : La probabilidad de ruina mide el riesgo de que pérdida excedan reservas, usando distribuciones agregadas.

Es actuarial, con foco en VAR y TVAR para primas.En el modelos hondureños, traduce shocks EVT en perdidas de producción simulando escenarios.

La probabilidad de ruina en modelos Cramer-Lundberg

$$\psi(u) = P(\tau < \infty | R(0) = u) \tag{3.26}$$

donde τ es el tiempo de ruina, R(t) el proceso de reserva.

Para VAR

$$VaR_q = inf x : P(L > x) \leq 1 - q \tag{3.27}$$

donde L es la pérdida agregada.

De Jiménez Hernándezet al [12] analiza probabilidad de ruina en Cramer Lundberg y considera la siguiente definición

Definición: El modelo de Creamer Lundberg modela flujos de primas y reclamos para riesgos actuariales.

Tambien asume procesos de Poisson para reclamos, con distribución exponencial. Aplica a simulaciones peŕdidas agricolas Hondureñas por extremos climaticos.

El proceso de reserva es el siguiente

$$R(t) = u + ct - S(t) \tag{3.28}$$

donde c es la tasa de primas, S(t) los reclamos acumulados, y la probabilidad de la ruina es

$$\psi(u) = \frac{\lambda}{\mu c} e^{(\mu c - \lambda)\frac{u}{\sigma^2}} \tag{3.29}$$

luego de Osepa et al [5] combina EVT con simulación para métricas de riesgo.Aplicando simulacion Monte Carlo genera distribuciones de pérdidas agregadas.Usa miles de iteraciones para estimar VaR/TVaR.Simula escenarios futuros con shocks EVT en producción Hondureña. Para la perdida agregada se definine de la siguiente manera

$$L = \sum_{i=1}^{N} X_i \tag{3.30}$$

donde X_i son pérdidas individuales, simuladas con EVT.

De Sarker et al [4] vincula extremos con impactos económicas para simulación. Eventos extremos generan shocks en variables económicas.

Con esta definición nos permite calcular EL , VAR. Integra con simulación para distribuciones de peŕdidas . La perdida esperada se define como

$$EL = E[L] \tag{3.31}$$

con VAR con en Dickson[10]

Resultados

En esta sección, se presentan los resultados empíricos obtenidos al aplicar el modelo actuarial econométrico basado en EVT a datos agrícolas hondureños. Los análisis se basan en series temporales de producción agrícola ejemplo, maíz y café, obtenidas de SENASA/IHCIT, variables climáticas precipitación y temperatura de COPECO, y el Índice de Riesgo Climático de los Actuarios (ACRI) de la American Academy of Actuaries (2024)[6]. Los datos se procesaron para homogeneizar series intertemporales, ajustando tendencias no climáticas y estacionalidades, como se describe en la metodología[8]. Los modelos se estimaron utilizando software estadístico ejemplo, R o Python, y los parámetros se validaron mediante pruebas de estacionariedad, cointegración y bondad de ajuste ejemplo, AIC, BIC y pruebas de Kolmogorov-Smirnov para EVT.

Estimación de parámetro econométrico y extremos. Aquí esta una subsección profundiza en las estimaciones de los parámetros cruciales, de modelos econométricos fundacionales, tales como ARIMA-X y cointegración, junto a los componentes EVT, todos ellos aplicados a las series de Honduras. Dichos hallazgos, combinan dinámicas temporales con eventos extremos, facultando la cuantificación precisa de la relación que existe, entre los factores climáticos y el rendimiento de la agricultura, como predijo la integración metodológica. Para el modelo econométrico central, se acomodó un ARIMA-X(1,1,1) empleando variables exógenas climáticas, un poco parecido a los planteamientos de Zulfiqar et al. (2024)[3] y Enders (2014)[8]. Se modeló la serie de producción agrícola Y_t, expresada en toneladas por hectárea, considerando la precipitación acumulada (P_t) y la temperatura media (T_t) como

variables exógenas, para capturar tanto tendencias estacionales como choques climáticos.

$$Y_t = 0{,}85 + 0{,}72Y_t - 1 + 0{,}45\epsilon_t - 1 + 0{,}28P_t - 0{,}15T_t + \epsilon_t \tag{3.32}$$

donde

- El coeficiente para P_t que es 0.28, sugiere una relación positiva entre las lluvias y la producción agrícola, donde un aumento del 1 % en la pluviosidad resulta en un alza del 0.28 % en la cosecha, esto muestra lo mucho que la agricultura de Honduras depende de esas lluvias de temporada [3].
- El coeficiente para T_t con -0.15, señala el efecto dañino de las temperaturas altas, algo bien común en estos tiempos de cambio climático[4].
- Para evitar problemas, se corrigió la raíz unitaria usando diferenciación (d=1), y comprobamos que todo estaba en orden con la prueba de Dickey-Fuller aumentada (p-valor < 0.05). El AIC del modelo alcanzó 125.4, más alto que los otros modelos que no incluían variables externas.

Para el análisis de cointegración, aplicaron el método de Ly y colegas (2024)[1]. Así se estudió la relación a largo plazo, entre producción agrícola (Y_t) y un índice mezclado de eventos extremos (EXT_t), este ultimo se saca de las variaciones en lluvias y temperatura. Johansen cointegration test usaron, estadístico de traza 18.5 y un p-valor $<$ a 0.01 comfirmo una relación de equilibrio

$$Y_t = 2{,}1 + 0{,}65EXT_t + \epsilon_t \tag{3.33}$$

donde ϵ_t es estacionario una prueba ADF en los residuos un p-valor $<$ que 0.05 esto indica que los impactos climáticos extremos como por ejemplo las sequías prolongadas generan desviaciones persistentes en la producción y de esta forma evitan regresiones espurias.[1].

En modelado de eventos extremos, usando EVT, el método Peaks Over Threshold (POT) se implementó para datos sobre umbrales, basándose en Coles (2001)[7] y Van Tassell (2024)[2]. Respecto a las sequías o precipitación mínima, un umbral u igual a 50 mm/mes fue definido, derivado de percentiles historicos de Honduras. Para los desbordamientos de lluvia, se aplicó un umbral u = 300 mm/mes.

$$\sigma(t) = 45{,}2 + 0{,}08t,\ \xi = 0{,}25 \tag{3.34}$$

donde

- $\xi = 0.25$ sugiere colas pesadas tipo Fréchet, perfectas para eventos extremos inusuales, pongamos por ejemplo, inundaciones en Honduras[7].
- $\sigma(t)$ crecé con el pasar del tiempo t, lo que revela la intensificación de eventos extremos causados por el cambio climático, con un asombroso aumento anual del 8 % en la escala[6].
- Se convalidó la exactitud del ajuste con el estadístico Kolmogorov-Smirnov (p-valor > 0.05) y el umbral óptimo se escogió con el método de Hill, asegurando firmeza en las colas[2].

Estos parámetros se integran con el ACRI (2024)[6], donde la frecuencia de extremos (FREQ) se estimó en 0.12 eventos/año ($\lambda \approx 0.12$), y la intensidad (INT) en 1.8 (basado en GPD), corroborando riesgos agrícolas hondureños.

Distribución de pérdida agregada y métricas de riesgo. La distribución de pérdidas agregadas, se expone aquí, se derivó de simulaciones Monte Carlo combinadas con EVT y el modelo actuarial, inspiradas en Klugman et al. de 2019[9] y Vanalle et al. del 2012[13]. Pérdidas que se miden como bajas en la producción agrícola a causa de eventos severos, expresadas como el porcentaje del rendimiento anticipado.

La pérdida agregada S se modela como una suma aleatoria de eventos extremos, donde la frecuencia $N \sim \text{Poisson}(\lambda = 0.12)$ y severidad $\xi \sim \text{GPD}(\xi = 0.25,\ \beta = 45.2)$, esto según residuos de ARIMA-X [9]. Para descubrir la distribución empírica de S, se llevaron a cabo, 10,000 simulaciones Monte Carlo, mostrando, sí una media de pérdidas anuales del 15 % en la producción agrícola hondureña.

Los momentos clave son[9]:

$$E[S] = \lambda E[X_i] \approx 0{,}12\ x\ 52{,}3 = 6{,}28\,\% \tag{3.35}$$

$$Var(S) = \lambda Var(X_i) + Var(N)(E[X_i])^2 \approx 0{,}12\ x\ 1200 + 0{,}12\ x\ (52{,}3)^2 \tag{3.36}$$

$$\approx 144 + 327 \approx 471\,\%$$

Donde E[X_i] y Var(X_i) surgen de la GPD, ilustrando la variabilidad extrema que enfrenta Honduras en su clima[7].

Las métricas de riesgo comprenden el Value at Risk (VaR) y el Tail Value at Risk (TVaR) las cuales son calculadas en el nivel q = 0.95(riesgo del 5 %):

$$VaR_{0,95} = \text{ínf}\, x : P(S > x) \leq 0{,}05 \approx 18{,}5\,\% \tag{3.37}$$

$$TVaR_{0,95} = E[S|S > VaR_{0,95}] \approx 25{,}2\,\% \tag{3.38}$$

Esas métricas sugieren, que en el 5 % de los escenarios más críticos, las pérdidas en la agricultura superan el 18.5 %, promediando condicionalmente un 25.2 %, revelando el golpe que asestan eventos como las sequías en Honduras[6]. La distribución empírica presenta asimetría positiva, un sesgo de 1.2, junto con colas pesada verificadas por EVT, esto respalda el empleo de GPD en lugar de distribuciones normales[2].

Calculo de prima actuarial. Esta sección evalúa primas actuariales equilibradas, fundadas en el modelo de riesgo total y el chance de insolvencia, basándose en Dickson (2016)[10] y Jiménez Hernández y Maldonado Santiago (2011)[12]. Las primas cubren perdidas anticipadas y riesgos considerables, combinando EVT con el modelo Cramér-Lundberg, esto para mantener la estabilidad financiera de los seguros agrícolas en Honduras.

El cálculo de la prima neta (c_n) requiere hallar el valor esperado de las pérdidas, y añadir un margen de seguridad, como el 10 % para cubrir esa variabilidad[9].

$$c_n = (1 + \theta) \times E[S] \approx 1{,}1 \times 6{,}28\,\% = 6{,}91\,\% \tag{3.39}$$

donde θ= 0.1 es el margen actuarial, reflejando incertidumbre climática[1].

Para primas brutas, se incorpora la probabilidad de ruina $\psi(u)$, en el modelo Cramér-Lundberg, con reservas iniciales u = 20 % basado en capital agrícola hondureño y primas constantes c[10]:

$$\psi(u) = \frac{\lambda\mu}{c} \exp\left(-\frac{u(c-\lambda\mu)}{\lambda\mu^2}\right) \tag{3.40}$$

Donde μ, qué resulta ser igual a E$[X_i]$ aproximadamentente a 52.3 %, y λ igual a 0.12. Para un c = 8 % la prima bruta ajustada, si tenemos que $\psi(u) \approx 0.03$ es un 3 % de probabilidad de la ruina, cumpliendo los criterios de solvencia ($\psi(u) < 0.05$),[12] . Esto, se confirma con simulaciones exhibiendo que primas menores del 7 % incrementan el peligro de insolvencia, ya en escenarios, extremos[5].

Integrando con ACRI [6] las primas escalan según exposición económica ,EXP $\approx$ 0.4, basado en el PIB agrícola hondureño, dando primas variadas, un 7.5 % en sitios muy peligrosos para el clima (ejemplo las costa) y 6.2 % donde no hay tanta amenaza. Estos cálculos respaldan seguros indexados climáticos minimizando los riesgos de productores hondureños[4].

Estos resultados claramente demuestran la eficacia del modelo, para evaluar los riesgos climáticos agrícolas en Honduras; así ofreciendo herramientas importantes para las políticas de mitigación y seguros. Sin embargo, las limitaciones surgen debido a la dependencia de datos históricos y supuestos de independencia. Próximamente, podrían usarse extensiones con machine learning, para ajustes dinámicos[5].

Concluciones

1. Combinando ARIMA-X, para comprender la dinámica temporal y la cointegración a largo plazo , y EVT para modelar eventos extremos tales como sequías e inundaciones , resulto en una robusta estimación de riesgos. Los parametros GPD ($\xi = 0.25$, σ(t) creciente) revelaron colas pesadas en las distribuciones de perdidas, confirmando la vulnerabilidad de la agricultura hondureña frente a eventos climáticos poco comunes pero con gran impacto, como aquellos ligados al cambio climático .
2. La simulaciones de pérdidas agregadas revelaron un promedio anual del 6.28 % en la producción agrícola, ademas un VaR al 95 % de 18.5 %, junto con un TVaR del 25.2 % . estos resultados revelan el efecto de los climas extremos, sobrepasando cálculos tradicionales que obvian colas pesadas, y justifican la importancia de los métodos EVT para escenarios no estáticos .
3. Las primas netas se calcularon al 6.91 %, pero subieron al 7-8 % como primas brutas, esto para conservar una chance de ruina menor al 3 % según el modelo Cramér-Lundberg . Así es más fácil crear seguros climáticos indexados, distintos por zonas como las costas con más riesgos de modo que se disminuye la vulnerabilidad económica de los agricultores hondureños .
4. Los resultados respaldan las estrategias de mitigación en Honduras, por ejemplo, inversiones en infraestructura que resiste eventos extremos (ejemplo sistemas de riego) y seguros subvencionados, todo esto en linea con el ACRI .Al combinar el EVT con la simulación Monte Carlo , el modelo ofrece herramientas para hacer pronósticos precisos, algo fundamental en un pais donde

la agricultura constituye una porción importante del PIB y es muy susceptible a las cambiantes condiciones climáticas.

5. Esta investigación profundiza la literatura usando EVT actuarial en sectores agrícolas en crecimiento,además de exhibir cómo residuos de modelos econométricos pueden alimentar análisis de extremos. Además, se atestiguan propiedades de EVT como invarianza y estabilidad de colas , además valida la integración metodológica para evitar regresiones en series no estacionarias .

Referencias

1. Ly, S., Riam, M., and Hizam, H. *Cointegration Analysis of Crop Yields and Extreme Weather Factors Using Actuaries Climate Index with Application of Bonus–Malus System*, Risks, Vol. 12, 2024.
2. Van Tassell, G. H. *Utilizing Extreme Value Theory to Uncover Yield Distributions from Farm and County Level Historical Corn Yields*, Ph.D. Thesis, University of Nebraska-Lincoln, 2024.
3. Zulfiqar, F., et al. *Agricultural Forecasting in a Changing Climate: ARIMA-X Model of Cereal Production in Tanzania*, ResearchGate (Preprint), 2024.
4. Sarker, S. A., and Sarker, M. A. R. *Extreme weather events, climate expectations, and agricultural export dynamics*, Economic Analysis and Policy, Vol. 83, pp. 696-708, 2024.
5. Osepa, E. R., and Mailafia, D. *Investment risk forecasting model using extreme value theory approach combined with machine learning*, AIMS Mathematics, Vol. 9, No. 8, pp. 19143-19172, 2024.
6. American Academy of Actuaries. *Actuaries Climate Risk Index (ACRI)*, Update Report, 2024.
7. Coles, S. *An Introduction to Statistical Modeling of Extreme Values*, Springer, London, 2001.
8. Enders, W. *Applied Econometric Time Series*, 4th Edition, John Wiley & Sons, New York, 2014.
9. Klugman, S. A., Panjer, H. H., and Willmot, G. E. *Loss Models: From Data to Decisions*, 5th Edition, John Wiley & Sons, New York, 2019.
10. Dickson, D. C. M. *Insurance Risk and Ruin*, 2nd Edition, Cambridge University Press, New York, 2016.
11. Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2019). *Loss Models From Data ro Decisions, 5 th., vol. 6, no. 1. Hoboken.*
12. Jiménez Hernández, J. D. C., & Maldonado Santiago, A. D. (2011). *Probabilidad de ruina en el modelo clásico de Cramer-Lundberg. REPOSITORIO NACIONAL CONACYT.*
13. Vanalle, R. M., Lucato, W. C., Vieira Júnior, M., & D Sato, I. (2012). *Uso de la Simulación Monte Carlo para la Toma de Decisiones en una Línea de Montaje de una Fábrica. Información tecnológica*, 23(4), 33-44.

Maestría en Matemática, Universidad Nacional Autónoma de Honduras.
Dirección de correo electrónico: axelcruzlopez@gmail.com

MODELOS VAR INTEGRADO CON VOLATILIDAD ESTOCASTICA MATRIZ EXPONENCIAL APLICADO A LOS TIPOS DE CAMBIO DE LA ALIANZA DEL PACIFICO

NELSON MOLINA MOLINA

Resumen. Los modelos autorregrsivos vectoriales (VAR) se emplean para capturar las relaciones dinámicas de series de tiempo multivariadas. Por otro lado, los modelos de volatilidad estocástica multivariada Matriz Exponencial (MESV) capturan la variabilidad cuando cambia en el tiempo , correlaciones dinámicas, y el efecto de apalancamiento. Por lo anterior, en la Tesis se propone la integración un modelo VAR y un modelo MESV (VAR-MESV). Para la elección del orden VAR-MESV mas adecuado se propone el uso del Criterio de Información de Desviación (DIC). Se presentarán formulas para estimar la asimetría de Mardia y la Curtosis de Koziol. Se hará una aplicación a los tipos de cambio de cuatro países de la Alianza del Pacifico (Chile, Colombia, México y Perú). Para estimar los paramétros se propone el uso de métodos de Monte Carlo vía Cadenas de Markov (MCMC).

Abstract. Vector autoregressive (VAR) models are used to capture the dynamic relationships of multivariate time series. On the other hand, Matrix Exponential Stochastic Volatility (MESV) models capture time-varying volatility, dynamic correlations, and leverage effects. Therefore, this thesis proposes the integration of a VAR model and a MESV model (VAR–MESV). To select the most suitable order of the VAR–MESV model, the Deviance Information Criterion (DIC) is proposed. Formulas will be presented to estimate Mardia's multivariate skewness and Koziol's kurtosis. An empirical application will be carried out using exchange rates from four Pacific Alliance countries (Chile, Colombia, Mexico, and Peru). For parameter estimation, Markov Chain Monte Carlo (MCMC) methods are proposed.

1. Introducción

Este trabajo tiene su motivación en el artículo de Cruz y Villafranca [1], en el cual integran un modelo autorregresivo vectorial (VAR) y un modelo de volatilidad estocástica multivariada con efecto de apalancamiento cruzado (MSV). En su propuesta, la parte VAR captura las relaciones dinámicas entre las series temporales multivariadas, mientras que la parte MSV captura la variabilidad de las series cuando cambia en el tiempo. Para estimar el modelo utilizan métodos de Monte Carlos Via Cadenas de Markov (MCMC).

Aunque el modelo propuesto por Cruz y Villafranca [1] es capaz de medir el efecto de apalancamiento cruzado —es decir, el efecto de un choque de las variables endógenas en el tiempo t sobre los choques de la volatilidad en el tiempo $t + 1$—,

Fecha: Agosto 2025.

Palabras y frases clave. Volatilidad estocástica, tipo de cambio, matriz exponencial, apalancamiento cruzado,dinámica multivariada.

asume correlaciones constantes en en choques de las variables endogenas. Sin embargo, en muchas aplicaciones es importante permitir que estas correlaciones varíen en el tiempo. Por ejemplo, las series de tiempo financieras tienden a moverse juntas en tiempos de crisis (alta correlación), mientras que en épocas de estabilidad suelen presentar menor correlación.

Por lo anterior, el primer objetivo de la Tesis es proponer un modelo VAR integrado con un modelo de volatilidad estocástica multivariada con efecto de apalancaminto cruzado que permita que las correlaciones de los choques de las varibles endogenas varíen en el tiempo. Para este fin, se integrará al modelo VAR el modelo de volatilidad estocástica matriz exponencial (MESV) propuesto por Ishihara, Omori y Asai [3]. El segundo objetivo es proporcionar algunas propiedades del modelo VAR-MESV, por ejemplo, la asimetría de Mardia y la curtosis de Koziol. Este objetivo es motivado por una conferencia de Cruz y Villafranca [4] en la que exponen propiedades de asimetría y curtosis de un modelo VAR-MSV integrado con una distribución t Student. El tercer objetivo es proporcionar una metodología con el fin de ajustar el modelo VAR-MESV para estimar los parámetros (se adaptara la metodología de Cruz y Villafranca [1]).

El cuarto objetivo de la Tesis es aplicar el modelo a datos simulados y reales con el propósito de responder las siguientes preguntas:

1. Se simularan datos con el modelo VAR-MESV en que haya periodos de alta correlación, periodos de baja correlación y periodos de correlación constante con las siguientes configuraciones: (a) Sin apalancamiento Cruzado, (b) Con apalancamiento Cruzado. Se estimaran los datos simulados con cuatro configuraciones de modelos: (a) Un modelo VAR-MESV sin apalancamiento cruzado, (b) Un modelo VAR-MESV con apalancamiento cruzado, (c) Un modelo VAR-MSV sin apalancamiento cruzado, (d) Un modelo VAR-MSV con apalancamiento cruzado. Esto se hara con el propósito de responder la pregunta ¿ Qué sucede si se estiman datos que fueron generados por un modelo VAR-MESV (con y sin apalancamiento cruzado) con un modelo VAR-MSV (con y sin apalancamiento cruzado)?.
2. Se estimaran datos reales consitentes en tipos de cambio de cuatro países de la Alianza del Pacífico (Chile, Colombia, México y Perú). La Alianza del Pacífico fue creada en el 2011 con el objetivo de impulsar un mayor crecimiento, desarrollo y competitividad de sus economías, promoviendo la libre circulación de bienes, servicios, capitales y personas. Por lo anterior, se ajustarán cuatro configuraciones de modelos: (a) Un modelo VAR-MESV sin apalancamiento cruzado, (b) Un modelo VAR-MESV con apalancamiento cruzado, (c) Un modelo VAR-MSV sin apalancamiento cruzado, (d) Un modelo VAR-MSV con apalancamiento cruzado. Esto se hará con el propósito de responder las preguntas ¿ Existe evidencia de un cambio en las relaciones a nivel de choques de los tipos de cambio antes y después del 2011? ¿ Existe evidencia de que los tipos de cambio pueden ser explicadas por observaciones pasadas? ¿ Existe evidencia de efecto de apalancamiento y apalancamiento cruzado en los tipos de cambio? ¿ Existe evidencia de que los tipos de cambio esten relacionados a nivel de variabilidad?.

Los objetivos antes mencionados se llevaran a cabo tomando como punto de partida los trabajos de Cruz y Villafranca [1, 4, 5]. Luego se adaptara al trabajo de Ishihara, Omori y Asai [3]. Todo lo expuesto en esta sección esta sujeto a cambios.

Línea de investigación

La investigación se enmarca en la línea de *Estadística multivariada y modelos lineales generalizados*, dado que integra modelos dinámicos multivariados (VAR y MESV) para analizar la evolución conjunta y la interdependencia de múltiples variables económicas en este caso, los tipos de cambio de los países de la Alianza del Pacífico. El estudio emplea herramientas propias de la estadística multivariada, como el análisis de covarianzas, medidas de asimetría y curtosis, y métodos bayesianos de estimación mediante cadenas de Markov Monte Carlo (MCMC). Además, el modelo propuesto contribuye al desarrollo de nuevas técnicas de inferencia en contextos multivariados dinámicos, fortaleciendo la investigación científica en modelización estadística dentro del eje prioritario "Cultura, ciencia y educación" de la Universidad Nacional Autónoma de Honduras (UNAH).

2. Justificación

El análisis de la volatilidad en los mercados financieros resulta fundamental para comprender la transmisión de choques económicos y la interacción entre activos en economías abiertas. En el contexto de los países de la Alianza del Pacífico (Chile, Colombia, México y Perú), el tipo de cambio desempeña un papel determinante en la competitividad, la estabilidad macroeconómica y la formulación de políticas monetarias. No obstante, los modelos tradicionales como los GARCH o los VAR con varianza constante presentan limitaciones al asumir correlaciones fijas y dinámicas simplificadas.

El modelo de Volatilidad Estocástica Matriz Exponencial (MESV) propuesto por Ishihara, Omori y Asai [3] ofrece una alternativa robusta al garantizar la positividad definida de las matrices de covarianza mediante una transformación exponencial matricial, permitiendo además capturar correlaciones dinámicas y efectos de apalancamiento cruzado. Integrar este modelo dentro de un marco autorregresivo vectorial (VAR-MESV) proporciona una herramienta flexible para analizar la evolución conjunta de los tipos de cambio y sus volatilidades, incorporando tanto los efectos contemporáneos como los retardos en las relaciones entre países.

En el ámbito aplicado, ofrece evidencia empírica sobre la dependencia dinámica y cambiaria en la Alianza del Pacífico, información clave para la gestión del riesgo financiero y la estabilidad cambiaria. Desde el punto de vista científico, el proyecto fortalece el campo de la modelación estadística y econométrica, al extender los modelos tradicionales de volatilidad estocástica hacia un marco multivariado, dinámico y bayesiano más general. La propuesta del modelo VAR–MESV representa una integración innovadora entre la dependencia temporal capturada por el modelo autorregresivo vectorial (VAR) y la estructura de correlaciones dinámicas y apalancamiento cruzado del modelo de Volatilidad Estocástica Matriz Exponencial (MESV). Esta combinación proporciona una herramienta metodológicamente sólida y flexible para la inferencia bayesiana en contextos financieros complejos,

permitiendo analizar simultáneamente la dinámica de los rendimientos y la evolución temporal de sus covarianzas.

En términos metodológicos, el estudio se desarrolla dentro del área de Estadística Multivariada y Series de Tiempo, con énfasis en el desarrollo y aplicación de modelos lineales y dinámicos para el análisis de fenómenos económicos y financieros. El enfoque integra técnicas de modelización multivariada, como los modelos VAR, con extensiones que incorporan volatilidad estocástica, correlaciones dinámicas y efectos de apalancamiento cruzado, contribuyendo así al avance de la investigación aplicada en economía y finanzas.

3. Antecedentes

Uhlig [7] introduce la volatilidad estocástica multivariada sin restricciones en el contexto de los modelos autorregresivos vectoriales. El modelo que propuso es de la siguiente manera

$$Y_t = A_0 V_t + B_1 y_{t-1} + \ldots + A_k y_{t-k} + R_t^{-1}\varepsilon_t, \qquad t = 1, \ldots, n, \tag{3.1}$$

$$H_{t+1} = \frac{1}{\lambda} R_t^T \Sigma_t R_t, \qquad t = 0, \ldots, n-1, \tag{3.2}$$

donde

$$\varepsilon_t \sim N(0, I_p), \qquad \Sigma_t \sim \beta_p \left(\frac{v + c + kp}{2}, \frac{1}{2} \right),$$

Y_t , $t = 1-k, \ldots, n$ de dimensión $p \times 1$ son datos observables. V_t de dimensión $c \times 1$ denota regresores deterministas como una constante y una tendencia de tiempo. La matriz de coeficientes B_0 es de dimensión $p \times c$. Las matrices de coeficientes B_i , $i = 1, \ldots, k$ son de dimensión $p \times p$. $v > p - 1$ y $\lambda > 0$ son parámetros. ε_t , $t = 1, \ldots, n$ son de dimensión $p \times 1$. Σ_t , $t = 1, \ldots, n$ son de dimensión $p \times p$ distribuidos independientemente. R_t denota la descomposición de Cholesky superior de H_b y $\beta_m(a, b)$ denota la distribución beta multivariada. Uhlig [7] escogió la distribución beta multivariante para explotar una conjugación entre esa distribución y la distribución Wishart para que la integración sobre el choque no observado en la matriz de precisión se puede realizar de forma cerrada, lo que lleva a una generalización de las fórmulas estándar de filtro de Kalman, el problema de filtrado no lineal. El estudio de los modelos autorregresivos vectoriales bayesianos con volatilidad estocástica (BVAR–SV) se origina a partir del reconocimiento de que las relaciones macroeconómicas y financieras varían en el tiempo y no pueden capturarse adecuadamente mediante modelos estáticos. Uhlig [7] escogió dicha distribución por su conjugación con la distribución Wishart, lo que permite realizar la integración sobre la matriz de precisión de forma cerrada y obtener una generalización del filtro de Kalman para el problema de filtrado no lineal.

Posteriormente, Cogley [8] propuso una estrategia de filtrado bayesiano para estimar la tendencia de crecimiento de la "nueva economía". Su modelo autorregresivo vectorial bayesiano con parámetros que varían en el tiempo se expresa como:

$$Y_t = X_t^T \beta_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \Sigma_t),$$

donde X_t incluye constantes y rezagos de Y_t, y β_t evoluciona como una caminata aleatoria:

$$\beta_t = \beta_{t-1} + W_t, \qquad W_t \sim N(0, Q).$$

Para modelar la varianza, Cogley adopta una versión multivariada del modelo de Jacquier et al. [10], donde:

$$\Sigma_t = B^{-1} H_t (B^{-1})^T, \qquad \log(h_{it}) = \log(h_{it-1}) + \sigma_i \eta_{it}.$$

Este esquema permite capturar la evolución temporal de la volatilidad bajo un enfoque bayesiano plenamente jerárquico.

A continuación, Cogley y Sargent [9] extendieron este marco a la política monetaria mediante un modelo autorregresivo vectorial con parámetros y volatilidades estocásticas variables, estimando densidades posteriores de interés para la inflación, el desempleo y la tasa de interés. Su enfoque demostró que los cambios estructurales en la política económica pueden representarse adecuadamente dentro de un VAR con parámetros dinámicos.

En una línea complementaria, Primiceri [11] estimó un modelo autorregresivo vectorial estructural con parámetros variando en el tiempo (TVP–SVAR) con el propósito de estudiar las causas del bajo desempeño económico de Estados Unidos en los años setenta y ochenta. Su especificación general es:

$$Y_t = V_t + A_{1,t} Y_{t-1} + \ldots + A_{k,t} Y_{t-k} + B_t^{-1} \Sigma_t \varepsilon_t,$$

donde los parámetros siguen procesos estocásticos:

$$\beta_t = \beta_{t-1} + w_t, \quad \lambda_t = \lambda_{t-1} + \varrho_t, \quad \log(\sigma_t) = \log(\sigma_{t-1}) + \eta_t.$$

Este diseño permitió identificar cambios en la conducta de la política monetaria y del sector privado, así como medir su impacto sobre la dinámica macroeconómica.

Benati [12] aplicó posteriormente un modelo TVP–SVAR bayesiano similar para investigar la llamada *Gran Moderación* en el Reino Unido, mostrando que la disminución de la volatilidad macroeconómica y los cambios en política monetaria explican la estabilidad inflacionaria observada en las décadas recientes. De igual forma, Galí y Gambetti [13] utilizaron un modelo estructural con parámetros y volatilidades estocásticas variables para analizar los cambios en la economía estadounidense posteriores a la Segunda Guerra Mundial, destacando la relevancia de los procesos de volatilidad temporal.

Basándose en estas contribuciones, Gambetti et al. [14] propusieron un modelo similar para realizar pronósticos en tiempo real de variables macroeconómicas (desempleo, inflación y tasa de interés), evaluando su desempeño mediante errores cuadráticos medios y puntuaciones logarítmicas. Encontraron que los modelos con parámetros y volatilidad estocástica variables mejoran significativamente la capacidad predictiva respecto a modelos tradicionales.

Clark [15] incorporó la volatilidad estocástica dentro de un VAR bayesiano para realizar pronósticos de densidad en tiempo real de variables macroeconómicas de Estados Unidos, tales como el crecimiento del producto, desempleo, inflación y tasa de fondos federales. Más adelante, Clark y Ravazzolo [16] compararon la precisión predictiva de diferentes configuraciones de volatilidad (constante, estocástica, estacionaria, con colas pesadas y GARCH), concluyendo que los modelos con volatilidad estocástica y parámetros dinámicos son los más robustos para la predicción y la inferencia.

De manera más reciente, Chiu et al. [17] propusieron un modelo autorregresivo vectorial con errores t de Student y volatilidad estocástica, que permite capturar tanto la heterocedasticidad de baja frecuencia como los episodios de alta volatilidad y valores extremos. Este modelo, definido como:

$$Y_t = V + A_1 Y_{t-1} + \ldots + A_k Y_{t-k} + \Sigma_t^{1/2}\varepsilon_t, \qquad \varepsilon_t \sim N(0, I_p),$$
$$\Sigma_t = B^{-1} H_t (B^{-1})^T, \qquad \ln(h_{it}) = \ln(h_{i,t-1}) + \eta_{it},$$

integra colas pesadas en la estructura de los choques, ofreciendo una representación más realista frente a valores atípicos y choques extremos.

Mumtaz [18, 19] desarrolló versiones generalizadas de los modelos VAR–SV, incluyendo volatilidad en la media y algoritmos MCMC optimizados, ampliando su aplicabilidad a contextos financieros internacionales. Del mismo modo, Ding et al. [20] emplearon un TVP–SVAR–SV para estudiar los efectos cambiantes de la incertidumbre financiera y geopolítica sobre los mercados de materias primas, destacando la utilidad de los enfoques bayesianos de volatilidad estocástica para analizar interdependencias complejas y no lineales.

El modelo de (MSV) propuesto por Ishihara y Omori [2] y adaptado por Cruz y Villafranca [1] permite capturar la heterocedasticidad condicional y las posibles no linealidades en las relaciones simultáneas entre variables endógenas (ver Primiceri [11]). Su formulación general es la siguiente:

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_k y_{t-k} + V_t^{1/2}\varepsilon_t, \qquad t = 1, \ldots, n, \tag{3.3}$$

$$\alpha_{t+1} = \Phi\alpha_t + \eta_t, \qquad t = 1, \ldots, n-1, \tag{3.4}$$

$$\alpha_1 \sim N_p(0, \Sigma_0), \qquad V_t = \operatorname{diag}\big(\exp(\alpha_{1t}), \ldots, \exp(\alpha_{pt})\big), \tag{3.5}$$

donde ν es un vector de interceptos de dimensión $p \times 1$, A_i son matrices de coeficientes $p \times p$, $\varepsilon_t \sim N(0, \Sigma_{\varepsilon\varepsilon})$ y los procesos de volatilidad $\alpha_t = h_t - \mu$ evolucionan según una caminata autorregresiva de primer orden con matriz de persistencia $\Phi = \operatorname{diag}(\phi_1, \ldots, \phi_p)$.

Los choques conjuntos $(\varepsilon_t, \eta_t)'$ siguen una distribución normal multivariada:

$$\begin{pmatrix}\varepsilon_t \\ \eta_t\end{pmatrix} \sim N\left(0, \Sigma = \begin{pmatrix}\Sigma_{\varepsilon\varepsilon} & \Sigma_{\varepsilon\eta} \\ \Sigma_{\eta\varepsilon} & \Sigma_{\eta\eta}\end{pmatrix}\right),$$

donde la matriz $\Sigma_{\varepsilon\varepsilon}$ captura la correlación entre los choques de las variables endógenas, $\Sigma_{\eta\eta}$ la correlación entre los choques de la volatilidad, y $\Sigma_{\varepsilon\eta}$ el *efecto de apalancamiento cruzado* (*cross leverage effect*) que vincula los choques contemporáneos de las variables con los de la volatilidad futura (ver Ishihara et al. [3]).

Finalmente, la condición de estacionariedad se garantiza mediante

$$\operatorname{vec}(\Sigma_0) = \big(I_{p^2} - \Phi \otimes \Phi\big)^{-1} \operatorname{vec}(\Sigma_{\eta\eta}),$$

lo cual asegura la existencia de una solución estable para la dinámica estocástica de la varianza. Asimismo, los trabajos de Ishihara, Omori y Asai [2, 3] contribuyeron decisivamente al desarrollo de los modelos *Matrix Exponential Stochastic Volatility* (MESV) con apalancamiento cruzado, consolidando la estimación bayesiana eficiente para sistemas multivariados de gran dimensión.

En conjunto, estas contribuciones constituyen la evolución metodológica que sustenta la presente investigación, la cual busca modelar la dinámica macroeconómica bajo un enfoque bayesiano con parámetros y volatilidades estocásticas variables,

incorporando los avances recientes de la literatura en estimación multivariada y errores con colas pesadas.

4. Modelo VAR-MESV

En esta sección se presenta el cuerpo central de este trabajo. El punto de partida de la tesis titulada *Modelos VAR Integrado con Volatilidad Estocástica Matriz Exponencial Aplicado a los Tipos de Cambio de la Alianza del Pacífico* es el trabajo de Cruz y Villafranca [1]; por lo tanto, en primer lugar se describe detalladamente el modelo VAR-MSV propuesto por Cruz y Villafranca. Posteriormente, se presenta la metodología que emplean para la estimación de los parámetros y la determinación del mejor orden del modelo VAR-MSV(k). Se expone el modelo de volatilidad estocastica matriz exponencial propuesto por Ishihara, Omori y Asai [3].Por ultimo,se precenta el modelo autorregresivo vectorial integrado con volatilidad estocástica matriz exponencial (VAR-MESV),junto con su metodología de estimación de parámetros y el procedimiento para seleccionar el mejor orden VAR-MESV(k).

4.1 Modelo VAR-MSV

El modelo propuesto por Cruz y Villafranca [1] es un modelo autorregresivo vectorial con volatilidad estocástica multivariada . La volatilidad estocástica modelada es la propuesta por Ishihara y Omori [2]. En este modelo, los choques de las variables endógenas están correlacionados, y se diseñó así para capturar las posibles relaciones lineales entre ellos. De igual manera, los choques de la volatilidad están correlacionados. Los choques de las variables endógenas en el tiempo t y los choques de la volatilidad en el tiempo $t+1$ están correlacionados y, de esta forma, medir el efecto de los choques de las variables endógenas en el tiempo t en los choques de la volatilidad en el tiempo $t+1$. De esta manera se puede medir el efecto de los choques económicos en la varianza condicional de las variables macroeconómicas .

Las matrices de coeficientes estan diseñadas para medir la dependencia lineal de las observaciones pasadas en las observaciones actuales,en otras palabras, miden la fuerza con las que las observaciones pasadas afectan las actuales. El modelo VAR-MSV es de la siguiente manera

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_k y_{t-k} + w_t, \quad w_t = V_t^{\frac{1}{2}} \varepsilon_t, \quad t = 1, \ldots, n, \tag{4.1}$$

$$\alpha_{t+1} = \Phi \alpha_t + \eta_t, \quad t = 1, \ldots, n-1, \tag{4.2}$$

$$\alpha_1 \sim N_p(0, \Sigma_0), \tag{4.3}$$

$$V_t^{\frac{1}{2}} = \operatorname{diag}\left(\exp\left(\frac{\alpha_{1t}}{2}\right), \ldots, \exp\left(\frac{\alpha_{pt}}{2}\right)\right), \tag{4.4}$$

$$\Phi = \operatorname{diag}(\phi_1, \ldots, \phi_p), \tag{4.5}$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim N_{2p}(0, \Sigma), \quad \Sigma = \begin{pmatrix} \Sigma_{\varepsilon\varepsilon} & \Sigma_{\varepsilon\eta} \\ \Sigma_{\eta\varepsilon} & \Sigma_{\eta\eta} \end{pmatrix}, \tag{4.6}$$

$$\operatorname{vec}(\Sigma_0) = (I_{p^2} - \Phi \otimes \Phi)^{-1} \operatorname{vec}(\Sigma_{\eta\eta}). \tag{4.7}$$

donde y_t, $t = -k+1, \ldots, 0, 1, \ldots, n$, son variables disponibles de dimensión $p \times 1$. ε_t, $t = 1, \ldots, n$, son choques gaussianos de dimensión $p \times 1$. El vector ν es un término de intercepción de dimensión $p \times 1$, mientras que A_i, $i = 1, \ldots, k$, son matrices

de coeficientes de dimensión $p \times p$. $\alpha_t = h_t - \mu_h$ es de dimensión $p \times 1$, donde h_t es el correspondiente vector de log volatilidad y μ_h es el vector media. El operador vec convierte una matriz $A = [a_1, \ldots, a_p]$, a_i es de dimensión $p \times 1$, $i = 1, \ldots, p$, en un vector $(a_1^T, \ldots, a_p^T)^T$ de dimensión $p^2 \times 1$. El operador $\otimes$ es el producto kronecker.

Los elementos de las matrices de coeficientes A_{ij}^l, $j = 1, \ldots, p$, $l = 1, \ldots, k$, denotan la dependencia lineal de y_{it} (valor actual de la serie i) en $y_{j,t-l}$, $j \neq i$, $l = 1, \ldots, k$ (valores pasados de las otras series) en la presencia de $y_{i,t-l}$, $l = 1, \ldots, k$ (valores pasados de la serie i). Por lo que, A_{ij}^l, $j = 1, \ldots, p$, $l = 1, \ldots, k$, es el efecto condicional de $y_{j,t-l}$, $j \neq i$, $l = 1, \ldots, k$, sobre y_{it} en presencia de $y_{i,t-l}$, $l = 1, \ldots, k$. Sí $A_{ij}^l = 0$ para todo l y $j \neq i$, entonces y_{it} no depende de $y_{j,t-l}$, $j \neq i$, $l = 1, \ldots, k$ pero si dé $y_{i,t-l}$, $l = 1, \ldots, k$. Por otro lado, sí $A_{ij}^l = 0$ para todo l y $j = i$, entonces y_{it} no depende de $y_{i,t-l}$, $l = 1, \ldots, k$ pero si dé $y_{j,t-l}$, $j \neq i$, $l = 1, \ldots, k$.

Los elementos de la matriz Φ en la ecuación 4.5 estan relacionados con la percistencia a los choques a la volatilidad y en su modelo $-1 < \phi_i < 1$, $i = 1, \ldots, p$. La percistencia de los choques a la volatilidad es el efecto del choque actual sobre el pronostico de la volatilidad y eventualmente desaparece.La vida media de un choque viene dada por $-\log(2)/\log|\phi_i|$, que en series de tiempo diarias,es el numero de dias transcurridos para que el impacto del choque se redusca a la mitad. Cuando ϕ_i es cercano a 1 y $\sigma_{ii,\eta\eta}$ es cercano a cero, la evolución de la volatilidad de una serie de tiempo es muy suave.

La volatilidad estocastica pretende capturar la posible heterosedastisidad de los choques y las posibles no linealidades en las relaciones simultaneas entre las varibles del modelo. En series de tiempo diarias, un día en el que $\alpha_t = 0$ puede ser visto como un día normal. Un día normal es uno en el que la velocidad de evolución de la volatilidad no es ni demaciada rápida ni demaciada lenta, en otras palabras, asume un valor promedio. Luego, $\sigma_{ii,\varepsilon\varepsilon}$ puede ser interpretado como la varianza condicional en un día normal. La varianza general de w_{it} es denotada por $\sigma_{ii,ww}$ y $100(1 - \sigma_{ii,\varepsilon\varepsilon}/\sigma_{ii,ww})$ es el porcentaje de la varianza que es atribuida a la presencia de heterocedasticidad en la serie temporal i. El flujo de la volatilidad de $w_{ii,ww}$ es dado por $\exp 0{,}5\sigma_{ii,\eta\eta}/(1 - \phi_i^2)$.

Para encontrar la función de verosimilitud del modelo dado por las ecuaciones 4.2. Las submatrices $\Sigma_{\varepsilon\varepsilon}$ y $\Sigma_{\eta\eta}$ se usan para capturar la posible correlación entre los choques de las variables endógenas y los choques de la volatilidad respectivamente. Además, la submatriz de covarianza $\Sigma_{\varepsilon\eta}$ se usa para calcular la posible correlación entre los choques de las variables endógenas en el mes actual y los choques de la volatilidad del siguiente mes.

Para encontrar la función verosimilitud del modelo de las ecuaciones (4.1)-(4.7) los autores realizaron una leve modificación a la verosimilitud propuesta por Ishihara y Omori [2], definiendo $Y_t = [1, y_{t-1}^T, \ldots, y_{t-k}^T]^T$ y $\beta = \text{vec}(v, A_1, \ldots, A_k)$ de dimensión $(kp + 1) \times 1$ y $(kp^2 + p) \times 1$ respectivamente. Luego, reescribieron el modelo de la ecuación (4.1) de la siguiente manera

$$y_t = (Y_t^T \otimes I_p)\beta + w_t \tag{4.8}$$

Luego se definen $\theta = (\phi, \Sigma, \beta, \nu)$, $\phi = (\phi_1, \ldots, \phi_p)^T$, $\alpha = (\alpha_1^T, \ldots, \alpha_n^T)^T$, $Y^n = (y_1, \ldots, y_n)$, $Y^k = (y_{-k+1}, \ldots, y_0)$ y $1_p = [1, \ldots, 1]^T$, y obtienen

$$\begin{aligned} f(\lambda, \alpha, Y^n|\theta, Y^k) &= f(Y^n, \alpha|\lambda, \theta, Y^k) f(\lambda|\theta, Y^k) \\ &\propto \exp\left\{ \sum_{t=1}^{n} \ell_t - \frac{1}{2}\alpha_1^T \Sigma_0^{-1} \alpha_1 - \frac{1}{2}\sum_{t=1}^{n-1} (\alpha_{t+1} - \Phi\alpha_t)^T \Sigma_{\eta\eta}^{-1} (\alpha_{t+1} - \Phi\alpha_t) \right\} \\ &\quad \times \left(\prod_{t=1}^{n} \lambda_t^{\frac{p+\nu}{2}-1} \right) |\Sigma_0|^{-1/2} |\Sigma|^{-(\frac{n-1}{2})} |\Sigma_{\varepsilon\varepsilon}|^{-1/2}, \end{aligned} \tag{4.9}$$

donde

$$\begin{aligned} \ell_t =& -\frac{1}{2}(y_t - ((Y_t^T \otimes I_p)\beta + \mu_t))^T \Sigma_t^{-1} (y_t - ((Y_t^T \otimes I_p)\beta + \mu_t)) \\ & - \frac{1}{2} 1_p^T \alpha_t + \text{const}, \end{aligned} \tag{4.10}$$

$$\mu_t = V_t^{\frac{1}{2}} m_t, \tag{4.11}$$

$$\Sigma_t = V_t^{\frac{1}{2}} S_t V_t^{\frac{1}{2}}, \tag{4.12}$$

$$m_t = \begin{cases} \Sigma_{\varepsilon\eta}\Sigma_{\eta\eta}^{-1}(\alpha_{t+1} - \Phi\alpha_t), & t < n, \\ 0, & t = n, \end{cases} \tag{4.13}$$

$$S_t = \begin{cases} \Sigma_{\varepsilon\varepsilon} - \Sigma_{\varepsilon\eta}\Sigma_{\eta\eta}^{-1}\Sigma_{\varepsilon\eta}, & t < n, \\ \Sigma_{\varepsilon\varepsilon}, & t = n. \end{cases} \tag{4.14}$$

4.2 Metodo de Estimación del Modelo VAR-MSV

Para estimar los parametros los autores usan inferencia Ballesiana calculando las distribuciones a posteriori por medio del algoritmo MCMC de seis bloques que es dado por

1. Inicializar $\alpha, \phi, \Sigma, \beta$.
2. Generar $\beta|\alpha, \phi, \Sigma, Y^n, Y^k$.
3. Generar $\alpha|\phi, \Sigma, \beta, Y^n, Y^k$.
4. Generar $\Sigma|\beta, \alpha, \phi, Y^n, Y^k$.
5. Generar $\phi|\Sigma, \beta, \alpha, Y^n, Y^k$.
6. Ir a 2.

Para generar β encuentran la función de densidad posterior en forma cerrada, escogiendo la distribución priori $f(\beta)$ de Litterman [23, 24], la cual corresponde a una distribución normal multivariante con media priori μ_β y matriz de covarianza priori Σ_β. Usan el algoritmo del muestreador de Gibbs para generar una muestra.

Para generar α aplican el método muestra de múltiples movimientos de Ishihara y Omori [2], sustituyen y_t por la serie transformada $y_t^* = y_t - (Y_t^\top \otimes I_p)\hat{\beta}$, donde $\hat{\beta}$ proviene del segundo paso del algoritmo MCMC de seis bloques. El método muestra de múltiples movimientos propuesto por Ishihara y Omori [2] divide $\alpha = (\alpha_1^T, \ldots, \alpha_n^T)^T$ en $K + 1$ bloques usando el algoritmo de Shephard y Pitt

[25]. Encuentran la distribución completa de densidad conjunta condicional de las perturbaciones del i-ésimo bloque y usan expansión de Taylor de segundo orden alrededor de la moda y la aproximan a una densidad normal que se usa para el algoritmo de Aceptación-Rechazo (AR). Como la dimensión de la matriz de covarianza crece cuando el tamaño de los bloques crece convierten la densidad normal aproximada en un modelo de espacios de estado auxiliar. Aplican el suavizador de perturbaciones de Koopman [26] repetidas veces al modelo de espacios de estados auxiliar para encontrar la moda y obtienen un modelo de espacios de estado gaussiano lineal aproximado. Por último, aplican un algoritmo de Metropolis-Hastings de Aceptación-Rechazo (AR-MH) en el que se utiliza un simulador de perturbaciones [27, 28] al modelo de espacios de estado gaussiano lineal aproximado para generar un candidato.

Para generar ϕ y Σ, sustituyen y_t por la serie transformada $y_t^* = y_t - (Y_t^\top \otimes I_p)\hat{\beta}$, donde $\hat{\beta}$ proviene del segundo paso del algoritmo MCMC de seis bloques. Las funciones de densidad a priori y las distribuciones posteriores condicionales de ϕ y Σ se toman de Ishihara y Omori [2]. Dado que las distribuciones condicionales completas no tienen forma cerrada, los autores emplean un paso de Metropolis–Hastings para generar las muestras correspondientes de ϕ y Σ.

4.3 Selección del Orden VAR-MSV

Para escoger el mejor modelo VAR-MSV usan la metodoligia propuesta por Ishihara y Omori [2]. Para cada modelo estimado, calculan el Criterio de Información de Desviación (DIC) de Spiegelhalter et al. [29]. La medida DIC es definida por

$$\text{DIC} = \mathbb{E}_{\theta|y^n}[D(\theta)] + P_D, \tag{4.15}$$

donde

$$P_D = \mathbb{E}_{\theta|y^n}[D(\theta)] - D\big(\mathbb{E}_{\theta|y^n}[\theta]\big), \quad D(\theta) = -2\log f(Y^n \mid \theta). \tag{4.16}$$

Para calcular $\mathbb{E}_{\theta|y^n}[D(\theta)]$, se puede aproximar mediante $\frac{1}{M}\sum_{m=1}^{M} D\big(\theta^{(m)}\big)$ donde $\theta^{(m)}$ son remuestreados a partir de la distribución posterior. El error estándar del estimador es obtenido estimando repetidamente $\mathbb{E}_{\theta|y^n}[D(\theta)]$. $D\big(\mathbb{E}_{\theta|y^n}[\theta]\big)$ es igual a $D(\theta)$ evaluado en la media posterior. Ishihara y Omori [2] configuraron $M = 100$, $I = 10000$ y repitieron 10 veces $\mathbb{E}_{\theta|y^n}[D(\theta)]$ para obtener el error estándar. Utilizan el filtro de partículas auxiliar propuesto por Shephard Pitt [30] para calcular la función verosimilitud ordinaria dado los parámetros $\log f(Y^n \mid \theta)$.

Para escoger el mejor orden VAR-MSV se aplican los siguientes pasos:

1. Suponiendo que se sabe que el orden VAR-MSV no puede exceder un entero K_1, se procede a estimar los modelos VAR-MSV comenzando desde 0 hasta K_1 y se almacenan sus parámetros estimados $\theta_0, \theta_1, \ldots, \theta_{K_1}$, donde θ_i son los parámetros estimados del modelo i.
2. Para cada modelo se sustituye y_t por $y_t^* = y_t - (Y_t^\top \otimes I_p)\hat{\beta}^i$, donde $\hat{\beta}^i$ son las matrices de coeficientes estimadas del modelo i. Luego se procede a calcular la correspondiente función verosimilitud ordinaria dado los parámetros $\log f(Y^n \mid \theta_i)$.
3. Se escoge el modelo que tenga la menor medida DIC.

4.4 Modelo de Volatilidad Estocastica Matriz Exponencial con Apalancamiento Cruzado

En esta sección se describe el modelo de Volatilidad estocastica matriz exponencial con efecto de apalancamiento cruzado (MESV) propuesto por Ishihara, Omori y Asai [3]. El modelo MESV se basa en la transformación exponencial matricial como se describe a continuación.

Sea $\mathbf{A}$ una matriz de dimensión $p \times p$, la exponencial de una matriz se define mediante el siguiente desarrollo en serie de potencias

$$\exp(\mathbf{A}) \equiv \sum_{s=0}^{\infty} \frac{1}{s!}\mathbf{A}^s,$$

donde la serie converge absolutamente si todos los autovalores de $\mathbf{A}$ son finitos. Para cualquier matriz simétrica real definida positiva $\mathbf{C}$,existe una matriz simétrica real $\mathbf{A}$ de dimensión $p \times p$ tal que $\mathbf{C} = \exp(\mathbf{A})$, y la matriz $\mathbf{A}$ se obtiene mediante la transformación logarítmica matricial. De forma recíproca, para cualquier matriz simétrica real $\mathbf{A}$, $\mathbf{C} = \exp(\mathbf{A})$ es una matriz simétrica definida positiva.

Si $\mathbf{A}$ es una matriz simétrica real de dimensión $p \times p$, entonces existe una matriz ortogonal $\mathbf{U}$ de dimensión $p \times p$ y una matriz diagonal $\mathbf{\Lambda}$ de dimensión $p \times p$ tal que $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$ y

$$\exp(\mathbf{A}) = \mathbf{U}\left(\sum_{s=0}^{\infty} \frac{1}{s!}\mathbf{\Lambda}^s\right)\mathbf{U}' = \mathbf{U}\exp(\mathbf{\Lambda})\mathbf{U}'.$$

Sea $\mathbf{y}_t = (y_{1t}, \ldots, y_{pt})'$ denotando el vector de retornos de activos de dimención $p \times 1$ en el tiempo t, y sea $\mathbf{H}_t$ denotando el logaritmo matricial de la matriz varianza-covarianza de $\mathbf{y}_t$. El modelo MESV con efecto de apalancamiento se define como

$$\mathbf{y}_t = \exp(\mathbf{H}_t/2)\,\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \text{i.i.d. } \mathcal{N}(0, \mathbf{I}_p), \quad t = 1, \ldots, n, \tag{4.17}$$

$$\mathbf{H}_{t+1} = \mathbf{M} + \tilde{\mathbf{\Phi}} \odot (\mathbf{H}_t - \mathbf{M}) + \mathbf{E}_t, \tag{4.18}$$

$$\begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{\eta}_t \end{pmatrix} \sim \text{i.i.d. } \mathcal{N}_{p+q}(0, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{I}_p & \mathbf{\Sigma}_{\varepsilon\eta} \\ \mathbf{\Sigma}_{\eta\varepsilon} & \mathbf{\Sigma}_{\eta\eta} \end{pmatrix}, \quad t = 1, \ldots, n-1, \tag{4.19}$$

$$\mathbf{h}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}, \mathbf{\Sigma}_0), \tag{4.20}$$

donde $\boldsymbol{\eta}_t = \text{vech}(\mathbf{E}_t)$, $q = p(p+1)/2$, $\mathbf{M} = \{\mu_{ij}\}$ y $\tilde{\mathbf{\Phi}} = \{\phi_{ij}\}$ son matrices simétricas $p \times p$ de parámetros, y $\odot$ denota el producto de Hadamard. Para la identificabilidad, fijamos la matriz de covarianza de $\boldsymbol{\varepsilon}_t$ igual a $\mathbf{I}_p$.

Si definimos $\mathbf{h}_t = \text{vech}(\mathbf{H}_t) = (h_{11,t}, h_{21,t}, \ldots, h_{p1,t}, h_{22,t}, \ldots, h_{pp,t})'$ como el vector columna apilado de los elementos de la parte triangular inferior de $\mathbf{H}_t$, entonces se tiene que

$$\mathbf{h}_{t+1} = \boldsymbol{\mu} + \mathbf{\Phi}(\mathbf{h}_t - \boldsymbol{\mu}) + \boldsymbol{\eta}_t, \tag{4.21}$$

donde $\boldsymbol{\mu} = \text{vech}(\mathbf{M}) = (\mu_{11}, \mu_{21}, \ldots, \mu_{p1}, \mu_{22}, \ldots, \mu_{pp})'$, $\mathbf{\Phi} = \text{diag}(\boldsymbol{\phi})$ (una matriz diagonal cuyos elementos diagonales son iguales a ϕ) y $\boldsymbol{\phi} = \text{vech}(\tilde{\mathbf{\Phi}}) = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1}, \phi_{22}, \ldots, \phi_{pp})'$.El número de parámetros en el modelo MESV es

$q(q+2p+3)/2$. La matriz de covarianza de la variable latente inicial, $\mathbf{\Sigma}_0$, se asume que satisface una condición de estacionariedad tal que

$$\text{vec}(\mathbf{\Sigma}_0) = (\mathbf{I}_{p^2} - \mathbf{\Phi} \otimes \mathbf{\Phi})^{-1}\text{vec}(\mathbf{\Sigma}_\eta) \tag{4.22}$$

. donde $\otimes$ es el producto kronecker.

Sea $\mathbf{\Sigma}_{\eta\eta} = \{\rho_{ij,\eta\eta}\, \sigma_{i,\eta\eta}\sigma_{j,\eta\eta}\}$, y $\mathbf{\Sigma}_{\varepsilon\eta} = \{\rho_{ij,\varepsilon\eta}\, \sigma_{j,\eta\eta}\}$ donde $\sigma_{i,\eta\eta}$ es la desviación estándar de η_{it} y $\rho_{ij,xy}$ es el coeficiente de correlación entre x_{it} y y_{jt}. Además, para mayor comodidad, utilizamos la notación $E(i,j) = k$ basada en la relación $\boldsymbol{\eta}_t = \text{vech}(\mathbf{E}_t)$, de modo que el elemento (i,j)-ésimo de $\mathbf{E}_t$, $\mathbf{E}_t(i,j)$, corresponde al elemento k-ésimo de $\boldsymbol{\eta}_t$, η_{kt}., es decir, $E(1,1) = 1,\ E(2,1) = 2, \ldots, E(p,1) = p,\ E(2,2) = p+1, \ldots, E(p,p) = p(p+1)/2$. Así, $\text{Cov}(\varepsilon_{lt}, \eta_{kt}) = \rho_{lk,\varepsilon\eta}\, \sigma_{k,\eta}$, es equivalente a $\text{Cov}(\varepsilon_{lt}, \mathbf{E}_t(i,j)) = \rho_{l\,E(i,j),\varepsilon\eta}\, \sigma_{E(i,j),\eta}$.

4.5 Modelo VAR-MESV

En esta sección se presenta el modelo autorregesivo vectorial integrado con volatilidad estocástica matriz exponencial con efecto de apalancamiento cruzado (VAR-MESV). La volatilidad estocástica matriz exponencial es la propuesta por Ishihara, Omori y Asai [3], la cual se describió en la subsección 4.4.

$$\mathbf{y}_t = \nu + \sum_{i=1}^{k} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{w}_t, \quad \mathbf{w}_t = \exp(\mathbf{H}_t/2)\, \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, n, \tag{4.23}$$

$$\mathbf{H}_{t+1} = \mathbf{M} + \tilde{\mathbf{\Phi}} \odot (\mathbf{H}_t - \mathbf{M}) + \mathbf{E}_t, \tag{4.24}$$

$$\begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{\eta}_t \end{pmatrix} \sim \text{i.i.d. } \mathcal{N}_{p+q}(0, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{I}_p & \mathbf{\Sigma}_{\varepsilon\eta} \\ \mathbf{\Sigma}_{\eta\varepsilon} & \mathbf{\Sigma}_{\eta\eta} \end{pmatrix}, \quad t = 1, \ldots, n-1, \tag{4.25}$$

$$\mathbf{h}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}, \mathbf{\Sigma}_0), \tag{4.26}$$

$$\text{vec}(\mathbf{\Sigma}_0) = (\mathbf{I}_{p^2} - \mathbf{\Phi} \otimes \mathbf{\Phi})^{-1}\text{vec}(\mathbf{\Sigma}_\eta) \tag{4.27}$$

4.6 Correlación Dinámica en Modelos Multivariados de Matriz Exponencial

En los modelos de volatilidad estocástica multivariada, la correlación dinámica se refiere a la evolución temporal de la dependencia entre los choques estructurales de un sistema multivariado. En el caso particular del modelo *Matrix Exponential Stochastic Volatility* (MESV), esta correlación surge directamente de la dinámica estocástica del logaritmo matricial H_t, el cual genera una matriz de covarianzas definida positiva en cada instante mediante la transformación exponencial:

$$\Sigma_t = \exp(H_t).$$

Dado que H_t evoluciona según un proceso estocástico matricial,

$$H_{t+1} = M + \widetilde{\Phi} \odot (H_t - M) + E_t,$$

la matriz Σ_t cambia continuamente en el tiempo, lo cual induce de forma natural una correlación que también varía dinámicamente. La correlación entre las series i

y j en el tiempo t se define como:

$$\rho_{ij,t} = \frac{\sigma_{ij,t}}{\sqrt{\sigma_{ii,t}\,\sigma_{jj,t}}}.$$

Por lo tanto, cualquier perturbación en la evolución de H_t —ya sea en su media, su persistencia o en los choques E_t— produce cambios simultáneos en las covarianzas y, en consecuencia, en las correlaciones.

Este mecanismo presenta varias ventajas metodológicas: (i) garantiza la positividad definida de Σ_t mediante la exponencial matricial; (ii) permite correlaciones completamente dinámicas sin imponer formas funcionales restrictivas; (iii) captura efectos de apalancamiento cruzado a través de la submatriz $\Sigma_{\varepsilon\eta}$, induciendo dependencia entre los choques contemporáneos y los de volatilidad futura; y (iv) representa adecuadamente fenómenos financieros como contagio, sincronización en crisis y divergencia en periodos estables.

En síntesis, en los modelos multivariados de matriz exponencial, la correlación dinámica no se especifica como un proceso separado, sino que emerge endógenamente de la evolución estocástica del logaritmo matricial H_t. Este enfoque es uno de los más flexibles y matemáticamente consistentes para modelar dependencia temporal en econometría financiera, y constituye la base teórica de modelos avanzados como el VAR–MESV.

4.7 Apalancamiento y Apalancamiento Cruzado.

El apalancamiento en modelos financieros describe el efecto mediante el cual los choques contemporáneos de una variable, especialmente los negativos, incrementan de forma desproporcionada su volatilidad futura. Este fenómeno surge típicamente a partir de una correlación negativa entre las innovaciones del retorno y las innovaciones del proceso de volatilidad, generando un aumento inmediato del riesgo ante movimientos adversos.

En el contexto multivariado, este mecanismo se extiende mediante el denominado apalancamiento cruzado, que caracteriza la interacción entre los choques contemporáneos de una variable y la volatilidad futura de otra diferente. Bajo esta estructura, un shock en la serie i puede modificar la evolución del proceso de volatilidad de la serie j, capturando efectos de transmisión, interdependencia y contagio entre activos o países. Este tipo de dependencia es fundamental en los modelos de volatilidad estocástica multivariada, pues permite representar de manera realista cómo perturbaciones en un mercado pueden influir en el nivel de incertidumbre de otros.

4.8 Trabajo futuro y consideraciones computacionales.

En el trabajo futuro se continuará profundizando en la aplicación del modelo a las series de tasas de cambio de los países de interés, así como en la realización de estudios de simulación que permitan evaluar el comportamiento del VAR–MESV bajo diferentes escenarios de volatilidad y dependencia dinámica. Sin embargo, es necesario señalar que este tipo de modelos posee una estructura altamente compleja, debido a la presencia de variables latentes, a la evolución matricial del proceso de volatilidad y a la necesidad de garantizar positividad definida mediante transformaciones exponenciales. Como resultado, el costo computacional asociado a su estimación es elevado, en particular cuando se utilizan métodos bayesianos basados

en algoritmos MCMC y cuando la dimensión del sistema aumenta. A pesar de estas exigencias computacionales, la riqueza estructural del modelo justifica el esfuerzo, ya que permite capturar de manera más realista la dinámica conjunta de las tasas de cambio y los mecanismos de transmisión entre economías interrelacionadas.

Conclusiones

A partir del análisis realizado, se concluye que la integración entre un modelo autorregresivo vectorial y un modelo de volatilidad estocástica matriz exponencial (VAR–MESV) constituye una estrategia metodológica robusta para estudiar la dinámica conjunta de variables financieras altamente interdependientes. Se evidencia que esta integración permite capturar no solo las dependencias temporales entre las series, sino también la evolución estocástica y flexible de sus covarianzas, aspecto fundamental en contextos donde las correlaciones cambian de forma significativa a lo largo del tiempo.

Se determina que los modelos VAR–MSV tradicionales presentan una limitación importante al asumir correlaciones constantes. A la luz de la teoría revisada, se reconoce que esta simplificación puede generar conclusiones sesgadas cuando se analizan fenómenos como los tipos de cambio, caracterizados por episodios de contagio, alta volatilidad y variaciones estructurales. En contraste, el modelo MESV, basado en la transformación exponencial matricial, confirma ser una alternativa más adecuada al garantizar positividad definida y permitir que las correlaciones evolucionen de manera coherente con los choques económicos.

Desde el punto de vista metodológico, se destaca el papel central de la inferencia bayesiana y del uso de algoritmos MCMC para la estimación del modelo, dado que permiten trabajar con variables latentes y estructuras altamente no lineales. De igual forma, se resalta la utilidad del criterio DIC como herramienta para la selección del orden óptimo del modelo, equilibrando adecuadamente la complejidad y la capacidad de ajuste.

Se proyecta que el trabajo futuro se orientará en tres direcciones principales. Primero, se plantea realizar simulaciones que permitan evaluar el comportamiento del VAR–MESV bajo distintos escenarios de correlación dinámica y apalancamiento cruzado. Segundo, se prevé aplicar el modelo a los tipos de cambio de los países de la Alianza del Pacífico, con el fin de identificar patrones de comovimiento, episodios de contagio y posibles quiebres estructurales antes y después de 2011. Tercero, se propone incorporar medidas adicionales como la asimetría de Mardia y la curtosis de Koziol, lo cual permitirá caracterizar con mayor precisión la distribución de los choques y evaluar la presencia de colas pesadas.

En síntesis, se confirma que el modelo VAR–MESV constituye una herramienta analítica avanzada y adecuada para estudiar fenómenos financieros complejos. La flexibilidad de su estructura, la capacidad para capturar correlaciones dinámicas y su fundamento bayesiano lo posicionan como un modelo idóneo para continuar la investigación, tanto a nivel teórico como aplicado, dentro del estudio de la dinámica cambiaria multivariada.

Referencias

[1] C. A. Cruz Torres and M. L. Villafranca Rivera, *Modelos autorregresivos vectoriales integrados con volatilidad estocástica multivariada aplicado a la economía de Estados Unidos en el periodo de 1948–2019*, Aglala, **15**(2), 116–142, 2024. Recuperado de `https://revistas.uninunez.edu.co/index.php/aglala/article/view/2523`.

[2] T. Ishihara and Y. Omori, *Efficient Bayesian estimation of a multivariate stochastic volatility model with cross leverage and heavy-tailed errors*, Computational Statistics & Data Analysis, **56**(11), 3674–3689, 2012. 1st issue of the Annals of Computational and Financial Econometrics, Sixth Special Issue on Computational Econometrics. Available at: `https://doi.org/10.1016/j.csda.2010.07.015`.

[3] T. Ishihara, Y. Omori, and M. Asai, *Matrix exponential stochastic volatility with cross leverage*, Computational Statistics & Data Analysis, **100**, 331–350, 2016. Available at: `https://doi.org/10.1016/j.csda.2014.10.012`.

[4] C. Cruz, M. Villafranca, *Modelo VAR Integrado con Volatilidad Estocastica Multivariada y Errores de Cola Pesada*. Disponible en: `https://matematica.unah.edu.hn/escuela/carreras/licenciaturas/semana-carrera-de-matematica/`

[5] C. Cruz, M. Villafranca, *Modelos autorregresivos vectoriales integrados con volatilidad estocástica multivariada*. [Tesis de Maestría, Universidad Nacional Autónoma de Honduras]. Disponible en: `https://mm.unah.edu.hn/tesis-defendidas/`

[6] R. F. Engle, *Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica*, **50**(4), 987–1007, 1982. Disponible en: `https://doi.org/10.2307/1912773`

[7] H. Uhlig, *Bayesian vector autoregressions with stochastic volatility. Econometrica*, vol. 65(1), pp. 59–72, 1997.

[8] T. Cogley, *How fast can the new economy grow? A Bayesian analysis of the evolution of trend growth. Macroeconomics*, vol. 27, pp. 179–207, 2005.

[9] T. Cogley and T. J. Sargent, *Drifts and volatilities: Monetary policies and outcomes in the post WWII US. Review of Economic Dynamics*, vol. 8, no. 2, pp. 262–302, 2005.

[10] E. Jacquier, N. G. Polson, and P. E. Rossi, *Bayesian analysis of stochastic volatility models. Journal of Business & Economic Statistics*, vol. 12(4), pp. 371–389, 1994.

[11] G. C. Primiceri, *Time varying structural vector autoregressions and monetary policy. The Review of Economic Studies*, vol. 72, no. 3, pp. 821–852, 2005.

[12] L. Benati, *The great moderation in the United Kingdom. Journal of Money, Credit and Banking*, vol. 40(1), pp. 121–147, 2008.

[13] J. Galí and L. Gambetti, *On the sources of the great moderation. American Economic Journal: Macroeconomics*, vol. 1(1), pp. 26–57, 2009.

[14] A. D'Agostino, L. Gambetti, and D. Giannone, *Macroeconomic forecasting and structural change. Journal of Applied Econometrics*, vol. 28(1), pp. 82–101, 2013.

[15] T. E. Clark, *Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. Journal of Business & Economic Statistics*, vol. 29(3), pp. 327–341, 2011.

[16] T. E. Clark and F. Ravazzolo, *Macroeconomic forecasting performance under alternative specifications of time-varying volatility. Journal of Applied Econometrics*, vol. 30(4), pp. 551–575, 2015.

[17] C. W. (J.) Chiu, H. Mumtaz, and G. Pintér, *Forecasting with VAR models: Fat tails and stochastic volatility. International Journal of Forecasting*, vol. 33, no. 4, pp. 1124–1143, 2017.

[18] H. Mumtaz, *A generalized stochastic volatility in mean VAR. Economics Letters*, vol. 173, pp. 10–14, 2018.

[19] H. Mumtaz, *A Generalised Stochastic Volatility in Mean VAR. An Updated Algorithm*. Working Paper No. 908, School of Economics and Finance, Queen Mary University of London, July 2020.

[20] Q. Ding, J. Huang, and H. Zhang, *The time-varying effects of financial and geopolitical uncertainties on commodity market dynamics: A TVP-SVAR-SV analysis. Resources Policy*, vol. 72, 2021.

[21] J. Nakajima, *Time-varying parameter VAR model with stochastic volatility: An overview of methodology and empirical applications. Monetary and Economic Studies*, vol. 29, pp. 107–142, 2011.

[22] K. Triantafyllopoulos, *Time-varying vector autoregressive models with volatility. Journal of Applied Statistics*, vol. 38(2), pp. 369–382, 2011.

[23] T. Doan, R. Litterman, and C. A. Sims, *Forecasting and conditional projection using realistic prior distributions*, Econometric Reviews, **1**, 1–100, 1984. Available at: https://doi.org/10.1080/07474938408800053.

[24] R. Litterman, *Forecasting with Bayesian Vector Autoregressions: Five Years of Experience*, Journal of Business and Economic Statistics, **1**, 25–38, 1986. Available at: https://doi.org/10.1080/07350015.1986.10509491.

[25] N. Shephard and M. K. Pitt, *Likelihood analysis of non-Gaussian measurement time series*, Biometrika, **84**, 653–667, 1997. Available at: https://doi.org/10.1093/biomet/84.3.653.

[26] S. J. Koopman, *Disturbance Smoother for State Space Models*, Biometrika, **80**, 117–126, 1993. Available at: https://doi.org/10.1093/biomet/80.1.117.

[27] P. Jong and N. Shephard, *The Simulation Smoother for Time Series Models*, Biometrika, **82**, 339–350, 1995. Available at: https://doi.org/10.1093/biomet/82.2.339.

[28] J. Durbin and S. J. Koopman, *A Simple and Efficient Simulation Smoother for State Space Time Series Analysis*, Biometrika, **89**, 603–616, 2002. Available at: https://doi.org/10.1093/biomet/89.3.603.

[29] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, *Bayesian measures of model complexity and fit*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **64**, 583–639, 2002. Available at: https://doi.org/10.1111/1467-9868.00353.

[30] N. Shephard and M. K. Pitt, *Time varying covariances: a factor stochastic volatility approach*, in *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), Oxford University Press, 1999, pp. 547–570.

[31] R. F. Engle, *Dynamic Conditional Correlation: A Simple Class of Multivariate GARCH Models*, *Journal of Business & Economic Statistics*, vol. 20, no. 3, pp. 339–350, 2002. Available at: https://doi.org/10.1198/073500102288618487.

Maestría en Matemática, Universidad Nacional Autónoma de Honduras.
Dirección actual: Tegucigalpa Hondonduras
Dirección de correo electrónico: nmolinam@unah.hn

Regresión Logística Robusta Basada en M-Estimadores Fundamentos Teóricos y Aplicaciones Prácticas

MAURICIO ARTURO MARTÍNEZ BACA

Resumen. En esta investigación se propone un modelo de regresión logística robusto que emplea M-estimadores para reducir la influencia de valores atípicos. La regresión logística es adecuada para problemas de clasificación con variables respuesta binarias, y su distribución presenta colas más pesadas que permiten manejar observaciones extremas.

El objetivo es obtener estimaciones precisas sin eliminar outliers, superando la sensibilidad de los métodos clásicos como la máxima verosimilitud. Se consideran dos M-estimadores: el de Huber y el de Bianco–Yohai, con el fin de mejorar la robustez y capturar la estructura de los datos de manera más adecuada.

El desempeño del modelo se evaluará mediante simulaciones, incluyendo escenarios con contaminación de datos, destacando su capacidad para identificar patrones relevantes sin perder información valiosa por la eliminación de valores extremos.

Resumen. This research proposes a robust logistic regression model using M-estimators to reduce the influence of outliers. Logistic regression is suitable for classification problems with binary response variables, and its distribution has heavier tails, allowing it to handle extreme observations.

The goal is to obtain accurate estimates without removing outliers, overcoming the sensitivity of classical methods such as maximum likelihood. Two M-estimators are considered: the Huber estimator and the Bianco–Yohai estimator, aiming to improve robustness and better capture the data structure.

The model's performance will be evaluated through simulations, including scenarios with data contamination, highlighting its ability to identify relevant patterns without losing valuable information due to extreme observations.

1. Introducción

La Estadística Robusta surge como una respuesta a las limitaciones de los métodos clásicos de análisis de datos frente a la presencia de valores atípicos o entornos con datos contaminados. Su propósito principal es ofrecer herramientas que permitan tratar de manera adecuada las características de los datos reales, donde las observaciones extremas no pueden ser simplemente ignoradas o eliminadas. En la actualidad, el análisis de datos constituye una herramienta esencial para la toma de decisiones y el desarrollo de investigaciones en áreas tan diversas como la biología, la medicina, la ingeniería, entre otras, donde la calidad y confiabilidad de la información resultan fundamentales.

Date: Octubre 2025.

Key words and phrases. regresión Logística, M-estimadores, estimación paramétrica,outliers, robustez.

Dentro de este contexto, los Modelos Lineales Generalizados (MLG) [6] se han consolidado como una metodología ampliamente utilizada para modelar relaciones entre variables en diferentes campos del conocimiento. Sin embargo, su desempeño puede verse afectado cuando los datos presentan valores atípicos. El objetivo de este artículo es integrar las características de un caso específico de los MLG con técnicas robustas, con el fin de desarrollar un modelo que incorpore la fortaleza de los M-estimadores como una función de pérdida como la que propuso Huber (1964) o los basados en estimadores redescendientes derivados del trabajo de Bianco y Yohai (1996). Para lograrlo, se propone la construcción de una regresión logística robusta, basada en dichos estimadores, que permita obtener inferencias más estables y precisas ante la presencia de datos anómalos.

El problema de los datos atípicos (outliers) representa un desafío recurrente en el análisis estadístico. Su presencia genera interrogantes sobre su tratamiento: algunos investigadores optan por eliminarlos, otros no logran detectarlos, o simplemente deciden ignorarlos. No obstante, en los conjuntos de datos reales es común encontrar observaciones que difieren significativamente del resto. Estas pueden deberse a errores de medición, condiciones experimentales excepcionales o incluso pertenecer a otra población [8]. En cualquier caso, su existencia puede distorsionar las estimaciones y deteriorar el ajuste del modelo, por lo que resulta de vital importancia contar con procedimientos robustos que garanticen resultados confiables y modelos capaces de adaptarse adecuadamente a este tipo de datos.

2. Justificación

Como finalidad principal de la Maestría en Matemática con Orientación en Estadística de la Universidad Nacional Autónoma de Honduras (UNAH), se establece que sus egresados deben ser capaces de analizar y resolver problemas presentes en las ciencias, contribuyendo al desarrollo del país mediante la aplicación rigurosa de herramientas estadísticas. Este propósito se alinea con los ejes primordiales de investigación de la UNAH, que promueven la generación de conocimiento útil para la toma de decisiones y la mejora de las condiciones de vida de la población.

En este sentido, el presente trabajo se enmarca dentro del eje de investigación "Población y condiciones de vida", específicamente en el tema "Cultura, ciencia y educación", contribuyendo al fortalecimiento de la investigación científica y la formación académica en el ámbito estadístico. Asimismo, dentro de las líneas de investigación de la Maestría, este estudio se ubica en la línea de Estadística multivariada y modelos lineales generalizados [7], al abordar el desarrollo y aplicación de técnicas robustas para el análisis de datos.

El análisis, detección y tratamiento de valores atípicos constituye una problemática relevante en el contexto nacional, ya que los datos obtenidos en distintas áreas como la ingeniería, la medicina, las ciencias sociales, la biología, la economía, entre otras suelen estar expuestos a errores de medición, registros irregulares o condiciones experimentales variables. Desarrollar modelos estadísticos que se adapten adecuadamente a estas características permite mejorar la calidad de las inferencias, optimizar la toma de decisiones y fortalecer la capacidad de respuesta en proyectos de investigación aplicada, lo que representa un aporte directo a la solución de problemas reales del país.

Los métodos clásicos para la estimación de parámetros, como el método de máxima verosimilitud o el método de los momentos, han demostrado ser eficientes

bajo condiciones ideales. En particular, el método de máxima verosimilitud es asintóticamente óptimo, consistente y presenta una tasa de convergencia mínima. No obstante, estos métodos suelen fallar cuando los datos presentan observaciones atípicas, ya que dichas condiciones rompen los supuestos de los modelos tradicionales. Ante esta limitación, los métodos robustos surgen como una alternativa necesaria, capaces de capturar las verdaderas características de los datos y adaptarse adecuadamente a la presencia de valores extremos. La incorporación de estas técnicas constituye, por tanto, un avance significativo en la búsqueda de modelos estadísticos más confiables y aplicables a las condiciones reales que enfrenta el país.

3. Antecedentes

El estudio de la estadística robusta surge de la necesidad de métodos estadísticos que no solo sean efectivos, sino también confiables frente a la presencia de valores atípicos, los cuales pueden afectar significativamente los resultados al analizar datos. Desde sus inicios, esta disciplina ha buscado desarrollar herramientas que permitan obtener inferencias más estables y resistentes a desviaciones de los supuestos del modelo.

Los primeros desarrollos importantes se centraron en los M-estimadores, introducidos por Hampel (1974) [1], que establecieron la base teórica para estimaciones resistentes a valores extremos. Posteriormente, Künsch, Stefanski y Carroll (1989)[2] propusieron los estimadores condicionalmente insesgados de influencia acotada, también basados en M-estimadores y denominados condicionalmente Fisher-consistentes, aplicables a Modelos Lineales Generalizados (MLG). Estos estimadores se obtienen como soluciones de problemas de optimización, similares a los planteados por Hampel.

En los años siguientes, se desarrollaron nuevas estrategias de estimación robusta para MLG. Maronna y Yohai (1993) [3] introdujeron los estimadores de proyección, que fueron aplicados posteriormente por Bergesio y Yohai (2001). Este enfoque incluyó la implementación de estimadores basados en la transformación integral de probabilidad (MI-estimadores), permitiendo construir modelos de regresión robustos, como la regresión beta, capaces de manejar datos contaminados o con valores atípicos.

Más recientemente, Abhik Ghosh [5] implementó métodos de inferencia robusta mediante estimadores robustos de divergencia mínima de potencia de densidad. Esta metodología demostró ventajas significativas frente a los métodos clásicos, como la estimación por máxima verosimilitud, especialmente en situaciones con valores atípicos o entornos de contaminación de datos.

Finalmente, Valdora (2014)[6] consolidó los avances de la estadística robusta aplicándolos a modelos lineales generalizados, incluyendo regresión de Poisson, regresión exponencial y regresión binomial. Su trabajo integró enfoques como M-estimadores, cuasiestimadores robustos y estimadores condicionalmente insesgados de influencia acotada, representando uno de los aportes más recientes y completos al desarrollo teórico de esta disciplina.

En conjunto, estos trabajos reflejan la evolución de la estadística robusta, desde sus fundamentos teóricos hasta las aplicaciones modernas en modelos lineales generalizados, mostrando cómo los métodos robustos han permitido realizar inferencias más confiables frente a la presencia de valores atípicos y entornos de datos complejos.

4. Marco Teórico

4.1. Distribución Logística. Dada una variable aleatoria X que sigue una distribución logística con parámetros α y β, su función de distribución acumulada (FDA) se define como:

$$F(x;\alpha,\beta) = \frac{1}{1+e^{-\frac{x-\alpha}{\beta}}} = \left(1+e^{-\frac{x-\alpha}{\beta}}\right)^{-1}, \tag{4.1}$$

donde α es el parámetro de *locación* y β el parámetro de *escala*.

Función de densidad. Derivando la FDA obtenemos la función de densidad de probabilidad:

$$f(x;\alpha,\beta) = \frac{e^{-\frac{x-\alpha}{\beta}}}{\beta\left(1+e^{-\frac{x-\alpha}{\beta}}\right)^2}. \tag{4.2}$$

Algunas de las propiedades clave de la distribución logística son:

- Media: La media de la distribución logística es igual a α.
- Varianza: La varianza es $\frac{\pi^2\beta^2}{3}$.
- Simetría: La distribución logística es simétrica respecto a α.
- Curtosis: La curtosis es 6. $\Rightarrow$ Es más "pesada"que una distribución normal.

4.2. Estimadores clásicos y sus limitaciones. En esta sección se estudiarán las características de los estimadores. Inicialmente, se analizarán algunos estimadores puntuales, como la media y la desviación estándar. Posteriormente, se abordarán métodos más generales de estimación, como el método de los momentos y el método de máxima verosimilitud. Se hará énfasis en que estos métodos pueden presentar debilidades frente a la presencia de valores atípicos o en escenarios con datos contaminados.

4.2.1. Media y desviación estándar. **Definición:** Sea $x = (x_1, x_2, \ldots, x_n)$ un conjunto de valores observados. La **media muestral** $\bar{x}$ y la **desviación estándar muestral** s se definen como:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.3}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{4.4}$$

4.2.2. Estimación por método de los momentos. El método de los momentos es un método de estimación puntual, al igual que los estimadores mencionados anteriormente. Para encontrar los estimadores usando este método se emplean el primer y el segundo momento.

$$\frac{1}{n}\sum_{i=1}^{n} x_i = E[X] \tag{4.5}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = E[X^2] \tag{4.6}$$

De la ecuación (4.5) se obtiene:

$$\bar{x} = E[X], \quad \text{y dado que } E[X] = \alpha, \text{ entonces } \hat{\alpha} = \bar{x} \tag{4.7}$$

Para la estimación de β, consideramos la varianza de la variable aleatoria X:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \frac{\pi^2\beta^2}{3} \tag{4.8}$$

De la ecuación (4.8) se despeja $E[X^2]$:

$$E[X^2] = \frac{\pi^2\beta^2}{3} + \alpha^2 = \frac{\pi^2\beta^2}{3} + \bar{x}^2 \tag{4.9}$$

Finalmente, se obtiene el estimador de β:

$$\hat{\beta} = \sqrt{\frac{3}{n\pi^2}\sum_{i=1}^{n} x_i^2 - \frac{3\bar{x}^2}{\pi^2}} \tag{4.10}$$

4.2.3. Estimación por máxima verosimilitud. La función de densidad de la distribución logística es:

$$f(x;\alpha,\beta) = \frac{e^{-\frac{x-\alpha}{\beta}}}{\beta\left(1+e^{-\frac{x-\alpha}{\beta}}\right)^2} \tag{4.11}$$

Si tomamos una muestra aleatoria de tamaño n, con observaciones $x_1, x_2, \ldots, x_n$ que siguen una distribución logística, la **función de verosimilitud** se expresa como:

$$L(\alpha,\beta;X) = \prod_{i=1}^{n} \frac{e^{-\frac{x_i-\alpha}{\beta}}}{\beta\left(1+e^{-\frac{x_i-\alpha}{\beta}}\right)^2} \tag{4.12}$$

El logaritmo de la función de verosimilitud es:

$$\log L(\alpha,\beta;X) = -n\log\beta + \sum_{i=1}^{n}\left(-\frac{x_i-\alpha}{\beta} - 2\log\left(1+e^{-\frac{x_i-\alpha}{\beta}}\right)\right) \tag{4.13}$$

Para obtener los estimadores de máxima verosimilitud, se derivan parcialmente respecto a α y β y se igualan a cero:

$$\frac{\partial \log L(\alpha,\beta;X)}{\partial \alpha} = \sum_{i=1}^{n}\left(\frac{1}{\beta} + \frac{2e^{-\frac{x_i-\alpha}{\beta}}}{1+e^{-\frac{x_i-\alpha}{\beta}}}\right) = 0 \tag{4.14}$$

$$\frac{\partial \log L(\alpha,\beta;X)}{\partial \beta} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n}\left((x_i-\alpha)\left[1 - \frac{2e^{-\frac{x_i-\alpha}{\beta}}}{1+e^{-\frac{x_i-\alpha}{\beta}}}\right]\right) = 0 \tag{4.15}$$

Notemos que al intentar despejar α y β a partir de las ecuaciones (4.14) y (4.15), se vuelve complejo resolverlas de forma analítica. Por esta razón, se recomienda utilizar métodos numéricos para estimar los parámetros.

En particular, se puede aplicar el método de Newton-Raphson, implementado con la ayuda del lenguaje de programación R, un entorno de software libre para análisis estadístico y visualización de datos.

4.3. Estadística robusta. La estadística robusta se utiliza para analizar datos que pueden verse afectados por errores de medición o por entradas incorrectas, así como por situaciones en las que los datos no cumplen con los supuestos clásicos del análisis estadístico, como la normalidad. Estos errores se manifiestan a menudo como observaciones que se encuentran alejadas del resto de los datos, denominadas valores atípicos o outliers. Sin embargo, estas observaciones pueden ser mediciones válidas que contienen información relevante, por lo que resulta necesario emplear métodos y modelos estadísticos capaces de capturar adecuadamente estas características sin verse fuertemente afectados por valores extremos. Para medir que tan lejos está una observación utilizando las variables explicadas con valores muy extremos o inuales, influyen Drásticamente en las estimaciones de los coeficientes, notando un gran impacto visual en la curva hacia ese punto.

4.4. Entorno de Contaminación. Según [10] si se considera la muestra

$$x_i = \mu + u_i, \quad i = 1, 2, \ldots, n, \tag{4.16}$$

donde los errores $u_1, u_2, \ldots, u_n$ son variables aleatorias que cumplen las siguientes condiciones:

- Tienen una función de distribución F_0.
- Son independientes.

Si se tienen $X = \{x_1, x_2, \ldots, x_n\}$ como variables independientes e idénticamente distribuidas (iid) con distribución

$$F(x) = F_0(x-\mu), \tag{4.17}$$

entonces la distribución de u_i y $-u_i$ es la misma, lo que implica que:

$$F_0(x) = 1 - F_0(-x). \tag{4.18}$$

Una forma de representar datos que se comportan normalmente es asumir que

$$F = D(x_i) = N(\mu, \sigma^2), \tag{4.19}$$

donde $D(x_i)$ denota la distribución de la variable aleatoria X.

La idea formal del entorno de contaminación considera que una proporción $1-\epsilon$ de los datos se comporta según la distribución esperada, mientras que una proporción ϵ de los datos se genera mediante un mecanismo desconocido. Esto se puede representar como:

$$F = (1-\epsilon)G + \epsilon H, \tag{4.20}$$

donde $G = N(\mu, \sigma^2)$ representa la distribución principal y H es alguna distribución desconocida.

Estas características se pueden trasladar a cualquier otra distribución G distinta de la normal.

Definition 4.1. Definición 3.1. La *tasa asintótica de contaminación* (*asymptotic contamination breakdown point*) del estimador $\hat{\theta}$ en F, denotada por $\varepsilon^*(\hat{\theta}, F)$, es el mayor valor $\varepsilon^* \in (0,1)$ tal que, para toda $\varepsilon < \varepsilon^*$, el valor límite $\hat{\theta}_\infty((1-\varepsilon)F + \varepsilon G)$ permanece acotado y alejado de la frontera de Θ para toda distribución G.

De manera intuitiva se considera el punto de ruptura como la proporción de contaminación que un estimador puede soportar antes de que sus valores sean extremadamente malos.

4.5. Estimadores M y funciones de pérdida. Consideremos nuevamente el modelo

$$x_i = \mu + u_i, \quad i = 1, 2, \ldots, n, \tag{4.21}$$

Supongamos que F_0, la función de distribución de u_i, tiene una densidad $f_0 = F_0'$. La densidad conjunta de las observaciones (la función de verosimilitud) es

$$L(x_1, \ldots, x_n; \mu) = \prod_{i=1}^{n} f_0(x_i - \mu) \tag{4.22}$$

El estimador de máxima verosimilitud (MLE) de μ es el valor $\hat{\mu}$, que depende de $x_1, \ldots, x_n$, que maximiza $L(x_1, \ldots, x_n; \mu)$:

$$\hat{\mu} = \hat{\mu}(x_1, \ldots, x_n) = \arg\max_{\mu} L(x_1, \ldots, x_n; \mu) \tag{4.23}$$

donde "`argmax`" significa el valor que maximiza la función.

Si conociéramos F_0 exactamente, el MLE sería "óptimo" en el sentido de alcanzar la varianza asintótica más baja posible dentro de una clase "razonable" de estimadores. Pero como solo conocemos F_0 aproximadamente, nuestro objetivo será encontrar estimadores que sean "casi óptimos" para las siguientes situaciones:

(A) cuando F_0 es exactamente normal
(B) cuando F_0 es aproximadamente normal (por ejemplo, normal contaminada)

Si f_0 es positiva en todo punto y dado que el logaritmo es una función creciente, (4.22) se puede reescribir como

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu) \tag{4.24}$$

donde

$$\rho = -\log f_0 \tag{4.25}$$

Si ρ es diferenciable, derivando (4.25) con respecto a μ se obtiene

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0 \tag{4.26}$$

donde $\psi = \rho'$.

Si ψ es discontinua, las soluciones de la ecuación (4.26) podrían no existir. En este caso, interpretaremos la ecuación como que el lado izquierdo cambia de signo en μ. Obsérvese que si f_0 es simétrica, entonces ρ es par y, por lo tanto, ψ es impar.

Un M-estimador introducidos por Hampel (1974) [1] minimiza una función de pérdida $\rho(r_i)$, para $i = 1, 2, ..., n$. donde $r_i = x_i - \mu$ son los residuos, luego

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(r_i) \tag{4.27}$$

- La función de influencia se relaciona con $\psi(r_i) = \rho'(r_i)$.

Se observa en 1 algunas funciones de pérdida consideradas en el estudio de estimadores en la teoría de estimadores robustos, luego se muestra en 2 una relación entre las funciones ρ y ψ con propiedades importantes como la **familia de funciones de Huber**. Además, cuando ψ en (4.26) no es monótona se llamarán *funciones redescendientes*. Por lo que un M-estimador que utiliza una función redescendiente, se le conoce como M-estimador redescendiente.

$$\rho_k(x) = \begin{cases} x^2 & \text{si } |x| \le k, \\ 2k|x| - k^2 & \text{si } |x| > k, \end{cases} \tag{4.28}$$

con derivada

$$\psi_k(x) = \begin{cases} x & \text{si } |x| \le k, \\ k\,\mathrm{sgn}(x) & \text{si } |x| > k, \end{cases} \tag{4.29}$$

donde la función signo se define como

$$\mathrm{sgn}(x) = \begin{cases} 1 & x > 0, \\ 0 & x = 0, \\ -1 & x < 0. \end{cases} \tag{4.30}$$

- Huber: lineal para residuos pequeños y constante para grandes.
- Más adelante se comentará sobre las ventajas de usar una función ρ acotada.

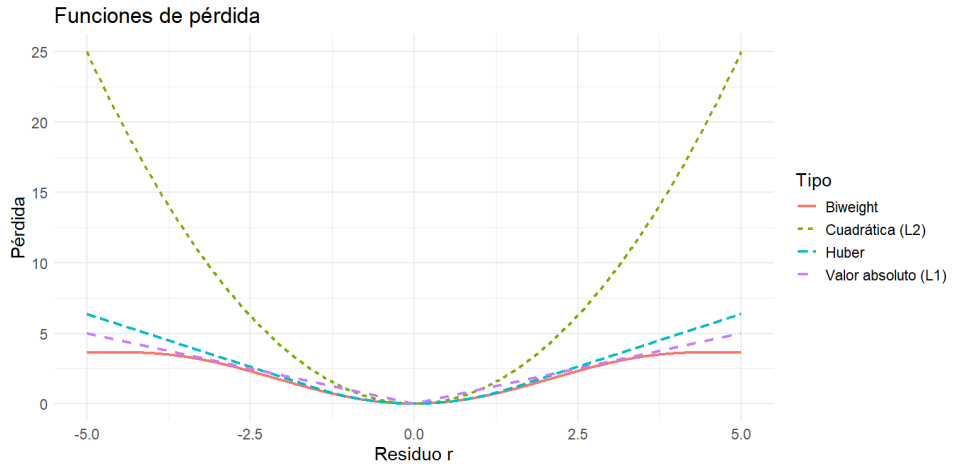


FIGURA 1. Algunas Funciones de pérdida utilizadas

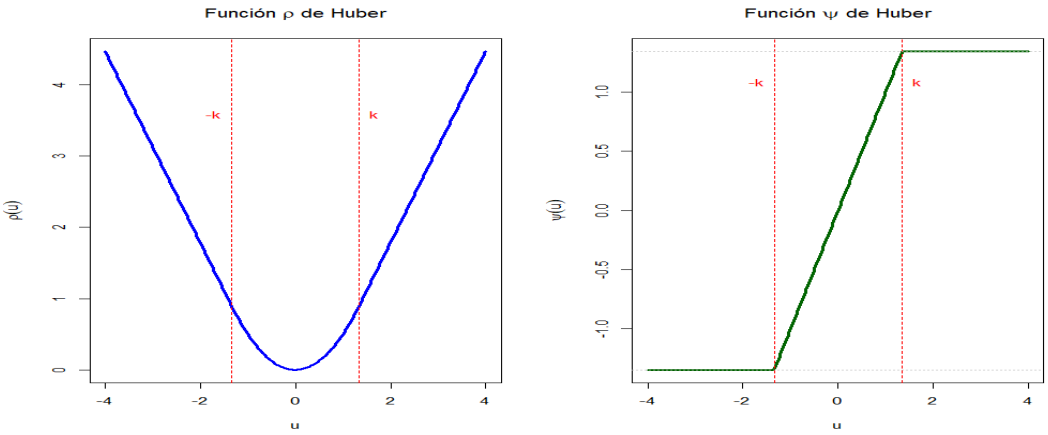


FIGURA 2. Funciones ρ y ψ de Huber.

Una elección popular de funciones ρ y ψ es la familia *bisquare* (también llamada *biweight*):

$$\rho(x) = \begin{cases} 1 - \left[1 - \left(\frac{x}{k}\right)^2\right]^3, & \text{si } |x| \leq k, \\ 1, & \text{si } |x| > k. \end{cases} \tag{4.31}$$

Su derivada viene dada por:

$$\rho'(x) = \frac{6}{k^2}\,\psi(x), \tag{4.32}$$

donde

$$\psi(x) = x\left[1 - \left(\frac{x}{k}\right)^2\right]^2 I(|x| \leq k), \tag{4.33}$$

y $I(\cdot)$ es la función indicadora.

- Tukey Biweight: redescendente, $\psi(r) \to 0$ para $|r| > c$.

Definición: Salvo que se indique lo contrario, una función ρ denotará una función ρ que cumple:

1. $\rho(x)$ es una función no decreciente de $|x|$.
2. $\rho(0) = 0$.
3. $\rho(x)$ es creciente para $x > 0$ y satisface $\rho(x) < \rho(\infty)$.
4. Si ρ está acotada, también se asume que $\rho(\infty) = 1$.

Definición: Una función ψ denotará una función ψ que es la derivada de una función ρ, lo cual implica en particular que, ψ es una función impar y se cumple $\psi(x) \geq 0$ para todo $x \geq 0$.

5. Métodos Robustos en Regresión

5.1. Regresión logística robusta. Según [10] estamos interesados en una y binaria (0–1) que puede representar por ejemplo la muerte o la supervivencia de un paciente después de una cirugía cardiaca. Aquí $y = 1$ representa muerte y $y = 0$ representa supervivencia. Queremos predecir este resultado utilizando distintos regresores, tales como $x_1 =$ edad, $x_2 =$ presión diastólica, etc.

Observamos los pares (x, y) donde $x = (x_1, \ldots, x_p)'$ es el vector de variables explicativas. Supondremos primero que x es fijo (no aleatorio). Para modelar la dependencia de y respecto de x, asumimos que $P(y = 1)$ depende de $\beta' x$ para algún vector desconocido $\beta \in \mathbb{R}^p$. Como $P(y = 1) \in [0, 1]$ y $\beta' x$ puede tomar cualquier valor real, hacemos la siguiente suposición adicional:

$$P(y = 1) = F(\beta' x), \tag{5.1}$$

donde F es una función de distribución continua. La función F^{-1} se denomina *función de enlace*. Si en cambio x es aleatorio, se asume que las probabilidades son condicionales; es decir,

$$P(y = 1 \mid x) = F(\beta' x). \tag{5.2}$$

En el caso común de un modelo con intercepto, la primera coordenada de cada x_i es uno, y la predicción puede escribirse como:

$$\beta' x_i = \beta_0 + x_i \beta_1, \tag{5.3}$$

donde x_i y β_1 son como en (4.6).

Las funciones F más populares son aquellas correspondientes a la distribución logística

$$F(y) = \frac{e^y}{1 + e^y}, \tag{5.4}$$

(modelo logístico), y a la distribución normal estándar $F(y) = \Phi(y)$ (modelo probit). Para el modelo logístico tenemos:

$$\log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta' x. \tag{5.5}$$

El lado izquierdo es llamado *log-odds* (logaritmo de la razón de chances), y es una función lineal de x.

Sea ahora $(x_1, y_1), \ldots, (x_n, y_n)$ una muestra del modelo (5.1), donde $x_1, \ldots, x_n$ son fijos. Para simplificar la notación escribimos:

$$p_i(\beta) = F(\beta' x_i). \tag{5.6}$$

Entonces, $y_1, \ldots, y_n$ son variables aleatorias que toman valores 1 y 0 con probabilidades $p_i(\beta)$ y $1-p_i(\beta)$, respectivamente, y por tanto su función de probabilidad está dada por:

$$p(y_i, \beta) = [p_i(\beta)]^{\,y_i}\,[1-p_i(\beta)]^{\,1-y_i}. \tag{5.7}$$

De esta manera, la log-verosimilitud de la muestra $L(\beta)$ viene dada por:

$$\log L(\beta) = \sum_{i=1}^{n} \Big[y_i \log p_i(\beta) + (1-y_i)\log(1-p_i(\beta)) \Big]. \tag{5.8}$$

Derivando (5.8) se obtienen las ecuaciones de estimación del estimador de máxima verosimilitud (MLE):

$$\sum_{i=1}^{n} \frac{y_i - p_i(\beta)}{p_i(\beta)\,(1-p_i(\beta))} F'(\beta' x_i)\, x_i = \mathbf{0}. \tag{5.9}$$

En el caso de x_i aleatorios, el modelo condicional (5.2) produce la log-verosimilitud:

$$\log L(\beta) = \sum_{i=1}^{n} \Big[y_i \log p_i(\beta) + (1-y_i)\log(1-p_i(\beta)) \Big] + \sum_{i=1}^{n} \log g(x_i), \tag{5.10}$$

donde $g(x_i)$ es la densidad de los regresores.

Separación perfecta y no existencia del MLE

Consideremos el modelo de regresión logística donde

$$p_i(\beta) = P(y_i = 1 \mid x_i) = F(\beta' x_i),$$

y deseamos estimar β por máxima verosimilitud. El problema aparece cuando los datos son *perfectamente separables*.

Se dice que hay separación perfecta si existen $\gamma \in \mathbb{R}^p$ y $\alpha \in \mathbb{R}$ tales que:

$$\gamma' x_i > \alpha \quad \text{si } y_i = 1,$$
$$\gamma' x_i < \alpha \quad \text{si } y_i = 0.$$

Esto implica la existencia de un hiperplano que separa completamente a los casos con $y = 1$ de los casos con $y = 0$. Si existe un hiperplano separador, existen

infinitos, ya que cualquier múltiplo escalar de γ también separa. Por ello se considera la secuencia:

$$\beta^{(k)} = k\gamma, \qquad k \to +\infty.$$

Por lo que podemos notar que no podemos encontrar un valor finito para el estimador.

El estimador robusto tipo Huber en regresión logística se define como sigue

$$d_i(\beta) = -\left[y_i \log p_i(\beta) + (1 - y_i) \log\big(1 - p_i(\beta)\big)\right], \tag{5.11}$$

Luego, el estimador Huber se obtiene minimizando la suma de la función de Huber aplicada a los deviances:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\big(d_i(\beta)\big), \tag{5.12}$$

con la función de Huber ρ definida en (4.28)

Carroll y Pederson (1993) propusieron una forma de convertir el MLE en un estimador con influencia acotada, reduciendo el peso de observaciones con alto leverage.

Leverage de una observación $\mathbf{x}$:

$$h_n(\mathbf{x}) = \sqrt{(\mathbf{x} - \hat{\mu}_n)' \hat{\Sigma}_n^{-1} (\mathbf{x} - \hat{\mu}_n)}, \tag{5.13}$$

con $\hat{\mu}_n$ y $\hat{\Sigma}_n$ robustos e invariantes bajo transformaciones afines.

Estimadores robustos:

$$\sum_{i=1}^{n} w_i \big[y_i \log p_i(\beta) + (1 - y_i) \log(1 - p_i(\beta))\big]. \tag{5.14}$$

Pregibon (1981) propuso estimadores M-robustos para el modelo logístico basados en minimizar:

$$M(\beta) = \sum_{i=1}^{n} \rho\big(d^2(p_i(\beta), y_i)\big), \tag{5.15}$$

donde $\rho(u)$ es una función que crece más lentamente que la función identidad, reduciendo así la influencia de observaciones discordantes.

Bianco y Yohai (1996) observaron que para x_i aleatorios estos estimadores no son *Fisher-consistent*

Para corregir esto propusieron estimar β minimizando:

$$M(\beta) = \sum_{i=1}^{n} \left[\rho\big(d^2(p_i(\beta), y_i)\big) + q\big(p_i(\beta)\big)\right], \tag{5.16}$$

donde $\rho(u)$ es no decreciente y acotada.

La función correctiva es:

$$q(u) = v(u) + v(1 - u), \tag{5.17}$$

con

$$v(u) = 2\int_0^u \psi(-2\log t)\,dt, \tag{5.18}$$

y

$$\psi = \rho' \tag{5.19}$$

donde $d(u, y)$ definido por

$$d(u,y) = \left\{-2\left[y\log(u) + (1-y)\log(1-u)\right]\right\}^{1/2}\operatorname{sgn}(y-u). \tag{5.20}$$

Esta expresión es una medida con signo de la discrepancia entre una variable Bernoulli y y su valor esperado u. Observe que

$$d(u,y) = \begin{cases} 0, & \text{si } u = y, \\ -\infty, & \text{si } u = 1,\ y = 0, \\ \infty, & \text{si } u = 0,\ y = 1. \end{cases} \tag{5.21}$$

En el modelo logístico, los valores $d(p_i(\beta), y_i)$ se denominan *residuos de desviancia*, y miden las discrepancias entre las probabilidades ajustadas por los coeficientes de regresión β y los valores observados.

6. Metodología

En este estudio se trabaja en el contexto de **regresión logística robusta** con el objetivo de comparar el desempeño de estimadores clásicos frente a estimadores robustos. La metodología seguida se describe a continuación:

6.1. Ejemplo Ilustrativo. Se considera un conjunto de 20 datos del contenido de hierro en agua (ppm) para observar el efecto de un solo outlier con relación a los estimadores puntuales como ser la media y a la desviación estándar.

6.2. Simulación de datos. Se generaron **datos simulados** con un tamaño de muestra $n = 100$ utilizando el lenguaje de programación **R**. Los predictores x se obtuvieron de una distribución normal estándar $N(0, 1)$. La variable respuesta y se simuló mediante una distribución Bernoulli con probabilidad

$$P(y = 1 \mid x) = \operatorname{plogis}(2x),$$

lo que corresponde a un modelo logístico con $\alpha = 0$ y $\beta = 2$.

6.3. Entorno de contaminación y generación de outliers. Se introdujeron valores atípicos para evaluar la robustez de los estimadores, usando un **entorno de contaminación** con $\epsilon = 0{,}1$. Se seleccionaron observaciones específicas del predictor $x_0 = (-3, 3)$ y se duplicaron ciertos valores de y para generar **outliers** controlados; en particular, se tomaron dos valores repetidos y tres valores repetidos de y como casos extremos.

6.4. Estimación de modelos. Se ajustaron diferentes modelos de regresión logística:

- **Modelo clásico:** estimación mediante máxima verosimilitud (GLM estándar).
- **Modelo robusto:** estimación utilizando **M-estimadores**, específicamente:
  - **Estimador de Huber**, basado en la función de pérdida propuesta por Huber (1964).
  - **Estimador de Bianco–Yohai (BY)**, especializado en regresión logística robusta (Bianco & Yohai, 1996).

6.5. Análisis comparativo. Para evaluar el desempeño de los modelos, se generaron **gráficas y tablas comparativas** mostrando las curvas y los valores de predicción del modelo clásico y de los modelos robustos. Esto permitió observar cómo los outliers afectan la estimación de parámetros y la capacidad de ajuste de cada modelo, destacando la ventaja de los estimadores robustos en presencia de valores extremos.

7. Resultados y Análisis

se desea observar el comportamiento que tienen los estimadores puntuales de la media y la varianza, para ello se analizará un conjunto de datos sin valores atípicos y con valores atípicos, mediante el siguiente ejemplo;

Ejemplo Ilustrativo: Contenido de hierro en agua (ppm)
Se midió el contenido de hierro (en partes por millón) en 20 muestras de agua:

1,8	2,0	2,1	2,2	2,3	2,4	2,5	2,5
2,6	2,6	2,7	2,8	2,8	2,9	3,0	3,1
3,1	3,2	3,3	**15,0**				

Observemos que el valor 15,0 se considera un **valor atípico (outlier)** ya que es un dato que se encuentra muy alejado de las demás observaciones. **Comparación de resultados**

Estadístico	**Sin outlier**	**Con outlier**
Media	2,626	3,425
Desviación estándar	0,42	2,80

la siguiente Figura 3 y muestra el comportamiento de los datos, la media sin el outlier y con el outlier notando el efecto y la poca robustez de los estimadores analizados.

Observando la figuras 4 y 5 podemos notar que el modelo GLM clásico con outliers se desvia bastante de la real la pendiente β_1 véase el tabla 1 y 2 es más baja porque los outliers tiran del ajuste hacia el centro, lo que denota sensibilidad a los valores atípicos.

En cambio la curva verde es la que mejor se ajusta a la curva real, la pendiente β_1 véase el cuadro 1 y 2 es más cercano al real, esto implica que sigue de cerca la curva real, aunque parece un poco más suavizada en los extremos, lo que refleja la propiedad redescendente del estimador BY que reduce la influencia de los valores atípicos más severos.

Media.png

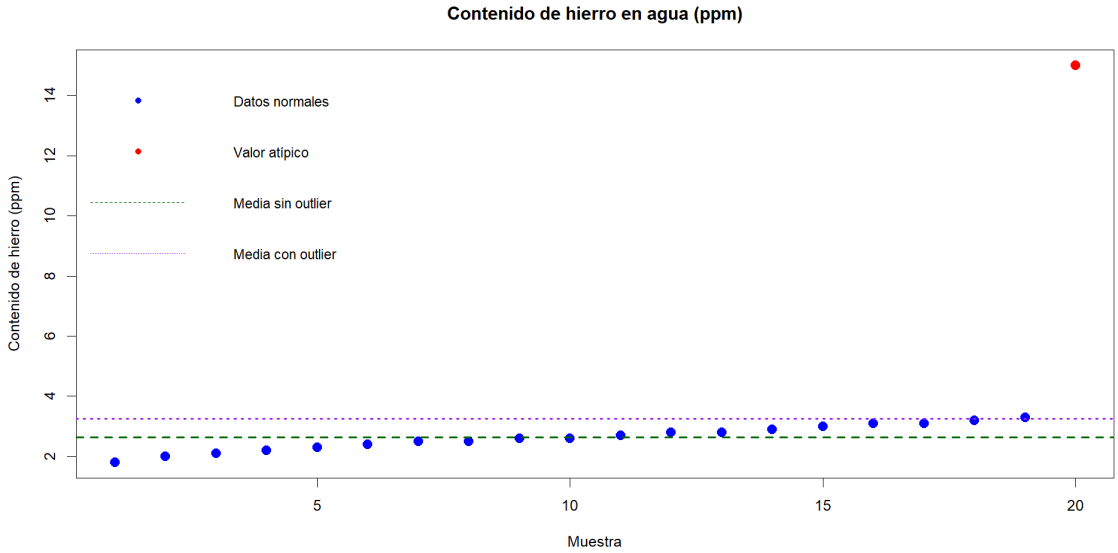


FIGURA 3. Contenido de hierro en 20 muestras de agua. El punto rojo indica un valor atípico (15.0 ppm).

TABLA 1. Comparación de modelos: coeficientes y métricas de ajuste

Modelo	**Coeficientes**	**AIC**	**Deviance**	**LogLik**
Valores reales	$(\beta_0 = 0, \beta_1 = 2)$	–	–	–
GLM sin outliers	$(\beta_0 = 0{,}311, \beta_1 = 1{,}993)$	93.72548	89.72548	-44.86274
GLM con outliers	$(\beta_0 = 0{,}199, \beta_1 = 0{,}925)$	126.69593	122.69593	-61.34797
Robusto Huber	$(\beta_0 = 0{,}258, \beta_1 = 1{,}812)$	137.48300	133.48300	–
Robusto BY	$(\beta_0 = 0{,}299, \beta_1 = 2{,}081)$	143.73029	139.73029	-69.86514

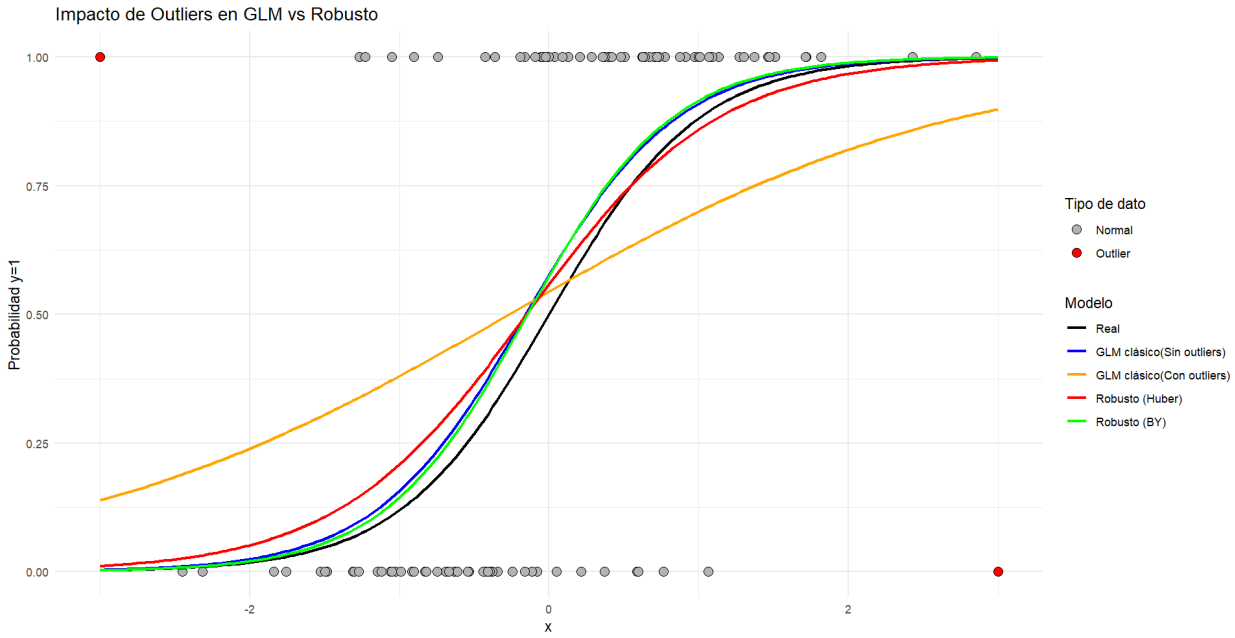


FIGURA 4. Comparación de modelos (2 outliers repetidos 2 veces).

Interpretación Tabla 1 :

- **GLM sin outliers:** Mejor ajuste a los datos limpios, con AIC = 93.73 y desviancia = 89.73 más bajos, y LogLik = -44.86 menos negativo.
- **GLM con outliers:** La presencia de valores extremos aumenta AIC = 126.70 y desviancia = 122.70, y disminuye LogLik = -61.35, indicando peor ajuste.
- **Robusto Huber:** Aunque AIC = 137.48 y desviancia = 133.48 son más altos y LogLik no está definido, las estimaciones son estables frente a outliers.
- **Robusto BY:** Similar al Huber; la robustez sacrifica el ajuste clásico (AIC = 143.73 y LogLik = -69.87 más altos/negativos), pero protege los coeficientes de la influencia de outliers.

TABLA 2. Comparación de modelos: coeficientes y métricas de ajuste

Modelo	Coeficientes	AIC	Deviance	LogLik
Valores reales	$(\beta_0 = 0, \beta_1 = 2)$	–	–	–
GLM sin outliers	$(\beta_0 = 0{,}311, \beta_1 = 1{,}993)$	93.72548	89.72548	-44.86274
GLM con outliers	$(\beta_0 = 0{,}178, \beta_1 = 0{,}668)$	136.52294	132.52294	-66.26147
Robusto Huber	$(\beta_0 = 0{,}235, \beta_1 = 1{,}573)$	151.76689	147.76689	–
Robusto BY	$(\beta_0 = 0{,}299, \beta_1 = 2{,}081)$	168.70511	164.70511	-82.35256

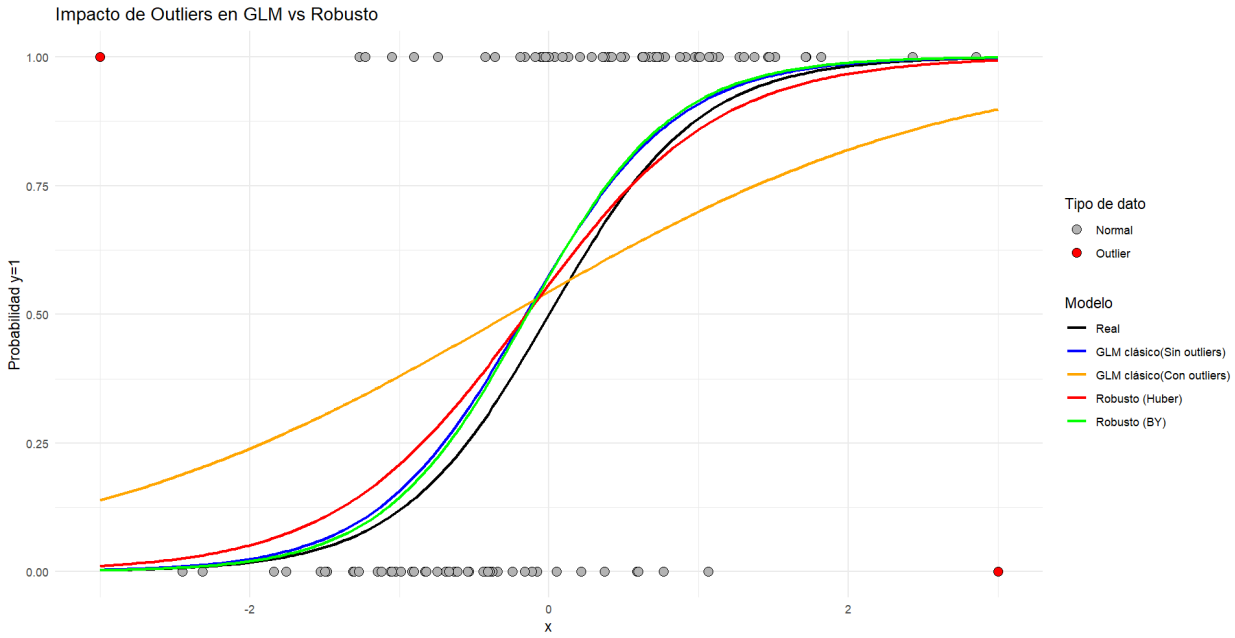


FIGURA 5. Comparación de modelos (2 outliers repetidos 3 veces).

Interpretación tabla 2

- **GLM sin outliers:** Mejor ajuste a los datos limpios, con AIC = 93.73 y desviancia = 89.73 más bajos, y LogLik = -44.86 menos negativo. Los coeficientes estimados ($\beta_0 = 0{,}311, \beta_1 = 1{,}993$) se aproximan bastante a los valores reales ($\beta_0 = 0, \beta_1 = 2$).
- **GLM con outliers:** La presencia de valores extremos aumenta AIC = 136.52 y desviancia = 132.52, y disminuye LogLik = -66.26, indicando un peor ajuste. Los coeficientes ($\beta_0 = 0{,}178, \beta_1 = 0{,}668$) se alejan significativamente de los valores reales.
- **Robusto Huber:** Aunque AIC = 151.77 y desviancia = 147.77 son más altos y LogLik no está definido, las estimaciones ($\beta_0 = 0{,}235, \beta_1 = 1{,}573$) muestran estabilidad frente a outliers, acercándose más a los valores reales que el GLM con outliers.
- **Robusto BY:** Similar al Huber; la robustez sacrifica el ajuste clásico (AIC = 168.71 y LogLik = -82.35 más altos/negativos), pero los coeficientes ($\beta_0 = 0{,}299, \beta_1 = 2{,}081$) se acercan mucho a los valores reales, indicando gran protección frente a la influencia de outliers.

Observando la figura 6 podemos notar que el modelo GLM clásico con outliers se desvia bastante de la real la pendiente β_1 véase el cuadro **??** es más baja porque los outliers tiran del ajuste hacia el centro, lo que denota sensibilidad a los valores atípicos.

En cambio la roja es la que mejor se ajusta a la curva real, la pendiente β_1 véase el cuadro 1 es más cercano al real, esto implica que sigue de cerca la curva real mostrando que Huber es bastante efectivo.

En cambio la curva verde se ajusta bien, aunque ligeramente más conservadora en valores extremos. Esto debido a los outliers son menos extremos o la proporción de contaminación es modesta ($\epsilon = 0{,}1$) por lo que Huber puede adaptarse mejor, conservando información relevante que BY podría descartar como ruido.

de outliers.png

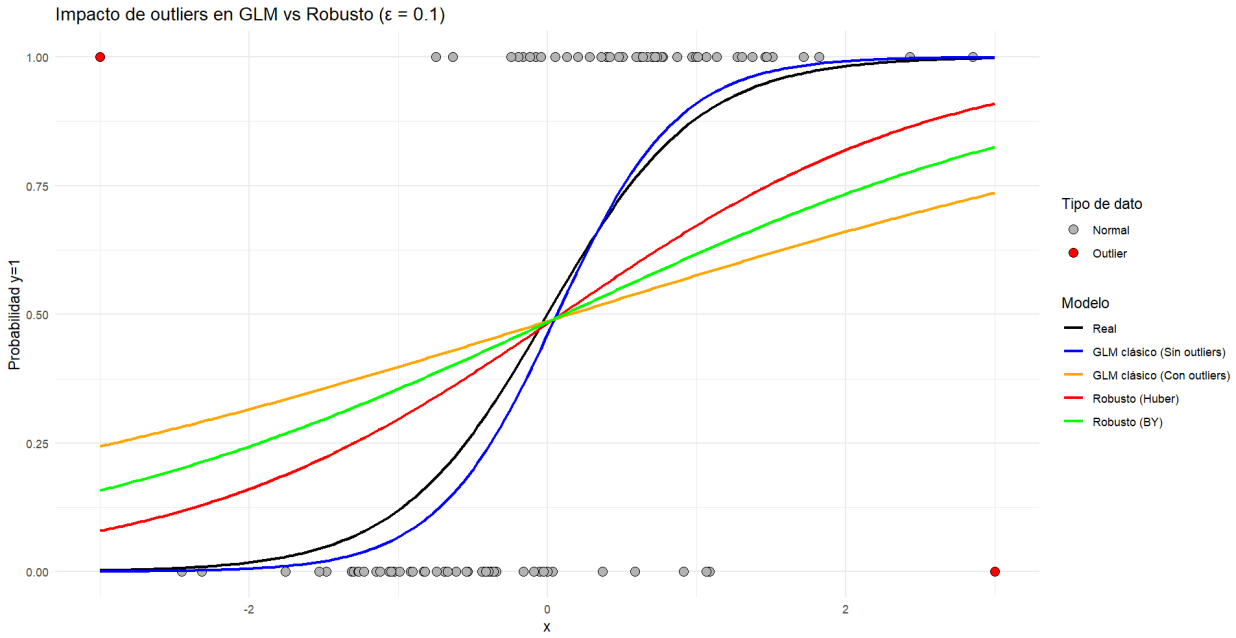


FIGURA 6. Comparación de modelos (usando $\epsilon = 0{,}1$).

Modelo	Coeficientes	AIC	Deviance	LogLik
GLM sin outliers	$(\beta_0 = -0{,}159, \beta_1 = 2{,}471)$	75.15171	71.15171	-35.57585
GLM con outliers	$(\beta_0 = -0{,}054, \beta_1 = 0{,}36)$	137.23405	133.23405	-66.61702
Robusto Huber	$(\beta_0 = -0{,}071, \beta_1 = 0{,}791)$	143.31220	139.31220	NA
Robusto BY	$(\beta_0 = -0{,}06, \beta_1 = 0{,}537)$	138.35826	134.35826	-67.17913

TABLA 3. Comparación de modelos: coeficientes y criterios de ajuste.

interpretación tabla 3 :

- **GLM sin outliers:** Mejor ajuste a los datos limpios, con AIC = 75.15 y desviancia = 71.15 más bajos, y LogLik = -35.58 menos negativo. Los coeficientes ($\beta_0 = -0{,}159, \beta_1 = 2{,}471$) se aproximan razonablemente a los valores reales ($\beta_0 = 0, \beta_1 = 2$), mostrando que el modelo captura bien la relación entre x y y cuando no hay contaminación.
- **GLM con outliers:** La presencia de outliers eleva AIC = 137.23 y desviancia = 133.23, y reduce LogLik = -66.62, mostrando un ajuste mucho peor comparado con el modelo sin outliers. Los coeficientes ($\beta_0 = -0{,}054, \beta_1 = 0{,}36$) se alejan significativamente de los valores reales, evidenciando la sensibilidad del GLM clásico ante datos atípicos.
- **Robusto Huber:** Aunque AIC = 143.31 y desviancia = 139.31 son más altos y LogLik no está definido, las estimaciones ($\beta_0 = -0{,}071, \beta_1 = 0{,}791$) permanecen relativamente estables frente a la contaminación $\epsilon = 0{,}1$. Esto muestra que el estimador robusto Huber protege los coeficientes frente a los outliers, aunque sacrifica algo de ajuste clásico.
- **Robusto BY:** La robustez sacrifica también el ajuste clásico (AIC = 138.36 y LogLik = -67.18), pero los coeficientes ($\beta_0 = -0{,}06, \beta_1 = 0{,}537$) se mantienen relativamente protegidos frente a la influencia de los outliers, demostrando la eficacia de la aproximación BY en situaciones con contaminación moderada.

8. Conclusiones

En el estudio de la regresión logística los modelos GLM clásicos se ajusta muy bien cuando tenemos datos sin presencia de valores atípicos, obteniendo valores de β_0 y β_1 cercanos a los verdaderos, además, de AIC, desviancia y LogLik óptimos. Pero, al considerar datos influenciados por valores atípicos se obtiene un aumento en el AIC y desviancia, y disminuye LogLik, afectando seriamente los coeficientes estimados.

Consideranado el análisis y los resultados de los estimadores robustos (Huber y BY) en la regresión logística se observa que los coeficientes están cercanos a los valores reales aun en presencia de outliers. Pero, cabe resaltar en este caso que el AIC y desviancia son más altos, en el caso de Huber el loglik es indefinido, por lo que se sacrifica de cierta manera el ajuste clásico para cuidar la estimación de parámetros cuando se considera la contaminación de los datos.

Al introducir un entorno de contaminación del 10 % el efecto en el modelo clásico es significativo, pero en los modelos robusto se puede notar una alta resistencia frente a esta contaminación. Por tanto un estudio con regresión logística robusta evita pérdida de información valiosa y malas interpretaciones de los resultados.

9. Trabajos Futuros

Analizar modelos de regresión logística robusta con más variables explicativas, con el fin de evaluar el comportamiento de los M-estimadores en dimensiones superiores, considerando otros M-estimadores, con diferentes niveles de contaminación, diferentes tamaños de muestra y estudio de diferentes criterios de comparación para

caso robusto.

Considerar la extensión del enfoque robusto hacia la regresión beta. La incorporación de estimadores robustos en este contexto podría mejorar el desempeño ante observaciones atípicas o distribuciones altamente asimétricas.

Referencias

[1] F. R. Hampel, *The Influence Curve and Its Role in Robust Estimation*, Journal of the American Statistical Association, 69(346), 383–393, 1974.

[2] H. R. Künsch, L. Stefanski, y R. Carroll, *Conditional Bias Reduction and Robust M-Estimation in Generalized Linear Models*, Journal of the American Statistical Association, 84(406), 621–632, 1989.

[3] R. A. Maronna y V. J. Yohai, *Robust Estimators of Multivariate Location and Scatter for High Breakdown Point*, The Annals of Statistics, 21(1), 283–293, 1993.

[4] L. Bergesio y V. J. Yohai, *Projection Estimators for Robust Generalized Linear Models*, Journal of Statistical Planning and Inference, 97(1-2), 57–79, 2001.

[5] A. Ghosh, *Robust Inference Using Minimum Density Power Divergence Estimators in Beta Regression*, [Revista o conferencia], Año.

[6] J. Valdora, *Estimadores robustos a transformaciones aplicados a modelos lineales generalizados*, [Revista o conferencia], 2014.

[7] Universidad Nacional Autónoma de Honduras, *Líneas de Investigación de la Maestría en Matemáticas*, Departamento de Postgrado, 2025.

[8] V. J. Hodge y J. Austin, *Robust Statistics for Outlier Detection*, Journal of Pattern Recognition Letters, 27(9), 861–874, 2006.

[9] A. M. Bianco, V. J. Yohai, *Robust estimation in the logistic regression model*, en Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday, pp. 17-34, Springer New York, 1996.

[10] R. A. Maronna, R. D. Martin, V. J. Yohai, *Robust Statistics: Theory and Methods*, Wiley Series in Probability and Statistics, 2006.

Maestría en Matemática, Universidad Nacional Autónoma de Honduras.
E-mail address: mauricio.martinez@unah.edu.hn

MÁS ALLÁ DE TENDENCIAS PARALELAS: UN ENFOQUE UNIVERSAL PARA ESTIMAR EFECTOS DISTRIBUCIONALES EN DID

SANCHEZ ANTHONY

Resumen. Este trabajo analiza la identificación causal en diseños Difference-in-Differences de dos periodos cuando el supuesto de Tendencias Paralelas puede no ser plausible, especialmente en entornos con resultados no lineales o discretos, o cuando el interés recae en parámetros distribucionales como el efecto cuantil en los tratados. En estos casos, las restricciones aditivas implícitas en el enfoque tradicional pueden fallar y alterar la comparación contrafactual entre grupos.

Para superar esta limitación, se introduce la condición de Odds Ratio Equi-Confounding, que describe la confusión en la escala de razón de probabilidades generalizada y permite una representación invariante a la escala del resultado potencial.

Abstract. This paper examines causal identification in two-period Difference-in-Differences settings when the usual Parallel Trends assumption may not be credible, particularly in applications where the outcome is nonlinear or discrete, or when interest lies in distributional parameters such as the quantile treatment effect on the treated. In such contexts, additive and scale-dependent restrictions underlying conventional DiD can fail, making the standard decomposition of counterfactual trends invalid.

To address this limitation, the analysis adopts the Odds Ratio Equi-Confounding condition, which characterizes confounding on the generalized odds-ratio scale and yields a scale-invariant representation of the counterfactual outcome distribution.

Fecha: Agosto 2025.

Palabras y frases clave. Inferencia causal, Diferencias en diferencias, OREC–UDiD, Efectos cuantilísticos (QTT), Confusión no observada.

1. Introducción

Los diseños de *difference-in-differences (DiD)* constituyen una de las metodologías centrales para la identificación de efectos causales en entornos observacionales. Su formulación clásica, sistematizada en [3], descansa en un esquema 2×2: dos periodos, un grupo tratado y un grupo de comparación. En este marco, la identificación del efecto medio del tratamiento sobre los tratados (ATT) se apoya en el supuesto de tendencias paralelas (PT), según el cual, en ausencia de tratamiento, los resultados potenciales no tratados habrían evolucionado, en promedio, de forma equivalente entre ambos grupos.
Sin embargo, el supuesto PT presenta limitaciones estructurales. Se formula en la escala aditiva, lo que puede entrar en tensión con restricciones naturales del soporte del resultado cuando éste es binario, discreto o mixto. Además, PT no es invariante frente a transformaciones monótonas, de modo que su plausibilidad depende de la escala en que se mida la variable de interés, aun cuando el parámetro causal subyacente sea invariante por construcción. Por otra parte, PT no se adapta de manera natural a parámetros no lineales, como efectos sobre la distribución o sobre cuantiles de los tratados, restringiendo el alcance inferencial del esquema DiD convencional.

En respuesta a estas limitaciones, la literatura reciente ha propuesto condiciones alternativas que modelan el sesgo de confusión de forma más flexible. Entre ellas, el marco *Odds Ratio Equi-Confounding (OREC)* desarrollado por Park y Tchetgen Tchetgen [20] reformula el problema en la escala del *odds ratio* generalizado, exigiendo estabilidad temporal de la asociación entre el tratamiento y el resultado potencial no tratado. Esta construcción es compatible con resultados continuos, discretos o mixtos, es invariante a transformaciones monótonas, permite la presencia de confusores no observados y admite una teoría completa de eficiencia semiparamétrica. En este trabajo se desarrolla un marco unificado para la identificación y estimación de efectos causales en diseños DiD bajo el supuesto OREC, se deriva la función de influencia eficiente, se construye un estimador $N^{1/2}$-consistente mediante técnicas de *cross-fitting* y se ilustra, mediante una simulación controlada, el comportamiento del enfoque en configuraciones en las que el supuesto de tendencias paralelas falla de manera sistemática.

2. Justificación

La investigación se motiva por la necesidad de disponer de herramientas formales que permitan evaluar efectos causales cuando las trayectorias de los grupos no son comparables y cuando la estructura del resultado exige ir más allá de los promedios. En muchos problemas económicos y de política pública, los grupos tratados y de comparación presentan diferencias previas al tratamiento, respuestas heterogéneas y posibles fuentes de confusión no observada. Bajo estas condiciones, los supuestos aditivos tradicionales dejan de ser una simple conveniencia técnica y se convierten en una restricción fuerte sobre el proceso generador de datos, especialmente cuando el resultado es discreto, acotado o presenta colas de particular interés.

En este contexto, la adopción del marco Universal Difference-in-Differences (UDiD) bajo el supuesto Odds Ratio Equi-Confounding (OREC) ofrece una justificación metodológica precisa: permite caracterizar el contrafactual de los tratados en una

escala invariante a transformaciones monótonas y compatible con parámetros distribucionales como el efecto cuantil en los tratados (QTT). Este tipo de parámetros es especialmente adecuado cuando el impacto de una política no se refleja de forma uniforme a lo largo de la distribución, sino que se concentra en determinados cuantiles o segmentos de la población. El uso de funciones de influencia eficientes y de técnicas de estimación no y semiparamétricas proporciona un marco en el que la presencia de confusión no observada puede tratarse explícitamente, sin renunciar a una teoría asintótica clara ni a condiciones de identificación transparentes.

Desde la perspectiva académica, el trabajo se inscribe de manera natural en la orientación en Estadística de la Maestría en Matemáticas, al combinar inferencia causal, modelación semiparamétrica y análisis distribucional dentro de un mismo esquema formal. La construcción del estimador, el estudio de sus propiedades de eficiencia y la representación de la confusión mediante razones de densidad y razones de momios generalizadas se alinean con las líneas de investigación en econometría y procesos estocásticos, donde la estructura matemática del modelo es tan importante como su interpretación aplicada. De este modo, la investigación contribuye a tender un puente entre la teoría estadística moderna y los problemas de identificación causal que surgen en el análisis de políticas económicas contemporáneas.

3. Antecedentes

Los diseños de DiD tienen una larga trayectoria en la evaluación de efectos causales en contextos observacionales, con aplicaciones que se remontan al siglo XIX y una consolidación moderna en economía aplicada y ciencias sociales [3]. En su formulación canónica, el diseño 2×2 considera dos periodos (pre y post) y dos grupos (tratado y de comparación), y define el estimador DiD como la diferencia entre el cambio promedio en el grupo tratado y el cambio promedio en el grupo de control. Bajo el supuesto de *tendencias paralelas* (PT), esto es, que en ausencia de tratamiento el cambio esperado en el resultado potencial no tratado hubiera sido igual en ambos grupos, dicho estimador coincide con el efecto medio del tratamiento sobre los tratados (ATT) [16, 1, 24]. Esta simplicidad conceptual explica en buena medida la enorme difusión del enfoque DiD en estudios empíricos recientes.

Con el tiempo, la práctica empírica dejó de restringirse al esquema 2×2 y evolucionó hacia configuraciones más complejas: múltiples periodos, adopción escalonada del tratamiento, tratamientos de intensidad variable, incorporación de covariables y heterogeneidad marcada en los efectos [3]. Durante años, el uso de modelos de regresión con efectos fijos de unidad y tiempo (especificaciones *two-way fixed effects*, TWFE) se convirtió en el estándar para implementar estos diseños, apoyándose en la equivalencia entre la regresión lineal y el estimador DiD en el caso básico. Sin embargo, investigaciones recientes han mostrado que, cuando los efectos del tratamiento son heterogéneos o la estructura del diseño se aparta del caso simple, los estimadores TWFE pueden producir combinaciones ponderadas difíciles de interpretar, con pesos negativos y, en casos extremos, estimaciones de signo contrario al efecto causal subyacente [15, 9, 28]. Estas evidencias han motivado un giro metodológico hacia marcos de análisis que "descomponen" cualquier diseño DiD complejo en bloques elementales 2×2 y construyen a partir de ellos los parámetros de interés mediante un enfoque de *forward-engineering*, es decir, fijando primero el parámetro

objetivo y derivando después el estimador adecuado [3].

En paralelo, la propia formulación del supuesto PT ha sido objeto de revisión crítica. En su versión estándar, PT es un supuesto aditivo sobre los resultados potenciales no tratados: exige la igualdad, entre grupos, de las diferencias esperadas $E(Y_1^{(0)} - Y_0^{(0)} \mid \cdot)$, lo cual es natural para resultados continuos sin restricciones de soporte, pero puede resultar problemático cuando la variable de interés es binaria, discreta o acotada. En estos casos, las extrapolaciones implícitas de PT pueden producir contra–ejemplos en los que el "contrafactual aditivo" sale fuera del rango posible del resultado, o bien resulta incompatible con la evidencia empírica aun cuando la dinámica verdadera sea razonable desde el punto de vista probabilístico. Además, PT es sensible a transformaciones monótonas del resultado: la plausibilidad del supuesto puede cambiar al pasar, por ejemplo, de niveles a logaritmos, aunque el objeto causal de interés (como un efecto sobre la distribución) no dependa de la escala particular en que se mida el resultado. Estas limitaciones se vuelven especialmente agudas cuando el interés se desplaza desde el ATT hacia parámetros distribucionales, como los efectos cuantilísticos en los tratados (QTT), para los que la estructura aditiva de PT no proporciona una ruta natural de identificación.

Para superar estas restricciones, la literatura ha propuesto múltiples extensiones y alternativas al supuesto PT clásico. Los modelos de *changes-in-changes* (CiC) introducen una estructura basada en transformaciones monótonas de un factor latente común, lo que permite identificar efectos distribucionales bajo una hipótesis de estabilidad en la distribución del "shock" subyacente [2]. Otros trabajos han formulado variantes no lineales de PT mediante la aplicación de funciones de enlace que restablecen la igualdad de tendencias en escalas transformadas, como en el caso de la *nonlinear parallel trends* (NPT) [21, 30]. En una dirección distinta, algunos enfoques han recurrido a condiciones sobre la función característica logarítmica del resultado potencial [5], o bien a suposiciones de estabilidad en la cópula que relaciona el resultado pretratamiento y el cambio del resultado a lo largo del tiempo [7, 6]. Finalmente, los esquemas de ignorabilidad secuencial extienden ideas de la evaluación de tratamientos en series temporales, imponiendo que, condicionando en resultados y covariables previas, no exista confusión no observada entre tratamiento y resultado potencial posterior [10].

No obstante, ninguna de estas alternativas constituye una solución universal al problema DiD. Los enfoques basados en PT (lineal o no lineal) y en funciones características suelen ser sensibles a la escala en que se mide el resultado; los modelos CiC[18] se apoyan en supuestos de monotonicidad y rank-preservation difíciles de verificar; las estrategias de cópulas requieren condiciones fuertes sobre la estructura de dependencia; y las formulaciones de ignorabilidad secuencial, si bien generan marcos potentes, descansan en la ausencia de confusores no observados, lo que rara vez es inocuo en aplicaciones económicas. Además, muchos de estos desarrollos carecen de una teoría completa de eficiencia semiparamétrica para parámetros distribucionales, lo que limita su capacidad para guiar el diseño de estimadores que aprovechen de forma óptima la información disponible en la muestra.

En este escenario surge el enfoque *Odds Ratio Equi-Confounding* (OREC), propuesto dentro del marco *Universal Difference-in-Differences* (UDiD) [20]. La idea central consiste en representar el sesgo de confusión —debido a variables no observadas que afectan simultáneamente el tratamiento y el resultado potencial libre de tratamiento— mediante funciones de razón de momios generalizadas que vinculan el tratamiento con el resultado potencial en cada periodo. El supuesto OREC postula que esa estructura de confusión, expresada en la escala del *odds ratio* generalizado, permanece estable entre el periodo pre y el periodo post. Esta formulación presenta varias ventajas acumulativas: es aplicable a resultados continuos, discretos o mixtos; es invariante a transformaciones monótonas del resultado; admite la presencia de confusores no observados siempre que su efecto sea estable en la escala de razón de momios; y permite derivar funciones de influencia eficientes y estimadores de raíz-N consistentes sin imponer parametrizaciones rígidas sobre las distribuciones subyacentes.

El marco UDiD, construido sobre OREC, proporciona así una síntesis entre la tradición DiD y la teoría moderna de inferencia semiparamétrica. Por un lado, preserva la lógica contrafactual de los diseños DiD, al centrarse en la identificación del resultado potencial no tratado de los grupos expuestos al tratamiento. Por otro, desplaza el análisis desde la escala aditiva hacia una escala log–odds que resulta compatible con distintos tipos de variables y con parámetros distribucionales como el QTT. Al incorporar funciones de influencia eficientes y técnicas de estimación basadas en *cross-fitting*, el enfoque UDiD ofrece un marco general para estudiar efectos causales en contextos donde los supuestos clásicos de tendencias paralelas se muestran frágiles, abriendo la puerta a aplicaciones en las que la heterogeneidad del efecto y la estructura del resultado son elementos centrales del problema empírico.

4. Diferencias en diferencias

Desde mediados del siglo XIX, el diseño de *Difference-in-Differences (DiD)* ha ocupado un lugar central en la estimación de efectos causales dentro de las ciencias sociales. Su esencia radica en comparar la evolución temporal de un grupo expuesto a un tratamiento con la de otro que permanece sin tratar, de modo que la inferencia no se base en niveles absolutos sino en cambios relativos. En su forma más elemental -con dos periodos y dos grupos- el estimador DiD se define como la diferencia entre las variaciones promedio del resultado en ambos grupos: la diferencia de dos diferencias.

El fundamento identificador de este esquema descansa en el supuesto de PT: en ausencia del tratamiento, las trayectorias promedio de ambos grupos habrían sido paralelas en el tiempo. Bajo esta condición, la comparación de diferencias permite recuperar el ATT.

Con el desarrollo de bases de datos más amplias y paneles de largo horizonte, los diseños DiD se extendieron a configuraciones más complejas. Las unidades pueden recibir el tratamiento en distintos momentos o intensidades, y las variables de control se incorporan para mejorar la comparabilidad entre grupos. En este contexto, la práctica empírica consolidó el uso de modelos de regresión lineal con efectos fijos por unidad y por tiempo -*el estimador conocido como Two-Way Fixed Effects (TWFE)*- cuya popularidad se sustentó en que, en el caso 2×2, reproduce exactamente el estimador clásico de DiD calculado a partir de medias muestrales.

Este señalamiento fue desarrollado con particular detalle por Baker, Larcker y Wang [4]. Mediante un extenso estudio de simulación, los autores evalúan el desempeño de siete métodos modernos de DiD bajo escenarios con efectos constantes y heterogéneos, mostrando que muchos de ellos presentan intervalos de confianza que no cubren el efecto promedio verdadero con la frecuencia nominal y que, además, sufren de baja potencia estadística.

Ante estas limitaciones, Baker, Callaway, Cunningham, Goodman-Bacon y Sant'Anna [3] proponen un marco unificado para los diseños DiD basado en los principios de la inferencia causal y la heterogeneidad del tratamiento. La propuesta se orienta hacia un marco unificado que reconcilia la diversidad de aplicaciones empíricas bajo los principios de la inferencia causal en presencia de heterogeneidad del efecto del tratamiento. En esencia, incluso los diseños más complejos pueden descomponerse en una colección de comparaciones elementales 2×2: pares de unidades en las que el tratamiento varía frente a otras en las que no lo hace. Cada uno de estos bloques constituye un "building block" identificador, cuya validez depende únicamente del supuesto de PT local a esa comparación.

Los autores denominan esta estrategia un enfoque de forward-engineering, pues parte de la definición clara de los parámetros de interés y, a partir de ellos, construye los métodos analíticos necesarios para su estimación. Este modo de proceder contrasta con la práctica habitual de reverse-engineering, que inicia desde especificaciones de regresión familiares y luego intenta derivar las condiciones bajo las

cuales podrían tener interpretación causal.

Al adoptar esta perspectiva, se evita la ambigüedad que genera el uso indiscriminado de modelos *Two-Way Fixed Effects*, cuya interpretación varía según la especificación y puede inducir confusión entre cambios en los supuestos de identificación y alteraciones en el parámetro objetivo. En cambio, el enfoque de forward-engineering ofrece una estructura metodológica en la que diferentes estimadores apuntan a un mismo parámetro, diferenciándose solo por la naturaleza explícita de los supuestos que los sostienen.

4.1. Diseño 2×2. El punto de partida de todo análisis es el diseño canónico 2×2, en él se consideran dos grupos -uno tratado y otro no tratado- y dos periodos de tiempo —uno previo y otro posterior a la introducción del tratamiento-.

Este supuesto establece que, en ausencia del tratamiento, ambas poblaciones habrían experimentado la misma variación promedio en el tiempo.

4.1.1. Efectos causales y parámetros objetivo. Todo análisis causal debe comenzar con la definición explícita de la cantidad de interés, o parámetro objetivo. Esta definición se formula naturalmente dentro del marco de resultados potenciales desarrollado por Rubin (1974) y Robins (1986)[23].

Definition 4.1. Sea $Y_{i,t}^0$ el resultado potencial de la unidad i en el periodo t si permaneciera sin tratamiento en ambos periodos, y $Y_{i,t}^1$ el resultado potencial si no recibiera tratamiento en el primer periodo pero sí en el segundo.

Dado que los resultados potenciales son mutuamente excluyentes, en la práctica solo se observa uno de ellos para cada unidad. El resultado observado puede expresarse como

$$Y_{i,t} = (1 - D_i)Y_{i,t}^0 + D_i Y_{i,t}^1, \tag{4.1}$$

donde la función de decisión

$$D_i = \begin{cases} 1, & \text{si la unidad (i) está expuesta al tratamiento en (t),} \\ 0, & \text{en caso contrario.} \end{cases}$$

Equivalentemente, $Y_{i,t}$ puede interpretarse como la realización efectiva de la función:

$$Y_{i,t} = Y_{i,t}^{D_i} = \begin{cases} Y_{i,t}^0, & \text{si } D_i = 0, \\ Y_{i,t}^1, & \text{si } D_i = 1, \end{cases}$$

Un supuesto central para la validez del diseño DiD es el de no anticipación (no anticipation). Este establece que el tratamiento no afecta los resultados antes de su implementación efectiva, es decir,

$$Y_{i,0}^1 = Y_{i,0}^0,$$

para todo i. Este supuesto garantiza que los resultados previos reflejan el estado no tratado y permite definir con precisión el momento en que el tratamiento surte efecto.

Supuesto 1. NA (No-Anticipation).

$$Y^1_{i,t} = Y^0_{i,t}, \quad \forall i \in D_i = 1, ; \forall t \text{ previo al tratamiento.}$$

El supuesto de *no anticipación* establece que, para todas las unidades tratadas y en todos los periodos previos a la intervención, los resultados potenciales bajo tratamiento y no tratamiento son idénticos.

Definition 4.2. Bajo el marco de resultados potenciales, el efecto causal individual se define como la diferencia

$$Y^1_{i,t} - Y^0_{i,t},$$

que representa el impacto del tratamiento sobre la unidad i en el periodo t.

Este marco permite la existencia de heterogeneidad arbitraria en los efectos del tratamiento entre unidades y a lo largo del tiempo, es decir, los efectos pueden diferir para cada i y t. Sin embargo, aprender sobre esta heterogeneidad completa requiere supuestos adicionales fuertes.

En la práctica, los diseños DiD no buscan identificar efectos individuales, sino promedios ponderados de ellos. En particular, el parámetro más comúnmente estimado es el efecto promedio del tratamiento sobre los tratados en el tiempo t, denotado ATT_t:

$$\begin{aligned} ATT_t &= \mathbb{E}_\omega\left[Y^1_{i,t} - Y^0_{i,t} \mid D_i = 1\right] \\ &= \mathbb{E}_\omega\left[Y^1_{i,t} \mid D_i = 1\right] - \mathbb{E}_\omega\left[Y^0_{i,t} \mid D_i = 1\right], \end{aligned} \tag{4.2}$$

donde $\mathbb{E}_\omega[\cdot]$ denota un promedio ponderado según un esquema de pesos ω_i.

La expresión 4.2 muestra que ATT_t compara el promedio ponderado de los resultados observados en el periodo posterior entre las unidades tratadas con el promedio ponderado de los resultados contrafactuales -no observados- que esas mismas unidades habrían tenido de no haber recibido el tratamiento.

Bajo el supuesto de no anticipación, se cumple además que $ATT_t = 0, \forall t$ previo al tratamiento, lo cual implica que las diferencias entre grupos antes de la intervención reflejan únicamente brechas en los resultados potenciales no tratados.
La inclusión de pesos ω_i no es un detalle técnico posterior, sino una decisión sustantiva que determina la población de referencia del efecto estimado.

En este sentido, el ATT_t ponderado y el no ponderado representan parámetros distintos. Mientras el primero describe el efecto promedio del tratamiento sobre una población definida por el esquema de pesos -por ejemplo, ponderada por tamaño o relevancia de las unidades-, el segundo se refiere al efecto promedio simple sobre las unidades tratadas. Así, las comparaciones entre estimaciones ponderadas y no ponderadas no reflejan diferencias de eficiencia estadística, sino variaciones en el propio parámetro objetivo.

Este punto cobra especial importancia en contextos con efectos heterogéneos del tratamiento, donde adoptar una estructura de ponderación destinada a mejorar la

precisión en presencia de heterocedasticidad -como en las regresiones de coeficientes constantes- puede alterar sustancialmente el parámetro identificado. Como advierten Solon, Haider y Wooldridge (2015)[27], cuando los efectos del tratamiento se correlacionan con los pesos, el parámetro ponderado puede diferir notablemente del no ponderado, lo que implica que ambos deben interpretarse como objetos causales distintos.

Cuando se reconoce la existencia de heterogeneidad en los impactos del tratamiento, resulta útil examinar no solo el efecto promedio, sino también cómo dicho impacto se distribuye a lo largo de los distintos puntos de la distribución de los resultados potenciales.

4.2. Identificación de supuestos: Tendencias paralelas. Todo diseño de investigación causal se sustenta en un conjunto de supuestos de identificación que permiten recuperar los parámetros objetivo a partir de datos observados. En el caso del diseño DiD, la identificación del contrafactual necesario para estimar ATT_t requiere establecer una relación entre los resultados observados y los potenciales no tratados.

En principio, existen múltiples supuestos que podrían identificar dicho contrafactual. Uno de ellos es la independencia en medias, que asumiría

$$\mathbb{E}_\omega[Y^0_{i,t} \mid D_i = 1] = \mathbb{E}_\omega[Y^0_{i,t} \mid D_i = 0],$$

lo que implica que, condicionalmente, el tratamiento es asignado de forma aleatoria. Bajo este supuesto, la diferencia transversal entre grupos en el periodo posterior identificaría directamente el ATT_t.

Otra posibilidad es la invariancia temporal de los resultados potenciales no tratados, que supone

$$\mathbb{E}_\omega[Y^0_{i,t} \mid D_i = 1] = \mathbb{E}_\omega[Y^0_{i,t-1} \mid D_i = 1],$$

en cuyo caso la variación temporal dentro del grupo tratado equivaldría al efecto del tratamiento.

No obstante, ambos supuestos son demasiado restrictivos en la práctica. El primero ignora las diferencias estructurales entre grupos antes del tratamiento, y el segundo desconoce la posibilidad de cambios temporales comunes que afectan a todas las unidades.

El diseño DiD se fundamenta, en cambio, en un supuesto más general: el de tendencias paralelas.

Definition 4.3. El supuesto de PT en el diseño 2×2 establece que el cambio promedio ponderado en los resultados potenciales no tratados es el mismo entre el grupo tratado y el grupo de comparación. Formalmente,

$$\underbrace{\mathbb{E}_\omega[Y^0_{i,t=2} - Y^0_{i,t=1} \mid D_i = 1]}_{\text{No observado. contrafactual}} = \mathbb{E}_\omega[Y^0_{i,t=2} - Y^0_{i,t=1} \mid D_i = 0]. \tag{4.3}$$

Si esta condición se cumple, es posible construir el resultado contrafactual promedio para las unidades tratadas en el periodo posterior, $\mathbb{E}_\omega[Y^0_{i,t=2} \mid D_i = 1]$, a partir de cantidades observables. En particular,

(4.4)
$$\mathbb{E}_\omega[Y^0_{i,t=2} \mid D_i = 1] = \mathbb{E}_\omega[Y^0_{i,t=1} \mid D_i = 1] + \Big(\mathbb{E}_\omega[Y^0_{i,t=2} \mid D_i = 0] - \mathbb{E}_\omega[Y^0_{i,t=1} \mid D_i = 0]\Big)$$

Sustituyendo la ecuación 4.4 en la definición de ATT_t y reemplazando los resultados potenciales no observados mediante los observados según 4.1, se obtiene el estimador DiD en términos de promedios poblacionales:

(4.5)
$$\begin{aligned} ATT_t &= \mathbb{E}_\omega[Y^1_{i,t} \mid D_i = 1] - \mathbb{E}_\omega[Y^0_{i,t} \mid D_i = 1] \\ &= \Big(\mathbb{E}_\omega[Y^1_{i,t} \mid D_i = 1] \underbrace{- \mathbb{E}_\omega[Y^0_{i,t-1} \mid D_i = 1]\Big) - \Big(\mathbb{E}_\omega[Y^0_{i,t} \mid D_i = 0] - \mathbb{E}_\omega[Y^0_{i,t-1} \mid D_i = 0]\Big)}_{\mathbb{E}_\omega[Y^0_{i,t} \mid D_i = 1]}. \end{aligned}$$

Esta expresión constituye el estimador canónico 2×2 DiD. La primera diferencia interna elimina sesgos invariables entre grupos, mientras que la segunda diferencia —entre las variaciones promedio de los grupos— captura el efecto causal medio del tratamiento bajo el supuesto de PT.

En la práctica, la decisión de tratamiento suele estar determinada por actores económicos o institucionales cuyas conductas pueden correlacionarse con las tendencias de los resultados no tratados. De ahí que las aplicaciones empíricas de DiD deban evaluar explícitamente la plausibilidad de este supuesto, tanto mediante evidencia empírica como a partir de modelos teóricos sobre el proceso de selección.

La literatura reciente ha profundizado en la relación entre los mecanismos de elección del tratamiento y las propiedades temporales de los resultados potenciales. Si los agentes conocen y actúan sobre los valores futuros de $Y^0_{i,t}$, el supuesto de PT solo podría sostenerse bajo condiciones muy restrictivas, como la constancia temporal de $Y^0_{i,t}$ salvo desplazamientos comunes[14].

En contextos más realistas, PT solo es válido si las variables que determinan la selección al tratamiento dependen de componentes permanentes de los resultados potenciales —por ejemplo, efectos fijos—, pero no de fluctuaciones transitorias. Si la selección también responde a choques de corto plazo, el supuesto requerirá restricciones temporales más fuertes sobre $Y_{i,t}0$ para mantenerse válido.

Otra implicación relevante es que PT no garantiza invariancia ante transformaciones funcionales del resultado. El supuesto se refiere a promedios de $Y^0_{i,t}$ en su forma específica, y no necesariamente se preserva bajo transformaciones como logaritmos o tasas. Roth y Sant'Anna [22] demuestran que la insensibilidad funcional de PT solo se cumple si el supuesto vale tanto entre grupos como a lo largo de toda la distribución de $Y^0_{i,t}$, lo cual equivale a un escenario de adopción aleatoria o estabilidad completa de la distribución. Cuando tales condiciones no son plausibles, la validez del supuesto puede depender de la elección de escala o forma funcional del resultado.

4.3. Estimación e inferencia. El paso de la expresión poblacional del ATT_t a su forma estimable en la muestra se obtiene reemplazando los valores esperados por sus análogos muestrales ponderados. Así, el estimador DiD en el diseño 2×2 se define como

$$\widehat{ATT}_t = \left(\overline{Y}^{\omega}_{D=1,t} - \overline{Y}^{\omega}_{D=1,t-1}\right) - \left(\overline{Y}^{\omega}_{D=0,t} - \overline{Y}^{\omega}_{D=0,t-1}\right), \tag{4.6}$$

donde $\overline{Y}^{\omega}_{D=g,t}$ representa la media ponderada de la variable de resultado Y para el grupo $g \in \{0,1\}$ en el periodo t, dada por

$$\overline{Y}^{\omega}_{D=g,t=t'} = \frac{\sum_{i=1}^{n} \mathbf{1}\{D_i = g, t_i = t'\}\omega_i Y_{i,t'}}{\sum_{i=1}^{n} \mathbf{1}\{D_i = g, t_i = t'\}\omega_i}.$$

La ecuación 4.6 constituye el estimador DiD clásico expresado como la diferencia de dos diferencias de medias muestrales, cada una calculada dentro de su respectivo grupo. Es una regla de estimación directa para el ATT_t y la base de la mayoría de implementaciones empíricas del método.

El mismo resultado expresado en la ecuación 4.6 puede obtenerse mediante la estimación por mínimos cuadrados ponderados del parámetro $\beta^{2\times2}$ en el siguiente modelo lineal, definido únicamente para los dos periodos de observación:

$$Y_{i,t} = \beta_0 + \beta_1 \mathbf{1}\{D_i = 1\} + \beta_2 \mathbf{1}\{t = t_2\} + \beta^{2\times2}\left(\mathbf{1}\{D_i = 1\} \times \mathbf{1}\{t = t_2\}\right) + \varepsilon_{i,t}, \tag{4.7}$$

donde $\varepsilon_{i,t}$ representa un término idiosincrático no correlacionado con D_i, y los coeficientes β son parámetros desconocidos.

En el caso no ponderado ($\omega_i = 1$), cada una de las cuatro medias muestrales involucradas en $\widehat{ATT}_t$ puede expresarse en función de los coeficientes estimados del modelo 4.6:

$$\begin{aligned}
\overline{Y}_{D=1,t_2} &= \widehat{\beta_0} + \widehat{\beta_1} + \widehat{\beta_2} + \widehat{\beta^{2\times2}}, \\
\overline{Y}_{D=1,t_1} &= \widehat{\beta_0} + \widehat{\beta_1}, \\
\overline{Y}_{D=0,t_2} &= \widehat{\beta_0} + \widehat{\beta_2}, \\
\overline{Y}_{D=0,t_1} &= \widehat{\beta_0}.
\end{aligned}$$

Sustituyendo estas expresiones en la definición del estimador DiD se obtiene directamente que:

$$\widehat{ATT}_t = \left[(\cancel{\widehat{\beta_0}} + \cancel{\widehat{\beta_1}} + \widehat{\beta_2} + \widehat{\beta^{2\times2}}) - (\cancel{\widehat{\beta_0}} + \cancel{\widehat{\beta_1}})\right] - \left[(\cancel{\widehat{\beta_0}} + \widehat{\beta_2}) - \cancel{\widehat{\beta_0}}\right] = \widehat{\beta^{2\times2}}.$$

Por tanto, el coeficiente de interacción $\beta_{2\times2}$ del modelo lineal coincide exactamente con el estimador DiD clásico.

5. Un marco no paramétrico universal para el análisis de diferencias en diferencias

Bajo el supuesto de PT, la variación esperada en el resultado potencial no tratado entre ambos periodos es la misma para los grupos tratado y de control, lo cual permite identificar el efecto promedio del tratamiento sobre los tratados.

Sin embargo, en la práctica, la validez del supuesto PT suele verse comprometida cuando existen diferencias sistemáticas en las covariables pretratamiento entre grupos, lo que introduce sesgos en la evolución contrafactual de los resultados. Este problema ha motivado una amplia literatura orientada a relajar o sustituir dicho supuesto, mediante variantes como el PT condicional en covariables[16, 1, 25] o transformaciones no lineales del resultado[30, 21]. A pesar de estos avances, tales enfoques presentan limitaciones: se restringen en su mayoría a resultados continuos y a efectos aditivos promedio, dependen de la escala de medición del resultado, presuponen ausencia de confusores no observados o carecen de una teoría de eficiencia semiparamétrica.

Frente a ello, surge una línea metodológica alternativa que replantea la identificación causal en términos de asociaciones invariantes entre el tratamiento y los resultados potenciales no tratados. En particular, el supuesto de *Odds Ratio Equi-Confounding (OREC)* introduce una representación del sesgo de confusión en la escala del *odds ratio generalizado*, concepto desarrollado por Chen (2007)[8] y Tchetgen Tchetgen et al. (2010)[29]. Bajo este enfoque, la relación entre el tratamiento y el resultado potencial libre de tratamiento puede expresarse mediante una función de razón de momios, lo que permite "depurar" dicho sesgo sin imponer restricciones sobre el tipo de variable de resultado ni sobre su transformación.

El supuesto OREC constituye, por tanto, una generalización natural del PT en la escala del *odds ratio*. No es estrictamente más fuerte ni más débil que los supuestos tradicionales de la literatura DiD, sino una condición alternativa de identificación que debe ser desarrollada y analizada de forma independiente. Su principal virtud radica en su universalidad: permite estimar efectos causales sobre los tratados en distintas escalas de interés -incluyendo ATT y QTT-, se aplica a resultados continuos, discretos o mixtos, es invariante ante transformaciones de escala y admite la presencia de confusores no observados.

Además, la estructura analítica derivada de OREC admite una teoría completa de eficiencia semiparamétrica, lo que lo posiciona como un marco general o "universal" para la inferencia causal en contextos DiD[12].

5.1. Configuración del modelo. Sea N el número de unidades observadas, indexadas por $i \in \{1, \ldots, N\}$. Para cada unidad, se observa un vector de variables aleatorias independientes e idénticamente distribuidas (i.i.d.)

$$O = (Y_t, Y_{t+1}, A, X),$$

$A \in \{0, 1\}$ denota la asignación al tratamiento entre ambos periodos; y $X \in \mathcal{X} \subseteq \mathbb{R}^d$ corresponde al conjunto de covariables observadas de dimensión d.

Denotemos por Y_t^a el resultado potencial que se habría observado si, posiblemente en contrafactual, el tratamiento hubiese sido fijado en $A = a$ en el periodo $t \in \{0, 1\}$. El parámetro de interés es el efecto promedio del tratamiento sobre los tratados ATT, definido como

$$\tau^* = \mathbb{E}[Y_1^1 - Y_1^0 \mid A = 1] = \tau_1^* - \tau_0^*,$$

donde $\tau_a^* = \mathbb{E}[Y_1^a \mid A = 1]$.

5.1.1. Densidades condicionales y función de propensión extendida. Para desarrollar el enfoque *Universal Difference-in-Differences* (UDiD) propuesto por Tchetgen Tchetgen et al. (2024a)[12], se introducen las siguientes notaciones de densidades condicionales.

Definition 5.1. Función de propensión extendida Sea

$$f_t^*(y \mid a, x), \quad f_t^*(y, x \mid a), \quad f_t^*(y, a, x)$$

las funciones de densidad de $Y_t^0 \mid (A = a, X = x)$, $(Y_t^0, X) \mid (A = a)$ y (Y_t^0, A, X), respectivamente. Asimismo, sea $e_t^*(a \mid y, x)$ la densidad condicional de A dado $(Y_t^0 = y, X = x)$.

Para garantizar identificabilidad, se impone la siguiente condición de soporte común:

Supuesto 2. Soporte. La densidad conjunta $f_t^*(y, a, x)$ tiene el mismo soporte para todos los periodos $t \in \{0, 1\}$ y para ambos estados de tratamiento $a \in \{0, 1\}$. Es decir, existe un conjunto

$$\mathcal{S} = \{(y, x) : f_t^*(y, a, x) \in (0, \infty)\},$$

tal que el soporte de las variables observadas es común entre los grupos tratados y de control, tanto en el periodo previo como posterior al tratamiento.

Este supuesto establece que las combinaciones posibles de valores de las variables de resultado (y) y de las covariables (x) deben tener presencia positiva en todos los grupos y periodos del análisis. En otras palabras, ningún valor de x relevante para los tratados puede estar completamente ausente en el grupo de control, ni viceversa.

El propósito de esta condición es garantizar que exista una base común de comparación entre las unidades tratadas y no tratadas, de modo que el efecto del tratamiento pueda identificarse a partir de diferencias observables.

Definition 5.2. Sea y_R un valor de referencia del resultado tal que $(y_R, x) \in S$. Definimos para cada $t \in \{0, 1\}$:

(5.1)

$$\beta_t^*(x) = \frac{e_t^*(1 \mid y_R, x)}{e_t^*(0 \mid y_R, x)}, \qquad \alpha_t^*(y, x) = \frac{f_t^*(y \mid 1, x)}{f_t^*(y \mid 0, x)} \frac{f_t^*(y_R \mid 0, x)}{f_t^*(y_R \mid 1, x)} = \frac{e_t^*(1 \mid y, x)}{e_t^*(0 \mid y, x)} \frac{e_t^*(0 \mid y_R, x)}{e_t^*(1 \mid y_R, x)}.$$

La función $\alpha_t^*(y, x)$ se conoce como función de razón de momios generalizada (*generalized odds ratio function*, Chen, 2007; Tchetgen Tchetgen et al., 2010).

Por definición, $\alpha_t^*(y, x) > 0$ para todo $(y, x) \in S$, y bajo ausencia de confusión no observada (exchangeabilidad), se cumple que $\alpha_t^*(y, x) = 1$ en todo su dominio y $\alpha_t^*(y_R, x) = 1$ para todo x.

Definimos además:

Definition 5.3. Función de resultado contrafactual promedio condicional.

$$\mu^*(x) = \mathbb{E}[Y_1^0 \mid A = 1, X = x],$$

de modo que el componente τ_0^* del ATT puede escribirse como $\tau_0^* = \mathbb{E}[\mu^*(X) \mid A = 1]$.

Finalmente, denotamos por $\text{logit}(v) = \log v/(1-v)$ y $\text{expit}(v) = 1/(1+e^{-v})$ las transformaciones logísticas estándar. Para un conjunto de índices $I \subseteq \{1, \ldots, N\}$, sea

$$P_I(V) = |I|^{-1} \sum_{i \in I} V_i$$

la media empírica de V sobre dicho subconjunto, y P la media empírica sobre toda la muestra. Usaremos la notación asintótica habitual: $V_N = O_P(r_N)$ si V_N/r_N es acotado en probabilidad, $V_N = o_P(r_N)$ si converge a cero en probabilidad, y $V_N \xrightarrow{D} W$ para convergencia débil. Finalmente, $V \mid Z \overset{D}{=} W \mid Z$ indica igualdad en distribución condicional en Z.

5.1.2. Revisión de enfoques en DiD. Antes de introducir el supuesto *OREC*, conviene revisar los supuestos tradicionales que sustentan la identificación causal en los modelos DiD.

Supuesto 3. Consistencia.

$$Y_t = Y_t^A \quad \text{casi seguramente, para todo } t \in \{0, 1\}.$$

El Supuesto de *consistencia* establece que el resultado observado coincide con el resultado potencial correspondiente al tratamiento efectivamente recibido. El Supuesto de no anticipación, por su parte, impone que la intervención no tiene efectos causales sobre los resultados antes de su implementación. En consecuencia, bajo ambos supuestos se cumple que $Y_0 = Y_0^0$ para todas las unidades, independientemente de su estado de tratamiento.

Bajo el Supuesto 3, el primer componente del parámetro $\tau^* = \mathbb{E}[Y_1^1 - Y_1^0 \mid A = 1]$ se identifica directamente como $\tau_1^* = \frac{\mathbb{E}[AY_1]}{\Pr(A=1)}$.
Por tanto, la identificación del ATT requiere únicamente establecer condiciones que garanticen la identificabilidad del segundo término, $\tau_0^* = \mathbb{E}[Y_1^0 \mid A = 1]$.

Para ello, consideremos el modelo clásico propuesto por Athey e Imbens (2006)[2] para el resultado potencial libre de tratamiento Y_t^0, suprimiendo las covariables para simplificar la exposición. Se asume que, para $t \in \{0, 1\}$, el proceso generador cumple:

(2)

$$\text{(Modelo DiD): } Y_t^0 = h_t(U_t), \quad h_t(u) = u + b_T t, \quad U_t = b_0 + b_A A + \varepsilon_t,$$

(3)

ε_t satisface ya sea invariancia temporal: $\varepsilon_1 \mid A \overset{D}{=} \varepsilon_0 \mid A$ o,

(4) invariancia respecto al tratamiento: $\varepsilon_t \mid (A = 0) \overset{D}{=} \varepsilon_t \mid (A = 1)$.

donde ε_t es un término de error no observado que se mantiene invariante en el tiempo o respecto al tratamiento.
Dicha estructura implica dos posibles restricciones estocásticas:

En este modelo, U_t es una variable no observada y Y_t^0 es una función determinista lineal de ella. El mecanismo de asignación A condicionado en U_t permanece, sin embargo, sin restricción funcional. Además, el modelo DiD clásico supone preservación de rangos (*rank preservation*), es decir, que no existen interacciones aditivas entre A y U_t en la determinación de Y_t^0.

Mediante un razonamiento algebraico sencillo, puede demostrarse que este modelo implica el conocido supuesto de tendencias paralelas, formulado condicionalmente en covariables como

$$\text{(PT)} \quad \mathbb{E}[Y_1^0 - Y_0^0 \mid A = 1, X] - \mathbb{E}[Y_1^0 - Y_0^0 \mid A = 0, X] \quad \text{c.s.}$$

Este supuesto establece que, en ausencia del tratamiento, las tendencias promedio de los resultados potenciales son equivalentes entre los grupos tratado y de control, una vez condicionadas en las covariables observadas.

Bajo los Supuestos 3, 1 y PT, se obtiene de manera directa la expresión de identificación del ATT:

$$\tau^* = \mathbb{E}\Big[\mathbb{E}(Y_1 \mid A = 1, X) - \mathbb{E}(Y_1 \mid A = 0, X) + \mathbb{E}(Y_0 \mid A = 0, X) - \mathbb{E}(Y_0 \mid A = 1, X) \Big| A = 1\Big],$$

lo cual justifica la construcción tradicional del estimador DiD bajo el supuesto de tendencias paralelas. Este supuesto puede interpretarse como una restricción sobre el grado de sesgo de confusión que afecta la asociación aditiva entre el tratamiento A y el resultado potencial no tratado Y_1^0. En efecto, el supuesto PT puede reescribirse de manera equivalente como $\mathbb{E}[Y_0^0 \mid A = 1, X] - \mathbb{E}[Y_0^0 \mid A = 0, X] - \mathbb{E}[Y_1^0 \mid A = 1, X] - \mathbb{E}[Y_1^0 \mid A = 0, X]$.
El lado derecho de la ecuación sería nulo en ausencia de confusión condicional en X; por tanto, cualquier desviación de cero cuantifica la magnitud del sesgo de confusión en la escala aditiva, aunque dicho sesgo no pueda observarse directamente. La igualdad anterior establece que el sesgo aditivo posterior al tratamiento es identificable a partir del sesgo aditivo previo al mismo.

En este sentido, el supuesto de PT es equivalente a la denominada condición de estabilidad del sesgo[16, 19], también conocida como supuesto de equi-confusión aditiva (additive equi-confounding assumption, [26]). Bajo esta formulación, el grado de confusión se evalúa en la escala aditiva, es decir, mediante la diferencia entre las medias condicionales contrafactuales según el estado observado del tratamiento.

A pesar de su utilidad y simplicidad, el supuesto de PT puede resultar incompatible con las restricciones naturales que presentan ciertos tipos de variables de resultado. En contextos donde los resultados son acotados o discretos -por ejemplo, binarios, de conteo o proporciones-, la formulación aditiva del PT puede inducir valores contrafactuales que exceden el dominio posible del resultado, comprometiendo así su plausibilidad empírica. Esta incompatibilidad pone de relieve que el PT, formulado en la escala lineal, puede no ser apropiado en situaciones donde la naturaleza del resultado impone límites estructurales.

Una limitación adicional del PT clásico es su falta de extensibilidad a medidas de efecto no lineales, como el efecto cuantil del tratamiento sobre los tratados (QTT). En estos casos, la interpretación aditiva del sesgo o de las diferencias medias resulta insuficiente, dado que la relación causal puede manifestarse de forma no lineal a lo largo de la distribución del resultado.

Con el fin de abordar estas limitaciones, Puhani (2012)[21] y Wooldridge (2022)[30] propusieron el denominado supuesto de tendencias paralelas no lineales (*Nonlinear Parallel Trends*, NPT). Este supuesto establece que las expectativas condicionales transformadas de los resultados potenciales satisfacen el principio de tendencias paralelas bajo una transformación monótona $L(\cdot)$, de manera que

$$\begin{aligned}\text{(NPT)}: & L\big(\mathbb{E}[Y_1^0 \mid A=1, X]\big) - L\big(\mathbb{E}[Y_0^0 \mid A=1, X]\big) \\ & L\big(\mathbb{E}[Y_1^0 \mid A=0, X]\big) - L\big(\mathbb{E}[Y_0^0 \mid A=0, X]\big).\end{aligned}$$

Esta formulación generaliza el supuesto PT al permitir una relación funcional no lineal entre el resultado y el tratamiento, preservando la estructura comparativa en una escala transformada por $L(\cdot)$.

Diversos enfoques alternativos han sido desarrollados para identificar efectos del tratamiento en entornos DiD no lineales. En particular, Athey e Imbens (2006) introducen el modelo de cambios-en-cambios (*Changes-in-Changes*, CiC) para resultados continuos, definido para $t \in \{0, 1\}$ por

$$\text{(CiC model)} Y_t^0 = h_t(U_t), \tag{5}$$

$$U_1 \mid A \stackrel{D}{=} U_0 \mid A. \tag{6}$$

En este modelo, U_t representa una variable no observada de distribución continua, y $h_t(\cdot)$ una transformación temporal estrictamente monótona (o no decreciente en el caso de resultados discretos). Bajo estas condiciones, la distribución contrafactual de $Y_1^0 \mid (A = 1)$ puede identificarse de manera no paramétrica a partir de los datos observados.

Para resultados discretos, la identificación puntual adicional requiere supuestos de independencia condicional tales como $U_t \perp A \mid Y_t$ para $t \in \{0, 1\}$, o formulaciones equivalentes del tipo $Y_t^0 = h_t(U_t, X)$ donde X es un conjunto continuo de covariables que satisface $U_t \perp X \mid A$.

Por su parte, Bonhomme y Sauder (2011) [5] consideran el caso de un resultado continuo generado por un modelo aditivo, en el cual las funciones características logarítmicas de $Y_t^0 \mid A$ satisfacen el supuesto de PT en la escala logarítmica. Bajo esta formulación, la función característica de $Y_t^0 \mid (A = 1)$ se identifica a partir de la condición PT en la escala logarítmica, y la distribución correspondiente puede recuperarse utilizando la relación biyectiva entre una distribución y su función característica.

A partir de esta idea, Fan y Yu (2012)[13] introducen una versión distributiva del supuesto DiD, postulando que la variación en los resultados potenciales libres de tratamiento a lo largo del tiempo es independiente del estado de tratamiento,

esto es, $Y_1^0 - Y_0^0 \perp A$. Este supuesto, conocido como *Distributional difference-in-differences*, se ha utilizado en trabajos posteriores (Callaway et al., 2018; Callaway y Li, [7]), aunque por sí solo resulta insuficiente para identificar la distribución contrafactual de $Y_t^0 \mid (A = 1)$.

Para lograr dicha identificación, Callaway et al. (2018) y Callaway y Li (2019) introducen el supuesto de estabilidad de cópulas (*copula stability assumption*), expresado en el diseño canónico como $C_{Y_0^0,,Y_1^0-Y_0^0|A=0} = C_{Y_0^0,,Y_1^0-Y_0^0|A=1}$, donde $C_{V,W|Z}$ denota la función cópula condicional de las variables aleatorias V y W dado Z. Este supuesto establece que la estructura de dependencia entre el resultado previo al tratamiento y el cambio temporal en el resultado potencial libre de tratamiento es invariante entre los grupos tratado y no tratado.

Finalmente, Ding y Li (2019)[11] aplican la ignorabilidad secuencial (*sequential ignorability*, véase Hernán y Robins, 2020[17]) al contexto DiD canónico como supuesto de identificación. En este caso, se asume la ausencia de confusores no observados que afecten simultáneamente la relación entre el resultado potencial posterior al tratamiento y la variable de tratamiento, una vez controlados el resultado previo y las covariables observadas, esto es: $Y_t^0 \perp A \mid (Y_0, X)$.

Estos desarrollos ilustran la diversidad de estrategias existentes para relajar el supuesto de PT abordando distintos aspectos del problema de identificación en presencia de no linealidad, heterogeneidad o confusión no observada. No obstante, cada uno de estos enfoques presenta limitaciones específicas -ya sea en términos de aplicabilidad, eficiencia semiparamétrica o robustez estructural-, lo cual motiva la introducción del supuesto Odds Ratio Equi-Confounding (OREC) como un marco alternativo, general e invariante a la escala de medición del resultado.

5.2. Modelo generativo. A fin de generalizar los modelos clásicos DiD y CiC, consideremos la formulación estructural introducida por Tchetgen Tchetgen. Su propósito es establecer un marco unificado de identificación causal aplicable a distintos tipos de resultados y medidas de efecto. Suponiendo, para simplificar la exposición, la ausencia de covariables explícitas, se define el siguiente modelo generativo:

$$\text{(UDiD model)} : Y_t^0 \perp A \mid U_t, \tag{7}$$

$$A \mid (U_1 = u) \overset{D}{=} A \mid (U_0 = u) \quad \text{para todo } u, \tag{8}$$

$$U_1 \mid (A = 0, Y_1 = y) \overset{D}{=} U_0 \mid (A = 0, Y_0 = y) \quad \text{para todo } y. \tag{9}$$

Comparado con los modelos DiD y CiC revisados en la sección anterior, el modelo UDiD introduce un conjunto de supuestos conceptualmente distintos y, en general, menos restrictivos.

En primer lugar, la condición (7) representa una ignorabilidad latente, en el sentido de que la independencia entre el tratamiento A y el resultado potencial no tratado Y_t^0 se cumple condicionalmente en una variable latente U_t. A diferencia de los modelos previos -(2) y (6)- donde Y_t^0 se especifica como una función determinista de U_t, la relación entre ambos no se impone estructuralmente en (7). En

consecuencia, el modelo UDiD relaja de forma sustantiva las restricciones funcionales de los marcos DiD y CiC.

En segundo lugar, la condición (8) establece la invariancia temporal del mecanismo de tratamiento, es decir, que la distribución condicional de A dado U_t permanece estable entre periodos. Este supuesto no está presente en los modelos DiD ni CiC, y representa una forma de estabilidad del proceso de asignación del tratamiento a lo largo del tiempo.

Finalmente, la condición (9) impone la estabilidad temporal de la distribución condicional de U_t dado el resultado observado entre las unidades no tratadas. Este supuesto guarda relación con la condición de invariancia temporal del error (3) en el modelo DiD y con la condición de estabilidad (6) en el modelo CiC, pero presenta diferencias conceptuales relevantes: (i) la estabilidad se requiere únicamente para el grupo no tratado, y (ii) la condición involucra el resultado observado Y_t en su argumento condicional.

Por tanto, las condiciones (3), (6) y (9) pueden entenderse como contrapartes marginales y condicionales de un mismo principio de estabilidad, aunque no son anidadas entre sí —del mismo modo que los supuestos de PT marginal y condicional tampoco lo son.

Al igual que DiD y CiC, UDiD permite selección en variables no observadas, pues la distribución de U_t puede diferir entre tratados y no tratados. No obstante, el modelo posee propiedades distintivas: (i) al igual que CiC, es invariante ante transformaciones monótonas del resultado, lo cual no ocurre en DiD; (ii) no impone estructura aditiva sobre la interacción $A \times U_t$; (iii) admite resultados continuos, discretos o mixtos; y (iv) exige invariancia temporal del mecanismo de tratamiento, lo que proporciona una base de identificación más robusta ante cambios en el tiempo.

5.3. Supuesto de Odds Ratio Equi-Confounding (OREC). Al igual que en los modelos DiD y CiC, el ATT es identificable bajo el modelo UDiD. No obstante, como señalan Tchetgen Tchetgen et al. (2024a), una condición más débil -implicada por dicho modelo- es suficiente para garantizar la identificación.

Supuesto 4. Odds Ratio Equi-Confounding.

$$\alpha_0^*(y,x) = \alpha_1^*(y,x) \quad \text{para todo } (y,x) \in S,$$

donde $\alpha_t^*(y,x)$ denota la función de razón de momios generalizada (*generalized odds ratio function*, Chen, 2007[8]; Tchetgen Tchetgen et al., 2010[29]) que caracteriza la asociación entre el tratamiento A y el resultado potencial no tratado Y_t^0.

La función $\alpha_t^*(y,x)$ proporciona una medida de la magnitud del sesgo de confusión en la escala del *odds ratio*. En particular, si $\alpha_t^*(y,x) = 1$ para todo (y,x), ello implica ausencia de asociación entre A y Y_t^0 condicionalmente en $X = x$, es decir, ausencia de confusión dada X. Por tanto, la igualdad $\alpha_0^*(y,x) = \alpha_1^*(y,x)$ establece que el sesgo de confusión -medido en la escala del *odds ratio*- es estable a lo largo del tiempo entre los periodos $t = 0$ y $t = 1$. De ahí su denominación como supuesto de OREC.

Este supuesto representa una generalización multiplicativa del principio de estabilidad del sesgo formulado en la escala aditiva bajo el supuesto PT. A diferencia de este último, el OREC es invariante ante transformaciones monótonas del resultado, lo que permite su aplicación en contextos con variables de tipo discreto, continuo o mixto. Además, no requiere que la distribución del resultado pertenezca a la familia exponencial, lo cual amplía sustancialmente su dominio de validez y lo convierte en un marco unificado para la corrección del sesgo de confusión en estimaciones causales sobre los tratados.

Si bien la condición OREC se deriva naturalmente del modelo estructural UDiD, es importante destacar que su formulación puede expresarse directamente en términos de resultados contrafactuales, sin necesidad de hacer referencia explícita al factor latente U_t que confunde la asociación entre el tratamiento A y el resultado potencial no tratado Y_1^0.

Para ilustrar este punto, considérese el supuesto de tendencias paralelas no lineales (*Nonlinear Parallel Trends*, NPT), en el cual el resultado es binario y se utiliza la función de enlace logit. Bajo este enfoque, el supuesto NPT equivale a imponer un modelo para el resultado en el periodo $t \in \{0, 1\}$ de la forma

$$\operatorname{logit}\left\{\mathbb{E}[Y_t^0 \mid A = a, X = x]\right\} = b_0(x) + a \cdot b_1(x) + t \cdot b_2(x),$$

donde b_0, b_1 y b_2 son funciones de las covariables. Bajo OREC, la asociación entre el tratamiento A y los resultados potenciales no tratados Y_t^0 se mantiene constante en el tiempo, pero sin depender de una función de enlace específica ni de la naturaleza del resultado.

Así, mientras el NPT se formula dentro de un marco paramétrico particular (logístico o exponencial), el supuesto OREC constituye una condición semiparamétrica de estabilidad del sesgo multiplicativo, válida para resultados discretos, continuos o mixtos, y que no requiere especificar una relación funcional entre el resultado y el tratamiento.

Más generalmente, el supuesto OREC puede interpretarse como una versión del supuesto de PT aplicada al mecanismo de exposición extendido en la escala logit. Tomando logaritmos en ambos lados de la igualdad definida por OREC, se obtiene la siguiente condición:

$$\operatorname{logit}\big(e_1^*(1 \mid y, x)\big) - \operatorname{logit}\big(e_1^*(1 \mid y_R, x)\big) = \operatorname{logit}\big(e_0^*(1 \mid y, x)\big) - \operatorname{logit}\big(e_0^*(1 \mid y_R, x)\big),$$

$\forall (y, x) \in S$. En palabras, el cambio en los log-odds asociados con la función de propensión extendida es constante en el tiempo para todo $(y, x) \in S$; es decir, existe una relación paralela en los log-odds de dicha función entre los periodos.

$$\Pr(A = 1 \mid Y_t^0, X) \quad t \in \{0, 1\}.$$

5.4. Propiedades del enfoque UDiD. Para concluir esta sección, resumimos las propiedades esenciales del enfoque desarrollado bajo el supuesto Odds Ratio Equi-Confounding (OREC).

El enfoque UDiD, fundamentado en el supuesto OREC, reúne un conjunto de propiedades que lo distinguen de los modelos tradicionales en diferencias en diferencias. En primer lugar, admite sin restricciones resultados continuos, discretos o mixtos, evitando las limitaciones estructurales presentes en enfoques como CiC o PT en transformaciones específicas. En segundo lugar, cuando el resultado pertenece a la familia exponencial, OREC adquiere una interpretación natural en términos de estabilidad temporal de los parámetros canónicos.

Una tercera propiedad clave es su invariancia de escala: las condiciones de identificación no dependen de transformaciones particulares del resultado, lo que elimina la necesidad de escoger un dominio funcional "correcto", un requisito habitual en modelos basados en PT o NPT. Además, el marco permite la presencia de confusores no observados siempre que su influencia sobre el mecanismo de tratamiento sea estable en la escala del *odds ratio*.

Desde el punto de vista teórico, UDiD es plenamente no paramétrico y cuenta con una caracterización explícita de la cota de eficiencia semiparamétrica para los efectos del tratamiento, junto con condiciones suficientes para que el estimador propuesto la alcance. Por combinar simultáneamente compatibilidad con diferentes tipos de resultado, invariancia de escala, robustez frente a confusión no observada y eficiencia semiparamétrica, el supuesto OREC configura un marco verdaderamente universal para la estimación de efectos causales en diseños DiD, tal como se resume en la Tabla 5.4.

	Rango resultados		Estimación		Eficiencia Semiparamétrica		Escala	Factor de
Supuestos	$\mathbb{R}$	$\{0,1\}$	ATT	QTT	ATT	QTT	Invariancia	confución
PT	✓	✓	✓	✗	✓	✗	✗	✓
NPT	✓	✓	✓	✗	✗	✗	✗	✓
CiC	✓	✓	✓	✓	✗	✗	✓	✓
PT con log	✓	✗	✓	✓	✗	✗	✗	✓
Copula invarianza	✓	✗	✓	✓	✗	✗	✗	✓
Ignorabilidad secuencial	✓	✓	✓	✓	✓	✓	✓	✗
OREC	✓	✓	✓	✓	✓	✓	✓	✓

CUADRO 1. Una comparación de enfoques para entornos de diferencias en diferencias. La marca de verificación 3 indica que se cumple un criterio bajo la suposición identificadora y las condiciones adicionales requeridas por trabajos previos. La cruz 7 indica que un criterio no se cumple.

6. SIMULACIÓN

Con el fin de evaluar el comportamiento en muestras finitas del estimador propuesto bajo el supuesto OREC, se implementó un estudio de simulación Monte Carlo en dos escenarios: uno con resultado continuo y otro con resultado binario. En ambos casos se construyó deliberadamente un diseño en el que la ignorabilidad condicional falla, OREC es válido y el PT es violado, de modo que la comparación con métodos estándar DiD sirve como prueba de estrés para el enfoque planteado.

Diseño: resultado continuo. En el primer escenario se consideró un resultado continuo. Para cada unidad se generaron dos covariables observadas $X = (X_1, X_2)$, con $X_1, X_2 \sim \mathcal{N}(0,1)$ independientes. El indicador de tratamiento A se simuló a partir de un modelo logístico

$$A \sim \text{Ber}\big(\,\text{expit}\{0{,}1(X_1 + X_2)\}\big),$$

de forma que la probabilidad de tratamiento depende de las covariables. Los potenciales resultados se especificaron como

$$Y_0^0 \mid (A, X) \sim \mathcal{N}\Big(3 + 0{,}01\big(5 + 2X_1 + 2X_2\big)A + 0{,}1(X_1 + X_2),\, 4\Big), \quad Y_0^1 = Y_0^0,$$

$$Y_1^{(a)} \mid (A, X) \sim \mathcal{N}\Big(3{,}5 + 0{,}5a + 0{,}01\big(5 + 2X_1 + 2X_2\big)A + 0{,}1(X_1 + X_2),\, 1\Big), \quad a \in \{0, 1\}.$$

La dependencia explícita de Y_t^0 con A implica que la ignorabilidad condicional respecto de X no se cumple. Al mismo tiempo, el sesgo de confusión puede representarse mediante una razón de momios generalizada que permanece estable en el tiempo, de modo que el supuesto OREC es válido con

$$\alpha_1^*(y, x) = \exp\big\{0{,}01\, y\,(5 + 2x_1 + 2x_2)\big\},$$

mientras que PT no se verifica. En este diseño el efecto medio del tratamiento sobre los tratados en el periodo post, ATT, es igual a 0,5.

Se consideraron tamaños muestrales $N \in \{500, 1000, 1500, 2000\}$. Para cada réplica se generaron los datos observados (Y_0, Y_1, A, X) a partir de los potenciales resultados y del mecanismo de tratamiento, y se estimaron dos cantidades:

- el estimador propuesto bajo OREC, $\hat{\tau}_{\text{OREC}}$, construido a partir de la función de influencia eficiente y el esquema de *cross-fitting* descrito en la Sección 5.2;
- un estimador DiD estándar basado en PT, $\hat{\tau}_{\text{PT}}$, implementado mediante el procedimiento de Sant'Anna y Zhao (2020) y Callaway y Sant'Anna (2021).

El desempeño de ambos estimadores se evaluó a partir de 1000 réplicas Monte Carlo para cada valor de N, analizando sesgo, error estándar empírico y cobertura de intervalos de confianza al 95 %.

Diseño: resultado binario. En el segundo escenario se consideró un resultado binario, manteniendo las mismas distribuciones para las covariables y el tratamiento. Los potenciales resultados se generaron según

$$Y_0^0 \mid (A, X) \sim \text{Ber}\Big(\,\text{expit}\{-0{,}75 + (1{,}5 - 0{,}2X_1 - 0{,}2X_2)A + 0{,}1X_1 + 0{,}1X_2\}\Big), \quad Y_0^1 = Y_0^0,$$

$$Y_1^{(a)} \mid (A, X) \sim \text{Ber}\Big(\,\text{expit}\{0{,}5 + (1{,}5 - 0{,}2X_1 - 0{,}2X_2)A + 0{,}1X_1 + 0{,}1X_2\}\Big), \quad a \in \{0, 1\}.$$

Este diseño mantiene la violación de ignorabilidad condicional y, al mismo tiempo, satisface OREC con

$$\alpha_1^*(y, x) = \exp\big\{y\,(1{,}5 - 0{,}2x_1 - 0{,}2x_2)\big\},$$

mientras que PT vuelve a fallar. Aquí las distribuciones de Y_1^0 y Y_1^1 coinciden para las unidades tratadas, de modo que la ATT verdadera es igual a cero. La estimación en este caso se realizó utilizando la versión binaria del procedimiento propuesto, que explota simplificaciones específicas de la escala Bernoulli. Se utilizaron los mismos tamaños muestrales y el mismo número de réplicas que en el escenario continuo.

Resultados principales. En ambos escenarios, el estimador basado en OREC mostró un sesgo empírico prácticamente nulo incluso en muestras moderadas, mientras que el estimador sustentado en PT exhibió sesgos sistemáticos, coherentes con la violación deliberada de dicho supuesto. A medida que N aumenta, las desviaciones estándar de $\hat{\tau}_{\text{OREC}}$ decrecen de forma compatible con un comportamiento de raíz-N, y los intervalos de confianza construidos a partir de la desviación estándar asintótica y de procedimientos de remuestreo tipo *multiplier bootstrap* alcanzan coberturas cercanas al nivel nominal del 95 %. Estas evidencias empíricas son consistentes con las propiedades asintóticas establecidas para el estimador eficiente bajo el supuesto OREC y respaldan su uso en contextos donde las hipótesis de tendencias paralelas resultan poco plausibles.

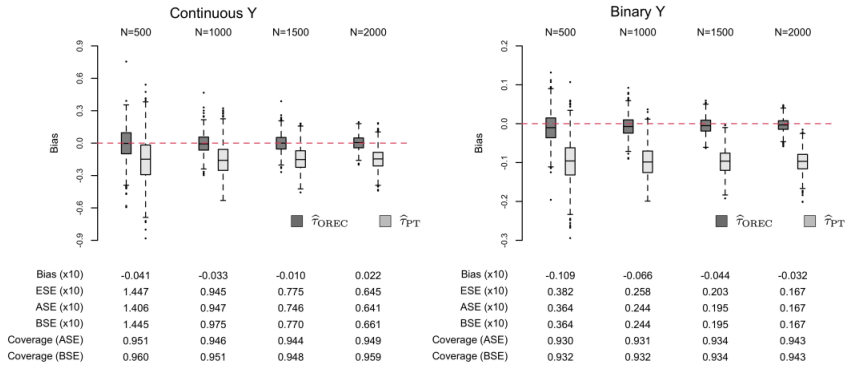


FIGURA 1. Resumen gráfico de los resultados de simulación. Los paneles izquierdo y derecho muestran los resultados para los casos con resultado continuo y binario, respectivamente. En la parte superior, cada columna presenta *boxplots* del sesgo de los estimadores $\hat{\tau}_{\text{OREC}}$ y $\hat{\tau}_{\text{PT}}$ para $N \in \{500, 1000, 1500, 2000\}$. La parte inferior reporta, para $\hat{\tau}_{\text{OREC}}$, el sesgo medio (Bias), el error estándar asintótico (ASE), el error estándar empírico (ESE), el error estándar por bootstrap (BSE) y la cobertura empírica de los intervalos de confianza al 95 % basados en ASE y BSE. Los valores de sesgo y errores estándar se muestran reescalados por un factor de 10.

7. CONCLUSIONES

El análisis desarrollado muestra que los diseños DiD clásicos descansan sobre una estructura aditiva que se vuelve frágil cuando los efectos del tratamiento son heterogéneos, las trayectorias entre grupos divergen antes o después de la intervención, o el resultado está restringido a un soporte acotado. En ese entorno, el supuesto de tendencias paralelas deja de ser una simplificación técnica y pasa a operar como una restricción estructural fuerte: condiciona la escala en la que debe medirse el resultado, limita el análisis a parámetros promedio y ofrece pocas garantías cuando el interés se desplaza hacia efectos distribucionales o cuantilísticos.

El marco OREC–UDiD desplaza el foco desde la igualdad de tendencias aditivas hacia la estabilidad temporal de la asociación entre tratamiento y resultado potencial en la escala del *odds ratio* generalizado. Esa reparametrización permite trabajar con resultados continuos, discretos o mixtos, mantiene la identificación frente a transformaciones monótonas del resultado y admite la presencia de confusión no observada siempre que su efecto actúe de forma estable en dicha escala. Bajo estas condiciones, parámetros como el ATT y el QTT se obtienen a partir de expresiones no paramétricas en las que el contrafactual de los tratados queda caracterizado de manera explícita.

Desde la perspectiva estadística, la derivación de la función de influencia eficiente y la construcción de un estimador basado en *cross-fitting* permiten combinar estimación flexible de densidades, razones de densidad y regresiones de resultado con propiedades asintóticas de raíz–(n). La estructura de sesgo mixto garantiza que la consistencia y la normalidad asintótica se preservan aun cuando no todos los componentes auxiliares convergen a la misma velocidad, siempre que un subconjunto suficiente de ellos lo haga a tasas adecuadas. Las simulaciones con resultados continuos y binarios confirman este comportamiento: el estimador OREC mantiene sesgos cercanos a cero, errores estándar acordes con la teoría y coberturas próximas al nivel nominal precisamente en configuraciones donde los estimadores basados en PT se desalinean de manera sistemática.

El marco abre varias extensiones naturales. Por un lado, la posibilidad de identificar QTT y otros funcionales distribucionales sugiere trabajar con curvas completas de efectos —por ejemplo, perfiles de impacto a lo largo de la distribución del ingreso, del riesgo o de la productividad— en lugar de concentrarse únicamente en promedios. Por otro lado, la presencia explícita de factores latentes y de estructuras de dependencia en el modelo de odds ratio ofrece un punto de encuentro con los modelos de ecuaciones estructurales (SEM), en los que tratamiento, covariables, resultados potenciales y confusores no observados pueden representarse de manera conjunta. Extender OREC–UDiD a paneles de múltiples periodos, patrones de tratamiento más generales y versiones relajadas del supuesto de equi-confusión —por ejemplo, permitiendo relaciones del tipo $\alpha_1 = \varphi(\alpha_0)$— abre un espacio prometedor donde la inferencia causal semiparamétrica y la modelación estructural pueden dialogar de forma más estrecha, manteniendo siempre visibles los supuestos que sostienen la interpretación causal de los parámetros estimados.

REFERENCIAS

[1] Alberto Abadie. "Semiparametric Difference-in-Differences Estimators". En: *The Review of Economic Studies* 72.1 (2005), págs. 1-19.

[2] Susan Athey y Guido W Imbens. "Identification and inference in nonlinear difference-in-differences models". En: *Econometrica* 74.2 (2006), págs. 431-497.

[3] Andrew Baker et al. "Difference-in-differences designs: A practitioner's guide". En: *arXiv preprint arXiv:2503.13323* (2025).

[4] Andrew C Baker, David F Larcker y Charles CY Wang. "How much should we trust staggered difference-in-differences estimates?" En: *Journal of Financial Economics* 144.2 (2022), págs. 370-395.

[5] Stéphane Bonhomme y Ulrich Sauder. "Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling". En: *Review of Economics and Statistics* 93.2 (2011), págs. 479-494.

[6] Brantly Callaway y Tong Li. "Quantile treatment effects in difference in differences models with panel data". En: *Quantitative Economics* 10.4 (2019), págs. 1579-1618.

[7] Brantly Callaway, Tong Li y Tatsushi Oka. "Quantile Treatment Effects in Difference in Differences Models under Dependence Restrictions and with Only Two Time Periods". En: *Journal of Econometrics* 206.2 (2018), págs. 395-413.

[8] Hua Yun Chen. "A semiparametric odds ratio model for measuring association". En: *Biometrics* 63.2 (2007), págs. 413-421.

[9] Clément De Chaisemartin y Xavier d'Haultfoeuille. "Fuzzy differences-in-differences". En: *The Review of Economic Studies* 85.2 (2018), págs. 999-1028.

[10] Peng Ding y Fan Li. "A Bracketing Relationship Between Difference-in-Differences and Lagged-Dependent-Variable Adjustment". En: *Political Analysis* 27.4 (2019), págs. 605-615.

[11] Peng Ding y Fan Li. "A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment". En: *Political Analysis* 27.4 (2019), págs. 605-615.

[12] Oliver Dukes et al. "Using negative controls to identify causal effects with invalid instrumental variables". En: *Biometrika* 112.1 (2025), asae064.

[13] Yanqin Fan y Zhengfei Yu. "Partial identification of distributional and quantile treatment effects in difference-in-differences models". En: *Economics Letters* 115.3 (2012), págs. 511-515.

[14] Dalia Ghanem, Désiré Kédagni e Ismael Mourifié. "Evaluating the impact of regulatory policies on social welfare in difference-in-difference settings". En: *arXiv preprint arXiv:2306.04494* (2023).

[15] Andrew Goodman-Bacon. "Difference-in-differences with variation in treatment timing". En: *Journal of econometrics* 225.2 (2021), págs. 254-277.

[16] James J. Heckman, Hidehiko Ichimura y Petra E. Todd. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme". En: *The Review of Economic Studies* 64.4 (1997), págs. 605-654.

[17] MA Hernan y JM Robins. *Causal inference: What if chapman hall/crc, boca raton*. 2020.

[18] Guido W Imbens y Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[19] Michael Lechner et al. “The estimation of causal effects by difference-in-difference methods”. En: *Foundations and Trends® in Econometrics* 4.3 (2011), págs. 165-224.
[20] Chan Park y Eric J. Tchetgen Tchetgen. “A Universal Nonparametric Framework for Difference-in-Differences Analyses”. En: *arXiv preprint arXiv:2212.13641v5* (2022). arXiv: 2212.13641 [stat.ME].
[21] Patrick A. Puhani. “The Treatment Effect, the Cross Difference, and the Interaction Term in Nonlinear “Difference-in-Differences” Models”. En: *Economics Letters* 115.1 (2012), págs. 85-87.
[22] Jonathan Roth. “Pretest with caution: Event-study estimates after testing for parallel trends”. En: *American Economic Review: Insights* 4.3 (2022), págs. 305-322.
[23] Donald B Rubin. “Estimating causal effects of treatments in randomized and nonrandomized studies.” En: *Journal of educational Psychology* 66.5 (1974), pág. 688.
[24] Pedro H. C. Sant’Anna y Jun Zhao. “Doubly robust difference-in-differences estimators”. En: *Journal of Econometrics* 219.1 (2020), págs. 101-122.
[25] Pedro HC Sant’Anna y Jun Zhao. “Doubly robust difference-in-differences estimators”. En: *Journal of econometrics* 219.1 (2020), págs. 101-122.
[26] Tamar Sofer et al. “On negative outcome control of unobserved confounding as a generalization of difference-in-differences”. En: *Statistical science: a review journal of the Institute of Mathematical Statistics* 31.3 (2016), pág. 348.
[27] Gary Solon, Steven J Haider y Jeffrey M Wooldridge. “What are we weighting for?” En: *Journal of Human resources* 50.2 (2015), págs. 301-316.
[28] Liyang Sun y Sarah Abraham. “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects”. En: *Journal of econometrics* 225.2 (2021), págs. 175-199.
[29] Eric J Tchetgen Tchetgen, James M Robins y Andrea Rotnitzky. “On doubly robust estimation in a semiparametric odds ratio model”. En: *Biometrika* 97.1 (2010), págs. 171-180.
[30] Jeffrey M Wooldridge. “Simple approaches to nonlinear difference-in-differences with panel data”. En: *The Econometrics Journal* 26.3 (2023), págs. C31-C66.

Maestría en Matemática, Universidad Nacional Autónoma de Honduras.
Dirección de correo electrónico: ajsanchezr@unah.hn

MODELACIÓN ESPACIAL Y VALIDACIÓN GEOESTADÍSTICA DE ESTIMACIONES SATELITALES CHIRPS CON DATOS DE ESTACIONES TERRESTRES EN CUENCAS HIDROGRÁFICAS DE HONDURAS

KEVIN FERNANDO VASQUEZ ZERONY[1] ANDRES FARALL[2]

Resumen. La modelación hidrológica depende en gran medida de una representación precisa de la precipitación, variable clave en la gestión del agua, la planificación territorial y la mitigación de desastres naturales. En Honduras, su análisis enfrenta limitaciones debido a la baja densidad, distribución irregular y discontinuidad de la red pluviométrica nacional. En este contexto, los datos satelitales CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data) ofrecen una alternativa valiosa al proporcionar cobertura casi global desde 1981 con resolución espacial de 0.05° (5 km); sin embargo, sus estimaciones presentan sesgos sistemáticos que requieren corrección local.
Este estudio tiene como objetivo **evaluar y corregir el sesgo de las estimaciones CHIRPS frente a datos de estaciones terrestres y, posteriormente, modelar su distribución espacial** mediante el uso de técnicas geoestadísticas avanzadas aplicadas en las principales cuencas hidrográficas de Honduras durante el período 1981–2023. **Para la corrección de sesgos se implementan métodos estadísticos reconocidos en la literatura, como el Escalamiento Lineal (Linear Scaling), la Transformación de Potencia (Power Transformation) y el Mapeo de Cuantiles (Quantile Mapping), empleados en estudios previos de validación de productos CHIRPS en África y América Central. Posteriormente, se emplean tres métodos geoestadísticos para estimar la precipitación en puntos de difícil acceso o en zonas sin pluviómetros: el Kriging Ordinario (OK), que considera la autocorrelación espacial; el Kriging Universal (UK), que incorpora covariables topográficas; y el Co-Kriging, que combina información satelital y observaciones terrestres aprovechando su correlación cruzada.** La validación se realiza mediante métricas estadísticas como r , R^2, NSE, RMSE, MAE y sesgo, utilizando validación cruzada con datos de estaciones meteorológicas nacionales. Los resultados permitirán generar campos de precipitación corregidos que mejoran la modelación hidrológica, el balance hídrico y el diseño de infraestructura hidráulica, fortaleciendo la gestión integrada de los recursos hídricos en Honduras.

Fecha: 19 Agosto 2025.

Abstract. Hydrological modeling strongly depends on an accurate representation of precipitation, a key variable for water resources management, land-use planning, and disaster risk reduction. In Honduras, precipitation analysis is constrained by the low density, irregular distribution, and discontinuity of the national rain-gauge network. In this context, CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data) provides a valuable near-global record since 1981 at 0.05° (5 km) spatial resolution; however, its estimates exhibit systematic biases that require local correction.

This study aims to **evaluate and correct the bias of CHIRPS estimates against ground stations and, subsequently, model their spatial distribution** using advanced geostatistical techniques across the main Honduran watersheds for 1981–2023. **For bias correction, well-established statistical approaches from the literature are implemented—Linear Scaling, Power Transformation, and Quantile Mapping—as used in previous CHIRPS validation studies in Africa and Central America. Afterwards, three geostatistical methods are applied to estimate precipitation in ungauged or hard-to-access areas: Ordinary Kriging (OK), which models spatial autocorrelation; Universal Kriging (UK), which incorporates topographic covariates; and Co-Kriging, which merges satellite estimates with ground observations by exploiting cross-correlation.** Validation is performed using R^2, NSE, RMSE, and bias through cross-validation with national meteorological stations. The results yield spatially corrected precipitation fields that enhance hydrological modeling, water balance estimation, and hydraulic infrastructure design, strengthening integrated water-resources management in Honduras.

Keywords: Geostatistics, Kriging, Co-Kriging, CHIRPS, satellite precipitation, validation, Honduras.

1. Introducción

La precipitación es un componente central del ciclo hidrológico por su impacto directo en la disponibilidad de agua, la agricultura, la planificación urbana y la reducción del riesgo de desastres. En Honduras, la complejidad topográfica y la alta variabilidad climática exigen información espacial y temporalmente consistente; sin embargo, la red pluviométrica nacional presenta baja densidad, distribución irregular y series incompletas, lo que limita la caracterización confiable de la lluvia.

Ante estas limitaciones, los productos satelitales CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data) constituyen una fuente continua desde 1981 (0.05°; 5 km), útil para complementar la observación terrestre. No obstante, sus estimaciones pueden exhibir sesgos sistemáticos en regiones con topografía compleja como Honduras, por lo que se requiere su validación y ajuste antes de su uso en aplicaciones hidrológicas e ingenieriles.

En este estudio, la corrección del sesgo entre CHIRPS y las observaciones de estaciones terrestres se realiza mediante métodos estadísticos reconocidos, tales como el Escalamiento Lineal (Linear Scaling), la Transformación de Potencia (Power Transformation) y el Mapeo de Cuantiles (Quantile Mapping), con el fin de obtener estimaciones de precipitación ajustadas localmente. Posteriormente, las técnicas geoestadísticas Kriging Ordinario (OK), Kriging Universal (UK) y Co-Kriging se emplean para representar la estructura espacial de la precipitación y **estimar valores en zonas sin pluviómetros o de difícil acceso**, integrando información satelital y terrestre.

Objetivo general: Validar y **evaluar el sesgo** de las estimaciones de precipitación CHIRPS frente a estaciones terrestres, aplicar **métodos de corrección estadística** y, posteriormente, **modelar espacialmente la precipitación corregida** mediante técnicas geoestadísticas (OK, UK y Co-Kriging) en las principales cuencas hidrográficas de Honduras (1981–2023).

Objetivos específicos:

1. Evaluar el desempeño del producto CHIRPS frente a los datos de estaciones terrestres mediante indicadores estadísticos de ajuste y precisión, como el coeficiente de correlacion de pearson (r), y de determinación (R^2), la eficiencia de Nash–Sutcliffe (NSE), la raíz del error cuadrático medio (RMSE), (MAE) y el sesgo medio, utilizando validación cruzada.
2. Aplicar métodos estadísticos de corrección de sesgo tales como; Escalamiento Lineal, Transformación de Potencia y Mapeo de Cuantiles.para ajustar las estimaciones de precipitación CHIRPS a las observaciones terrestres y obtener series corregidas localmente.
3. Modelar la estructura espacial de la precipitación corregida mediante técnicas geoestadísticas (Kriging Ordinario, Kriging Universal y Co-Kriging) para **estimar valores en zonas sin pluviómetros o de difícil acceso**, evaluando el aporte de covariables topográficas en la interpolación.
4. Generar mapas continuos de precipitación corregida y evaluar su desempeño en la mejora de la modelación hidrológica, el balance hídrico y el diseño de obras hidráulicas a escala de cuenca.

Estas herramientas permiten generar campos de precipitación corregidos, espacialmente continuos y más representativos de la realidad climática del territorio

hondureño. De esta manera, el estudio contribuye a mejorar la calidad y cobertura espacial de la información de lluvia, fortaleciendo la toma de decisiones en gestión del agua, el diseño de infraestructura hidráulica —como puentes y drenajes— y los sistemas de alerta temprana ante eventos extremos. Asimismo, proporciona una herramienta técnica robusta para estimar la precipitación en zonas sin observaciones directas o de difícil acceso, ampliando la cobertura de datos y optimizando la planificación y gestión sostenible de los recursos hídricos del país.

2. Justificación

En Honduras, los estudios hidrológicos y de ingeniería civil dependen en gran medida de la información pluviométrica para estimar caudales, diseñar drenajes, puentes y sistemas de control de inundaciones. Sin embargo, la red nacional de estaciones meteorológicas presenta limitaciones históricas en cobertura, mantenimiento y continuidad de datos, lo que genera vacíos espaciales y temporales que reducen la representatividad espacial y la precisión de los modelos hidrológicos.

La ausencia de registros consistentes en muchas zonas del país ha obligado a los profesionales a recurrir a estimaciones generalizadas o valores promedios regionales, reduciendo la precisión de los diseños hidráulicos y aumentando la incertidumbre en la gestión y planificación de los recursos hídricos. Frente a esta situación, los productos satelitales como CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data) representan una herramienta valiosa para suplir esta carencia, pues ofrecen registros continuos desde 1981 con una resolución espacial de 0.05° (5 km). No obstante, estas fuentes presentan sesgos sistemáticos que deben corregirse mediante métodos estadísticos especializados, tales como el Escalamiento Lineal (Linear Scaling), la Transformación de Potencia (Power Transformation) y el Mapeo de Cuantiles (Quantile Mapping), antes de su aplicación en el ámbito técnico.

En este contexto, la validación de los datos CHIRPS y su posterior modelación espacial mediante técnicas geoestadísticas constituyen una alternativa metodológica robusta para mejorar la calidad y cobertura de la información de precipitación. A través de métodos como el Kriging Ordinario, Kriging Universal y Co-Kriging, es posible integrar información satelital con observaciones de estaciones terrestres y realizar interpolaciones en puntos donde no existen pluviómetros o en zonas de difícil acceso, lo que permite estimar la precipitación en áreas sin mediciones directas y generar registros históricos continuos en todo el territorio nacional.

El producto *CHIRPS* fue desarrollado por el *Climate Hazards Group* de la University of California, Santa Barbara (UCSB), en colaboración con el United States Geological Survey (USGS/EROS), con el propósito de apoyar el sistema de alerta temprana para sequías del *Famine Early Warning Systems Network* (FEWS NET) de USAID. El conjunto de datos fue presentado oficialmente por [Funk et al., 2015] en la revista *Scientific Data*, y constituye un registro de precipitación cuasi-global (50°S – 50°N), con resolución de 0,05° y disponibilidad desde 1981.

Este trabajo se enmarca dentro de la línea de investigación **Estadística Espacial** de la orientación en Estadística de la Maestría en Matemáticas de la UNAH, ya que utiliza herramientas de análisis espacial para representar y modelar fenómenos geográficos relacionados con la precipitación. De acuerdo con los ejes de investigación institucionales, este estudio se vincula con los temas prioritarios de

la **infraestructura y desarrollo territorial**, al generar información de soporte para el diseño de obras hidráulicas y planificación urbana; y con el eje de **cambio climático y vulnerabilidad**, al mejorar la comprensión de la variabilidad espacial de la lluvia y su impacto sobre la gestión de los recursos hídricos.

Finalmente, el estudio se justifica porque permitirá disponer de una base de datos de precipitación corregida mediante métodos estadísticos y modelada espacialmente con técnicas geoestadísticas, adecuada para su uso en proyectos de modelación hidrológica, balance hídrico y diseño de infraestructura hidráulica. Además, aportará una herramienta estadística replicable que podrá aplicarse en otras regiones del país y de Centroamérica, contribuyendo al fortalecimiento de la gestión del agua, la adaptación al cambio climático y la planificación territorial sustentable.

3. Antecedentes

El uso de datos satelitales para la estimación y validación de la precipitación ha cobrado gran relevancia en los estudios hidrológicos de América Latina, especialmente en regiones con limitada cobertura de estaciones meteorológicas. Entre los productos más utilizados se encuentra **CHIRPS** (*Climate Hazards Group InfraRed Precipitation with Station data*), Estos datos han sido aplicados con éxito en la caracterización de lluvias, análisis de sequías y validación climática en contextos de topografía compleja. Sin embargo, aunque su uso se ha expandido considerablemente en los últimos años, la precisión y aplicabilidad de este producto puede variar según las condiciones climáticas, fisiográficas y el nivel de densidad de estaciones disponibles en cada región. Por ello, resulta fundamental revisar antecedentes científicos que evalúen su desempeño bajo diferentes contextos geográficos, climáticos y metodológicos.

El conjunto de datos *CHIRPS* (Climate Hazards Group InfraRed Precipitation with Stations) fue desarrollado como una herramienta para el monitoreo de sequías y cambios ambientales sobre superficie terrestre. Recientes esfuerzos de validación en Sudamérica han evaluado su capacidad para reproducir los principales patrones espaciales y temporales de la precipitación. No obstante, se ha avanzado poco en determinar su capacidad para evaluar condiciones húmedas y secas, particularmente en áreas con registros pluviométricos escasos. En este estudio, se investigó el desempeño de *CHIRPS* para monitorear eventos húmedos y secos en la región semiárida del centro-oeste de Argentina. Mediante el Índice Estandarizado de Precipitación (SPI), se comparó la base de datos *CHIRPS* con registros provenientes de 49 estaciones meteorológicas durante el período 1987–2016. Los resultados indicaron que *CHIRPS* reprodujo adecuadamente la variabilidad temporal del SPI en múltiples escalas (1, 3 y 6 meses), especialmente en la región dominada por precipitación de temporada cálida. Sin embargo, se observó una sobrestimación considerable en la precipitación estacional en la región dominada por lluvias de temporada fría, lo que introduce errores reflejados en el desempeño de *CHIRPS* en el sector occidental del área de estudio. Además, aunque *CHIRPS* reprodujo con precisión la frecuencia de clases húmedas y secas en escalas superiores a un mes, el sesgo húmedo (*wet bias*) produjo una subestimación de la frecuencia de valores cero, afectando la clasificación de condiciones extremas en eventos secos (1998) y húmedos (2016). Los autores concluyeron que *CHIRPS* es una herramienta adecuada para la evaluación de condiciones secas y húmedas en escalas superiores a un mes, pudiendo

apoyar procesos de toma de decisiones en agencias hidrometeorológicas regionales ([Rivera et al., 2019]).

De manera similar El estudio [Al-Shamayleh et al., 2024]) evaluó la capacidad del producto CHIRPS con resolución espacial de 0.05° para estimar precipitación mensual y anual en la cuenca Wala, Jordania, durante el período 1987–2017 mediante una comparación punto-a-píxel y utilizando once índices extremos recomendados por el ETCCDI. Los resultados mostraron una correlación moderada en la estimación mensual (r = 0.50–0.73), pero un bajo desempeño en la detección de eventos extremos, con tendencia a sobreestimar valores bajos y subestimar valores altos de precipitación, especialmente en años hidrometeorológicos extremos. Además, el producto presentó subestimación en indicadores CDD, CWD, R10, R20 y R30, mientras que sobreestimó R95p, R99p y Rx1day, lo cual evidencia limitaciones en la representación de extremos pluviométricos. La prueba de Wilcoxon indicó falta de equivalencia estadística con los registros observados, concluyendo que es necesaria una corrección de sesgo antes de emplear CHIRPS en análisis extremos o aplicaciones hidrológicas.

En Honduras, ([Pichardo, 2024]) desarrolló un estudio pionero titulado *"Validación de precipitación en la subcuenca del Lago de Yojoa: datos satelitales versus observados"*, donde comparó los productos **CHIRPS v2.0** y **CMORPH** con observaciones de diez estaciones hidroclimatológicas de la Empresa Nacional de Energía Eléctrica (ENEE). El estudio reportó un bajo ajuste en la escala diaria (R^2 entre 0.02 y 0.07), pero un desempeño considerablemente mejor a nivel mensual (R^2 entre 0.6 y 0.85), destacando una fuerte correlación (($\rho > 0{,}85$)) y eficiencia de Nash-Sutcliffe (NSE>0.70) en estaciones como El Mochito y Santa Elena. Para la corrección de sesgos, se aplicaron los métodos de *Escalamiento Lineal (LS)* y *Transformación de Potencias (PT)*, logrando ajustar los datos de CHIRPS a las observaciones en tierra y realizar relleno de series históricas de precipitación entre 1981–2023. Aunque el estudio demostró la validez del uso de CHIRPS a escala mensual, no incorporó técnicas geoestadísticas ni análisis espacial continuo, limitándose al ámbito local de la subcuenca del Lago de Yojoa

([Bollat Flores, 2023]) desarrolló en Guatemala un *análisis comparativo de datos CHIRPS con registros pluviométricos locales* en el departamento de Chiquimula, aplicando **interpolación espacial mediante Kriging Ordinario**. Su investigación logró una correspondencia espacial del 80 % y una correlación positiva de 0.84, demostrando la eficacia del Kriging para ajustar las diferencias entre estimaciones satelitales y observaciones de superficie en regiones montañosas del corredor seco centroamericano. Este enfoque permitió generar mapas continuos de precipitación corregida y evidenció el potencial de las técnicas geoestadísticas para mejorar la precisión de los productos satelitales.

De Manera complementaria En Ghana ([Atiah et al., 2023]), donde la red de pluviómetros presenta un continuo deterioro, se evaluó el desempeño del producto satelital CHIRPS-v2 mediante un proceso de corrección de sesgo utilizando el enfoque *Bias Correction and Spatial Disaggregation* (BCSD). El estudio analizó el impacto de dicha corrección sobre la identificación de la estacionalidad y de los índices extremos de precipitación. Los resultados mostraron que, tras la aplicación del método BCSD, los patrones estacionales y anuales fueron mejor representados

y se obtuvo una mayor correspondencia con los datos de estaciones, especialmente en las fechas de inicio y fin de la temporada lluviosa. El estudio concluye que el enfoque BCSD mejora tanto la estimación de la precipitación como la identificación de índices de estacionalidad, sugiriendo su aplicación en la corrección de otros productos satelitales utilizando registros históricos de largo plazo.

Los antecedentes revisados conforman una base conceptual y metodológica relevante para el presente estudio; sin embargo, también evidencian brechas investigativas tanto en la escala de aplicación como en la combinación metodológica. A nivel nacional, Pichardo (2024) validó los datos CHIRPS únicamente a escala local sin incorporar modelación espacial, mientras que Bollat Flores (2023) aplicó Kriging en Guatemala para la interpolación de precipitación sin considerar procesos de corrección estadística previos. De manera complementaria, estudios internacionales han demostrado que CHIRPS requiere una corrección de sesgo antes de su aplicación hidrológica o espacial (Rivera et al., 2019; Atiah et al., 2023; Al-Shamayleh et al., 2024), particularmente para la representación de eventos extremos y estacionalidad.

En función de estas brechas, la presente investigación propone una ampliación metodológica a escala nacional en Honduras, integrando dos etapas complementarias: (1) corrección estadística del sesgo mediante *Linear Scaling*, *Power Transformation* y *Quantile Mapping*; y (2) modelación espacial mediante Kriging Ordinario, Kriging Universal y Co-Kriging, combinando datos satelitales *CHIRPS* con observaciones de estaciones terrestres. En el caso del Co-Kriging, se incorporarán covariables físico-ambientales espacialmente continuas tales como topografía (DEM), temperatura del aire, velocidad/dirección del viento, distancia al litoral u otras variables climáticas relacionadas, siempre que presenten correlación significativa con la precipitación y mejoren la capacidad predictiva del modelo. Este enfoque permitirá obtener estimaciones corregidas localmente y generar una representación espacial continua en zonas sin cobertura instrumental, fortaleciendo la disponibilidad de información pluviométrica para aplicaciones hidrológicas, gestión de riesgos y diseño de infraestructura hidráulica en Honduras.

4. MARCO TEORICO

4.1. Fuentes de informacion y Datos utilizados.

4.1.1. Área de estudio. El estudio se desarrollará en las principales cuencas hidrográficas de Honduras, las cuales presentan variaciones espaciales y temporales significativas en la precipitación debido a características topográficas, climáticas y oceánicas. La presencia de cadenas montañosas, valles intermontanos, planicies costeras y la influencia tanto del océano Pacífico como del mar Caribe genera gradientes pluviométricos marcados, lo que requiere integrar datos satelitales y observaciones terrestres mediante técnicas estadísticas y espaciales.

4.1.2. Datos satelitales (CHIRPS-v2). Se utilizará el producto satelital *Climate Hazards Group InfraRed Precipitation with Station data* (CHIRPS-v2), el cual integra estimaciones infrarrojas con información de estaciones meteorológicas mediante un proceso de interpolación inteligente, con resolución espacial de 0,05° (aproximadamente 5 km) y cobertura temporal desde 1981 hasta la actualidad, lo que permite construir un registro histórico de alta continuidad espacial.

El producto CHIRPS fue desarrollado por el *Climate Hazards Group* de la *University of California, Santa Barbara (UCSB)* en colaboración con el *United States Geological Survey (USGS/EROS)*, con el propósito de apoyar el sistema de alerta temprana ante sequías del *Famine Early Warning Systems Network (FEWS NET)* de USAID. El conjunto de datos fue presentado oficialmente por ([Funk et al., 2015]) en la revista *Scientific Data*, y constituye un registro de precipitación cuasi-global (50°S – 50°N), con una resolución de 0,05° y disponibilidad desde 1981, lo cual lo convierte en una fuente adecuada para estudios hidrológicos en regiones con limitada cobertura instrumental.

4.1.3. Datos de estaciones meteorológicas terrestres. Para la comparación, validación y corrección del sesgo se utilizarán registros provenientes de estaciones meteorológicas ubicadas dentro de las cuencas de estudio. En Honduras, la disponibilidad y densidad espacial de estaciones es limitada, especialmente en zonas rurales, montañosas y de difícil acceso. Además, una parte considerable de las estaciones sólo cuentan con registros mensuales y presentan lagunas temporales, periodos de inactividad y series históricas incompletas, lo cual dificulta su uso directo en análisis hidrológicos detallados. Por esta razón, se vuelve necesario complementar estas mediciones con productos satelitales y aplicar técnicas de corrección estadística antes de realizar la modelación espacial.

4.1.4. Covariables ambientales. Con el propósito de mejorar la representación espacial de la precipitación, se evaluará la incorporación de covariables físico-ambientales en el modelo de *Co-Kriging*, siempre que estas demuestren correlación estadística significativa y coherencia físico-climática. Entre las variables candidatas se consideran:

- Elevación y pendiente (DEM),
- Temperatura del aire superficial,
- Distancia al litoral,
- Velocidad y dirección del viento,
- Índices de vegetación o humedad del suelo.

La decisión de integrar cada covariable se basará en análisis estadístico preliminar y revisión de literatura con el fin de optimizar la capacidad predictiva y estabilidad del modelo.

4.2. Metodología. La metodología propuesta se estructura en cuatro fases principales: (i) preparación y depuración de los datos disponibles, (ii) comparación estadística punto–píxel entre observaciones terrestres y estimaciones satelitales, (iii) aplicación de técnicas estadísticas de corrección de sesgo reportadas en la literatura científica, y (iv) modelación espacial mediante métodos geoestadísticos. El propósito de este enfoque metodológico es integrar información satelital y registros provenientes de estaciones meteorológicas, con el fin de generar estimaciones continuas y espacialmente coherentes de precipitación corregida en las principales cuencas hidrográficas de Honduras. Cabe señalar que estas etapas representan un esquema metodológico planificado para su implementación en el desarrollo de la presente investigación.

4.2.1. Fase I: Preparación y depuración de datos. En esta etapa se realizó la recopilación, estandarización y verificación de calidad de los datos satelitales y terrestres. Las actividades consideradas se detallan a continuación:

1. Descarga y organización de la serie satelital CHIRPS-v2 para el periodo definido en el estudio.
2. Obtención de los registros de precipitación provenientes de estaciones meteorológicas ubicadas dentro de las cuencas seleccionadas.
3. Aplicación de control de calidad de datos mediante verificación de valores extremos, duplicados, discontinuidades temporales y consistencia interna.
4. Ajuste de la resolución temporal entre ambas fuentes (mensual o diaria según disponibilidad).
5. Unificación de formatos, unidades y estructuras de archivo para su tratamiento estadístico.

4.2.2. Fase II: Comparación estadística punto-píxel. En la fase de evaluación se propone utilizar indicadores estadísticos para cuantificar el ajuste entre la precipitación estimada por CHIRPS-v2 y las observaciones en estaciones meteorológicas. En particular, se emplearán el Sesgo (BIAS), el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE), el coeficiente de correlación de Pearson (r) y la eficiencia de Nash–Sutcliffe (NSE). Las expresiones propuestas se describen a continuación.

Sesgo medio (BIAS)

Sea $\{x_i\}_{i=1}^n$ la secuencia de valores observados y sea $\{y_i\}_{i=1}^n$ la secuencia correspondiente de valores estimados. El *sesgo medio* (BIAS) se define como

$$\text{BIAS} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i). \tag{4.1}$$

Un valor positivo de BIAS indica una sobreestimación sistemática de los valores estimados respecto a los observados, mientras que un valor negativo indica una subestimación sistemática. En el caso ideal, un valor de BIAS cercano a cero sugiere ausencia de sesgo promedio.

Error Absoluto Medio (MAE)

Sea $\{x_i\}_{i=1}^n$ la secuencia de valores observados y sea $\{y_i\}_{i=1}^n$ la secuencia correspondiente de valores estimados. El *error absoluto medio* (MAE) se define como

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - x_i \,|. \tag{4.2}$$

El MAE mide la magnitud promedio del error entre las observaciones y las estimaciones, sin considerar su signo.

Raíz del Error Cuadrático Medio (RMSE)

Sea $\{x_i\}_{i=1}^n$ la secuencia de valores observados y sea $\{y_i\}_{i=1}^n$ la secuencia correspondiente de valores estimados. La *raíz del error cuadrático medio* (RMSE) se

define como

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}. \tag{4.3}$$

El RMSE penaliza con mayor peso las discrepancias grandes entre estimaciones y observaciones y constituye una medida estándar de la precisión de un modelo.

Interpretación:

- Un RMSE **cercano a cero** indica un buen ajuste entre los valores estimados y observados.
- Un RMSE **elevado** señala una mayor discrepancia entre ambos conjuntos de valores.
- El RMSE conserva las mismas unidades que la variable analizada.

Coeficiente de correlación de Pearson (r)

Sea $\{x_i\}_{i=1}^n$ la secuencia de valores observados y sea $\{y_i\}_{i=1}^n$ la secuencia correspondiente de valores estimados. Denote por

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

las medias respectivas.

El *coeficiente de correlación lineal de Pearson* se define como

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{4.4}$$

Este coeficiente mide el grado de asociación lineal entre ambas series.

Interpretación:

- r cercano a 1 indica una fuerte relación lineal positiva.
- r cercano a -1 indica una fuerte relación lineal negativa.
- r cercano a 0 sugiere ausencia de relación lineal.

Eficiencia de Nash–Sutcliffe (NSE)

Sea $\{x_i\}_{i=1}^n$ la secuencia de valores observados y sea $\{y_i\}_{i=1}^n$ la secuencia correspondiente de valores estimados. Denote por

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

la media de los valores observados. La *eficiencia de Nash–Sutcliffe* se define como

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{4.5}$$

El NSE evalúa la capacidad de un modelo para reproducir los valores observados, comparándolo con el desempeño obtenido al usar la media observada como estimador.

Interpretación:

- Valores de NSE **cercanos a 1** indican un desempeño alto.
- Valores de NSE **cercanos a 0** sugieren que el modelo no mejora respecto a usar la media de los datos observados.

- Valores de NSE **negativos** indican que el modelo tiene un desempeño peor que la media observada.

4.2.3. Fase III: Corrección estadística del sesgo. Con el propósito de ajustar la serie satelital a las condiciones reales medidas por estaciones terrenas, se aplicarán tres métodos de corrección estadística: *Linear Scaling (LS)*, *Power Transformation (PT)* y *Quantile Mapping (QM)*. A continuación, se describen las expresiones matemáticas de cada técnica.

Método Linear Scaling (LS). Sea $\{x(t)\}$ la serie de valores observados y sea $\{y(t)\}$ la serie correspondiente de valores estimados para cada tiempo t. Denote por

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x(t), \qquad \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y(t)$$

las medias respectivas. El *método Linear Scaling* corrige cada valor estimado mediante el factor

$$y_{\text{corr}}(t) \;=\; y(t)\,\frac{\bar{x}}{\bar{y}}. \tag{4.6}$$

Este método aplica un factor multiplicativo constante basado en la razón entre la media observada y la media estimada. Es adecuado cuando el sesgo es proporcional y se manifiesta principalmente en la magnitud promedio de la serie.

Interpretación de las variables:

- $y_{\text{corr}}(t)$: valor corregido en el tiempo t.
- $y(t)$: valor estimado en el tiempo t.
- $\bar{x}$: media de la serie observada.
- $\bar{y}$: media de la serie estimada.

(b) Método Power Transformation (PT).

$$P_{\text{corr}}(t) = \left(P_{\text{sat}}(t)\right)^{\lambda} \tag{4.7}$$

Este procedimiento ajusta la distribución mediante una transformación potencial controlada por el parámetro λ, modificando la asimetría y mejorando la representación de valores extremos.

Donde:
$P_{\text{corr}}(t)$: valor corregido de precipitación en el tiempo t
$P_{\text{sat}}(t)$: valor satelital en el tiempo t
λ: parámetro de transformación obtenido mediante calibración estadística

(c) Método Quantile Mapping (QM).

$$P_{\text{corr}}(t) = F_{\text{obs}}^{-1}\left(F_{\text{sat}}\big(P_{\text{sat}}(t)\big)\right) \tag{4.8}$$

Este método realiza la corrección mediante el emparejamiento de cuantiles entre las distribuciones satelital y observada, logrando ajustar no sólo la media y la varianza, sino la forma completa de la distribución.

Donde:
$P_{\text{corr}}(t)$: valor corregido para el tiempo t
$P_{\text{sat}}(t)$: valor satelital estimado para el tiempo t
$F_{\text{sat}}(\cdot)$: función de distribución acumulada (CDF) del satélite
$F_{\text{obs}}^{-1}(\cdot)$: función inversa de la CDF de las observaciones

4.2.4. Fase IV: Modelación espacial mediante geoestadística. Posteriormente, se aplicarán métodos geoestadísticos con el fin de estimar la distribución espacial continua de la precipitación corregida. El procedimiento comenzó con el cálculo del semivariograma experimental y posteriormente se utilizaron los métodos de Kriging Ordinario (OK), Kriging Universal (UK) y Co-Kriging (CoK).

A continuacion se describen algunos Fundamentos teóricos de la dependencia espacial:
En diversos fenómenos naturales, las variables de interés se observan a través del tiempo, del espacio o en una combinación espacio-temporal. Esta característica implica que su análisis no puede abordarse únicamente mediante los métodos tradicionales de la estadística clásica, pues los supuestos que dichos métodos requieren especialmente el de independencia entre observaciones rara vez se cumplen en estos casos.

En el ámbito espacial, este comportamiento ha sido ampliamente documentado. La denominada *primera ley de la Geografía*, atribuida a Waldo Tobler ([Tobler, 1970]) , establece que "todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes". Esta afirmación resume el principio fundamental de la **autocorrelación espacial**, según el cual las observaciones geográficamente próximas tienden a presentar valores similares. En consecuencia, los datos espaciales no son independientes: cada medición está influenciada por su entorno, y en procesos multivariados puede existir además **correlación cruzada** entre distintas variables medidas en un mismo espacio geográfico.

Por ello, el análisis estadístico de fenómenos espaciales requiere métodos que incorporen explícitamente esta estructura de dependencia. La estadística espacial y espacio-temporal constituye el marco teórico que permite describir, modelar y predecir procesos que varían en el territorio, proporcionando herramientas para:

- estimar valores en lugares sin observaciones directas,
- caracterizar cómo cambia la relación entre puntos conforme aumenta la distancia,
- extender los modelos de regresión al caso en que las observaciones están correlacionadas espacialmente,
- analizar patrones de ocurrencia y variación de fenómenos geográficos.

Este enfoque es indispensable en estudios hidrológicos y climáticos, donde variables como la precipitación presentan dependencia espacial marcada y estructuras de correlación complejas. En este trabajo, dicha dependencia constituye un elemento central, ya que la construcción de campos continuos y coherentes de precipitación corregida requiere modelos capaces de representar la autocorrelación inherente al proceso, garantizando estimaciones más precisas y consistentes en áreas sin estaciones meteorológicas.

Definición 1 (Proceso espacio-temporal). Un proceso espacio-temporal es un proceso estocástico denotado por

$$\{Z(s,t) : (s,t) \in D_s \times D_T\},$$

donde $D_s \subset \mathbb{R}^d$ representa el conjunto índice correspondiente a la ubicación espacial s, y $D_T \subset \mathbb{R}$ es el conjunto índice asociado al tiempo t. Por lo tanto, cada par (s,t) pertenece al dominio espacio–temporal $\mathbb{R}^d \times \mathbb{R}$, y el producto $D_s \times D_T \subset \mathbb{R}^d \times \mathbb{R}$ constituye el dominio índice completo del proceso.

Los conjuntos D_s y D_T pueden ser continuos o discretos, fijos o aleatorios, según el fenómeno bajo estudio y el diseño de muestreo disponible. Este marco general permite modelar variables que presentan variación simultánea en el espacio y en el tiempo, incorporando su estructura conjunta de dependencia.

Definición 2 (Proceso espacial). Sea Z la variable de interés, y sea s la ubicación espacial donde existe Z. Así, el proceso espacial es el proceso estocástico

$$\{Z(s) : s \in D_s\},$$

donde D_s está formado por todas las ubicaciones s y es su conjunto índice. La ubicación espacial s puede estar en una, dos o más dimensiones. Cuando s es un vector, al proceso espacial se le suele llamar *campo aleatorio*. Vease con mas detalle en el siguiente cuadro.

ID	Spatial location	t
A	$s_1 = (x_1, y_1)$	t_1
B	$s_2 = (x_2, y_2)$	t_2
C	$s_3 = (x_3, y_3)$	t_3
D	$s_4 = (x_4, y_4)$	t_4
E	$s_5 = (x_5, y_5)$	t_5

CUADRO 1. Notación para las coordenadas espacio–temporales de un proceso espacio-tiempo.

Definición 3 (Proceso temporal). Sea Z la variable de interés, y sea t el momento del tiempo en el que ocurre Z. Así, el proceso temporal es el proceso estocástico

$$\{Z(t) : t \in D_T\},$$

donde $D_T \subset \mathbb{R}$ es el conjunto de todos los tiempos y constituye su conjunto índice.

Así, tanto el proceso espacial como el proceso temporal son casos particulares del proceso espacio–tiempo. El conjunto índice de un proceso temporal tiene una sola dimensión. Sin embargo, uno de los objetivos principales en este caso es encontrar pronósticos, lo que en general difiere de los objetivos perseguidos con datos espaciales.

Clases de datos espaciales

Los métodos estadísticos aplicados a datos espaciales varían según las características del dominio espacial o conjunto índice D_s. A partir de estas características, surgen tres grandes ramas de la estadística espacial: **geoestadística**, **datos de área** y **procesos espaciales puntuales**.

Geoestadística. Es el conjunto de métodos aplicados a datos espaciales con variación continua, donde D_s es un subconjunto fijo de $\mathbb{R}^d$; esto es, D_s es continuo y fijo y $Z(s)$ es una variable aleatoria con ubicación s, $(s \in D_s)$. puede ser observada en cualquier punto del dominio. Este enfoque es apropiado para fenómenos que pueden considerarse como campos continuos, tales como precipitación, temperatura o humedad del suelo.

Datos de área. Son los datos espaciales con variación espacial discreta. D_s es un subconjunto contable y fijo de $\mathbb{R}^d$; esto es, D_s es discreto y fijo y $Z(s)$ es una variable aleatoria con ubicación s, $(s \in D_s)$.

Procesos espaciales puntuales. En esta categoría, las observaciones no se registran en puntos fijos, sino que corresponden a la ubicación donde ocurre un evento de interés. El conjunto X es un conjunto de puntos definidos en un subconjunto generalmente aleatorio de $\mathbb{R}^d$. Estos procesos modelan fenómenos como sismos, incendios, delitos o eventos biológicos registrados mediante su localización.

En este trabajo se hará énfasis en la geoestadística, debido a que la precipitación presenta variación continua en el espacio.

Geoestadística

El valor observado en cada punto $s = (x_i, y_i)$ se considera como la realización $z(s)$, de una variable aleatoria $Z(s)$. En términos matemáticos, la familia de todas estas variables aleatorias se denomina una función aleatoria, proceso estocástico o campo aleatorio. Un campo aleatorio es caracterizado por su distribución de probabilidad finito dimensional, es decir, la distribución de probabilidad conjunta de un conjunto de variables $Z(s_1), Z(s_2), \ldots, Z(s_n)$ para todo n y para todos los puntos $s_1, s_2, \ldots, s_n$. Un proceso estocástico está dotado de los siguientes elementos:

- **Función de distribución finito dimensional.** Para cualesquiera n puntos $s_1, s_2, \ldots, s_n$, el vector aleatorio

$$Z = \begin{pmatrix} Z(s_1) \\ Z(s_2) \\ \vdots \\ Z(s_n) \end{pmatrix}$$

  se caracteriza por su función de distribución n-dimensional:

$$F_{s_1,s_2,\ldots,s_n}(z_1, z_2, \ldots, z_n) = P[Z(s_1) \leq z_1,\ Z(s_2) \leq z_2,\ \ldots,\ Z(s_n) \leq z_n].$$

- **Función de media.** El momento de primer orden es la esperanza matemática definida como:

$$E[Z(s)] = \mu(s)$$

  A veces también llamada la función de media, la deriva o la tendencia del proceso.

- **Función de varianza.** La varianza o momento de segundo orden de $Z(s)$ respecto a $\mu(s)$ es:

$$\sigma^2(s) = \mathrm{Var}[Z(s)] = E[(Z(s) - \mu(s))^2]$$

En general, estas funciones pueden depender de la posición s de manera determinística.

- **Función de autocovarianza.** La autocovarianza de un proceso espacial $Z(s)$ es en general una función de las ubicaciones espaciales s_i y s_j, con $s_i, s_j \in \mathbb{R}^d$, para todo $i, j \in \mathbb{Z}_+$. La covarianza $\text{Cov}(Z(s_i), Z(s_j))$ se define como:

$$\text{Cov}(Z(s_i), Z(s_j)) = C(s_i, s_j) = E[(Z(s_i) - \mu(s_i))(Z(s_j) - \mu(s_j))]$$

donde $C(\cdot)$ es una función definida positiva para garantizar una varianza de error de predicción no negativa. Esto es, para cualquier número finito m de ubicaciones espaciales $s_1, s_2, \ldots, s_m$ y cualquier conjunto de números reales $\{a_1, a_2, \ldots, a_m\}$ con $m \in \mathbb{Z}_+$, C debe satisfacer:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j C(s_i, s_j) \geq 0 \tag{4.9}$$

Nótese que $C(s_i, s_i) = \text{Var}(Z(s_i)) = \sigma_Z^2$.

- **Función de autocorrelación** La autocorrelación de dos de las variables aleatorias $Z(s_i)$ y $Z(s_j)$, $\rho(s_i, s_j)$, definida como:

$$\rho(s_i, s_j) = \frac{C(s_i, s_j)}{\sigma(s_i)\sigma(s_j)}$$

Es en general una función de s_i y s_j. Esta es la función de autocorrelación del proceso.

- **Función de semivarianza** El semivariograma $\gamma(s_i, s_j)$ que se define como:

$$\gamma(s_i, s_j) = \frac{1}{2} E\left[(Z(s_i) - Z(s_j))^2\right]$$

El variograma es por tanto $2\gamma(s_i, s_j)$. Aunque, se usan ambos términos indistintamente para referirse a la función $\gamma(s_i, s_j)$. Nótese que el semivariograma estima la varianza espacial para distancias específicas, por lo tanto es una función positiva.

Supuesto de Estacionariedad

Un proceso es estacionario, si las relaciones entre cualquier subconjunto de puntos son iguales independientemente del lugar donde residen los puntos en el espacio. La estacionariedad puede pensarse como la propiedad que posee la función aleatoria de que muchas realizaciones de la misma función aleatoria proporcionan la misma información. Se distinguen tres tipos de estacionariedad:

- **Estacionariedad fuerte o de primer orden:** En términos de funciones de distribución.
- **Estacionariedad débil o de segundo orden:** En términos de los momentos media y covarianza.
- **Estacionariedad intrínseca o de incrementos:** En términos de media y varianza de los incrementos del proceso.

Supuesto de isotropía

Si $C(\cdot)$ y/o $\gamma(\cdot)$ son funciones únicas de la magnitud $\|h\|$, esto es,

$$\text{Cov}\left(Z(s), Z(s+h)\right) = C(\|h\|) \quad \text{y/o} \quad \frac{1}{2}\text{Var}\left(Z(s+h) - Z(s)\right) = \gamma(\|h\|)$$

el proceso posee función de covarianza y/o semivarianza isotrópica.

La estacionariedad permite combinar pares de datos con la misma diferencia de coordenadas, pero si además, los vectores de diferencias pueden ser reemplazados con distancias escalares, por ejemplo una distancia euclidiana, entonces el campo aleatorio se dice isotrópico. Esto es, la correlación entre los datos no depende de la dirección en la que ésta se calcula.

Así, un campo aleatorio que es estacionario pero no isotrópico se desarrolla de manera diferente según las distintas direcciones del espacio; no solo basta con conocer cuánto están separados un par de puntos, sino también se necesita conocer la orientación de dicha distancia; estos se conocen como campos aleatorios anisotrópicos. Entonces, hay anisotropía, si la dependencia espacial entre $Z(s)$ y $Z(s+h)$ es una función tanto de la magnitud como de la dirección del vector h.

En términos geométricos, la estacionariedad y la isotropía son propiedades de invarianza; la estacionariedad es invarianza bajo traslación y la isotropía es invarianza bajo rotaciones y reflexiones.

Semivariograma

El semivariograma es una función que describe cómo cambia la variabilidad espacial de una variable conforme aumenta la distancia entre dos ubicaciones.

El semivariograma $\gamma(h)$ se define como la función de varianza de la variable incrementos, es decir:

$$\gamma(h) = \frac{1}{2}\operatorname{Var}\left(Z(s+h) - Z(s)\right)$$

Es por esto que, el semivariograma de un proceso estacionario de segundo orden es de soporte compacto o tiene una asíntota en $C(0)$ cuando se incrementa la separación de los puntos. Si el semivariograma no se estabiliza, sino que continúa creciendo, la varianza de la variable incrementos no es finita, pero aún puede ser al menos intrínsecamente estacionario si cumple que:

$$\frac{\gamma(h)}{\|h\|^2} \to 0 \quad \text{cuando } h \to \infty$$

Esto es, el semivariograma no debe crecer más rápido que una ecuación de segundo grado.

Los parámetros de los cuales depende un semivariograma de un proceso estacionario de segundo orden son los siguientes (ver figura 1):

Silla: Es la cota superior de la semivarianza o la asíntota superior del semivariograma. Únicamente los procesos estacionarios de segundo orden tienen silla. En estos casos la silla es $C(0)$; también es conocida como meseta.

Rango: Es la distancia a la cual los puntos ya no se consideran correlacionados espacialmente. Los puntos separados por una distancia inferior al rango se consideran espacialmente correlacionados; observaciones espaciadas por más que el rango se consideran independientes o al menos aproximadamente independientes. Algunos procesos alcanzan correlación cero solo asintóticamente, mientras que otros tienen un rango finito.

Efecto pepita: De la definición de semivariograma, se puede ver que para $h = 0$, debería ocurrir que $\gamma(h) = 0$. Sin embargo, en general se presenta el comportamiento

observado en la Figura 1, existiendo una discontinuidad en el origen, $\gamma(h) \to c_0$ cuando $h \to 0$.

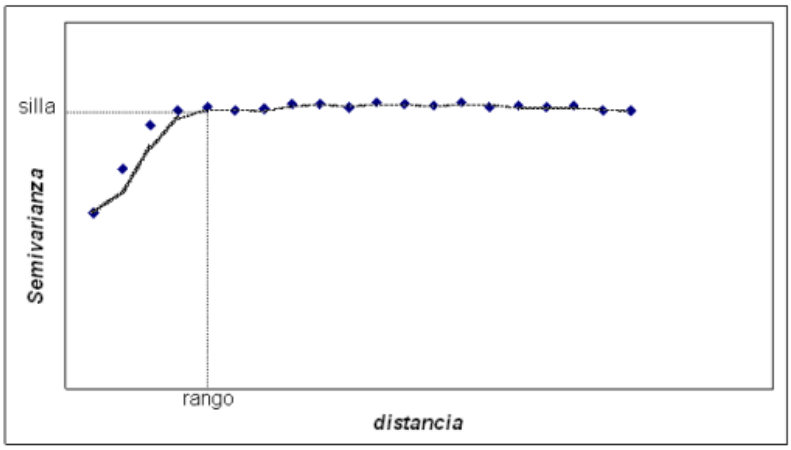


FIGURA 1. Pepita, silla y rango en presencia de estacionariedad de segundo orden.

Como el semivariograma $\gamma(h)$ es la varianza de la variable incrementos, como se muestra en (1), un estimador muy natural es el conocido como el estimador clásico, y consiste de la estimación de esta varianza por el método de los momentos:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} \left(Z(s+h) - Z(s)\right)^2, \qquad h \in \mathbb{R}^d,$$

donde

$$N(h) \equiv \{(s_i, s_j) : s_i - s_j = h\}.$$

$N(h)$ es el conjunto de todos los pares de ubicaciones cuya separación corresponde a un vector h y $|N(h)|$ es el cardinal de $N(h)$.

Una vez definida la estructura matemática del semivariograma y descritos sus parámetros fundamentales (pepita, silla y rango), es necesario introducir los modelos teóricos que permiten ajustar el semivariograma experimental obtenido a partir de los datos.

Los modelos teóricos son funciones válidas que cumplen las propiedades de no negatividad y definida-positividad, y que representan distintos comportamientos de la variabilidad espacial. Su selección es un paso esencial para la posterior aplicación de métodos de interpolación como el Kriging. Algunos modelos capturan estructuras suaves, otros representan comportamientos asintóticos, y algunos permiten incluso patrones oscilatorios.

La Figura 2 ilustra varios de los modelos teóricos más

Estos modelos permiten capturar diferentes formas de dependencia espacial y serán evaluados en la fase de modelación para seleccionar aquel que mejor reproduzca la estructura observada en los datos de precipitación.

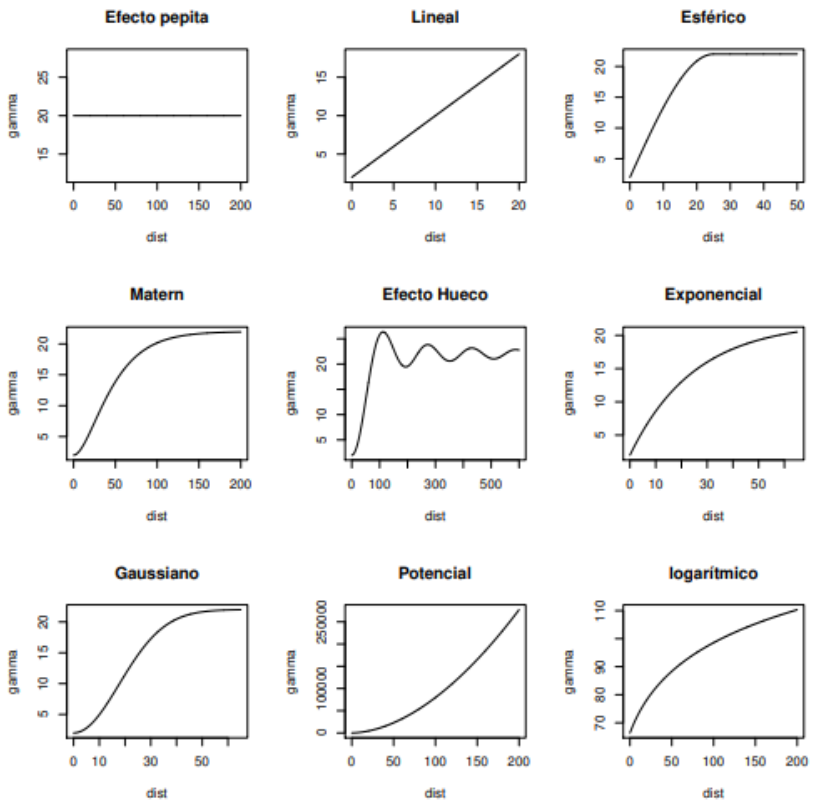


FIGURA 2. Ejemplos de modelos teóricos de semivariograma.

Una vez definido y modelado el semivariograma, se cuenta con la estructura necesaria para realizar predicción espacial. Con este modelo como base, se introduce el método de Kriging, que permite estimar valores en puntos no muestreados de manera óptima.

Kriging

Uno de los objetivos principales del analisis estadístico de datos espaciales en dominio continuo es la prediccion en lugares no muestreados. Así, se ha observado el campo aleatorio

$$\{Z(s) : s \in D \subset \mathbb{R}^d\}$$

en las ubicaciones $s_1, s_2, \ldots, s_n$ y se desea predecir la variable aleatoria espacial no observada $Z(s_0)$ con base en los valores observados $z(s_1), z(s_2), \ldots, z(s_n)$ utilizando su estructura de autocorrelacion espacial. Aunque existen muchos metodos determinísticos para obtener valores en lugares no muestreados, usar los metodos estadísticos de prediccion espacial presentan una gran ventaja y es que ademas de la predicción se obtiene la estimación de la varianza del error de prediccion, En particular el predictor kriging es insesgado y de mínima varianza. Los mapas de prediccion generados con kriging se acompañan, de los respectivos mapas de residuos para poder determinar cuales zonas tienen predicciones mas precisas. Ademas,

se usan las medidas generales para calidad de prediccion, tales como los estadísticos de los residuos, el MAPE, el CME, el coeficiente de correlacion lineal entre los valores observados y sus respectivas predicciones.

Se requiere una forma de predecir valores en puntos intermedios o en el caso de bloques, por ejemplo, estimar el promedio sobre el bloque. La precision del predictor usado depende de varios factores:

- El numero de muestras tomadas. Debido a la existencia de autocorrelación, los datos espacio temporales presentan redundancia. Por lo tanto, una muestra de tamano n de datos independientes tiene mayor cantidad de informacion que una muestra de tamaño n de datos autocorrelacionados.
- La calidad de la medicion en cada punto. Aunque el parámetro conocido como efecto pepita permite cuantificar el error de medicion, esto aumenta la incertidumbre en el modelo de dependencia espacial y por lo tanto en la prediccion.
- Las ubicaciones de las muestras en la zona; si las muestras son tomadas de acuerdo a un diseno de muestreo optimo los resultados son mucho mejores, las predicciones son mas precisas ya que la varianza es menor, se evita la redundancia espacial y ademas se optimizan los recursos.

Kriging ordinario

El kriging ordinario se usa cuando la variable es al menos estacionaria intrínseca y tiene media constante pero desconocida. Es decir, se asume que el proceso espacial se puede descomponer de la siguiente forma:

$$Z(s) = \mu + e(s) \qquad s \in D$$

Donde $E[Z(s)] = \mu \ \forall s \in D$, $\mu \in \mathbb{R}$ pero es necesario estimarla y por lo tanto no se puede trabajar directamente con la variable centrada.

$$E\left(\sum_{i=1}^{n} \lambda_i Z(s_i)\right) = E(Z(s_0))$$

Tomando esperanzas se obtiene:

$$\sum_{i=1}^{n} \lambda_i \mu = \mu$$

De donde se concluye que para que se cumpla la propiedad de insesgamiento se requiere que:

$$\sum_{i=1}^{n} \lambda_i = 1$$

Kriging Universal

Sea el modelo geoestadístico

$$Z(s) = \mu(s) + \varepsilon(s)$$

y $E(\varepsilon(s)) = 0$. Se conoce como kriging universal al caso en el que $\mu(s)$ es desconocida y localmente puede expresarse como una combinacion lineal de funciones $f_k(s)$. Es

decir, la media en general es es una combinacion lineal de términos de la forma $x^p y^q$, $p, q \in \mathbb{N}$. Esto es:

$$\mu(s) = \sum_{k=1}^{K} \beta_k f_k(s) \qquad s = (x, y)$$

Por ejemplo, si el modelo para la media es: $\mu(s) = \beta_1 + \beta_2 x + \beta_3 y$

$$f_1(s) = 1, \qquad f_2(s) = x, \qquad f_3(s) = y.$$

El predictor es el usual, y lo expresamos en términos de la media:

$$Z^*(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

$$Z^*(s_0) = \sum_{i=1}^{n} \lambda_i \left(\mu(s_i) + \varepsilon(s_i)\right)$$

$$Z^*(s_0) = \sum_{i=1}^{n} \lambda_i \sum_{k=1}^{K} \beta_k f_k(s_i) + \sum_{i=1}^{n} \lambda_i \varepsilon(s_i)$$

$$E(Z^*(s_0)) = E\left(\sum_{i=1}^{n} \lambda_i \sum_{k=1}^{K} \beta_k f_k(s_i)\right)$$

Ahora, para garantizar el insesgamiento se requiere que

$$E(Z^*(s_0) - Z(s_0)) = 0$$

Entonces,

$$E\left(\sum_{i=1}^{n} \lambda_i Z(s_i)\right) = E(Z(s_0))$$

Así,

$$\sum_{i=1}^{n} \lambda_i \mu(s_i) = \mu(s_0)$$

Reemplazando la expresión de la media:

$$\sum_{i=1}^{n} \lambda_i \sum_{k=1}^{K} \beta_k f_k(s_i) = \sum_{k=1}^{K} \beta_k f_k(s_0)$$

$$\sum_{k=1}^{K} \beta_k \sum_{i=1}^{n} \lambda_i f_k(s_i) = \sum_{k=1}^{K} \beta_k f_k(s_0)$$

$$\sum_{i=1}^{n} \lambda_i f_k(s_i) = f_k(s_0), \qquad k = 1, \ldots, K$$

Restricciones

$$\sum_{i=1}^{n} \lambda_i f_k(s_i) = f_k(s_0), \qquad k = 1, \ldots, K$$

La expresión a minimizar queda:

$$Q = E(Z^*(s_0) - Z(s_0))^2 - 2\sum_{k=1}^{K} \delta_k \left(\sum_{j=1}^{n} \lambda_j f_k(s_j) - f_k(s_0)\right)$$

$$\frac{\partial Q}{\partial \lambda_i} = 2\gamma(s_i - s_0) - 2\sum_{j=1}^{n} \lambda_j \gamma(s_i - s_j) - 2\sum_{k=1}^{K} \delta_k f_k(s_i), \qquad i = 1, \ldots, n$$

sujeta a las restricciones:

$$\sum_{i=1}^{n} \lambda_i f_k(s_i) = f_k(s_0), \qquad k = 1, \ldots, K$$

Por ejemplo para $k = 1$,

$$\lambda_1 f_1(s_1) + \lambda_2 f_1(s_2) + \cdots + \lambda_n f_1(s_n) = f_1(s_0)$$

En general:

$$\sum_{j=1}^{n} \lambda_j \gamma(s_i - s_j) + \sum_{k=1}^{K} \delta_k f_k(s_i) = \gamma(s_i - s_0), \qquad i = 1, \ldots, n$$

Las ecuaciones generales del kriging universal en términos del semivariograma y en forma matricial están dadas por:

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & f_1^1 & f_2^1 & \cdots & f_K^1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & f_1^2 & f_2^2 & \cdots & f_K^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & f_1^n & f_2^n & \cdots & f_K^n \\ f_1^1 & f_1^2 & \cdots & f_1^n & 0 & 0 & \cdots & 0 \\ f_2^1 & f_2^2 & \cdots & f_2^n & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & 0 & \cdots & 0 \\ f_K^1 & f_K^2 & \cdots & f_K^n & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \delta_1 \\ \vdots \\ \delta_K \end{pmatrix} = \begin{pmatrix} \gamma(s_1 - s_0) \\ \gamma(s_2 - s_0) \\ \vdots \\ \gamma(s_n - s_0) \\ f_1^0 \\ f_2^0 \\ \vdots \\ f_K^0 \end{pmatrix}$$

Se evidencia el requerimiento de que las f_k sean linealmente independientes:

$$\sum_{k=1}^{K} c_k f_k(s_i) = 0 \Longleftrightarrow c_k = 0, \ k = 1, \ldots, K$$

La varianza del Kriging Universal queda:

$$E(Z^*(s_0) - Z(s_0))^2 = 2\sum_{i=1}^{n} \lambda_i \gamma(s_i - s_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j)$$

Sustituyendo:

$$E(Z^*(s_0) - Z(s_0))^2 = \sum_{i=1}^{n} \lambda_i \gamma(s_i - s_0) + \sum_{k=1}^{K} \delta_k f_k(s_0)$$

Note que si $K = 1$ y $f_1(s) = 1$:

$$E(Z^*(s_0) - Z(s_0))^2 = \sum_{i=1}^{n} \lambda_i \gamma(s_i - s_0) + \delta,$$

y se obtiene la varianza del kriging ordinario como un caso particular.

Cokriging

El cokriging consiste en encontrar predicciones de una variable de interes en un lugar s_0 utilizando la informacion dada por covariables. No es indispensable que tanto la variable de interes como las covariables sean medidas en el mismo lugar.

$$Z_1^*(Z_1, Z_2; s_0) = \sum_{i=1}^{n_1} \lambda_{1i} Z_1(s_i) + \sum_{j=1}^{n_2} \lambda_{2j} Z_2(s_j) = \boldsymbol{\lambda}_1^0 Z_1(s) + \boldsymbol{\lambda}_2^0 Z_2(s)$$

Supuestos

$$E[Z_1(s)] = \mu_1$$

y

$$E[Z_2(s)] = \mu_2, \qquad \forall s \in D$$

$$\mathrm{Cov}(Z_1(s), Z_1(s+h)) = C_1(h)$$

$$\mathrm{Cov}(Z_2(s), Z_2(s+h)) = C_2(h)$$

$$\mathrm{Cov}(Z_1(s), Z_2(s+h)) = C_{12}(h)$$

Hasta ahora se ha llevado a cabo la prediccion espacial de un proceso $Z(s)$ utilizando unicamente su propia informacion. Sin embargo, los fenomenos del mundo real son en general multivariados. Este capítulo describe como llevar a cabo la prediccion espacial de un proceso $Z_1(s_0)$ utilizando su propia informacion, así como la de covariables que se encuentren espacialmente correlacionadas con este, esto es, utilizando $Z_1(s), \ldots, Z_P(s)$. Este método es conocido como cokriging.

Para la aplicacion de este método no es necesario que todas las variables esten medidas en las mismas ubicaciones espaciales. Si todos los datos se encuentran medidos en la misma grilla de n ubicaciones espaciales, los datos son $P \times 1$-vectores que forman una matriz $n \times P$, con (i, j)-ésimo elemento $Z_p(s_i)$, $i = 1, \ldots, n$, $p = 1, \ldots, P$. La i-esima fila de la matriz de datos corresponde a las mediciones de todas las variables en la ubicacion s_i:

$$Z(s_i) = (Z_1(s_i), Z_2(s_i), \ldots, Z_P(s_i))^{\mathrm{t}}$$

El interes es predecir el vector

$$Z(s_0) \equiv (Z_1(s_0), \ldots, Z_P(s_0))^{\mathrm{t}}$$

La prediccion es realizada para una variable a la vez.

Si todas las variables son medidas en las mismas n ubicaciones la matriz de covarianza completa $\mathrm{Cov}(Z) = \Sigma$ con todas las variables y ubicaciones observadas es

$$\Sigma = \begin{pmatrix} \Sigma(s_1, s_1) & \Sigma(s_1, s_2) & \cdots & \Sigma(s_1, s_n) \\ \Sigma(s_2, s_1) & \Sigma(s_2, s_2) & \cdots & \Sigma(s_2, s_n) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma(s_n, s_1) & \Sigma(s_n, s_2) & \cdots & \Sigma(s_n, s_n) \end{pmatrix}$$

La cual debe ser una matriz definida positiva, Vease con mas detalle en ([Bohórquez, 2024]).

5. Ejemplos

5.1. Ejemplos ilustrativo de Métodos de Corrección de Sesgo y Métricas. Con el fin de ilustrar el procedimiento de corrección de sesgo aplicado a productos satelitales de precipitación, se construyó un ejemplo simulado utilizando una serie mensual de longitud $n = 120$. La precipitación observada se generó mediante una distribución Gamma con parámetros *shape* $= 2$ y *rate* $= 1/80$ (media ≈ 160 mm), mientras que la serie satelital se definió como $P_{\text{sat}} = 1{,}2\, P_{\text{obs}} + \varepsilon$, con $\varepsilon \sim N(0, 20^2)$, reproduciendo una sobreestimación sistemática típica de algunos productos satelitales.

Para evaluar el desempeño del satélite se calcularon cinco métricas: BIAS, MAE, RMSE, la correlación temporal (r) y la eficiencia de Nash–Sutcliffe (NSE). Los resultados iniciales mostraron un sesgo positivo elevado y errores significativamente superiores a los de la serie observada, aunque con una correlación temporal alta.

A la serie satelital se le aplicaron tres métodos de corrección: *Linear Scaling* (LS), *Power Transformation* (PT) y *Quantile Mapping* (QM). LS eliminó casi por completo el sesgo medio; PT produjo mejoras similares mediante un ajuste no lineal; y QM ofreció el mejor rendimiento global al corregir tanto la media como la forma de la distribución, reduciendo MAE y RMSE y mejorando la NSE.

Evaluación antes y después de la corrección de sesgo. El cuadro 2 resume el desempeño de la serie satelital antes y después de aplicar los métodos de corrección. La serie original presenta un sesgo positivo considerable (BIAS = 27.64 mm), lo que indica una sobreestimación sistemática respecto a la precipitación observada. Tanto el MAE (31.77 mm) como el RMSE (38.72 mm) muestran errores elevados, aunque la correlación temporal es muy alta ($r \approx 0{,}98$), señal de que la variabilidad temporal está bien representada. La eficiencia de Nash–Sutcliffe (NSE = 0.81) confirma una capacidad predictiva moderada.

Tras aplicar el método *Linear Scaling* (LS), el sesgo se reduce prácticamente a cero y los errores disminuyen de forma notable (MAE = 14.66 mm; RMSE = 17.99 mm), manteniendo la misma correlación. De forma similar, la *Power Transformation* (PT) elimina el sesgo y mejora significativamente los errores, obteniendo valores comparables a los de LS.

Por otro lado, el método *Quantile Mapping* (QM) alcanza la discrepancia promedio más baja (MAE = 13.99 mm), ajustando mejor la distribución de los datos y conservando una correlación elevada ($r \approx 0{,}98$), aunque mantiene un sesgo residual pequeño (BIAS = 1.93 mm).

En conjunto, los tres métodos mejoran de manera importante la representación de la precipitación, siendo LS y PT los más efectivos para corregir el sesgo medio, mientras que QM proporciona el mejor ajuste de la distribución al lograr el MAE más bajo. Estos resultados demuestran la eficacia de las técnicas de corrección de sesgo para mejorar productos satelitales como CHIRPS.

La Figura 3 presenta los diagramas de dispersión entre la precipitación observada y las distintas versiones de la serie satelital: original y corregidas mediante *Linear Scaling* (LS), *Power Transformation* (PT) y *Quantile Mapping* (QM).

Métrica	Original	LS	PT	QM
BIAS (mm)	27.64	0.02	0.01	1.93
MAE (mm)	31.77	14.66	15.02	13.99
RMSE (mm)	38.72	17.99	18.44	17.21
r	0.98	0.98	0.98	0.98
NSE	0.81	0.94	0.93	0.95

CUADRO 2. Métricas antes y después de la corrección de sesgo

En el panel correspondiente a la serie satelital original, se aprecia una dispersión considerable y una tendencia sistemática por encima de la línea de referencia, reflejando la sobreestimación capturada por el BIAS positivo. Tras aplicar LS, los puntos se acercan notablemente a la diagonal, lo que indica que el ajuste de escala corrige efectivamente el sesgo medio. La transformación PT produce un patrón similar, con una alineación estrecha a lo largo de toda la distribución.

Por su parte, el método QM genera la mayor adherencia a la línea 1:1, mostrando una corrección más completa que abarca tanto la media como la forma de la distribución. Los puntos se distribuyen de manera más compacta y cercana a la diagonal, especialmente en los valores medios y altos de precipitación.

En conjunto, la figura evidencia que todos los métodos mejoran la correspondencia entre ambas series, siendo QM el que logra la corrección más uniforme sobre toda la gama de valores.

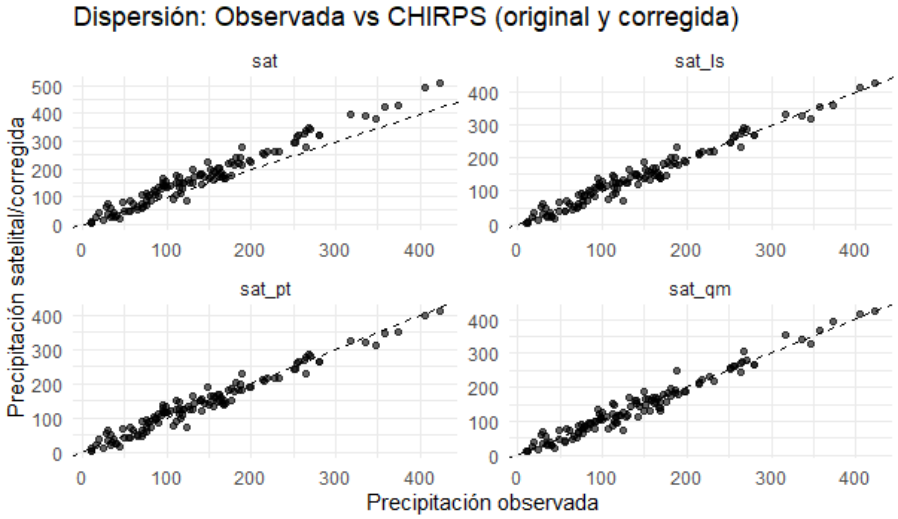


FIGURA 3. Diagramas de dispersión entre la precipitación observada y la satelital, antes y después de aplicar cada método de corrección (LS, PT y QM).

Las Figuras 4, 5 y 6 muestran la comparación temporal entre la serie observada y las versiones corregidas del producto satelital mediante los métodos Linear Scaling (LS), Power Transformation (PT) y Quantile Mapping (QM), respectivamente.

En el caso de Linear Scaling, la serie corregida reproduce con gran fidelidad la dinámica temporal de la precipitación, ya que el método ajusta únicamente la escala y conserva la forma original de la serie satelital. Esto se refleja en una alineación estrecha entre ambas curvas a lo largo de todo el periodo.

El método Power Transformation produce un comportamiento similar, manteniendo la coherencia temporal pero aplicando un ajuste adicional según el parámetro de potencia λ, lo que permite modificar levemente la forma de los valores extremos.

Por su parte, Quantile Mapping muestra la mayor correspondencia con la serie observada, ya que corrige no solo la media y la escala sino también la distribución completa. Esto se evidencia en una coincidencia más precisa de los picos altos y de los valores bajos de precipitación.

En conjunto, las tres gráficas confirman visualmente la mejora lograda con los distintos métodos de corrección, siendo QM el que presenta el ajuste más completo en términos de magnitud y comportamiento temporal.

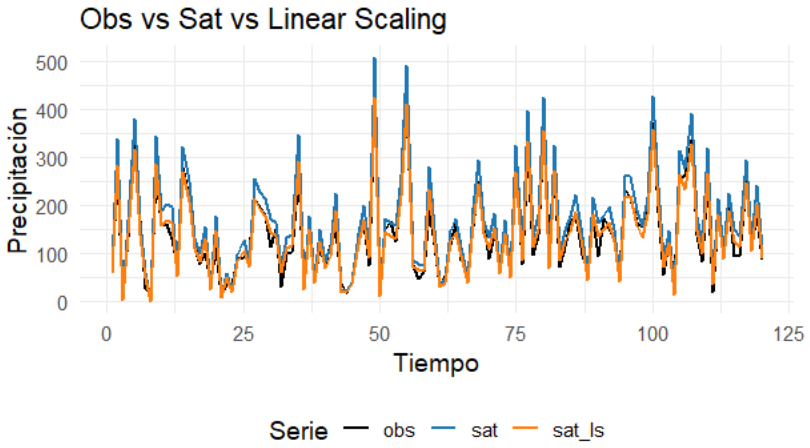


FIGURA 4. Comparación temporal entre la precipitación observada, satelital y corregida mediante Linear Scaling.

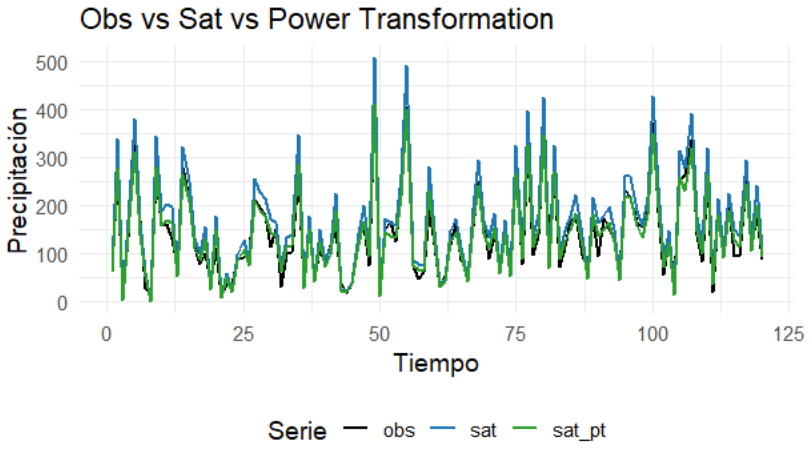


FIGURA 5. Comparación temporal entre la precipitación observada, satelital y corregida mediante Power Transformation.

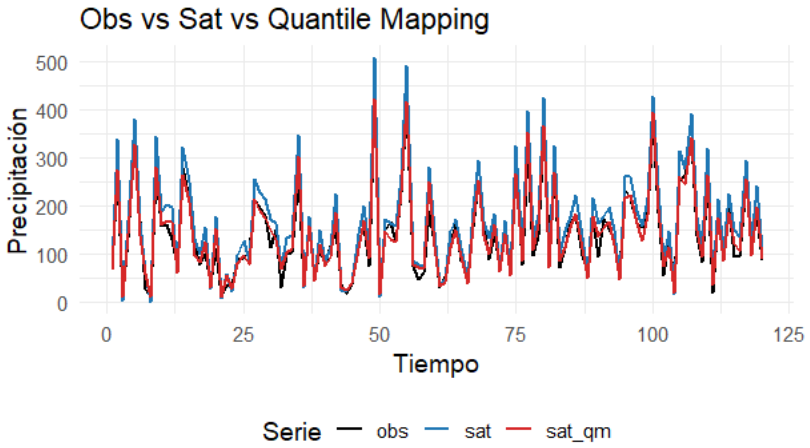


FIGURA 6. Comparación temporal entre la precipitación observada, satelital y corregida mediante Quantile Mapping.

5.2. Ejemplo ilustrativo Método Kriging con datos simulados. Con el objetivo de mostrar de forma clara el procedimiento de interpolación espacial aplicado en este estudio, se desarrolló un ejemplo utilizando datos simulados. Se generaron 15 estaciones ubicadas dentro del rango geográfico aproximado de Honduras y se asignaron valores de precipitación simulada mediante una distribución Gamma con parámetros *shape* $= 2$ y *scale* $= 60$ (equivalente a *rate* $= 1/60$). Esta parametrización produce valores positivos y una dispersión amplia, con una media cercana a 120 mm, lo cual resulta adecuado para representar variabilidad típica de precipitaciones intensas. Aunque los datos no corresponden a mediciones reales, permiten visualizar de manera fiel el flujo de modelación utilizado en el análisis principal.

5.3. Semivariograma experimental y selección del modelo. Tras transformar las coordenadas geográficas al sistema UTM Zona 16N, se construyó el semivariograma experimental para describir la variabilidad espacial de la precipitación simulada. El semivariograma mostró un aumento rápido de la dispersión para distancias cortas, seguido de una estabilización progresiva al incrementarse la distancia. Este patrón es característico de procesos que presentan correlación espacial hasta un rango finito.

Con base en este comportamiento, se ajustaron varios modelos teóricos (Exponencial, Gaussiano y Esférico). El modelo esférico presentó el ajuste más coherente, ya que:

- reproduce adecuadamente el incremento inicial de variabilidad,
- captura la meseta (sill) observada en el semivariograma experimental,
- y presenta un rango consistente con la separación máxima entre las estaciones simuladas.

La Figura 7 muestra el semivariograma experimental junto con el modelo esférico ajustado, evidenciando su adecuación para este ejemplo.

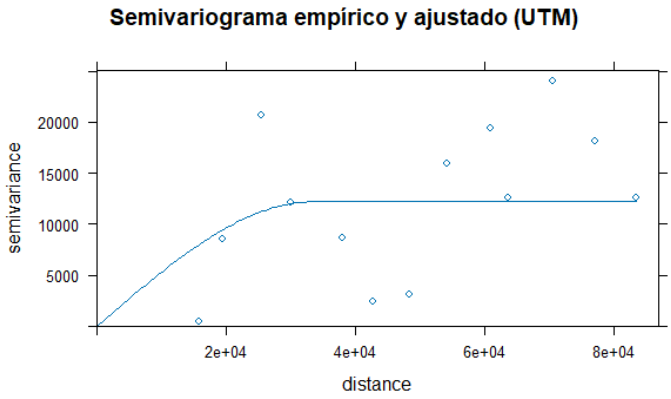


FIGURA 7. Semivariograma experimental y modelo esférico ajustado para los datos simulados.

5.4. Aplicación del Kriging Ordinario. Una vez seleccionado el modelo esférico y definido el grid de interpolación, se aplicó Kriging Ordinario para obtener la superficie continua de precipitación simulada. El resultado (Figura 8) muestra una distribución espacial claramente diferenciada por intervalos de valores, los cuales aparecen en la leyenda del mapa. Estos rangos corresponden a la precipitación predicha en milímetros y abarcan desde aproximadamente **30 mm** hasta cerca de **335 mm**.

En el mapa se observan cinco clases de color, cada una asociada a un intervalo específico:

- **Azul oscuro (30–91 mm):** representa las zonas con la precipitación estimada más baja, generalmente influenciadas por estaciones que en los datos simulados tenían valores pequeños.
- **Morado (91–152 mm):** indica valores bajos a moderados, formando una transición alrededor de las zonas más frías.

- **Rosado (152–213 mm):** corresponde a valores intermedios de precipitación y cubre gran parte de la superficie, reflejando la suavidad típica del kriging.
- **Naranja (213–273 mm):** identifica regiones donde el modelo predice precipitaciones relativamente altas.
- **Amarillo (273–334 mm):** marca los valores máximos predichos, ubicados en áreas cercanas a estaciones con altos valores simulados.

Cada núcleo o mancha circular de color corresponde a la *zona de influencia* de una estación, lo cual es característico del Kriging cuando se trabaja con un conjunto reducido de puntos. Las transiciones entre colores son suaves, lo que confirma que la interpolación respeta la estructura espacial definida por el semivariograma ajustado: la predicción coincide estrechamente con los valores de las estaciones en zonas cercanas y se suaviza progresivamente a medida que aumenta la distancia.

En conjunto, la figura evidencia que el modelo geoestadístico genera una superficie continua, coherente y estructuralmente consistente con los datos simulados, permitiendo identificar con claridad zonas de mayor y menor precipitación dentro del área evaluada.

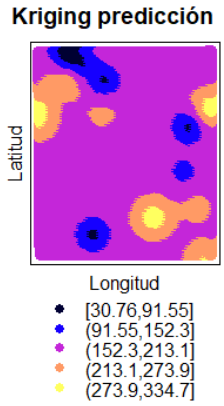


FIGURA 8. Predicción de precipitación obtenida mediante Kriging Ordinario en el ejemplo simulado.

6. CONCLUSIONES

El análisis geoestadístico realizado con datos simulados permitió verificar la coherencia del proceso de construcción y ajuste del semivariograma, así como la idoneidad del modelo esférico seleccionado. El semivariograma experimental presentó una meseta en torno a 12 000 mm^2 y un rango aproximado de 30 000 metros, parámetros que permitieron generar mediante Kriging Ordinario una superficie de predicción suave y espacialmente consistente con la disposición de las estaciones simuladas. Este resultado demuestra la estabilidad y representatividad del enfoque utilizado.

Complementariamente, el ejemplo de corrección de sesgo mostró que los métodos LS, PT y QM mejoran sustancialmente la correspondencia entre la precipitación satelital y la observada, reduciendo el sesgo y los errores asociados, y preservando la estructura temporal de la serie. Entre ellos, Quantile Mapping destacó por ofrecer la corrección más completa al ajustar tanto la media como la forma de la distribución.

En conjunto, ambos ejercicios ilustran de manera clara la validez de los procedimientos empleados en esta investigación, tanto para la modelación espacial mediante kriging como para la corrección estadística de productos satelitales, lo que respalda su aplicación sobre datos reales en el análisis principal.

Referencias

[Al-Shamayleh et al., 2024] Al-Shamayleh, S., Tan, M. L., Samat, N., Rahbeh, M., and Zhang, F. (2024). Performance of chirps for estimating precipitation extremes in the wala basin, jordan. *Journal of Water and Climate Change*, 15(3):1349–1363. Open Access, CC BY 4.0.

[Atiah et al., 2023] Atiah, W. A., Johnson, R., Muthoni, F. K., Mengistu, G. T., Amekudzi, L. K., Kwabena, O., and Kizito, F. (2023). Bias correction and spatial disaggregation of satellite-based data for the detection of rainfall seasonality indices. *Heliyon*, 9(e17604):1–15. Open Access, CC BY-NC-ND 4.0.

[Bohórquez, 2024] Bohórquez, M. (2024). Estadística espacial y espacio-temporal para campos aleatorios escalares y funcionales: Notas de clase. Apuntes de clase. Universidad (curso de Estadística Espacial y Espacio-Temporal).

[Bollat Flores, 2023] Bollat Flores, P. L. (2023). Análisis comparativo de datos de precipitación chirps con datos pluviométricos locales en el departamento de chiquimula, guatemala. Tesis de Licenciatura, Universidad de San Carlos de Guatemala, Facultad de Ingeniería. Se aplicó interpolación espacial mediante Kriging Ordinario para comparar datos CHIRPS con registros locales, obteniendo una correlación positiva de 0.84 y correspondencia espacial del 80 %.

[Funk et al., 2015] Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J. (2015). The climate hazards infrared precipitation with stations (chirps)—a new environmental record for monitoring extremes. *Scientific Data*, 2(150066):1–21.

[Pichardo, 2024] Pichardo, D. (2024). Validación de precipitación en la subcuenca del lago de yojoa: datos satelitales versus observados. Tesis de Licenciatura, Universidad Nacional Autónoma de Honduras (UNAH). Se evaluaron los productos CHIRPS v2.0 y CMORPH frente a datos de estaciones hidroclimatológicas de la ENEE durante 1981–2023, aplicando corrección mediante escalamiento lineal y transformación de potencias.

[Rivera et al., 2019] Rivera, J. A., Hinrichs, S., and Marianetti, G. (2019). Using chirps dataset to assess wet and dry conditions along the semiarid central-western argentina. *Advances in Meteorology*, 2019:1–18. Open Access, CC BY 4.0.

[Tobler, 1970] Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(sup1):234–240.

[1] Maestría en Matemática,Orientación en estadística, Universidad Nacional Autónoma de Honduras

Dirección de correo electrónico: kvasquezz@unah.hn