



Maestría en

INTELIGENCIA ARTIFICIAL APLICADA

**Trabajo previo a la obtención de título de
Magister en Inteligencia Artificial Aplicada**

AUTOR/ES:

Byron Daniel Velasco Cueva

Alex Javier Pérez Vilañez

Christian Patricio Vaca Benalcázar

TUTOR/ES:

Bryan Steven Vinueza Bustamante

Karla Estefanía Mora Cajas

TEMA: Aplicación de Modelos Estadísticos Tradicionales y
de Machine Learning para predecir el crecimiento poblacional
utilizando datos censales escasos.

Certificación de autoría

Nosotros, **Byron Daniel Velasco Cueva, Alex Javier Pérez Vilañez, y Christian Patricio Vaca Benalcázar**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Firma
Byron Daniel Velasco Cueva



Firma
Alex Javier Pérez Vilañez



Firma
Christian Patricio Vaca Benalcázar

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Byron Daniel Velasco Cueva, Alex Javier Pérez Vilañez, y Christian Patricio Vaca Benalcázar**, en calidad de autores del trabajo de investigación titulado **Aplicación de Modelos Estadísticos Tradicionales y de Machine Learning para predecir el crecimiento poblacional utilizando datos censales escasos**, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, (julio 2026)



Firma

Byron Daniel Velasco Cueva



Firma

Alex Javier Pérez Vilañez



Firma

Christian Patricio Vaca Benalcázar

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López** Director EIG y **Karla Estefanía Mora Cajas** Coordinadora UIDE, declaramos que: **Byron Daniel Velasco Cueva, Alex Javier Pérez Vilañez, y Christian Patricio Vaca Benalcázar** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López Director de la
Maestría en Inteligencia Artificial
Aplicada



Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Inteligencia Artificial Aplicada

DEDICATORIA

A mi amada esposa Raquel por ser un pilar fundamental en cada proyecto que inicio, a mis pequeños hijos Neidan y Francisco por permitirme usar los fines de semana de juego y salidas al parque para tomar las clases y elaborar tareas.

Christian Vaca Benalcázar

Para Dios, Giu y Gael. Si el amor fuera como la gravedad, ustedes serían mi constante.

Alex Javier Pérez Vilañez

A mis padres y hermanos, por sostenerme en cada etapa de este camino. A mis hijos, por acompañarme y darme fuerza incluso sin saber bien qué hacía su papá tantas noches frente a la computadora. Esta maestría es apenas un paso más de un camino que dejé pendiente hace años, y que hoy retomo con la misma convicción de entonces. Esta tesis es para ustedes: para que aprendan que no hay victoria que celebrar sin antes haber sacrificado algo por ella. Den siempre lo mejor de sí mismos, la vida sabe recompensar a quien se esfuerza.

Byron Daniel Velasco Cueva

AGRADECIMIENTOS

A Dios por siempre bendecirme con los recursos económicos y de tiempo para cubrir esta inversión, a mis colegas Byron y Alex por su compromiso en los trabajos que desarrollamos durante los diferentes módulos donde tuve la oportunidad de conocerlos como personas y estrechar el lazo de amistad, a la UIDE por crear este tipo de programas que permiten incrementar el conocimiento para hacer un mejor Ecuador.

Christian Vaca Benalcázar

Agradecido con Dios, Giu, Lito, Anahichis y mis padres. Su tiempo, que siempre avanza, supo detenerse para darme la oportunidad de crecer. Siempre les estaré en deuda.

Alex Javier Pérez Vilañez

A Dios, por ayudarme a poner en orden mis prioridades y darme la disciplina para seguir estudiando. A mis padres y hermanos, por su apoyo constante y por acompañarme sabiendo desde el inicio lo que este camino significaba. A mi hermano James, en particular, por motivarme con su ejemplo a seguir preparándome y a ser cada día mejor. A Renatha Valentina y Daniel Alejandro, mis hijos, por su compañía en tantas horas de trabajo que para ellos eran solo eso: horas en que su papá estaba ocupado, sin saber por qué.

Ellos no saben que curso esta maestría. No ha sido un secreto por falta de confianza, sino porque hay logros que se agradecen mejor con hechos que con palabras, y este quería dárselo así: como una sorpresa, el día que me vean graduarme. Este trabajo es un agradecimiento silencioso, escrito antes de que ellos sepan que existía algo que agradecer.

Byron Daniel Velasco Cueva

RESUMEN

La predicción del crecimiento poblacional se considera un insumo esencial para la planificación territorial, la asignación de recursos y las políticas públicas. Sin embargo, Ecuador y gran parte de América Latina y el Caribe, enfrentan una limitación estructural: la escasez de datos temporales, con apenas cuatro censos disponibles (1990, 2001, 2010 y 2022) en 32 años. Esta investigación evalúa comparativamente cinco familias de modelos para proyectar la población de los 221 cantones del Ecuador, integrando los microdatos censales del Instituto Nacional de Estadística y Censos (INEC) con registros vitales anuales de nacidos vivos (ENV) y defunciones generales (EDG) del período 1990-2024. Además, se contrastan modelos de referencia (Tasa de Crecimiento Medio Anual, TCMA), el método demográfico Hamilton-Perry amortiguado, los modelos de aprendizaje automático basados en árboles (LightGBM global, con efectos aleatorios por cantón y por tipología demográfica), una arquitectura de aprendizaje profundo (Temporal Fusion Transformer, TFT) y ensambles híbridos. La evaluación emplea una partición temporal, entrenamiento con los censos 1990-2010 y prueba exclusiva sobre el censo 2022, y el Error Porcentual Absoluto Medio (MAPE) como métrica principal, examinada en tres niveles de agregación (cantonal, provincial y nacional) y bajo dos marcos de referencia diferenciados: el dato censal real y las proyecciones oficiales EPP INEC 2024. Los resultados muestran que el nivel de agregación geográfica determina el modelo óptimo: el Hamilton-Perry amortiguado es el más preciso a escala cantonal (MAPE 7,79 %) y provincial (≈ 8 %), mientras que el LightGBM global solo lo supera a escala nacional, donde reduce su sesgo a +0,49 % frente al +4,58 % de sobreestimación de la proyección oficial. Los seis modelos propios superan el baseline de TCMA, pero ninguna arquitectura compleja, incluido el TFT, mejora al método demográfico a nivel cantonal, confirmando, lógicamente, que la limitación es la escasez de datos y no la capacidad de los modelos. El hallazgo central, la inexistencia de un modelo completamente óptimo en ambas instancias y la dependencia de la escala territorial, constituye el aporte distintivo del presente trabajo de titulación, acompañado de un pipeline reproducible y proyecciones prospectivas a 2030 y 2035.

Palabras Claves: proyección poblacional, datos censales escasos, método Hamilton-Perry, aprendizaje automático, nivel de agregación geográfica, Ecuador.

ABSTRACT

Population growth projections are considered an essential input for spatial planning, resource allocation, and public policy. However, in Ecuador, as in much of Latin America, there is a structural limitation: a lack of time-series data, with only four censuses available (1990, 2001, 2010, and 2022) over a 32-year period. This study comparatively evaluates five model families to project the population of Ecuador's 221 cantons, integrating census microdata from the National Institute of Statistics and Censuses (INEC) with annual vital records of live births (ENV) and total deaths (EDG) for the period 1990–2024. In addition, the study compares reference models (Average Annual Growth Rate, TCMA), the damped Hamilton-Perry demographic method, tree-based machine learning models (global LightGBM, with random effects by canton and demographic type), a deep learning architecture (Temporal Fusion Transformer, TFT), and hybrid ensembles. The evaluation employs a temporal split, with training using the 1990–2010 censuses and testing exclusively on the 2022 census, and the Mean Absolute Percentage Error (MAPE) as the primary metric, examined at three levels of aggregation (cantonal, provincial, and national) and under two distinct reference frameworks: actual census data and the official INEC 2024 EPP projections. The results show that the level of geographic aggregation determines the optimal model: the damped Hamilton-Perry model is the most accurate at the cantonal (MAPE 7.79%) and provincial ($\approx 8\%$) levels, while the global LightGBM model outperforms it only at the national level, where it reduces its bias to +0.49% compared to the official projection's overestimation of +4.58%. The six in-house models outperform the TCMA baseline, but no complex architecture—including TFT—improves the demographic method at the cantonal level, logically confirming that the limitation lies in data scarcity rather than the models' capabilities. The central finding—the absence of a fully optimal model in both cases and the dependence on the territorial scale—constitutes the distinctive contribution of this thesis, accompanied by a reproducible pipeline and prospective projections for 2030 and 2035.

Keywords: population projection, sparse census data, Hamilton-Perry method, machine learning, geographic aggregation level, Ecuador.

TABLA DE CONTENIDOS (Índice)

CAPITULO 1:	1
1. INTRODUCCIÓN	1
1.1. Definición del proyecto	1
1.2. Justificación e importancia del trabajo de investigación	4
1.3. Alcance	6
1.4. Objetivos	6
CAPITULO 2:	8
2. REVISIÓN DE LITERATURA	8
2.1. Estado del Arte	8
2.2. Marco Teórico	12
2.3. Terminología	24
CAPITULO 3:	28
3. DESARROLLO	28
3.1. Desarrollo del Trabajo	28
3.2. Descripción del Data Set	31
3.3. Preparación y Procesamiento de Datos	36
3.4. Metodología de Análisis o Modelo	43
CAPITULO 4:	49
4. ANÁLISIS DE RESULTADOS	49
4.1. Pruebas de Concepto	49
4.2. Análisis de Resultados	67
CAPITULO 5:	79
5. CONCLUSIONES Y RECOMENDACIONES	79
5.1. Conclusiones	79
5.2. Recomendaciones	81
Referencias Bibliográficas	84
ANEXOS	88

LISTA DE TABLAS (Índice de tablas)

Tabla	Pág.
Tabla 1: Resumen de los problemas técnicos, metodológicos y de gestión identificados	2
Tabla 2: Relevancia e impacto académico, tecnológico y social	5
Tabla 3: Dimensiones de Alcance de Análisis del proyecto	6
Tabla 4: Clasificación de métodos tradicionales	10
Tabla 5: Comparación de los tipos de modelos de Proyección	12
Tabla 6: Componentes para la dinámica de crecimiento poblacional	13
Tabla 7: Métricas de evaluación para modelos de proyección	23
Tabla 8: Tipos de Data Leakage y estrategias de Mitigación	25
Tabla 9: Síntesis del diseño metodológico	31
Tabla 10: Variables del conjunto de datos por bloque de disponibilidad	33
Tabla 11: Población nacional del Ecuador por censo	37
Tabla 12: Correspondencia entre fases metodológicas, scripts y productos generados	42
Tabla 13: Aporte de las decisiones metodológicas a los objetivos específicos	43
Tabla 14: Comparación de modelos de referencia candidatos (Marco A)	52
Tabla 15: Versiones del Temporal Fusion Transformer evaluadas	54
Tabla 16: Variables de mayor importancia global según SHAP	56
Tabla 17: Valor óptimo de d según el periodo intercensal observado	67
Tabla 18: Validación retrospectiva 2010-2022 (Marco B)	69
Tabla 19: Cantones con mayor error de predicción (outliers)	74
Tabla 20: MAPE por modelo y nivel geográfico	76
Tabla 21: Proyecciones 2030 y 2035, frente a EPP INEC 2024	78

LISTA DE FIGURAS (Índice de figuras)

Figura	Pág.
Figura 1: Pasos para la estandarización de indicadores	11
Figura 2: Principios de la verdad demográfica	14
Figura 3: Evolución de los algoritmos basados en árboles de decisión	20
Figura 4: Pirámide poblacional del Ecuador, Censo 2022	35
Figura 5: Cantones más poblados y relación precariedad-servicios básicos, Censo 2022	35
Figura 6: Serie temporal cantonal del Ecuador, 1990-2022	36
Figura 7: Comparación gráfica de los modelos de referencia candidatos (Script 10)	53
Figura 8: Población real frente a predicha, LightGBM global (Script 22)	55
Figura 9: Comparación del ensamble híbrido frente a LightGBM y TCMA por separado (Script 23)	56
Figura 10: Importancia global de variables según SHAP, LightGBM global	57
Figura 11: Selección del número de clústeres: inercia, silhouette y Davies-Bouldin por valor de k (Script 26)	59
Figura 12: Distribución geográfica de las tres tipologías de cantones (Script 26)	60
Figura 13: Clústeres territoriales en el espacio de componentes principales (Script 26)	60
Figura 14: Perfil de variables por tipología de cantón (Script 26)	61
Figura 15: Clústeres espaciales LISA del crecimiento poblacional real, 2010-2022 (Script 25)	62
Figura 16: Distribución de efectos aleatorios por cantón (Script 27b)	62
Figura 17: Población real frente a predicha, LightGBM por tipología corregido (Script 28)	63
Figura 18: MAPE del ensamble según el peso asignado a Hamilton-Perry (Script 30)	65
Figura 19: MAPE de validación cruzada frente a MAPE de prueba, por peso del ensamble (Script 31)	65
Figura 20: Distribución del factor de amortiguación d^* calibrado por cantón (Script 32)	66
Figura 21: Valor óptimo de d según el periodo intercensal observado (Script 29)	68
Figura 22: Comparación de MAPE cantonal por modelo	70
Figura 23: Sesgo sistemático nacional por modelo	71
Figura 24: MAPE por quintil de tamaño poblacional	71
Figura 25: Cantones ganados por Hamilton-Perry frente a los modelos de ML, por provincia	72
Figura 26: Hamilton-Perry frente a ML, por quintil de tamaño poblacional	73
Figura 27: Desempeño en las provincias creadas después de 1990	73
Figura 28: Mapa coroplético de error de predicción por provincia	74
Figura 29: Heatmap de MAPE por provincia y modelo	75
Figura 30: Hamilton-Perry frente a ML, por tipología demográfica	75
Figura 31: Curva de agregación geográfica del error	76
Figura 32: Sesgo direccional por provincia: Hamilton-Perry frente a LightGBM	77
Figura 33: Ranking de desempeño por modelo y nivel geográfico	78
Figura 34: Proyecciones propias frente a EPP INEC 2024	79

CAPITULO 1:

1. INTRODUCCIÓN

1.1. Definición del proyecto

La predicción precisa del crecimiento poblacional constituye la base fundamental para la planificación territorial, la asignación de recursos públicos y la formulación de políticas de desarrollo. En Ecuador, el Instituto Nacional de Estadística y Censos - INEC (INEC, 2024) es el organismo encargado de realizar y publicar las estimaciones y proyecciones de población (EPP) siguiendo la recomendaciones de Organismos Internacionales como: la Organización de Naciones Unidas – ONU, la Comisión Económica para América Latina y el Caribe - CEPAL, la Ley de Estadística (publicada el 7 de mayo de 1976 según Decreto Supremo 323, Registro Oficial 82), y el ordenamiento del Sistema Estadístico Nacional - SEN.

Según el INEC en Ecuador se han realizado 8 censos de población y 7 de vivienda desde 1950, siendo el más reciente el EPP 2022 (INEC, 2023) que aplicó la metodología del derecho que consiste en obtener datos de las personas que residen en el hogar, aunque no estén presentes durante el empadronamiento, esta metodología ya se ha utilizado en países como Brasil, México y Argentina, hace varias décadas. (CEPAL, 2024)

El censo tradicional de población y vivienda sigue siendo una herramienta válida en la era digital, aunque si debería ser reemplazado por registros administrativos – RRAA y big data. En este estudio identifica problemas metodológicos, técnicos y conceptuales en los censos de hoy. (Durr J.-M, 2021)

Los censos presentan tres retos importantes: i) implementar nuevas tecnologías, ii) mejorar la cobertura y la calidad de los datos, y iii) insuficiencia de recursos financieros e incremento de costos en operaciones censales. (Mrkić S, 2020)

En el seminario regional sobre innovaciones y desafíos en encuestas de hogares realizado por la CEPAL del 19 al 21 de noviembre 2024, se presentó un breve panorama de los censos

de población y vivienda de la ronda 2020 en América Latina y el Caribe y sus principales desafíos hacia la ronda 2030, donde la omisión censal es uno de los principales problemas y en el caso de Ecuador presenta valores de 5% en base a la población publicada y 9,3% en base a población censada. (CEPAL, 2024)

A continuación, en la Tabla 1 se resumen los problemas técnicos, metodológicos y de gestión identificados en las fuentes de investigación:

Tabla 1

Resumen de los problemas técnicos, metodológicos y de gestión identificados

Categoría del Problema	Problema Principal identificado	Descripción del Problema
Técnicos y tecnológicos	Implementación de nuevas tecnologías sin suficiente madurez	Uso de CAPI, autocuestionarios en línea, sistemas de georreferenciación y monitoreo en tiempo real, sin que todos los países cuenten con la infraestructura digital necesaria ni con personal capacitado.
	Dificultad en la integración de múltiples modos de recolección (PAPI - pen-and-paper interview, CAPI - computer assisted personal interviewing, y web self-enumeration)	Duplicidad, inconsistencia en el formato de los datos y problemas de relación entre registros generados de los distintos modos de recolección.
	Limitación en la gestión de los datos y de los sistemas de información	El alto volumen de información necesita de arquitecturas robustas que pueden generar el riesgo de pérdida o corrupción de datos.
Metodológicos y de calidad	Errores de sub cobertura y sobre cobertura	Las poblaciones históricamente invisibilizadas (migrantes, personas sin domicilio fijo, grupos étnicos, zonas rurales remotas) no se reflejan adecuadamente.
	Errores de medición	Variables sensibles como edad, etnia, migración, discapacidad, o situación económica a causa del diseño de cuestionarios o falta de capacitación del personal no permiten obtener información correcta.
	Pruebas y encuestas de medición de cobertura insuficientes	Falta de pruebas piloto que permitan corregir errores, a pesar de las recomendaciones metodológicas de la ONU.
Organizativos, logísticos y de recursos	Presión por costos y limitación en recursos financieros	Limitación de presupuesto para operaciones censales más complejas que obliga a realizar recortes en: personal, pruebas, marketing o estrategias de recuperación de cobertura.
	Planificación y cronología con dificultades	Preparación insuficiente, pruebas y ajustes metodológicos con decisiones de última hora que afectan la calidad de los resultados.
	Problemas de reclutamiento de personal calificado	Inadecuados procesos de reclutamiento, retención, capacitación y pago de censistas y supervisores.
Transición metodológica	Transición lenta de modelos “puros” a enfoques combinados	Se siguen utilizando enfoques tradicionales o combinan registros administrativos (RRAA) con encuestas, sin una metodología clara, lo que genera problemas de comparabilidad y consistencia.
	Sobre uso de imputación y supuestos para cubrir vacíos	Se recurre demasiado a la imputación de datos (por ejemplo, en viviendas sin entrevista), lo que afecta la validez estadística.
	Resistencia al cambio	Organismos nacionales de estadística con estructuras y procedimientos tradicionales se

		adaptan con lentitud a enfoques digitales o basados en registros.
Impacto de contextos externos	Factores externos inesperados	Crisis sanitarias, políticas de movilidad, crisis económicas, inseguridad), que después se confirman como problemas adicionales de implementación (rehacer cronogramas, cambios de metodología, aumento de no-respuesta, etc.)

En el caso particular de Ecuador, el último censo genero controversias como lo señalan medios de comunicación como Primicias en su nota de prensa “El censo poblacional, entre críticas, dudas técnicas y retrasos” de 15 de enero de 2023, donde entre otros aspectos se señala: terminación del censo fuera de la fecha planificada, inconsistencia técnica y diferencias en el costo, personas no censadas por aproximadamente 10 millones que corresponderían a 500.000 viviendas. Así también se menciona en la nota de prensa “No se debe repetir el censo, para nada” de 21 de marzo de 2024 una omisión censal del 4,2% que representaría 738.000 personas, reducción de la población afroecuatoriana, y confiabilidad de los resultados. (Primicias, 2022).

El problema radica en la metodología utilizada por el INEC en el último censo, que requiere datos anuales de fecundidad, mortalidad y migración, los cuales presentan importantes limitaciones de disponibilidad y calidad a nivel cantonal, especialmente en territorios rurales y amazónicos.

Esta investigación aborda el problema de construir modelos de predicción poblacional a nivel cantonal para los 221 cantones del Ecuador, utilizando exclusivamente datos de los cuatro censos nacionales disponibles (1990, 2001, 2010 y 2022) complementados con registro estadístico de defunciones generales -EDG y estadísticas de nacidos vivos - ENV generados por el INEC de 1990-2024. (INEC, 2024)

La escasez de datos temporales (cuatro puntos de medición en 32 años) representa una limitación estructural que afecta a la mayoría de los países en Latinoamérica, cuyos sistemas estadísticos están aún en desarrollo, por lo que nuestra propuesta de estudio tendría relevancia no solo para el caso de Ecuador sino para otros países que tengan variables similares de estudio. (Villacís B, 2019; Wilson et al., 2022)

Se analiza la implementación y evaluación comparativa de cinco modelos: i) baseline de referencia (tasa de crecimiento medio anual constante y otros), ii) Temporal Fusion Transformer - TFT como modelo de aprendizaje profundo (Lim et al., 2021), iii) LightGBM como modelo de gradient boosting (Ke et al., 2017), iv) Método demográfico Hamilton-Perry (Hamilton & Perry, 1962) amortiguado, y v) un ensamble híbrido (Grossman et al., 2022). Los resultados se evalúan en tres niveles geográficos (cantonal, provincial y nacional), y se validan externamente contra las estimaciones y proyecciones de población - EPP del INEC para el año 2024, que cubren el periodo 2010-2035. (INEC, 2024)

1.2. Justificación e importancia del trabajo de investigación

La proyección poblacional es una herramienta de gestión para que los gobiernos puedan diseñar, planificar y ejecutar políticas públicas relacionadas con la educación, salud, vivienda, generación de empleo entre otras; sin embargo, la generación de proyecciones poblacionales enfrenta desafíos metodológicos severos cuando existen condiciones de escasez de datos e incertidumbre. Al revisar la literatura científica sobre métodos de pronóstico demográfico, se identificaron convergencias, divergencias, limitaciones y vacíos que justifican la propuesta de este estudio.

Metodológicamente, la viabilidad de esta investigación está dada en los siguientes términos: (a) los datos censales INEC son de acceso público; (b) los registros vitales del INEC (ENV y EDG 1990–2024) están disponibles como microdatos; (c) las proyecciones EPP INEC 2024 están publicadas y se pueden integrar al análisis; (d) se cuenta con la infraestructura computacional para implementar los modelos que se proponen.

Tanto los demógrafos clásicos y científicos de datos plantean la necesidad de transitar de modelos deterministas como las extrapolaciones lineales, exponenciales y logísticas clásicas hacia enfoques probabilísticos. Existe una coincidencia en que los supuestos de crecimiento constante basados en parámetros fijos son insuficientes para capturar la complejidad de la dinámica demográfica y la incertidumbre a mediano y largo plazo. Tanto el análisis

probabilístico de Raftery y Ševčíková (2023) como los enfoques de aprendizaje automático sistematizados por Wang y Lee (2021) coinciden en que la modelación debe ser dinámica y adaptativa para con la finalidad de cuantificar el riesgo y el sesgo de las predicciones.

El Método de Componentes por Cohortes (CCM) se considera el estándar global. Sin embargo, como menciona Vanella et al. (2020), el CCM requiere de una alta cantidad de datos continuos de fecundidad, mortalidad y migración, y depende de la subjetividad del especialista para definir las tasas de transición futuras.

Por otro lado, los algoritmos basados en árboles de decisión (como *LightGBM* o *XGBoost*) descubren relaciones no lineales complejas a partir de los datos históricos. Sin embargo, la literatura relacionada con Inteligencia Artificial ignoran las variables de estructura demográfica tradicional (como las cohortes de edad), presentando a la población como una "caja negra" encaminada exclusivamente a la optimización de métricas de error global.

La relevancia de esta experimentación está dada desde 3 puntos de vista; académico, tecnológico y social como se muestra en la Tabla 2:

Tabla 2
Relevancia e impacto académico, tecnológico y social

Impacto y Relevancia	Base Teórica	Solución Propuesta
Académico	Según Wilson et al., 2022, Grossman et al., 2022, y Baker et al., 2023 identifican explícitamente la escasez de datos como la principal limitación metodológica en países en desarrollo, y señala el machine learning como un área de investigación con alto potencial pero escaso desarrollo empírico para contextos censales escasos.	Contribuir a cerrar la brecha identificada aportando evidencia empírica sobre el comportamiento de los modelos cuando se evalúan en múltiples niveles de agregación geográfica.
Tecnológico		Implementar un pipeline reproducible y completamente documentado que integra microdatos censales SPSS, datos vitales del registro civil, y proyecciones EPP como referencia de validación externa.
Social	Ecuador cuenta con 221 cantones con realidades demográficas muy heterogéneas: La disponibilidad de proyecciones cantonales precisas y metodológicamente sólidas tiene impacto directo en la planificación de infraestructura educativa, sanitaria y de servicios básicos a nivel local.	Este pipeline puede ser adoptado directamente por el INEC o replicado por otros institutos estadísticos de la región para la actualización de proyecciones intermunicipales en contextos de baja frecuencia de datos.

1.3.Alcance

El alcance se delimita en las dimensiones descritas en la Tabla 3:

Tabla 3

Dimensiones de Alcance de Análisis del proyecto

Dimensión	Incluye	No incluye
Datos	Censos INEC 1990, 2001, 2010, 2022. Registros de ENV y EDG anuales 1990–2024. EPP INEC del 2024.	Datos de movilidad, encuestas intercensales, datos de salud desagregados.
Unidades de análisis	221 cantones del Ecuador continental e insular.	Parroquias, barrios, zonas censales.
Modelos	TFT, LightGBM global, LightGBM por tipología, Hamilton-Perry amortiguado, Ensamble híbrido, TCMA baseline.	Modelos bayesianos probabilísticos, modelos de micro simulación.
Niveles de evaluación	Cantonal (221 unidades), provincial (24 unidades), nacional (total Ecuador).	Proyecciones por sexo y grupos de edad.
Horizonte de proyección	Validación retrospectiva 2010–2022. Proyecciones prospectivas 2030 y 2035.	Proyecciones más allá de 2035.
Geografía	Ecuador (221 cantones según División Político-Administrativa - DPA vigente).	Comparaciones con otros países de la región.

1.4. Objetivos

1.4.1. Objetivo general

Evaluar modelos de predicción de crecimiento poblacional a múltiples niveles geográficos para los 221 cantones del Ecuador utilizando datos censales escasos complementados con registros vitales anuales, con la finalidad de la identificación del enfoque de mayor desempeño según la escala territorial de análisis.

1.4.2. Hipótesis

La aplicación de modelos de machine learning y métodos demográficos sobre datos censales escasos permitirá identificar que el nivel de agregación geográfica es el determinante principal de la selección del modelo óptimo para la predicción subnacional de población.

1.4.3. Objetivo específico

- a. Realizar un estudio del estado del arte sobre los métodos tradicionales y con uso de Inteligencia artificial para la predicción de población por medio de una revisión sistemática de la literatura que permita soportar de manera bibliográfica la propuesta de investigación.
- b. Construir un data set que consolide los 221 cantones del Ecuador con microdatos de los cuatro censos INEC (1990–2022) y registros vitales anuales del registro civil.

- c. Evaluar los modelos de predicción poblacional (TFT, LightGBM, Hamilton-Perry, Ensamble híbrido y baselines) mediante un split temporal estricto que reserva el año censal 2022 como conjunto de prueba.
- d. Analizar el comportamiento del error de predicción en tres niveles de agregación geográfica: cantonal (221 unidades), provincial (24 unidades) y nacional (total Ecuador).
- e. Validar las proyecciones 2030–2035 de los modelos propios contra las proyecciones oficiales EPP INEC 2024 con la finalidad de identificar el modelo que menor distorsión tiene respecto a resultados oficiales.
- f. Documentar los pasos efectuados con la finalidad de que el ejercicio sea reproducible y sirva de guía en la selección de modelos por nivel geográfico aplicable a contextos de datos censales escasos.

CAPITULO 2:

2. REVISIÓN DE LITERATURA

2.1. Estado del Arte

Las estimaciones y proyecciones poblacionales constituyen un insumo necesario para los planes de desarrollo económico y social de los países, así como también para monitorear el cumplimiento de los Objetivos de Desarrollo Sostenible ODS – así como la Agenda 2030. Contar con información de la población con granularidad lo más fina posible como: sexo, edad, lugar de residencia, escolaridad, pertenencia a grupos étnicos, participación en la actividad económica, permiten elaborar una proyección con la menor cantidad de sesgo.

La demografía es una disciplina dinámica indispensable en la planificación económica y social. Según un estudio base para realizar proyecciones elaborado por el Centro Latinoamericano de Demografía (CELADE) en 1984, el propósito vital trasciende el acto de censar pues este busca comprender la evolución de la estructura social para que los Gobiernos puedan anticipar necesidades y diseñar políticas públicas eficientes para sus sistemas de salud, educación y empleo. Este estudio plantea tres objetivos primordiales:

- **La actualización de estimaciones:** Revisando periódicamente de las cifras de población incorporando nuevos censos, encuestas y estadísticas vitales.
- **La diseminación de métodos:** Compartiendo y aplicando los procedimientos técnicos más robustos para garantizar que los datos sean comparables internacionalmente.
- **La validación de resultados:** Comparando las proyecciones que realizan sus Departamentos de Estadísticas y Censos, con cálculos realizados por Organismos Internacionales como las Naciones Unidas.

Es importante distinguir la proyección poblacional del pronóstico poblacional, la proyección es un ejercicio de simulación condicionado a supuestos específicos mientras que el pronóstico poblacional tiene por objetivo predecir la evolución correcta de la dinámica demográfica. Sin embargo, esta capacidad de previsión enfrenta el desafío de la incertidumbre inherente que aumenta de forma exponencial conforme el horizonte de tiempo se aleja del presente,

motivando a los investigadores a realizar una transición de modelos deterministas simples a arquitecturas probabilísticas y algorítmicas que logren capturar la complejidad de las variables vitales. (Vanella et al. 2020)

Los inicios demográficos modernos se construyeron utilizando funciones matemáticas deterministas. Estos modelos asumen que el futuro poblacional es una extrapolación lógica de tendencias pasadas, capturando la esencia del crecimiento mediante parámetros fijos que establecieron los marcos iniciales de la cuantificación demográfica. Según la revisión de la literatura publicada por Román Sánchez et al. (2023) y Vanella et al. (2020), los modelos clásicos se agrupan en:

- **Funciones Lineales y Exponenciales:** Usadas bajo la premisa de crecimiento constante. La función lineal asume incrementos absolutos uniformes, mientras que la exponencial proyecta un crecimiento proporcional al volumen poblacional preexistente.
- **Función Logística:** Establece un crecimiento de naturaleza sigmoide. Este modelo es importante debido a la introducción de la capacidad de carga de un territorio. Autores como Keyfitz (1979) destacan el uso de asíntotas (cotas superiores e inferiores) para modelar poblaciones que, tras una fase de aceleración, se estabilizan al aproximarse a límites de saturación territorial. Sin embargo, dichos modelos enfrentan limitaciones estructurales como la rigidez de sus hipótesis y la dependencia de datos censales introducen errores significativos al no considerar cambios bruscos en las tasas vitales.

La evolución metodológica hacia la desagregación permitió transitar de modelos que trataban a la población como una caja negra hacia un análisis de las variables que configuran su dinámica: tasa de fecundidad, tasa de mortalidad y migración.

El Método de Componentes por Cohortes (CCM) se ha consolidado como el estándar de oro en la práctica mundial. Este enfoque simula cada componente por separado antes de integrarlas para generar proyecciones de la estructura por edad y sexo (Vanella et al., 2020).

En México, el Consejo Nacional de Población (CONAPO) ha adoptado formalmente este método bajo la perspectiva propuesta por Rogers (1995), la cual permite integrar supuestos diferenciados para diversas unidades territoriales (García, 2013).

En la Tabla 4 se resumen la clasificación de métodos tradicionales basada en González y Torres (2012):

Tabla 4

Clasificación de métodos tradicionales

Método	Ejemplos	Principales Características	Aplicabilidad
Basados en funciones matemáticas	Lineal, Exponencial, Logística, Incrementos relativos.	Extrapolación de tendencias históricas totales.	Áreas pequeñas o con severa limitación de datos.
Basados en variables sintomáticas	Tasas vitales, Razón censal, Correlación de tasas y razones.	Uso de indicadores indirectos (nacimientos, defunciones) para estimar cambios.	Estimaciones intercensales en niveles subnacionales.

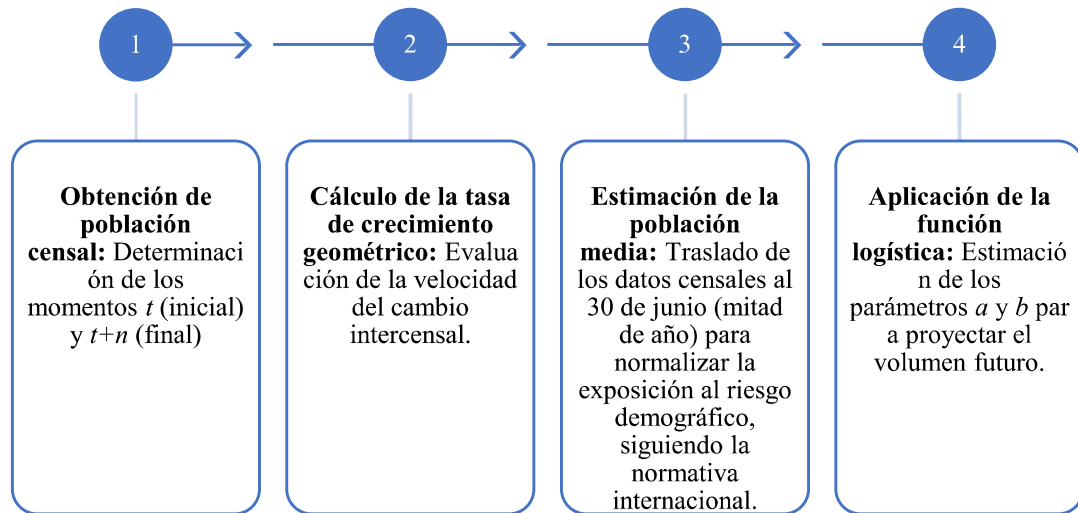
A pesar de la robustez del método de las componentes demográficas - MCD este depende de la subjetividad del especialista al definir las tasas futuras, situación que ha impulsado al desarrollo de métodos estocásticos y de aprendizaje automático para manejar la incertidumbre de forma objetiva.

La función logística es de importancia crítica cuando la ausencia de datos granulares impide la aplicación del MCD. Arriaga (2001) subraya que las asíntotas logísticas permiten establecer límites de crecimiento realistas para proyecciones de largo plazo en áreas locales.

Román Sánchez et al. (2023) sistematizan este enfoque en un procedimiento de cuatro pasos, explicado en el Manual de las Naciones Unidas (1978) para la estandarización de indicadores como se muestra en la Figura 1:

Figura 1

Pasos para la estandarización de indicadores



La integración de la Inteligencia Artificial (IA) representa la respuesta contemporánea a los sesgos presentes en los modelos tradicionales. Mientras los modelos matemáticos clásicos se basan en límites predefinidos, los algoritmos de Machine Learning (ML) descubren el peso real de las variables sin sesgos de preconcepción (Wang & Lee, 2021).

De acuerdo con Wang & Lee (2021), dos arquitecturas destacan en la frontera predictiva:

- **XGBoost Regression:** Algoritmo de boosting que destaca por su precisión técnica, empleando el Similarity Score y el Information Gain para optimizar la construcción de árboles y evitar el sobreajuste. El desempeño es significativamente superior a los modelos tradicionales, registrando un Error Porcentual Absoluto Medio (MAPE) significativamente menor al de Regresión Lineal y al Long Short-Term Memory - LSTM.
- **Long Short-Term Memory - LSTM:** Es una red neuronal recurrente RNN que se plantea como solución moderna al problema de la desaparición del gradiente en series temporales de largo alcance, permitiendo capturar dependencias históricas que los modelos lineales ignoran.

Houssein, E.H., Mohamed, M., Younis, E.M.G. et al. (2025), presento una visión de las perspectivas futuras para las proyecciones poblacionales, como se muestra en la Tabla 5:

Tabla 5

Comparación de los tipos de modelos de proyección

Modelo	Fortaleza	Debilidad
Estadístico	– Simple implementación e interpretación. – Computacionalmente eficiente. – Adecuada para tendencias de corto plazo y lineales.	– Limitada capacidad para captar patrones no lineales o complejos. – Requiere fuertes supuestos de estacionariedad. – Poco eficaz para datos multivariantes a gran escala.
Modelos de aprendizaje automático	– Gestiona eficazmente las relaciones no lineales. – Es flexible ante diferente tipo de datos. – Costo de entrenamiento moderado.	– Requiere de ingeniería de características robusta. – Interpretabilidad limitada sin métodos post hoc. – Degradación del rendimiento con dependencias temporales de alta dimensionalidad.
Modelos de aprendizaje profundo	– Capturar dependencias temporales complejas y no linealidades. – Escalable a grandes conjuntos de datos y series multivariantes. – Rendimiento consistente en la proyección a múltiples horizontes.	– Alto costo computacional y energético, – Tiempos de entrenamiento más largos y con mayor latencia. – Menor interpretabilidad.
Modelos Híbridos	– Combina las fortalezas de diferentes paradigmas. – Mayor robustez y precisión predictiva. – Flexible para adaptarse a requisitos específicos.	– Modelo más complejo. – Altos costos de mantenimiento y entrenamiento. – Puede reducir la interpretabilidad.

2.2. Marco Teórico

Según la página oficial del Fondo de Población de las Naciones Unidas (UNFPA), la población mundial para el año 2025 es de 8.232 millones de personas lo que evidencia un crecimiento exponencial.

Según Adrian E. Raftery, Hana Ševčíková. (2023) la dinámica en el crecimiento de una población es el resultado de la interacción de tres componentes: Fecundidad, Mortalidad y Migración, mismos que actúan de forma combinada por medio de la ecuación compensadora, determinando así la fisonomía y el ritmo de crecimiento de una sociedad, a continuación, en la Tabla 6 se resume su descripción y aplicación.

Tabla 6

Componentes para la dinámica de crecimiento poblacional

Componente	Característica	Aplicación	Fórmula
Fecundidad (Fertility rates)	Mide el nivel y frecuencia de nacimientos (mediante la TFR – Total Fertility Rate).	Suma personas a la población de partida.	$TFR = n \sum_{\text{todos los grupos de edad } [x, x+n]} n^F x$ [1] Donde: N= número de años en cada grupo de edad (intervalos de 1 o 5 años) $n^F x$ = tasa de fecundidad específica por edad para el intervalo. Esta tasa se calcula dividiendo el número esperado de nacimientos de mujeres en ese rango de edad por el número de años-persona vividos por las mujeres en ese mismo rango.
Mortalidad (Mortality rates)	Mide la frecuencia de defunciones según la edad y el sexo (indicador del nivel de salud).	Resta personas a la población base.	$n^S x = \exp(-n x n m_x)$ [2] Donde: n= número de años del intervalo de edad. $n m_x$ = tasa de mortalidad específica para el grupo de edad correspondiente.
Migración	Mide los movimientos de entrada y salida a través de las fronteras territoriales.	Suma o Resta personas según el saldo neto migratorio.	$G_t = I_t - E_t$ [3] Donde: I_t = Número de inmigrantes E_t = Número de emigrantes

Bajo condiciones en las que se cuenta con información demográfica suficiente, la población es el resultado de la conciliación de las distribuciones por sexo y grupos atareos en los censos de población realizados por cada gobierno con las estimaciones de fecundidad, mortalidad y migración de cada periodo intercensal.

Según la CELADE se define la conciliación demográfica como el conjunto de procedimientos demográficos por medio del cual se logra:

1. Evaluar el grado de cobertura de cada uno de los censos de población.
2. Corregir las distribuciones por sexo y edad de los censos, en los que exista falta de cobertura y mala declaración de edad.
3. Buscar coherencia de los censos corregidos con las estimaciones de mortalidad, fecundidad y migración.
4. Establecer una población base para la proyección de la población, de preferencia la que se acepte como la de mejor calidad.

La vinculación de los resultados entre dos momentos en el tiempo (usualmente dos censos), los demógrafos utilizan ecuación compensadora. Esta verifica que el cambio en el volumen de población sea coherente con los hechos vitales ocurridos en el intervalo y está dada por la fórmula [4] (CELADE,1984):

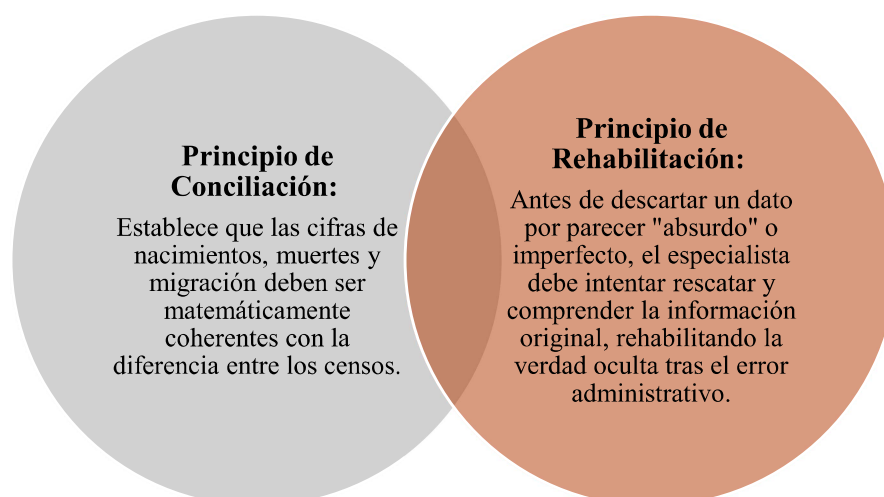
$$N_t = N_0 + B - D + I - E \quad [4]$$

Donde:

La aplicación de esta fórmula debe cumplir con los principios ético-técnicos de conciliación y rehabilitación que se explican en la Figura 2:

Figura 2:

Principios de la verdad demográfica



Dentro de los retos al realizar pronósticos está el manejo de la incertidumbre, misma que aumenta conforme el horizonte del pronóstico se aleja.

En el estudio realizado por Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2022), se analizan varios métodos para la proyección de poblaciones en áreas pequeñas, sin embargo, se destaca que no existe una definición única de área pequeña, aunque normalmente se hace referencia a poblaciones de menos de 100.000 habitantes. A continuación, se describen los métodos usualmente utilizados para la proyección poblacional:

a. Métodos comparativos y de extrapolación

Se aplica usualmente para proyectar poblaciones totales sin considerar segregación por edad y sexo.

i. Métodos de extrapolación

Se basan exclusivamente en la observación de tendencias del pasado de la población del área de estudio para proyectarlas hacia el futuro. Consideradas como modelos atóricos debido a que no asumen causas diferentes a las propias del cambio poblacional como modificaciones en la economía o migración, entre otros. Dentro de esta clasificación se puede mencionar:

- **Lineal y Exponencial:** Método común para evitar resultados inadmisibles. Se aplica la extrapolación lineal si el crecimiento pasado fue positivo y una extrapolación exponencial si este crecimiento fue negativo.
- **Curva Logística:** El uso del modelo Pearl-Reed permite modelar el crecimiento de la vivienda, pues este modelo imita la saturación natural de las áreas urbanas a medida que se pueblan nuevos terrenos.
- **Series de Tiempo:** Es uno de los más complejos, necesita de gran cantidad de datos históricos y tiempo de implementación; sin embargo, ofrece una precisión comparable a los modelos de componentes de cohorte. A pesar del crecimiento de enfoques basados en inteligencia artificial, el modelo Autorregresivo Integrado de Media Móvil (ARIMA) ofrece una ventaja importante, en especial en contextos donde la interpretabilidad y el bajo requerimiento computacional son factores determinantes.

ii. Métodos comparativos

Estos métodos crean la proyección de un área pequeña basándose en su relación con una proyección independiente de un área geográfica mayor como un estado o un país. Dentro de estos grupos se puede mencionar:

- **Participación Constante (Constant Share):** Postula que el área de proyección mantendrá siempre la misma proporción de la población total del

área mayor; este método ha demostrado ser uno de los más precisos en diversas investigaciones.

- **Participación del Crecimiento (Share-of-growth):** Proyecta la población basándose en la proporción que el área de estudio aporta al crecimiento total de la región. Este modelo presenta la variante conocida como el modelo de participación de crecimiento modificada, en la que se ajustan a cero las participaciones negativas para evitar desniveles imaginarios.
- **Desplazamiento de Participación (Shift-share):** Aplica una participación proyectada que puede cambiar con el tiempo; sin embargo, a largo plazo puede generar problemas como poblaciones negativas.

Entre los principales beneficios de estos métodos son: requerimiento mínimo de datos, simplicidad, precisión. Sin embargo, también presenta las siguientes limitaciones: falta de detalle, inestabilidad a largo plazo, sensibilidad al ruido.

iii. Método de Componentes de Cohorte (CCR - Cohort Change Ratio)

Calcula la población futura basándose en los componentes que impulsan el cambio poblacional explicados en párrafos anteriores. Sin embargo, a diferencia de los métodos de extrapolación simples que utilizan solo el total de habitantes, este desglosa la población por cohortes de edad y sexo, y proyecta cada uno de los componentes de manera individual. Una de las limitaciones más relevantes en este modelo es su dificultad para aplicarlo en poblaciones menores a 100.000 habitantes, debido a que los datos demográficos suelen presentar baja calidad, presentar ruido estadístico o tienen valores nulos.

iv. Método Hamilton-Perry (Componentes de Cohorte Simplificado)

Es una versión simplificada del modelo de componentes de cohorte que ha tomado popularidad recientemente.

A diferencia del modelo de componentes tradicional, Hamilton-Perry proyecta la población por cohortes - grupos de edad, pero sin desglosar los componentes demográficos individuales como: nacimientos, muertes y migración.

Utiliza la relación de Cambio de Cohorte (CCR), donde la población de una cohorte al inicio de un intervalo se multiplica por esta relación para obtener la población de esa misma cohorte ya avejentada al final del intervalo. Para proyectar a la población infantil que no existía al inicio del intervalo utiliza la relación niño-mujer.

Al igual que otros modelos presentan limitaciones, como la falta de desglose y sin un adecuado ajuste produce resultados demasiado altos o bajos en áreas con cambios volátiles.

Se han desarrollado mejoras al modelo mediante varias técnicas, así se puede citar:

- ponderación espacial: que promedia los pronósticos de un área con los vecinos.
- restricción (constraining): es más preciso cuando los resultados por edad se ajustan para sumarlos al total de la proyección mediante métodos de extrapolación.
- enfoque compuesto: algunos autores plantean el uso de relaciones de cambio para poblaciones en declive y diferencias absolutas de cohorte para poblaciones en crecimiento, lo que evita crecimientos exponenciales irreales.
- suavizado: combina las relaciones locales con las de una región más grande, como una prefectura o cantón, ayudando a reducir el ruido en los datos de áreas más pequeñas.

v. Método de la Unidad de Vivienda (Housing-unit Method)

Establece un vínculo directo entre el tamaño de la población y el número de viviendas disponibles en el área geográfica de estudio, tiene por premisa que la población total puede derivarse de la cantidad de unidades habitacionales.

Tiene por beneficio que no requiere de alto procesamiento computacional, es útil siempre y cuando se conozca con una alta certeza el número de viviendas que se construirán en los próximo 5 a 10 años.

Su limitación radica principalmente en que genera solo totales poblacionales; es decir no proporciona detalles por edad o sexo. Así como también tiene una alta dependencia de la

confiabilidad de los supuestos como el tamaño de los hogares, tasas de ocupación, información que se ha levantado recién en los últimos censos poblacionales.

vi. Promediado y Combinación de Modelos

Este enfoque tiene por premisa que combinar los pronósticos reduce los errores de predicción, pues permite aprovechar la base de información generada y ayudar a que los errores en direcciones opuestas de los modelos individuales se compensen entre sí. Algunas combinaciones pueden ser:

- Media Simple: Calcula el promedio de los resultados de todos los modelos individuales evaluados.
- Media Recortada (Trimmed Mean): Excluye los resultados más altos y bajos de los modelos individuales antes de calcular el promedio, evitando proyecciones extremas que distorsionen el resultado final.
- Modelos Compuestos: Aplican diferentes métodos a áreas específicas según las características de tamaño poblacional o tasa de crecimiento.
- Combinación con Restricciones (Constraining): se ajusta para que su total poblacional coincida con el resultado obtenido de un promedio de modelos de extrapolación, lo que aumenta de manera importante la precisión de los datos por edad y sexo.

La principal ventaja de este modelo radica en no depender de un solo algoritmo, reduce la probabilidad de una tendencia errática en un modelo específico afecte la proyección final.

vii. Métodos de "Downscaling" y Desagregación

Toma proyecciones poblacionales realizadas a gran escala (nacionales o regionales) y las distribuye en unidades geográficas mucho más pequeñas, como áreas locales o celdas de cuadrícula (grid cells), mientras otros métodos se enfocan en la planificación local aplicada, este tiene una visión global y utiliza supuestos que simplifican considerablemente la realidad para manejar grandes volúmenes de datos. Existen varias técnicas para la aplicación de este modelo así tenemos:

- modelos de gravedad: distribuyen las proyecciones estatales en celdas de cuadrícula, especificando el comportamiento de forma separada para áreas urbanas y rurales.
- coeficiente de desarrollo potencial: asigna el crecimiento poblacional a las celdas

basándose en variables como el uso del suelo, la proximidad a carreteras, la accesibilidad urbana y la población actual.

- análisis de "Shift-Share" adaptado: desglosa los pronósticos regionales de edad y sexo mediante una función que combina los efectos de crecimiento total regional, cambio regional por grupo de edad y crecimiento específico local por edad.
- métodos dasimétricos: distribuye el crecimiento regional basándose en categorías de densidad de vivienda.
- modelos de uso de Suelo: aplican el Waikato Integrated Scenario Explorer (WISE) o modelos de desarrollo urbano para distribuir la población según la disponibilidad de tierra residencial y la accesibilidad a servicios básicos.

Las principales limitaciones de este modelo radican en que no han sido validados de manera exhaustiva y que para su aplicación se requieren datos espaciales como carreteras, límites de uso de suelo, que no siempre están levantados o disponibles en cada país.

Por otro lado, Hajirahimova, M. S., & Aliyeva, A. S. (2023), plantea el uso de cuatro algoritmos de aprendizaje automático para la predicción de la población, así podemos mencionar:

- **Regresión Lineal (LR):** Encuentra la relación lineal entre variables independientes y dependientes.
- **K-vecinos más cercanos (KNN):** método de aprendizaje no paramétrico basado en muestras que se aplica con éxito al pronóstico de series temporales; identifica puntos de datos con valores o características similares que se encuentran cerca de un punto específico en la base de datos. Para medir la distancia relativa entre puntos, se utilizan métricas como la distancia Euclidiana, Minkowski, Manhattan o Chebychev.
- **Árbol de Decisión (DT):** algoritmo no paramétrico compuesto por un nodo raíz, ramas, nodos internos y nodos hoja. Para seleccionar el mejor atributo en cada nodo durante la construcción del modelo, se utilizan métricas populares como el Índice de Gini y la Ganancia de Información.
- **Bosque Aleatorio (RF):** enfoque de aprendizaje por conjuntos (ensemble learning) donde se genera una gran cantidad de árboles de decisión. Cada árbol individual se construye utilizando únicamente un subconjunto de atributos de entrada seleccionados

de forma aleatoria, en lugar de usar todos los atributos disponibles. Para obtener el resultado final se combinan los resultados de los múltiples árboles de decisión generados, lo que mejorará significativamente la precisión de la predicción.

Espinosa-Zúñiga, Javier Jesús. (2020), resume la evolución de los algoritmos basados en árboles de decisión como se muestra en la Figura 3:

Figura 3

Evolución de los algoritmos basados en árboles de decisión



Gradient Boosting es la base de XGBoost y LightGBM. Este funciona mediante un ensamblado secuencial de árboles de decisión, donde cada nuevo árbol se añade para aprender de los resultados de los árboles anteriores y corregir el error producido mediante la técnica de gradiente descendente hasta que el error no pueda reducirse más.

Para el cálculo del Gradient Boosting se aplica la fórmula [5]:

$$F_m(x) = F_{m-1}(x) + v * h_m(x) \quad [5]$$

Donde:

$F_m(x)$ = modelo resultante de la iteración actual m.

$F_{m-1}(x)$ = modelo acumulado de la iteración anterior.

$h_m(x)$ = nuevo árbol de decisión añadido para corregir el residuo actual.

v = tasa de aprendizaje - learning rate. Valor entre 0 y 1 que controla la contribución de cada árbol para evitar el sobreajuste.

Por otro lado, el XGBoost (Extreme Gradient Boosting), una implementación avanzada, y eficiente del gradiente boosting, lo que la hace más escalable en conjuntos de datos tabulares. Se diferencia de otros modelos donde los árboles crecen hasta su máxima extensión pues aquí el usuario define la profundidad aplicando técnicas de regularización para evitar el sobreajuste, procesamiento en paralelo y manejo automático de valores perdidos.

- Finalmente, tenemos el LightGBM (Light Gradient Boosting Machine) reconocido por su velocidad y precisión, lo que lo hace que sea la opción más popular para conjuntos de datos complejos y a gran escala. Este introduce dos técnicas para optimizar el proceso de entrenamiento:
 - i) GOSS (Gradient-based One-Side Sampling), permite acelerar el entrenamiento al seleccionar solo una fracción de los puntos de datos en cada iteración, priorizando los gradientes más grandes, reduciendo de manera significativa el costo computacional sin perder precisión.
 - ii) EFB (Exclusive Feature Bundling): permite reducir el uso de memoria al combinar características en paquetes, reduciendo de esta manera el número total de variables procesadas sin perder información.

viii. **Modelo Temporal Fusion Transformer (TFT)**

Según Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, Tomas Pfister (2021), el TFT es una arquitectura de red neuronal profunda utilizada para pronosticar series temporales de horizontes múltiples combinando un alto rendimiento con una interpretabilidad inherente.

Funciona integrando numerosos tipos de entradas: covariables estáticas, entradas futuras conocidas y series temporales observadas en el pasado, mediante componentes especializados que se describen a continuación:

- **Mecanismos de Compuerta (Gating):** utiliza el gated Residual Networks - GRN dado por la fórmula [6], para excluir componentes innecesarios de la arquitectura, permitiendo una profundidad adaptativa según la complejidad del conjunto de datos.

$$GRN_{\omega}(a, c) = \text{Lawyer Norm} (a + GLU_{\omega}(n_1)) \quad [6]$$

Donde: n_1 y n_2 son capas intermedias y GLU es la Unidad Lineal de Compuerta

$$GLU_{\omega}(\gamma) = \sigma(W_{4,\omega}\gamma + b_{4,\omega}) \odot (W_{5,\omega}\gamma + b_{5,\omega}) \quad [7]$$

Donde: $\sigma(\cdot)$ es la función sigmoide y es $\odot$ el producto de Hadamard elemento a elemento.

- **Redes de Selección de Variables:** selecciona las características más relevantes en cada paso de tiempo, eliminando entradas ruidosas que pueden perjudicar el rendimiento del modelo.
- **Codificadores de Covariables Estáticas:** integra metadatos estáticos que permite condicionar las dinámicas temporales mediante vectores de contexto.
 - **Procesamiento Temporal:** mediante: i) capa secuencia a secuencia (Seq2Seq), donde aprovecha el LSTMs para el procesamiento local de las entradas, capturando relaciones de corto plazo, ii) atención de múltiples cabezales interpretable, en este paso una capa de autoatención temporal aprende las dependencias, permitiendo identificar pasos de tiempo significativos en el pasado, la fórmula de cálculo está dada por la siguiente fórmula:

$$InterpretableMultiHead(Q, K, V) = (\tilde{H}W_H) \quad [8]$$

Donde: H se calcula como el promedio de los pesos de atención de cada cabezal aplicado a valores compartidos

$$\tilde{H} = \left\{ \frac{1}{m_H} \sum_{h=1}^{m_H} A(QW_Q^{(h)}, KW_K^{(h)}) \right\} VW_V \quad [9]$$

- **Pronóstico por Cuantiles:** en lugar de una predicción única se generan intervalos de confianza en cuantiles para representar la incertidumbre del pronóstico, utilizando la fórmula [10]:

$$QL(y, \hat{y}, q) = (y - \hat{y})_+ + (1 - q)(\hat{y} - y)_+ \quad [10]$$

Donde: $(x)_+ = \max(0, x)$ y $\hat{y}$ es la predicción del cuantil q

b. Métricas utilizadas para evaluar la precisión de los modelos de proyección.

Hajirahimova, M. S., & Aliyeva, A. S. (2023), propone 4 métricas para evaluar la precisión de los modelos de proyección poblacional, como se muestran en la Tabla 7:

Tabla 7

Métricas de evaluación para modelos de proyección

Métrica	Descripción	Rango	Fórmula	
Error Absoluto Medio - MAE	promedio de la diferencia absoluta entre los valores reales y los valores predichos por el modelo de regresión.	Valores oscilan entre 0 e infinito, y un valor más bajo indica que el modelo predice mejor.	$MAE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)$	[11]
Error Porcentual Absoluto Medio - MAPE	promedio de todos los errores porcentuales absolutos entre los valores predichos y los reales	Un valor menor representa mayor precisión, es aconsejable no usarlo si los valores reales están cerca de cero	$MAPE = \frac{1}{N} \sum_{i=1}^N \left(\frac{ y_i - \hat{y}_i }{y_i} \right)$	[12]
Error Cuadrático Medio - MSE	promedio de los cuadrados de las diferencias entre los valores reales y los predichos	Un valor menor indica un alto rendimiento del modelo	$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$	[13]
Raíz del Error Cuadrático Medio - RMSE	Raíz cuadrada del promedio de las diferencias cuadráticas (o la media cuadrática de estas diferencias) entre los valores predichos y observados. Comparar errores de pronóstico de diferentes modelos sobre un mismo conjunto de datos	Un valor inferior constante indica un mejor desempeño	$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$	[14]
Coefficiente de determinación - R-Square	Determina como se ajusta el modelo a un conjunto de datos, representando la proporción de la varianza de la variable dependiente que es predecible a partir de las independientes	Un valor más cercano a 1 (o al 100%) se considera un mejor rendimiento	$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$	[15]

En el caso de Ecuador los censos de población y vivienda se realizan cada diez años situación que permite entre otras cosas la actualización de las estimaciones-proyecciones de población (EPP) llevadas a cabo por el Instituto Nacional de Estadística y Censos – INEC. Instituto Nacional de Estadística y Censos (2024)

2.3. Terminología

En el desarrollo práctico del modelo de proyección utilizado existen algunos términos que se deben tener en cuenta, así podemos mencionar:

- a. **Projection baseline:** punto de partida temporal o año base a partir del cual se calcula y extiende una proyección demográfica hacia el futuro. (Marquez et al., 2024).
- b. **Dataset (Conjunto de datos):** conjunto de datos estructurado de forma secuencial en el tiempo (series temporales), compuesto por variables específicas (como nombre del país/ciudad, año y volumen de población) recopiladas a lo largo de varias décadas para identificar patrones históricos y entrenar modelos. (Houssein et al., 2025)
- c. **Benchmark (Evaluación comparativa):** método de comparación utilizado de para contrastar, validar y demostrar si nuevos modelos de aprendizaje automático realmente aportan una mejora en la precisión de las predicciones. (Wilson et al., 2022)
- d. **Split temporal (División temporal):** procedimiento aplicado en series de tiempo que consiste en dividir el conjunto de datos de manera meticulosa en conjuntos de entrenamiento y de prueba manteniendo la integridad de su orden cronológico, con la finalidad de preservar las dependencias temporales intrínsecas del conjunto de datos antes de entrenar y evaluar el modelo. (Tanmoy, et al., 2025).
- e. **Fuga de datos (data leakage):** Sasse, L., Nicolaisen-Sobesky, E., Dukart, J. et al. (2025), define la fuga de datos cuando “...se produce contaminación —es decir, una distribución inadecuada de las muestras entre los conjuntos de entrenamiento y de prueba—, ciertas características pueden replicar involuntariamente el comportamiento de la variable objetivo...”

M. Gupta, J. Christopher and K. Ramisetty, (2025), resumen los tipos de fuga de datos y sus estrategias de mitigación, como se muestran en la Tabla 8:

Tabla 8

Tipos de Data Leakage y estrategias de Mitigación

No.	Fase del ML	Tipo de Data Leakage	Mitigación
1	Recolección de Datos Crudos	Fuga de datos interna (Inner data leakage)	Colaboración entre expertos en el dominio y analistas de datos

2	Preprocesamiento Conjunto de Datos	del	Fuga por división tardía (Late-split leakage)	Realizar la partición de datos antes de llevar a cabo el preprocesamiento
3	Partición de Datos		Contaminación entrenamiento-prueba (Train-test contamination)	Descarte de solapamiento (Overlap Dropping)
4	Ingeniería Características	de	Fuga de características (Feature leakage)	Validar las razones detrás de la elección de diferentes enfoques de ingeniería de características
5	Muestreo de Datos		Fuga de datos sintéticos (Synthetic data leakage)	Descarte de solapamiento (Overlap Dropping)
6	Entrenamiento del Modelo		Fuga de objetivo (Target leakage)	Descarte de características (Feature Dropping)
7	Validación del Modelo		Fuga por división incorrecta (Wrong-split leakage)	Descarte de solapamiento (Overlap Dropping)
8	Prueba del Modelo		Interpretabilidad (para confirmar conocimiento correcto)	Métodos explicables post-hoc
9	Analítica de Rendimiento		Consistencia (para confirmar rendimiento consistente)	Experimentación con conjuntos de datos públicos
10	Mantenimiento del Modelo		Fuga de concepto (Concept leakage)	Actualización del conocimiento de manera incremental

f. Clustering (Agrupamiento): método donde los algoritmos de clustering agrupan puntos de datos similares con el fin de simplificar conjuntos de datos complejos o identificar patrones espaciales (como zonas de alta delincuencia o puntos críticos de congestión de tráfico en las ciudades). (Cina et al., 2025).

g. Ensemble (Métodos de ensamble): enfoque que destaca por su capacidad para combinar las fortalezas de múltiples modelos individuales con el objetivo de mejorar la precisión global de las predicciones. Se fundamentan en que la combinación y el promedio de modelos reducen el error de predicción pues extraen una mayor cantidad de información y permiten que las desviaciones individuales en diferentes direcciones se compensen entre sí. (Hajirahimova & Aliyeva, 2023; Houssein et al., 2025; Wilson et al., 2022)

h. Estimación Poblacional: Según Adrian E. Raftery, Hana Ševčíková. (2023) se refiere generalmente a cálculos realizados sobre la población de años pasados o periodos

actuales, es decir, basan en datos históricos (como censos y registros vitales) aunque existe incertidumbre sobre si los registros del pasado son de calidad. Por ejemplo, las Naciones Unidas publica sus estimaciones oficiales de la población mundial desde el año 1950.

i. Proyección Poblacional: Según Adrian E. Raftery, Hana Ševčíková. (2023) son cálculos sobre el futuro, que se realizan bajo supuestos específicos. En demografía, una proyección se convierte en un pronóstico o forecast cuando los supuestos están diseñados para ser lo más realistas posibles. Byron Villacis, BID (2019) lo define de manera más simple como estimaciones de escenarios hipotéticos de la evolución demográfica de un país o una región determinada.

j. Tasa de Crecimiento Medio Anual (TCMA): indicador estadístico utilizado por los Organismos de Control de cada país, el INEC para el caso de Ecuador, que permite medir el ritmo al que aumenta o disminuye la población de un país.

k. Incertidumbre: es la falta de conocimiento exacto sobre los componentes (fecundidad, mortalidad y migración) que determinan el cambio poblacional en el pasado como para el futuro. Adrian E. Raftery, Hana Ševčíková. (2023)

l. Tasa de migración neta: Según las Naciones Unidas la tasa de migración neta de un país en un periodo de cinco años como la migración neta dividida por la población promedio del mismo periodo.

m. SHAP (SHapley Additive exPlanations): Molnar Christoph (2025), lo define como “...un método para explicar predicciones individuales. SHAP se basa en los valores de Shapley óptimos en teoría del juego, y tiene por objetivo es explicar la predicción de una instancia x calculando la contribución de cada característica a la predicción...”.

n. Shapley: Según Molnar Christoph (2025) se define “...mediante una función valor **val** de jugadores en **S**. El valor de Shapley de un valor de característica es su contribución al pago, ponderado y sumado sobre todas las posibles combinaciones de valores de características...”. Este valor es el único método de atribución que satisface las

propiedades Eficiencia, Simetría, Dummy y Aditividad.

o. Decision Rules (Reglas de Decisión): Son funciones de decisión que tienen como objetivo proporcionar una regla para seleccionar una decisión particular dentro de un espacio de decisiones posibles, basándose en un punto muestral observado. Básicamente, asocian cada resultado experimental con una acción o decisión específica. Wald, A. (1950)

p. Función de pérdida y riesgo:

- **Pérdida (Loss):** función no negativa de la distribución verdadera F y la decisión tomada d , representa el perjuicio sufrido al tomar la decisión d cuando F es la distribución real. Wald, A. (1950).
- **Riesgo (Risk),** función que representa el valor esperado de la pérdida más el costo esperado de la experimentación cuando se utiliza una regla de decisión determinada. Wald, A. (1950).

q. Minimax Solutions (Soluciones Minimax): es una regla de decisión η_0 que minimiza el riesgo máximo posible. Matemáticamente, se define como aquella regla donde el riesgo máximo sobre todas las posibles distribuciones no es mayor que el riesgo máximo de cualquier otra regla de decisión. Wald, A. (1950).

CAPITULO 3:

3. DESARROLLO

3.1.Desarrollo del Trabajo

El desarrollo del presente trabajo se organiza en torno a la construcción de un conjunto de datos poblacional, a nivel de cantones, del Ecuador. Se describe el conjunto de datos consolidado a partir de fuentes como el Instituto Nacional de Estadística y Censos (INEC), cuatro censos de población, considerando proyecciones poblacionales. Además, se documenta el proceso de preparación y procesamiento que transforma los datos crudos en un dataset analítico apto para ser modelado, haciendo énfasis en la partición temporal de los datos. Este desarrollo ejemplifica el objetivo específico de este proyecto: consolidar los 221 cantones del Ecuador, así como documentar el paso a paso con motivo de presentar un trabajo reproducible.

a. Diseño metodológico

La investigación adopta un enfoque cuantitativo, de diseño no experimental y longitudinal retrospectivo, con alcance descriptivo, comparativo y predictivo. No se manipulan variables independientes: se observan los valores efectivamente registrados por el sistema estadístico nacional en cuatro cortes censales y se contrasta la capacidad de distintas familias de modelos para reproducir un corte reservado de antemano. El carácter comparativo proviene de que el objeto de estudio no es un modelo en particular, sino el desempeño relativo de cinco familias de modelos evaluadas bajo un protocolo común; el carácter predictivo, de que el criterio de comparación es el error de proyección fuera de muestra y no la bondad de ajuste dentro de la muestra.

La unidad de análisis es el cantón, identificado mediante el código de la División Político-Administrativa (DPA) vigente en 2022 y homologado entre censos para asegurar la trazabilidad longitudinal. La población de estudio corresponde a la totalidad de los 221 cantones del Ecuador continental e insular. Al tratarse de datos censales agregados, el estudio opera sobre el universo y no aplica muestreo probabilístico, por lo que no procede el cálculo de error muestral; la incertidumbre relevante no es de muestreo sino de proyección. El panel resultante tiene una estructura de 884 registros (221 cantones $\times$ 4

censos), con la salvedad de que en 1990 únicamente 173 cantones disponen de dato por haberse creado 48 cantones con posterioridad a ese censo.

El horizonte temporal comprende dos tramos. El tramo de validación retrospectiva abarca 1990-2022 y se emplea para medir el error de cada familia de modelos contra un dato observado. El tramo prospectivo corresponde a los años 2030 y 2035, cuyos resultados no admiten verificación empírica a la fecha y se presentan, por tanto, como contraste entre estimaciones frente a las Estimaciones y Proyecciones de Población (EPP) del INEC 2024, y no como una validación en sentido estricto.

El procedimiento analítico se desarrolla en cinco fases encadenadas:

- a. construcción del conjunto de datos, mediante agregación cantonal de los microdatos de los cuatro censos e integración de los registros de nacidos vivos (ENV) y defunciones generales (EDG);
- b. unificación e imputación, que consolida la serie longitudinal y trata de forma explícita los tres tipos de ausencia identificados;
- c. ingeniería de características y normalización, con estimación de los parámetros de escala exclusivamente sobre el subconjunto de entrenamiento;
- d. partición temporal estricta, que asigna los censos de 1990 y 2001 al entrenamiento, el de 2010 a validación y reserva el de 2022 como conjunto de prueba nunca observado por los modelos; y
- e. evaluación comparativa de las cinco familias de modelos en tres niveles de agregación territorial.

El criterio de contrastación se define antes de ejecutar los modelos. La métrica principal es el Error Porcentual Absoluto Medio (MAPE) calculado a nivel cantonal, y es la única que se reporta de forma sistemática para todos los modelos, acompañada del error relativo a escala nacional. La raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2) cumplen un papel auxiliar: se calculan en la selección de los modelos de referencia y en experimentos puntuales, donde ayudan a interpretar aspectos del error que el MAPE por sí solo no revela, como la penalización de los cantones más poblados. No se recalculan para el conjunto completo de modelos ni en

todos los niveles de agregación, porque el ordenamiento comparativo se resuelve con el MAPE.

Dado que la proposición sometida a contraste se refiere al papel del nivel de agregación geográfica, el MAPE se vuelve a calcular sobre las 221 unidades cantonales, las 24 provinciales y el total nacional, de modo que la comparación entre modelos siempre ocurra dentro de un mismo nivel. Los resultados obtenidos frente al censo observado de 2022 y los obtenidos frente a las EPP INEC 2024 se reportan por separado y no se promedian entre sí, porque entre ambas referencias existe una brecha sistemática que proviene de la fuente contra la que se compara y no del desempeño de los modelos.

Finalmente, la reproducibilidad se incorpora como requisito de diseño y no como resultado accesorio. Cada fase del procedimiento se materializa en un script versionado y ejecutable en secuencia, cuyos productos intermedios quedan persistidos en disco, de manera que cualquier resultado del Capítulo 4 pueda rastrearse hasta la transformación que lo originó. La correspondencia entre cada fase metodológica y su artefacto computacional se consolida en una tabla de trazabilidad al cierre de la sección 3.3, con el propósito de que la comprensión del método no dependa de la lectura del código fuente. La Tabla 9 sintetiza los elementos del diseño metodológico descritos en este apartado.

Tabla 9

Síntesis del diseño metodológico

Elemento	Definición adoptada
Enfoque	Cuantitativo
Diseño	No experimental, longitudinal retrospectivo
Alcance	Descriptivo, comparativo y predictivo
Unidad de análisis	Cantón (código DPA 2022 homologado entre censos)
Población de estudio	221 cantones del Ecuador; universo, sin muestreo probabilístico
Estructura del panel	884 registros (221 cantones $\times$ 4 censos); 173 cantones con dato en 1990
Fuentes	Censos INEC 1990, 2001, 2010 y 2022; registros ENV y EDG 1990-2024; EPP INEC 2024 como referencia de contraste
Partición de los datos	Temporal estricta: entrenamiento 1990-2001, validación 2010, prueba 2022
Métrica principal	MAPE cantonal, reportado para todos los modelos en los tres niveles. RMSE, MAE y R^2 se calculan de

	forma puntual en la selección de los modelos de referencia y en experimentos específicos.
Niveles de evaluación	Cantonal (221 unidades), provincial (24 unidades) y nacional (1 unidad)
Horizonte prospectivo	2030 y 2035, presentados como contraste frente a las EPP INEC 2024

Nota: Elaboración propia.

3.2.Descripción del Data Set

El conjunto de datos de este estudio integra cuatro fuentes de información oficiales del Instituto Nacional de Estadística y Censos (INEC) del Ecuador, consolidadas a nivel cantonal para el período 1990-2022. La presente sección describe las fuentes que lo componen, la unidad de análisis, las variables disponibles y los principales hallazgos del análisis exploratorio de datos (EDA) realizado durante su construcción.

a. Fuentes de datos

El conjunto se construye a partir de las siguientes cuatro fuentes:

- i) **Censos de Población y Vivienda (CPV):** Microdatos individuales de los cuatro censos nacionales disponibles: 1990, 2001, 2010 y 2022. El censo de 1990 se distribuye en formato SPSS y los censos de 2001, 2010 y 2022 en formato CSV. A partir de estos microdatos se derivan, por agregación cantonal, las variables demográficas del estudio.
- ii) **Estadísticas Vitales:** Registros de nacidos vivos (ENV) y de defunciones generales (EDG) a nivel cantonal, provenientes del registro civil del INEC para el período 1990-2024. Estas fuentes aportan quince variables vitales que se integran en la fase anual del pipeline.
- iii) **Estimaciones y Proyecciones de Población (EPP) INEC 2024:** Proyecciones poblacionales cantonales oficiales del INEC, accesibles mediante el aplicativo web de proyecciones (<https://cubos.inec.gob.ec/appproyecciones/>). Esta fuente no interviene en el entrenamiento de los modelos; se emplea como referencia de contraste frente a la metodología demográfica oficial del país.
- iv) **Unidad de análisis y cobertura:** La unidad de análisis es el cantón,

identificado mediante el código de la División Político-Administrativa (DPA) de 2022, homologado a cuatro dígitos para asegurar la trazabilidad entre censos. El conjunto consolida los 221 cantones del Ecuador a lo largo de cuatro puntos censales (1990, 2001, 2010 y 2022), conformando una serie temporal en formato largo de hasta 884 observaciones.

Una característica determinante del conjunto es la escasez de datos temporales pues se dispone únicamente de cuatro mediciones en 32 años. Adicionalmente, 48 cantones fueron creados con posterioridad a 1990, por lo que carecen de registro en ese año y presentan valores faltantes (NaN) en la primera observación de su serie. Esta escasez y estos vacíos condicionan tanto la estrategia de imputación (véase la sección Preparación y Procesamiento de Datos) como la elección metodológica de los modelos (véase la sección Metodología de Análisis o Modelo).

b. Variables del conjunto

Las variables se organizan en tres bloques según su disponibilidad temporal, como se resume en la Tabla 10.

Tabla 10

Variables del conjunto de datos por bloque de disponibilidad

Bloque	Variables (nombre técnico)	Disponibilidad
Demográficas comparables	total_personas, total_mujeres, pct_mujeres, pct_urbano, edad_promedio, pct_analf, pct_nbi	Los cuatro censos
Socioeconómicas de hogar y vivienda	pct_hacinamiento, pct_nuclear, pct_unipersonal, pct_digna, pct_luz, pct_agua_pub, pct_bas_carro, idx_servicios, idx_precariedad, ratio_pob_viv, ratio_pob_hog	Solo 2022 (covariables estáticas)
Objetivo y derivadas	total_personas (nivel), tcma (tasa), log_poblacion	Calculadas

Dos variables (escolaridad promedio y porcentaje de discapacidad) se calcularon para 2022 pero se excluyeron de la serie temporal por cambios metodológicos entre censos que las hacen no comparables. La variable de necesidades básicas insatisfechas (pct_nbi) no se

registra en el censo de 2001 y se estima por interpolación lineal, según se detalla en la sección siguiente.

La variable objetivo del estudio es la población cantonal ($total_personas$), modelada tanto en su nivel, expresado en escala logarítmica, como a través de su tasa de crecimiento media anual (TCMA), calculada de forma intercensal como

$$\left(\left(\frac{P_t}{P_0} \right)^{\frac{1}{n}} - 1 \right) \cdot 100, [16]$$

Donde n es el número de años del intervalo.

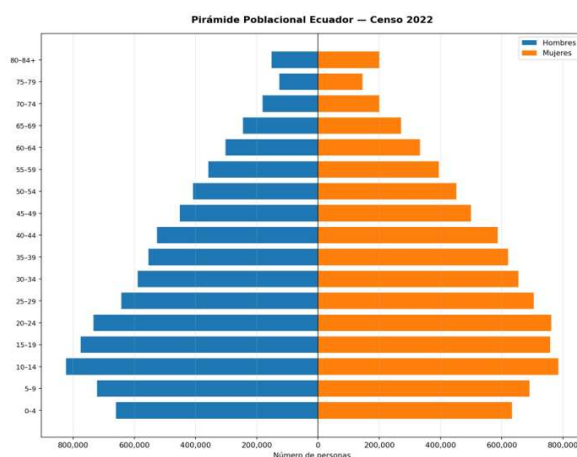
c. Análisis exploratorio de datos (EDA)

El proceso de ingesta de cada censo incorpora un análisis exploratorio que caracteriza la estructura poblacional y valida la calidad de la agregación cantonal. Sus hallazgos principales son los siguientes:

- **Volumen y evolución nacional:** la población del Ecuador registrada por los censos crece de forma sostenida en el período de estudio, pasando de 9.697.979 habitantes en 1990 a 16.938.986 en 2022 (INEC, 2023). Su evolución se aprecia en la Figura 6 (panel superior izquierdo), y los totales por censo se detallan en la Tabla 11.
- **Estructura por edad y sexo:** según el Censo 2022, el 51,3 % de la población son mujeres y el 48,7 % hombres, con una edad mediana nacional de 29 años (INEC, 2023). La pirámide poblacional generada evidencia un estrechamiento de la base, consistente con la transición demográfica del país (Figura 4).

Figura 4

Pirámide poblacional del Ecuador, Censo 2022

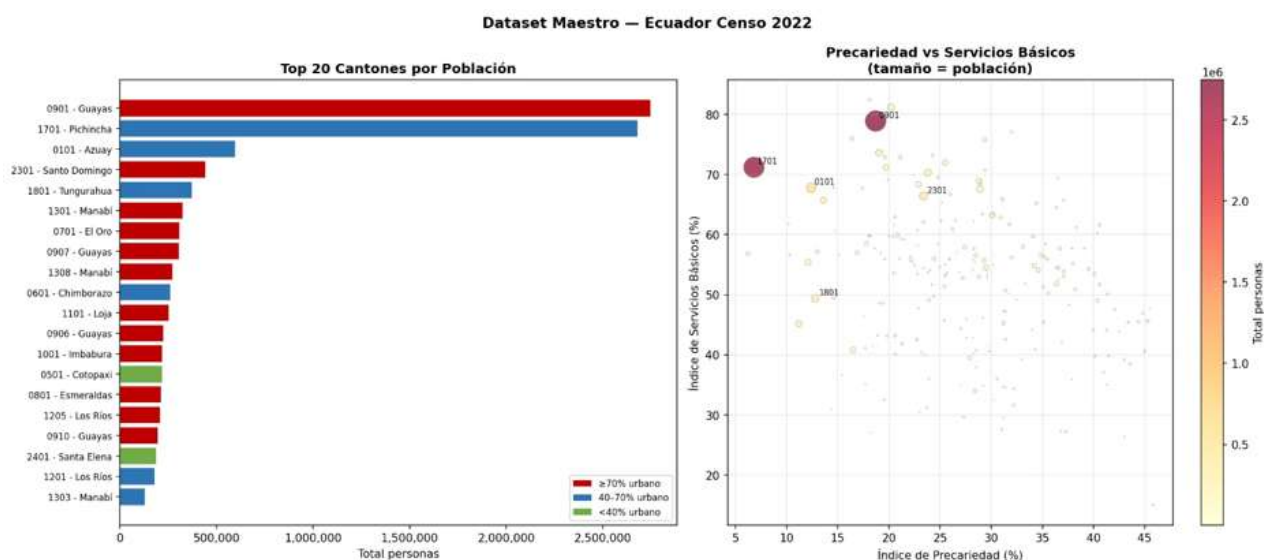


Nota: Elaboración propia a partir de los microdatos del Censo de Población y Vivienda 2022 del INEC (INEC, 2023).

– **Concentración poblacional:** la población se concentra en un número reducido de cantones; los de mayor tamaño, entre ellos Guayaquil y Quito, agrupan una fracción sustancial del total nacional (Figura 5).

Figura 5

Cantones más poblados y relación precariedad-servicios básicos, Censo 2022



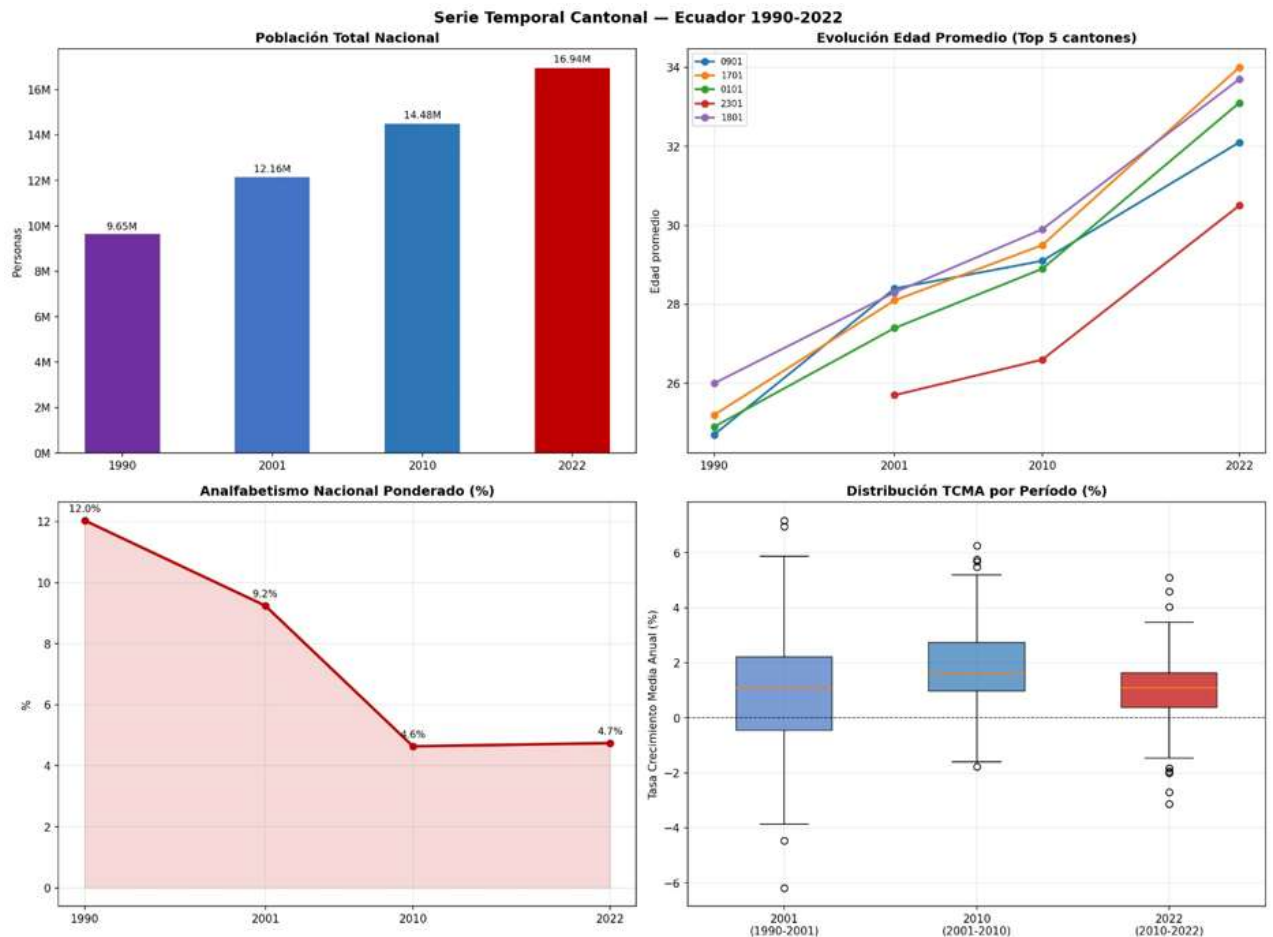
Nota: Elaboración propia a partir de los microdatos del Censo de Población y Vivienda 2022 del INEC (INEC, 2023).

– **Reducción del analfabetismo:** el analfabetismo nacional ponderado desciende de manera sostenida a lo largo de los cuatro censos (Figura 6, panel inferior izquierdo).

– **Heterogeneidad del crecimiento:** la distribución de la TCMA por período intercensal muestra una fuerte dispersión entre cantones y la presencia de valores extremos, lo que anticipa la dificultad de un modelo único para todos los territorios (Figura 6, panel inferior derecho).

Figura 6

Serie temporal cantonal del Ecuador, 1990-2022



Nota. Elaboración propia a partir de la serie temporal cantonal 1990-2022 construida con los microdatos censales del INEC. Los cuatro paneles muestran, en sentido de lectura: población total nacional, evolución de la edad promedio en los cinco cantones más poblados, analfabetismo nacional ponderado y distribución de la TCMA por período intercensal.

En la Tabla 11, se muestra un resumen del crecimiento poblacional desde 1990 hasta 2022 y la variación de los cantones creados.

Tabla 11

Población nacional del Ecuador por censo

Censo	Población total nacional	Cantones con dato
1990	9.697.979	173
2001	12.156.608	221
2010	14.483.499	221
2022	16.938.986	221

Nota: Totales nacionales oficiales de los censos de población del INEC (INEC, 2023). El número de cantones con dato en 1990 es inferior porque 48 cantones fueron creados con posterioridad a ese año. El INEC ajustó posteriormente la cifra de 2022 a aproximadamente 17,7 millones de habitantes para corregir la omisión censal; en este estudio se emplea la población efectivamente empadronada por constituir la base de la agregación cantonal. Los paneles del EDA (Figura 6) reflejan los totales agregados a partir de los microdatos; para 1990 dicho agregado (9,65 millones) difiere levemente de la cifra oficial debido a la homologación de los 173 cantones existentes en ese año.

3.3.Preparación y Procesamiento de Datos

La preparación de los datos transforma los cuatro resúmenes en un único conjunto analítico en formato apto para modelado. El proceso comprende cuatro etapas encadenadas y reproducibles:

- a. La unificación de la serie temporal
- b. El tratamiento de los datos faltantes
- c. La ingeniería de características
- d. La normalización y partición temporal del conjunto

Cada etapa se implementa en un script versionado del repositorio, lo que garantiza la trazabilidad y la reproducibilidad de las decisiones metodológicas.

a. Unificación de la serie temporal

Los resúmenes cantonales de 1990, 2001, 2010 y 2022 se integran verticalmente en una serie temporal única identificada por dos llaves: el código de división político-administrativa del cantón (DPA 2022) y el año censal. Con motivo de comparar de forma longitudinal, únicamente se conservan las variables presentes y que se puedan homologar entre censos; siendo así que la escolaridad y el porcentaje de discapacidad se excluyen de la serie por corresponder a cambios metodológicos que impiden su comparación. El conjunto resultante alcanza hasta 884 registros ($221 \text{ cantones} \times 4 \text{ censos}$) en formato largo (08_serie_temporal_unificada.py, línea 21).

b. Tratamiento de datos faltantes (imputación)

La escasez de datos característica del problema, solo cuatro observaciones temporales por cantón, exige un tratamiento explícito de tres tipos de ausencias:

- **Cantones creados con posterioridad a 1990:** Los 48 cantones constituidos

después del censo de 1990 carecen de registro en ese año. Se generan filas con valores nulos para preservar la estructura del panel, heredando de 2022 únicamente los metadatos administrativos como el código y el nombre de la provincia.

- **Necesidades Básicas Insatisfechas (NBI) en 2001:** El indicador de NBI no se encuentra disponible para el censo de 2001. Su valor se imputa mediante interpolación lineal a partir de la pendiente intercensal 2010-2022: se estima la variación anual del indicador como $(NBI\ 2022 - NBI\ 2010) / 12$ y se proyecta nueve años hacia atrás desde 2010, acotando el resultado al rango admisible de 0 a 100. Cada valor imputado se señala con la bandera `flag_nbi_imputado`, de modo que el modelo pueda distinguir los registros estimados de los observados. El NBI de 1990 se mantiene como faltante por no disponerse de un dato base que sustente su imputación.
- **Quiebre metodológico de la urbanización en 1990:** Dado que la definición de área urbana en 1990 difiere de la utilizada en censos posteriores, todos los registros de ese año se marcan con la bandera `flag_urb_quiebre` para advertir del cambio de criterio sin descartar el dato.

Como cierre de esta etapa se calcula la Tasa de Crecimiento Media Anual (TCMA) intercensal de cada cantón respecto al censo previo, mediante la fórmula de crecimiento geométrico compuesto, variable que constituye uno de los objetivos de modelado.

c. Ingeniería de características

Sobre la serie unificada se derivan nuevas variables orientadas a capturar la dinámica poblacional. Entre ellas se incluyen la transformación logarítmica de la población:

$$\log_{poblacion} = \log(1 + total_{personas}) \quad [17]$$

Esto permite reducir la asimetría de la distribución y estabiliza la varianza; los rezagos temporales (`lag_poblacion`, `lag_log_pob`), que incorporan el valor del censo anterior de cada cantón como predictor autorregresivo; términos de interacción y vulnerabilidad, como la interacción entre NBI y urbanización (`interac_nbi_urb`), la brecha de género respecto a la

paridad (brecha_genero) y un índice proxy de envejecimiento derivado de la edad promedio; una variable de tiempo relativo (tiempo_rel) que sitúa cada censo en una escala continua entre 0 (1990) y 1 (2022) para modelar la tendencia secular; y la codificación entera de la provincia y el cantón (provincia_id, canton_id) para su uso como variables categóricas y como identificador de grupo en los modelos temporales.

Adicionalmente, las covariables socioeconómicas disponibles únicamente en el censo de 2022, como los indicadores de servicios básicos, precariedad habitacional y estructura del hogar, se propagan como atributos estáticos hacia los años anteriores de cada cantón, al considerarse rasgos estructurales de lenta variación.

d. Normalización y prevención de fuga de información

Las veintitrés variables numéricas del conjunto se estandarizan mediante StandardScaler. Como salvaguarda metodológica frente a la fuga de información (data leakage), los parámetros de estandarización, la media y la desviación típica de cada columna, se estiman exclusivamente con los datos de entrenamiento, correspondientes a los censos de 1990 y 2001, y posteriormente se aplican a los conjuntos de validación y prueba, evitando que información de estos últimos contamine el ajuste (09_feature_engineering.py, líneas 169-183). Dichos parámetros se conservan en el archivo scaler_params.csv para permitir la inversión de la transformación durante la inferencia.

- **Partición temporal.** En coherencia con la naturaleza longitudinal del problema y con el objetivo de proyectar hacia el futuro, la partición del conjunto es temporal y no aleatoria (09_feature_engineering.py, líneas 189-199). El conjunto de entrenamiento reúne los censos de 1990 y 2001, para el aprendizaje de los patrones históricos; el de validación corresponde al censo de 2010, empleado en el ajuste de hiperparámetros; y el de prueba se reserva al censo de 2022, utilizado como evaluación final sobre datos nunca vistos por los modelos. Este esquema reproduce las condiciones reales de la proyección poblacional, es decir, estimar un censo futuro a partir de los precedentes, y evita el optimismo que introduciría una partición aleatoria al mezclar años dentro de un mismo cantón.

- **Ruta de datos anual complementaria.** De forma paralela a la ruta censal descrita, el proyecto construye un segundo conjunto de datos de resolución anual (1990-2022) destinado a los modelos temporales. A partir del conjunto maestro (dataset_modelo_final.csv), la población se interpola a frecuencia anual y las variables censales se propagan mediante un esquema de escalón. Sobre esta estructura se integran los registros vitales cantonales, nacidos vivos (ENV) y defunciones generales (EDG), junto con las tasas demográficas derivadas. En esta ruta la normalización emplea MinMaxScaler en lugar de StandardScaler, ajustado igualmente solo sobre el subconjunto de entrenamiento (años anteriores a 2022) para prevenir la fuga de información; una etapa específica re-normaliza las variables heredadas y verifica la ausencia de fuga. El conjunto resultante (dataset_fase6_final.csv) alimenta el modelo temporal anual, cuya metodología se detalla en la sección 3.4.

e. Salvaguardas y limitaciones de la preparación de datos

La preparación descrita incorpora salvaguardas explícitas frente a la fuga de información (data leakage), cuyo alcance conviene delimitar con precisión, así como tres limitaciones que la escasez de datos obligó a asumir y que se declaran aquí de forma expresa.

- **Salvaguardas.** Los parámetros de normalización se estiman únicamente sobre el subconjunto de entrenamiento, tanto en la ruta censal como en la anual. La partición es temporal y no aleatoria, de modo que el censo de 2022 no interviene en ninguna etapa del ajuste. La variable objetivo, la TCMA, se calcula siempre entre censos observados y nunca sobre valores interpolados: los extremos de cada intervalo intercensal son años censales, por lo que la interpolación anual no alcanza a la variable que los modelos deben predecir. Por último, las covariables socioeconómicas de 2022 propagadas hacia atrás como atributos estáticos no forman parte del conjunto de predictores del LightGBM global, cuyas variables corresponden al inicio de cada intervalo y a los agregados de registros vitales del propio intervalo; su uso queda circunscrito a la ruta censal y a los modelos temporales.
- **Imputación del NBI de 2001.** El valor imputado se obtiene de la pendiente

intercensal 2010-2022, por lo que incorpora información posterior al año que describe y participa como predictor en los intervalos de entrenamiento. Se mantiene porque constituye la única forma de conservar el indicador para un censo completo y porque cada registro estimado queda señalado mediante la bandera correspondiente; debe considerarse, sin embargo, una fuente potencial de optimismo en el ajuste.

- **Denominador de las tasas vitales.** Las tasas de natalidad y de mortalidad de la ruta anual emplean la población interpolada como denominador. Dado que la interpolación del intervalo 2010-2022 está anclada en el censo de 2022, dichas tasas incorporan información del año reservado para prueba en los años intermedios del intervalo. Los conteos absolutos de nacimientos y defunciones, en cambio, no se ven afectados.
- **Naturaleza de la interpolación anual.** La población intercensal se interpola de forma lineal y no geométrica, por lo que subestima levemente el efecto del crecimiento compuesto en los años centrales de cada intervalo. Su alcance se limita a la ruta anual y a las variables derivadas de ella, sin afectar a los valores censales ni a la variable objetivo.
- **Alcance de estas limitaciones.** Las tres afectan a los modelos que incorporan covariables socioeconómicas o de registros vitales (las familias de aprendizaje automático y los modelos temporales) y no al método Hamilton-Perry, que opera únicamente sobre la población censada. Por consiguiente, los errores de los primeros que se reportan en el Capítulo 4 deben interpretarse como una cota optimista. Corregir estas fuentes tendería a ampliar, y no a revertir, la diferencia observada en los niveles subnacionales entre el método demográfico y los modelos de aprendizaje automático.

f. Trazabilidad computacional

El procedimiento descrito en las secciones 3.2 y 3.3, junto con los modelos que se desarrollan en la sección 3.4, se materializa en una secuencia de scripts numerados que se ejecutan en orden y que persisten sus productos intermedios en disco. La Tabla 12 consolida esa correspondencia en un único lugar, de manera que la trazabilidad computacional quede documentada sin que la lectura metodológica del capítulo dependa

del código fuente. En el cuerpo del texto se conservan referencias puntuales a líneas de código únicamente en aquellos pasajes donde respaldan una cifra reportada o una salvaguarda explícita frente a la fuga de información.

Tabla 12

Correspondencia entre fases metodológicas, scripts y productos generados

Fase metodológica	Script	Producto principal
Agregación cantonal del Censo 2022 (personas, hogar y vivienda)	01-04	dataset_maestro_canton_2022.csv
Agregación cantonal de los censos 2010, 2001 y 1990	05-07	resumen_canton_2010.csv; resumen_canton_2001.csv; resumen_canton_1990.csv
Unificación de la serie temporal, imputación y cálculo de la TCMA	08	serie_temporal_canton.csv
Ingeniería de características, normalización y partición temporal	09	dataset_modelo_final.csv; scaler_params.csv; train_set.csv; val_set.csv; test_set.csv
Modelos de referencia (baseline)	10	resultados_baseline.csv
Temporal Fusion Transformer, versión censal y versión anual	11 (v3) y 19	metricas_tft_v3.csv; metricas_tft_v5.csv
Ruta anual e integración de los registros vitales (ENV y EDG)	15-18 y 18b	dataset_fase6_final.csv
Incorporación de las EPP INEC 2024 como referencia de contraste	20a	epp_cantonal_2024.csv
LightGBM global	22	script22_metrics.csv; script22_feature_importance.csv
Ensamblados híbridos y selección del peso por validación cruzada	23, 30 y 31	script23_metrics.csv; script30_mejor_ensamble.csv; script31_comparacion_final.csv
Segmentación en tipologías y modelos por tipología	26 y 28	script26_cluster_assignments.csv; script28_model_metrics.csv
Efectos aleatorios por cantón	27b	script27b_metrics.csv; script27b_random_effects.csv
Hamilton-Perry amortiguado	29	script29_hp_retroccion_2022.csv; script29_hp_proyecciones.csv
Variante con factor de amortiguación aprendido (HP-ML-d)	32	script32_comparacion_final.csv
Evaluación por nivel geográfico y proyecciones 2030-2035	33-38	Resultados presentados en el Capítulo 4

Nota: Elaboración propia. Los identificadores corresponden al prefijo numérico de cada script en el repositorio público del proyecto (véase Anexos). Las filas siguen el orden de ejecución de los scripts en el repositorio.

g. Correspondencia con los objetivos de la investigación

Las decisiones descritas a lo largo de este capítulo no son neutrales respecto de las preguntas que la investigación se plantea: cada una habilita o condiciona el cumplimiento de un objetivo específico concreto de los enunciados en la sección 1.4.3. La Tabla 13 hace explícita esa correspondencia, señalando para cada decisión metodológica el objetivo al que sirve y la razón por la cual resulta necesaria; los objetivos se identifican con la letra que reciben en dicha sección.

Tabla 13

Aporte de las decisiones metodológicas a los objetivos específicos

Decisión metodológica (sección)	Objetivo	Contribución al cumplimiento del objetivo
Homologación del código DPA 2022 entre los cuatro censos (3.2)	b	Permite seguir al mismo cantón a lo largo de 32 años pese a los cambios de la división político-administrativa; sin ella el panel no sería longitudinalmente comparable
Integración de los registros vitales ENV y EDG 1990-2024 (3.2 y 3.3)	b	Incorpora la única fuente de frecuencia anual disponible, que compensa parcialmente la escasez de los cuatro puntos censales
Conservación de los 48 cantones creados después de 1990 mediante filas nulas (3.3)	b	Mantiene completa la estructura del panel en lugar de descartar unidades, lo que evitaría un sesgo hacia los cantones de creación más antigua
Imputación del NBI de 2001 y señalización mediante banderas (3.3)	b	Recupera una covariable socioeconómica para un censo completo y deja marcado el dato estimado para que pueda distinguirse del observado
Exclusión de escolaridad y discapacidad por quiebre metodológico (3.3)	b	Antepone la comparabilidad longitudinal al número de variables disponibles
Estimación de los parámetros de estandarización solo con los censos de entrenamiento (3.3)	c	Impide que la media y la desviación de los censos de validación y prueba influyan en el ajuste, condición necesaria para que el error medido sobre 2022 sea interpretable
Partición temporal no aleatoria con el censo de 2022 reservado (3.3)	c	Reproduce las condiciones reales de una proyección y evita el optimismo de una partición aleatoria que mezclaría años de un mismo cantón
Cálculo de la TCMA exclusivamente entre censos observados (3.3 y 3.4)	c	Evita que valores interpolados intervengan en la construcción de la variable objetivo
Adopción de un factor de amortiguación fijo $d = 0,5$ bajo criterio minimax (3.4)	c	El valor no se optimiza sobre el conjunto de prueba. Los dos periodos intercensales disponibles arrojan óptimos muy distintos entre sí, por lo que ninguno permite estimar un d confiable. Ante esa

		inestabilidad se adopta el punto medio del intervalo $[0,1]$, que acota el error máximo posible y preserva la comparabilidad con los modelos de aprendizaje automático, los cuales tampoco acceden al conjunto de prueba durante el ajuste.
Evaluación de las cinco familias bajo un mismo protocolo y una misma métrica (3.4)	d, hipótesis	Aísla el efecto atribuible al modelo, de modo que las diferencias observadas no provengan de tratamientos desiguales
Recálculo de las métricas en los niveles cantonal, provincial y nacional (3.4)	d, hipótesis	Constituye el instrumento directo de contrastación: sin medir el error por nivel no sería posible sostener ni refutar que la escala determina el modelo preferible
Uso de las EPP INEC 2024 únicamente como referencia externa y separación de los dos marcos (3.2 y 3.4)	e	Mantiene la independencia entre lo estimado y aquello con lo que se contrasta, y evita atribuir a los modelos la brecha que corresponde a la referencia
Ejecución secuencial de scripts con persistencia de los productos intermedios (3.3)	f	Hace auditable cada transformación y permite reejecutar el estudio completo a partir de las fuentes públicas

Nota: Elaboración propia. Las letras de la columna central corresponden a los objetivos específicos enunciados en la sección 1.4.3.

En cuanto a la hipótesis, según la cual el nivel de agregación geográfica constituye el determinante principal de la selección del modelo, su contrastación descansa sobre dos decisiones en particular. La primera es la evaluación de las cinco familias bajo un protocolo y una métrica idénticos, que aísla el efecto atribuible al modelo. La segunda es el recálculo de esa misma métrica en los tres niveles de agregación, que aísla el efecto atribuible a la escala territorial. Sin ambas condiciones, cualquier diferencia observada entre niveles podría explicarse por un tratamiento desigual entre modelos y no por el nivel de agregación, con lo que la hipótesis quedaría fuera del alcance de una contrastación objetiva.

3.4. Metodología de Análisis o Modelo

La metodología de análisis se articula en torno a las dos rutas de datos preparadas en la sección 3.3. Los modelos de referencia y de aprendizaje automático de nivel cantonal se ajustan sobre la ruta censal, el conjunto maestro de los cuatro censos, mientras que los modelos temporales operan sobre la ruta anual complementaria, que incorpora los registros

vitales (ENV y EDG) a frecuencia anual. El desarrollo detallado de cada familia de modelos se completa en esta sección.

a. Estrategia de evaluación comparativa

El estudio adopta un diseño de evaluación comparativa entre cinco familias de modelos, con el propósito de contrastar enfoques estadísticos tradicionales y de aprendizaje automático bajo un mismo protocolo. La validación funciona de manera retrospectiva, cada modelo estima la población cantonal del Censo 2022 utilizando exclusivamente la información disponible en los censos previos de 1990, 2001 y 2010, lo que reproduce las condiciones reales de una proyección. La métrica principal es el Error Porcentual Absoluto Medio (MAPE) a nivel cantonal, que es la que se reporta para todos los modelos y en los tres niveles de agregación. El RMSE, el MAE y el coeficiente de determinación (R^2) intervienen únicamente en la comparación de los modelos de referencia y en experimentos específicos, de modo que sirven como apoyo interpretativo y no como base del ordenamiento final.

El desempeño se examina en tres niveles de agregación territorial (cantonal, provincial y nacional), dado que el nivel de agregación condiciona el error de cada método. La comparación se efectúa dentro de dos marcos de referencia diferenciados: frente al dato censal real de 2022 y frente a las Estimaciones y Proyecciones de Población (EPP) del INEC 2024, sin combinar ambos por existir una brecha sistemática entre el censo real y la proyección oficial. Los resultados numéricos de esta evaluación se presentan en el Capítulo 4.

b. Familia 1: Modelos de referencia (baseline)

Establece la base de comparación mediante cinco modelos de referencia:

- modelo naive que proyecta el último valor conocido;
- modelo de crecimiento constante basado en la TCMA histórica promedio;
- regresión lineal; regresión Ridge con regularización L2 (Hoerl & Kennard, 1970); y un Random Forest (Breiman, 2001) como cota superior no lineal del conjunto de

referencia.

Estos modelos permiten cuantificar la ganancia real que aportan los enfoques más complejos frente a extrapolaciones simples.

c. Familia 2: Método demográfico de pequeñas áreas: Hamilton-Perry amortiguado

Corresponde al método Hamilton-Perry (Hamilton & Perry, 1962), estándar demográfico para la proyección de áreas pequeñas, aquí en una variante amortiguada que modera el crecimiento extremo conforme al principio de regresión a la media. El método extrapola la población mediante la expresión

$$P(t + n) = P(t) \cdot \left[\frac{P(t)}{P(t-k)} \right]^{\frac{n}{k^d}}, [18]$$

Donde: d es un factor de amortiguación comprendido entre 0 y 1. Con $d = 1$ el procedimiento equivale algebraicamente a una TCMA constante, mientras que el valor $d = 0,5$, adoptado en este estudio, modera los crecimientos extremos conforme al principio de regresión a la media demográfica, según el cual los cantones de crecimiento acelerado tienden a desacelerarse y los de decrecimiento a recuperarse.

En esta implementación el método se aplica sobre la población cantonal total, por lo que se reduce a una extrapolación amortiguada de la razón de crecimiento intercensal. Cabe destacar que $d = 0,5$ no se optimizó sobre el conjunto de prueba: es un parámetro fijo adoptado por criterio minimax (Wald, 1950), tras comprobar que el valor óptimo de d no es estable entre los periodos intercensales disponibles (ver Capítulo 4, sección 4.1.5), lo que preserva la validez de la comparación con los modelos de aprendizaje automático. Como extensión, se explora una variante en la que el factor de amortiguación se personaliza por cantón mediante aprendizaje automático (HP-ML- d), calibrando un d óptimo para cada cantón según sus características demográficas.

d. Familia 3: Modelos de aprendizaje automático basados en árboles

Agrupar modelos de potenciación del gradiente (gradient boosting) y sus extensiones, entrenados para predecir la tasa de crecimiento cantonal y reconstruir a partir de ella la población objetivo. Comprende tres variantes:

- LightGBM global:** Modelo de potenciación del gradiente (Ke et al., 2017) entrenado sobre veintiocho covariables demográficas y de registros vitales, con validación cruzada grupal por cantón (GroupKFold) que impide que un mismo cantón aparezca a la vez en entrenamiento y validación. Conviene precisar que esta validación cruzada agrupada opera únicamente como control interno de generalización sobre la partición de entrenamiento, o sea los intervalos intercensales 1990-2001 y 2001-2010, y no es la fuente de la métrica de error definitiva; esta última procede de la partición temporal estricta descrita para el conjunto de datos, en la que el modelo se entrena con dichos intervalos y se evalúa sobre el intervalo 2010-2022, reservado como prueba, del cual se reporta el MAPE cantonal en el Capítulo 4 (script_22_lgbm_global_v1.py, líneas 363-370, 494-522). Ambos esquemas son, por tanto, complementarios y no contradictorios: GroupKFold verifica la robustez del ajuste, mientras que la retrodicción temporal cuantifica el desempeño final. Debe señalarse, además, que estas veintiocho covariables no coinciden con las veintitrés variables numéricas estandarizadas en la sección 3.3: aquellas corresponden a la ruta censal, mientras que el LightGBM global incorpora adicionalmente variables derivadas de los registros vitales (ENV y EDG) y, por tratarse de un modelo basado en árboles, no requiere la estandarización previa. La configuración final emplea 400 árboles, una tasa de aprendizaje de 0,03, quince hojas por árbol, regularización L1 y L2 y submuestreo del 90 % de filas y columnas. El procedimiento recalcula las TCMA históricas únicamente con censos reales y descarta los objetivos nulos para prevenir la fuga de información.
- Efectos aleatorios por cantón (MERF):** Reformulación de datos de panel que añade al modelo global un intercepto aleatorio por cantón, de la forma

$$TCMA_pred = F(X) + b_cantón + error \quad [19]$$

El término $b_cantón$ se estima como el mejor predictor lineal insesgado (BLUP), equivalente al paso de esperanza-maximización del algoritmo Mixed-Effects Random Forest (Hajjem et al., 2014; Bandara et al., 2022), y captura la desviación sistemática

de cada cantón respecto al modelo global.

- **Modelos por tipología demográfica:** Se segmenta los cantones en tipologías mediante aprendizaje no supervisado, k-means sobre variables estandarizadas, seleccionando el número de grupos por el coeficiente de silueta, y luego ajusta un modelo específico por tipología. La segmentación distingue tres grupos: cantones rurales de baja dinámica, metrópolis nacionales (con retorno a la TCMA por su reducido número) y cantones semi-urbanos de alta expansión. El entrenamiento utiliza los períodos intercensales 1990-2001 y 2001-2010, reservando 2010-2022 como prueba fuera de muestra.

e. Familia 4: Aprendizaje profundo temporal: Temporal Fusion Transformer

Corresponde al Temporal Fusion Transformer (TFT; Lim et al., 2021), arquitectura de aprendizaje profundo para el pronóstico de series temporales con múltiples entidades. Su implementación fue iterativa: una primera versión (v1) se descartó al detectarse una fuga de información en la partición temporal, ya que incluía el censo de 2022 en el conjunto de entrenamiento, y una versión intermedia (v2) introdujo correcciones parciales, hasta consolidar la configuración corregida que aquí se describe; el Capítulo 4 documenta en detalle esa trayectoria y sus implicaciones.

Se implementan dos configuraciones:

- La versión censal opera sobre los puntos censales de cada cantón, con una ventana de codificación de tres pasos y un horizonte de predicción de un paso, un tamaño oculto de 32 unidades, dos cabezas de atención y una función de pérdida por cuantiles, entrenada durante 50 épocas.
- La versión anual (TFT v5) aprovecha la ruta de datos anual descrita en la sección 3.3 (treinta y tres puntos temporales por cantón e indicadores de natalidad y mortalidad), lo que aporta mayor densidad temporal y una mejor regularización.

En ambas configuraciones el normalizador de grupo se ajusta exclusivamente con los años de entrenamiento (1990, 2001 y 2010) para evitar la fuga de estadísticas futuras.

f. Familia 5: Modelos de ensamble híbrido

Evalúa si la combinación de un método demográfico con un modelo de aprendizaje automático mejora la precisión.

Se construyó ensambles lineales ponderados de la forma $\text{pred} = w \cdot A + (1-w) \cdot B$, combinando la TCMA con el LightGBM global, y el Hamilton-Perry amortiguado con el LightGBM global.

Para evitar el sesgo optimista que supondría seleccionar el peso w sobre el mismo conjunto de prueba, una variante estima dicho peso mediante validación cruzada temporal de tipo dejar-un-período-fuera sobre un conjunto de validación independiente del de prueba.

En conjunto, este diseño permite contrastar de forma sistemática los cinco enfoques bajo idénticas condiciones de datos y de evaluación, y examinar cómo varía el modelo de menor error según el nivel de agregación territorial.

El análisis comparativo de los resultados obtenidos se desarrolla en el Capítulo 4.

CAPITULO 4:

4. ANÁLISIS DE RESULTADOS

4.1. Pruebas de Concepto

En este capítulo se describe el procesamiento, entrenamiento y evaluación de los seis modelos de predicción poblacional aplicados a los 221 cantones del Ecuador, a partir de los censos INEC de 1990, 2001, 2010 y 2022.

El pipeline completo (Scripts 10 al 38) se organiza en dos bloques: primero la etapa de experimentación y selección de modelos (secciones 4.1.1 a 4.1.5), y después el análisis consolidado de resultados finales (sección 4.2), que incluye la validación retrospectiva 2010-2022, el análisis por nivel de agregación geográfica y las proyecciones 2030-2035.

Las pruebas de concepto del proyecto consistieron en entrenar y validar de forma independiente cada familia de modelos bajo un mismo esquema de partición temporal: entrenamiento con los censos 1990, 2001 y 2010, y prueba exclusivamente sobre el censo 2022, nunca visto durante el entrenamiento.

Este esquema se aplicó de manera uniforme a los seis enfoques evaluados, ya que comparar modelos bajo condiciones distintas de validación habría invalidado cualquier conclusión posterior sobre cuál es el más preciso.

Se documenta a continuación, en orden cronológico de ejecución, tanto los modelos que se mantuvieron en la versión final de resultados como los que fueron descartados por errores metodológicos detectados en el camino, ya que ese proceso de corrección forma parte del aporte de esta tesis.

4.1.1. Selección del modelo de referencia

Para decidir qué modelo emplear como benchmark, en el Script 10 se comparó cinco candidatos bajo el mismo esquema de validación:

- Un modelo naive que replica el último valor censal conocido (2010).
- La Tasa de Crecimiento Medio Anual (TCMA) constante.

- Una regresión lineal.
- Una regresión Ridge (Hoerl & Kennard, 1970).
- Un Random Forest (Breiman, 2001) como cota superior no lineal.

En la Tabla 14 resume el resultado, evaluado contra el censo real de 2022 (Marco A).

Tabla 14

Comparación de modelos de referencia candidatos (Marco A)

Modelo	RMSE	MAE	MAPE	R ²
B1 Naive (valor 2010)	43,001	11,781	13.77%	0.973
B2 TCMA constante	18,559	5,884	8.65%	0.995
B3 Regresión lineal	47,584	14,865	27.16%	0.967
B4 Ridge (alfa=10)	70,808	15,916	23.47%	0.928
B5 Random Forest	86,939	13,907	12.98%	0.891

La TCMA constante obtuvo el menor error de los cinco (MAPE=8.65%), por debajo incluso del Random Forest (12.98%), pese a ser el modelo más simple del grupo.

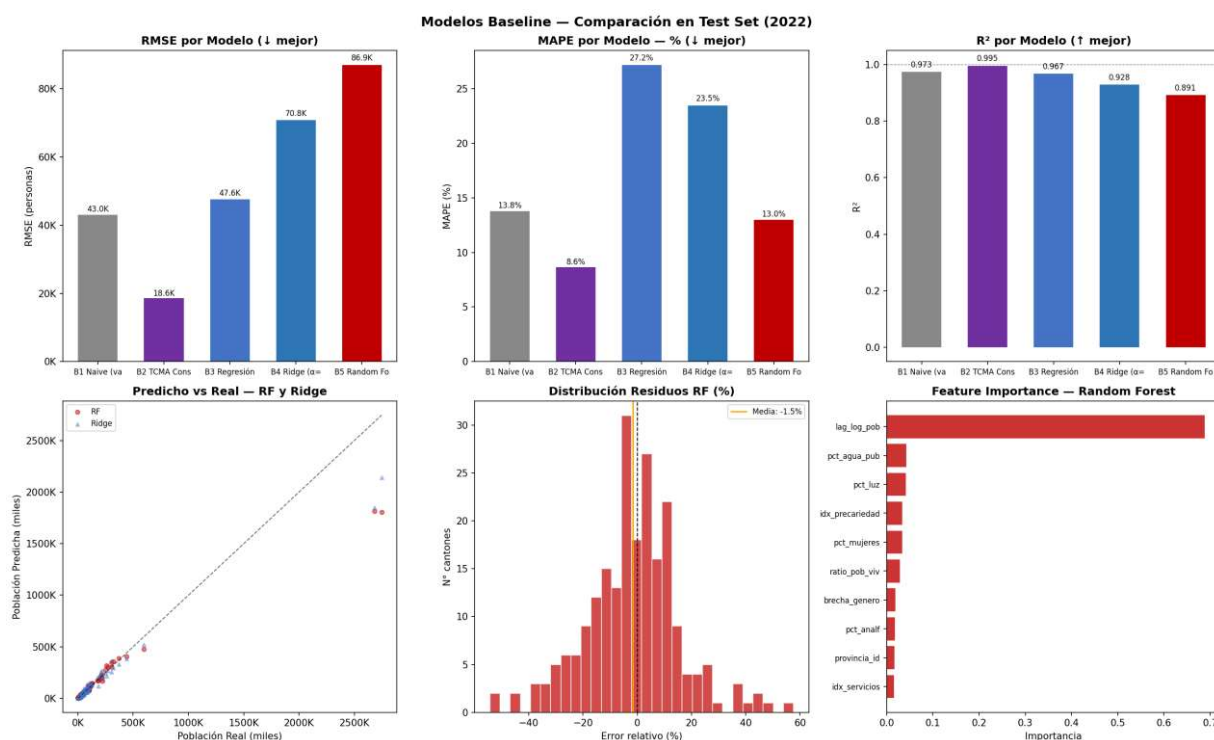
El R² sigue un patrón distinto, el Random Forest queda último (0.891) porque esta métrica penaliza con fuerza los errores grandes en los pocos cantones más poblados, algo que el MAPE, al trabajar en términos porcentuales, no castiga de la misma manera; por eso ambas métricas no siempre ordenan a los modelos igual, y el MAPE es el criterio que se usa de aquí en adelante por ser comparable entre cantones de tamaño muy distinto.

Que un extrapolador de una sola tasa supere a un ensamble de árboles permite anticipar, en esta etapa temprana del pipeline, la idea que con solo cuatro puntos censales por cantón, sumar complejidad no garantiza mejor desempeño. Por esa razón, la TCMA constante queda como benchmark de referencia para todos los modelos evaluados en las secciones siguientes.

La Figura 7 presenta el comparativo completo generado por el Script 10: RMSE, MAPE y R² por modelo, el ajuste de Random Forest y Ridge frente a los valores reales, la distribución de residuos del Random Forest y la importancia de variables de este último.

Figura 7:

Comparación gráfica de los modelos de referencia candidatos (Script 10)



Esta comparación corresponde al Marco A (contra el censo real de 2022). El benchmark de TCMA se vuelve a evaluar más adelante bajo el Marco B (contra las proyecciones oficiales EPP INEC 2024, MAPE=14.29%), marco que se usa de forma consistente en el resto del capítulo; los otros cuatro candidatos (naive, lineal, Ridge, Random Forest) no se evaluaron bajo ese segundo marco, pues quedaron descartados en esta etapa de selección.

4.1.2. Modelo de referencia y exploración con redes neuronales

El Script 10 estableció la Tasa de Crecimiento Medio Anual (TCMA) constante como benchmark: cada cantón proyecta su crecimiento futuro extrapolando la tasa observada entre los dos últimos censos.

Contra el censo real 2022 (Marco A) esta referencia obtiene un MAPE de 8.65%; contra las proyecciones oficiales EPP INEC 2024 (Marco B, usado de aquí en adelante en este capítulo salvo que se indique lo contrario) el MAPE sube a 14.29%.

La diferencia entre ambos valores no es un error de cálculo, sino el reflejo de que el propio INEC sobreestimó la población de 2022 en sus proyecciones; esto se explica con más detalle en la sección 4.2.1.

Para explorar si una arquitectura de aprendizaje profundo especializada en series temporales podía superar este benchmark, se implementaron tres versiones del Temporal Fusion Transformer (Scripts 11 y 11_v2/v3) y una cuarta variante a resolución anual (Script 19). La Tabla 15 resume el resultado de cada intento.

Tabla 15

Versiones del Temporal Fusion Transformer evaluadas

Versión	MAPE	Observación
TFT v1	≈0%	Descartada: incluía el censo 2022 en el entrenamiento por error de partición
TFT v2	n/d	Corrección del split temporal, usada como base para v3
TFT v3	13.52%	Split correcto; supera el benchmark por 0.77pp, el margen más estrecho de todos los modelos retenidos
TFT v5 (anual)	474%	Descartada: entrenada sobre datos interpolados linealmente, no sobre observaciones reales

La primera versión fue la más reveladora, un MAPE cercano a 0% no significa que el modelo hubiera aprendido a predecir, sino que el conjunto de entrenamiento incluía por error observaciones del propio año que se le pedía predecir.

Una vez corregido ese split, el resultado real del TFT cayó a 13.52%, apenas por debajo del benchmark. La versión anual (v5) falló de forma más severa aún, porque al entrenar sobre una serie interpolada linealmente entre censos, la red terminó aprendiendo el patrón de la interpolación en sí, no la dinámica demográfica real.

Estos tres resultados, en conjunto, apuntan a una misma causa, con solo tres periodos intercensales disponibles, no hay suficientes observaciones para que una red profunda generalice de forma confiable, sin importar cuánto se ajuste su arquitectura.

4.1.3. LightGBM global, ensamble inicial e interpretabilidad

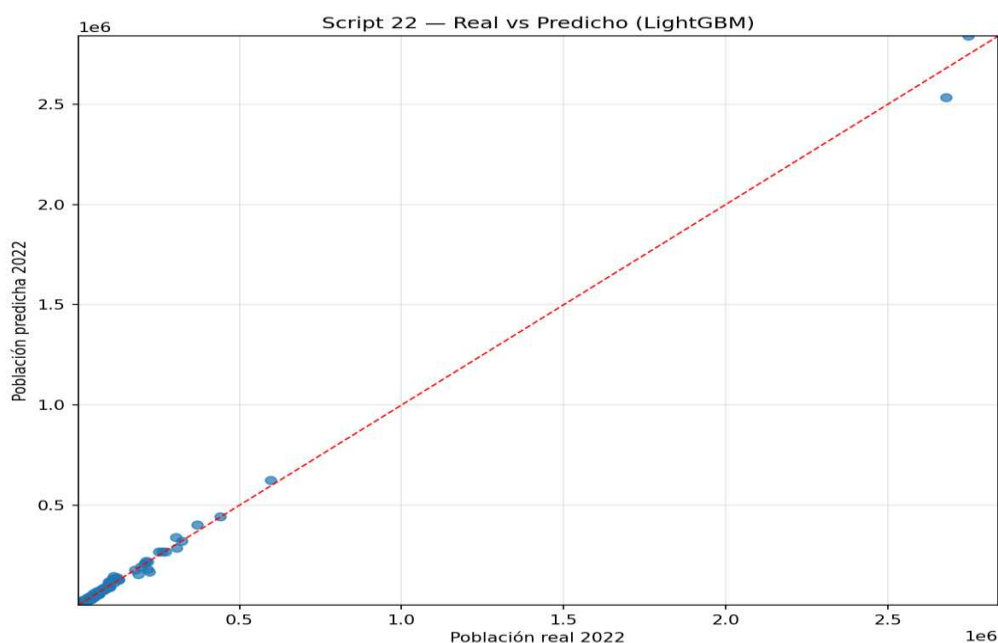
El Script 22 entrenó un único modelo LightGBM para los 221 cantones, usando variables demográficas derivadas de los censos.

En el periodo de prueba 2010-2022 alcanzó un MAPE cantonal de 9.43%. Su resultado más llamativo aparece al agregar las predicciones a nivel nacional, el error de sesgo total es de apenas +0.49% (82,563 habitantes sobre 16,938,986 reales), lo que lo vuelve más preciso a esa escala que las propias proyecciones oficiales del INEC, que sobreestimaron la población nacional en +4.58%.

La Figura 8 compara, cantón por cantón, la población real de 2022 contra la predicha por este modelo.

Figura 8

Población real frente a predicha, LightGBM global (Script 22)



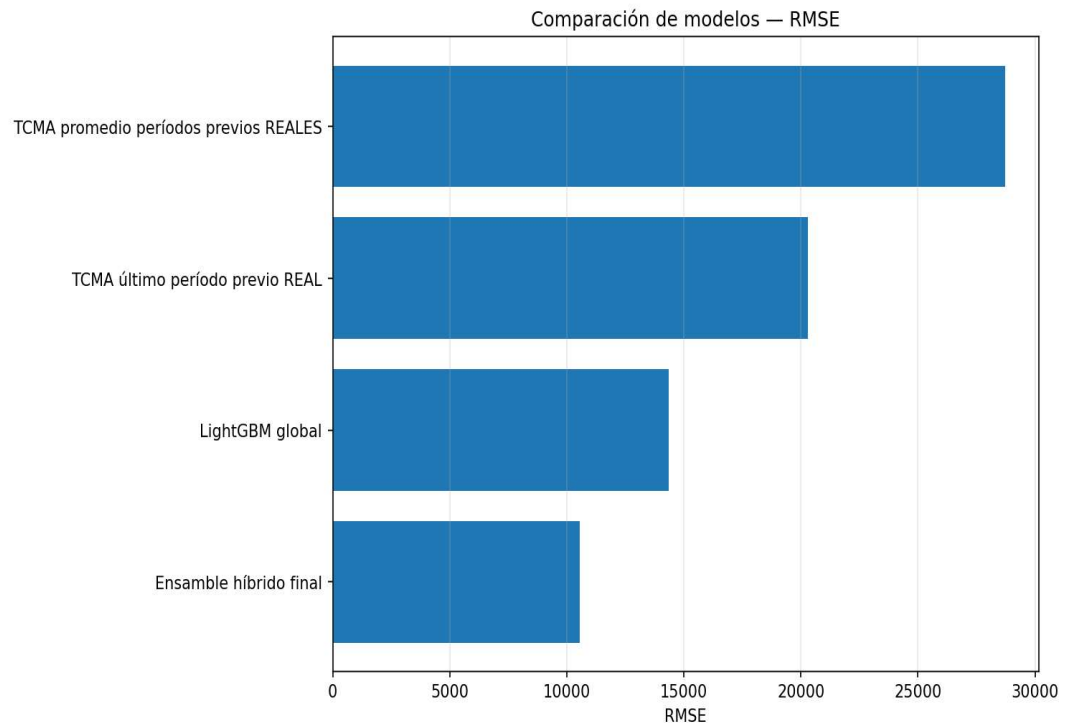
El Script 23 combinó este LightGBM con la TCMA del periodo previo mediante una búsqueda en grilla sobre el peso de cada componente.

El peso óptimo resultó 0.65 para LightGBM y 0.35 para la TCMA, con un MAPE de 9.56%, igual al del LightGBM individual, pero con menor RMSE (10,546 frente a 14,354). Este ensamble es el que aparece como "Ensamble híbrido" en la tabla maestra de la sección 4.2.2.

La Figura 9 muestra el resultado de esa búsqueda de peso frente a los dos modelos por separado.

Figura 9

Comparación del ensemble híbrido frente a LightGBM y TCMA por separado (Script 23)



Antes de avanzar a modelos más especializados, el Script 24 aplicó un análisis de valores SHAP sobre el LightGBM global, con el fin de verificar que las predicciones se apoyaran en variables demográficamente coherentes y no en artefactos del dataset.

La Tabla 16 y la Figura 10 muestran el top diez de variables de mayor peso.

Tabla 16

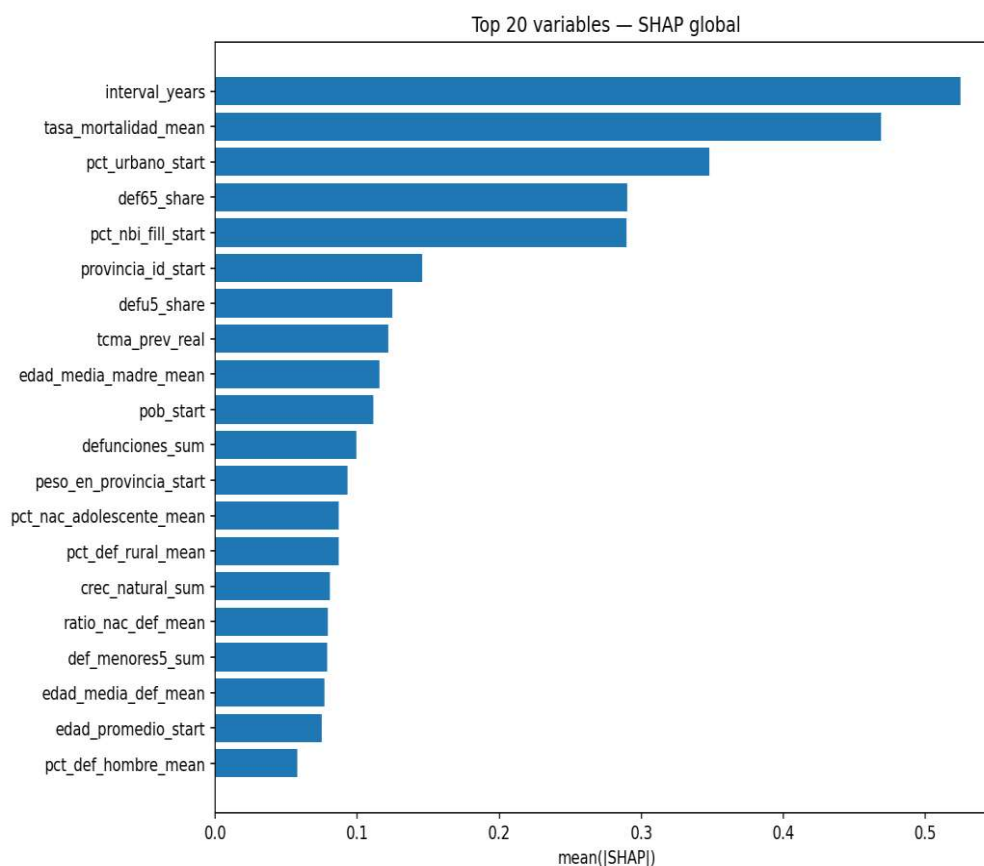
Variables de mayor importancia global según SHAP

Variable	SHAP medio
interval_years (años entre censos)	0.52
tasa_mortalidad_mean	0.47
pct_urbano_start	0.35
def65_share (participación de adultos mayores)	0.29
pct_nbi_fill_start (necesidades básicas insatisfechas)	0.29

provincia_id_start	0.15
defu5_share (mortalidad infantil)	0.12
tcma_prev_real	0.12
edad_media_madre_mean	0.12
pob_start	0.11

Figura 10

Importancia global de variables según SHAP, LightGBM global



La variable de mayor peso es el número de años entre censos, algo esperable dado que el modelo predice una tasa de crecimiento y ese intervalo determina directamente la magnitud del cambio proyectado.

Le siguen la mortalidad, el porcentaje de población urbana y la participación de adultos mayores, todas con respaldo en la literatura de transición demográfica. La identidad de la provincia también pesa de forma notable lo que sugiere efectos regionales no capturados del todo por las variables observables.

4.1.4. Segmentación tipológica y modelos especializados por cantón

El Script 26 agrupó los 221 cantones mediante clustering no supervisado, evaluando silhouette, Calinski-Harabasz y Davies-Bouldin para valores de k entre 3 y 7.

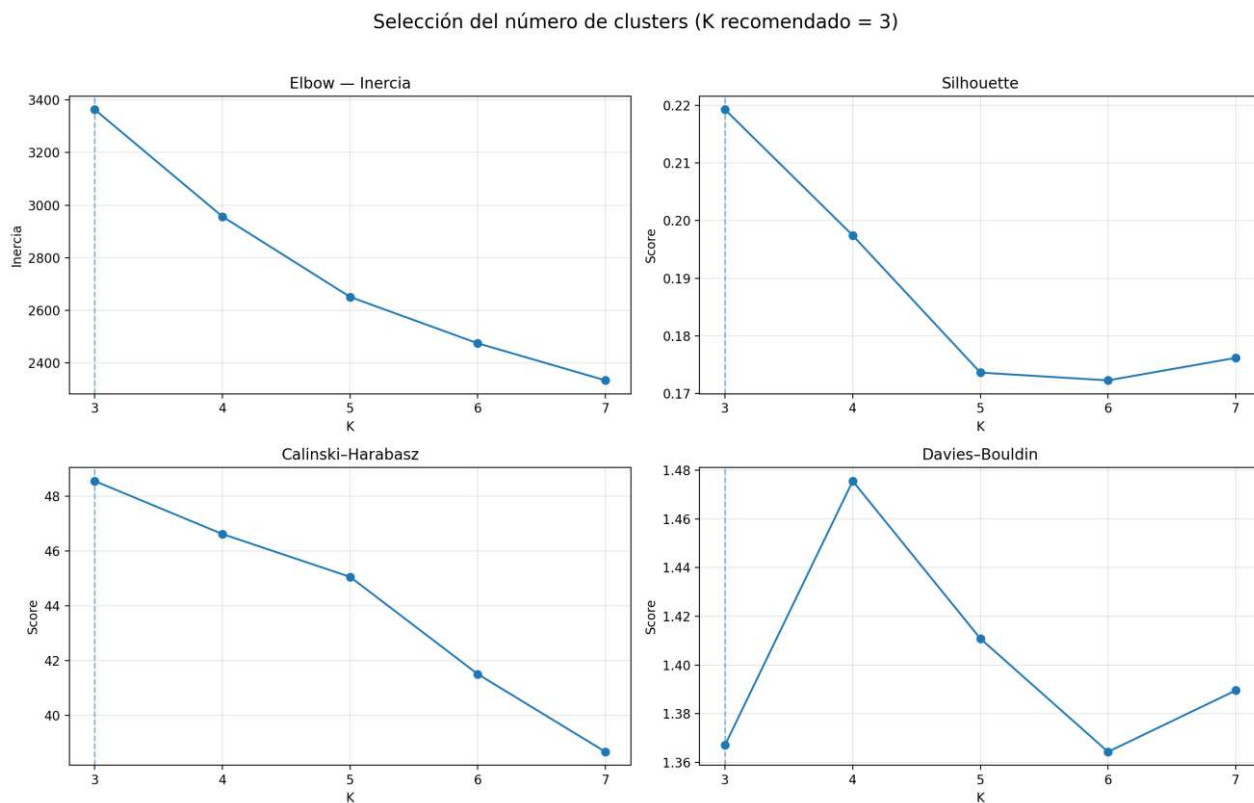
La configuración óptima fue $k=3$ (silhouette=0.22), y dio lugar a tres tipologías: cantones rurales de baja dinámica ($n=130$, crecimiento mediano 7.27%), metrópolis nacionales, Quito y Guayaquil ($n=2$, crecimiento mediano 18.25%), y cantones semi-urbanos de alta expansión ($n=89$, crecimiento mediano 22.91%).

La Figura 11 muestra que la curva de codo (inercia) no presenta un quiebre marcado que sugiera un valor de k , por lo que la decisión se apoyó en dos métricas:

- El índice silhouette que alcanza su valor más alto en $k=3$ (0.22) y decae de forma sostenida a partir de ahí.
- El índice de Calinski-Harabasz que sigue el mismo patrón, con su máximo también en $k=3$ (48.55 frente a 46.62 en $k=4$).
- El índice Davies-Bouldin es la única métrica que favorece, por un margen mínimo, a $k=6$ (1.364) sobre $k=3$ (1.367); dado que esa diferencia es prácticamente despreciable y las otras dos métricas señalan a $k=3$ con un margen bastante más amplio, se mantiene $k=3$ como configuración final.

Figura 11

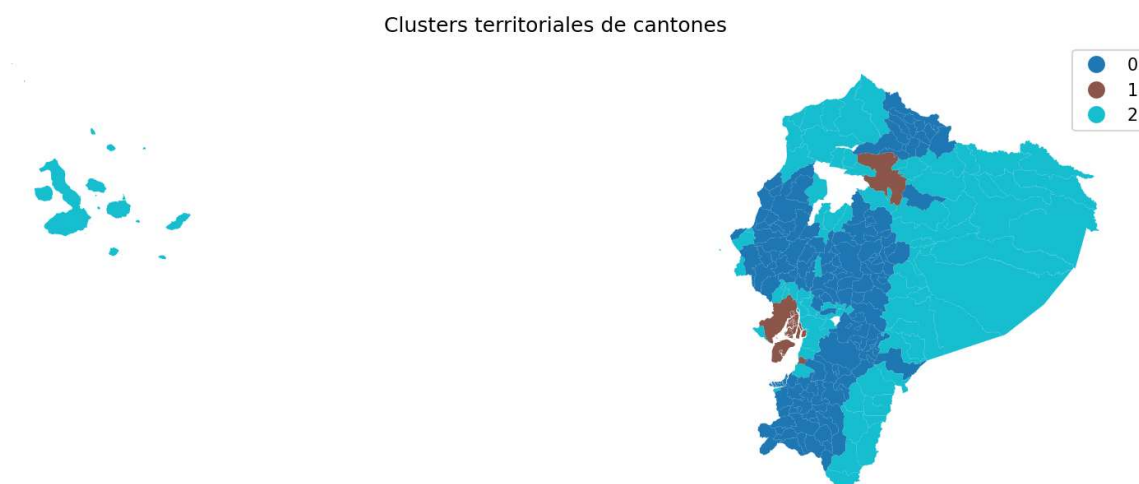
Selección del número de clústeres: inercia, silhouette y Davies-Bouldin por valor de k (Script 26)



La Figura 12 ubica geográficamente las tres tipologías resultantes. El mapa confirma que la tipología rural de baja dinámica domina la mayor parte del territorio continental, mientras que los cantones semi-urbanos de alta expansión se concentran alrededor de las dos metrópolis y en la franja costera e insular.

Figura 12

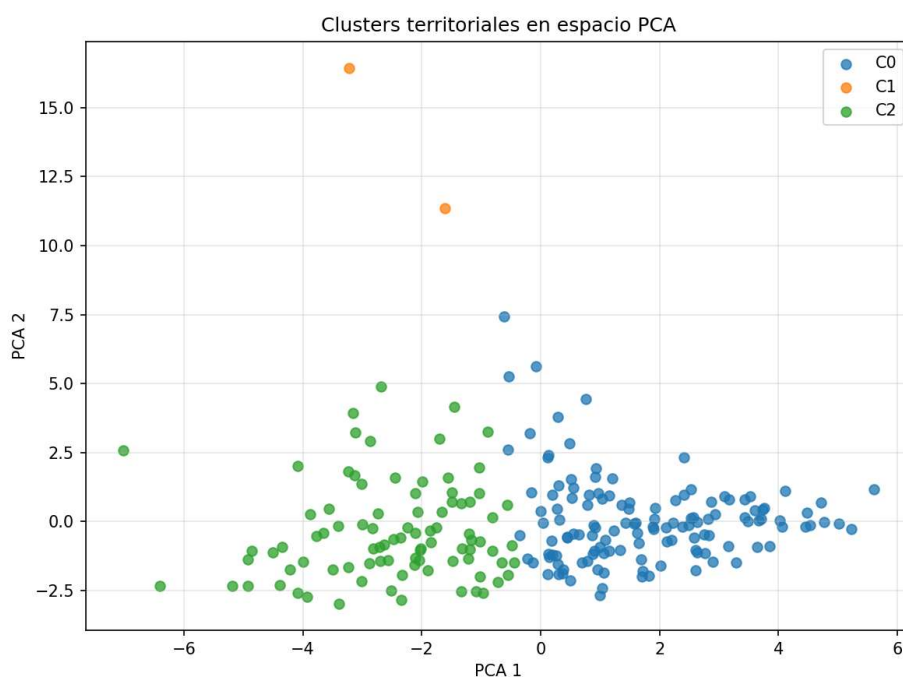
Distribución geográfica de las tres tipologías de cantones (Script 26)



La Figura 13 confirma esta separación en el espacio de variables, no solo en el mapa: los tres grupos ocupan regiones distinguibles al proyectar las variables originales sobre sus dos primeros componentes principales.

Figura 13

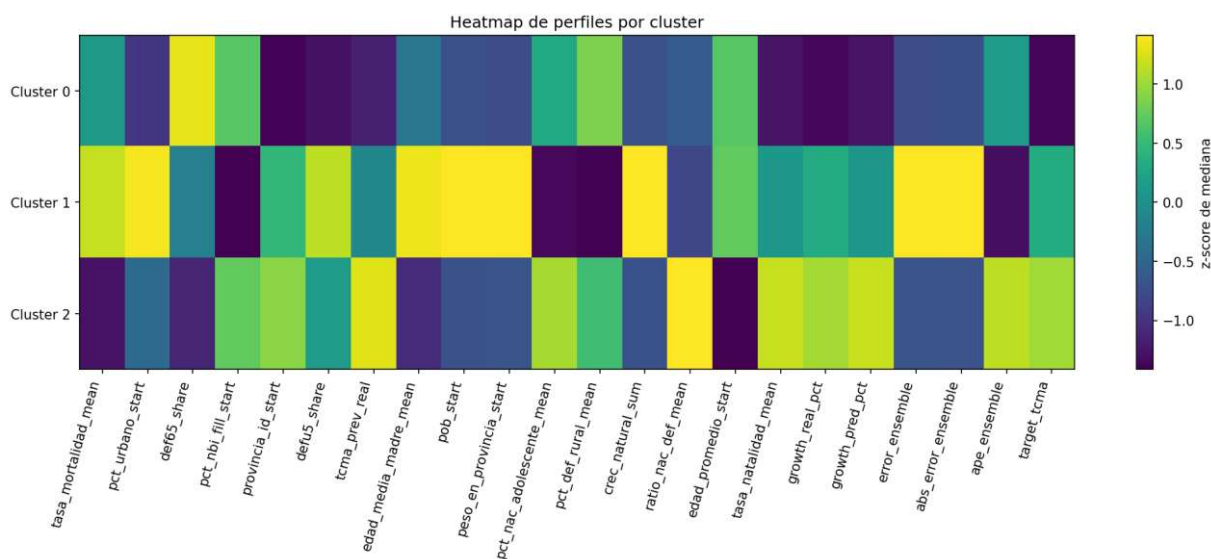
Clústeres territoriales en el espacio de componentes principales (Script 26)



La Figura 14 completa el panorama mostrando qué variables caracterizan a cada tipología: el clúster de metrópolis nacionales destaca por su mortalidad y participación de adultos mayores más bajas, mientras que el clúster rural de baja dinámica muestra el patrón opuesto.

Figura 14

Perfil de variables por tipología de cantón (Script 26)

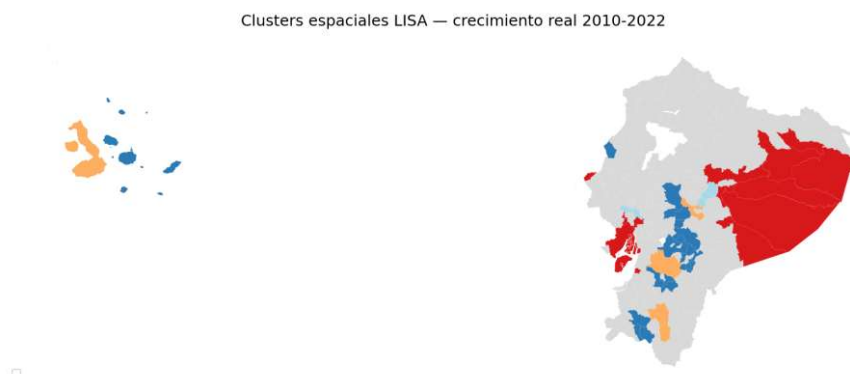


El Script 25, por su parte, midió la autocorrelación espacial mediante el índice de Moran. Para la tasa de crecimiento real, el resultado fue significativo ($I=0.36$, $p=0.001$), es decir, los cantones con crecimiento similar tienden a agruparse geográficamente. Para el error del ensamble híbrido, en cambio, el I de Moran cayó a 0.0175 y dejó de ser significativo ($p=0.28$): el error se distribuye de forma espacialmente aleatoria, sin formar clústeres. Este dato anticipa el hallazgo de la sección 4.2.5, porque un error aleatorio es justamente el tipo de error que se cancela al agregar cantones en provincias.

La Figura 15 ubica los clústeres LISA sobre el mapa del país: los High-High (rojo) se concentran principalmente en la Amazonía, y los Low-Low (azul) en la sierra centro, lo que sugiere que la dinámica de crecimiento no solo se agrupa estadísticamente, sino que responde a patrones regionales reconocibles.

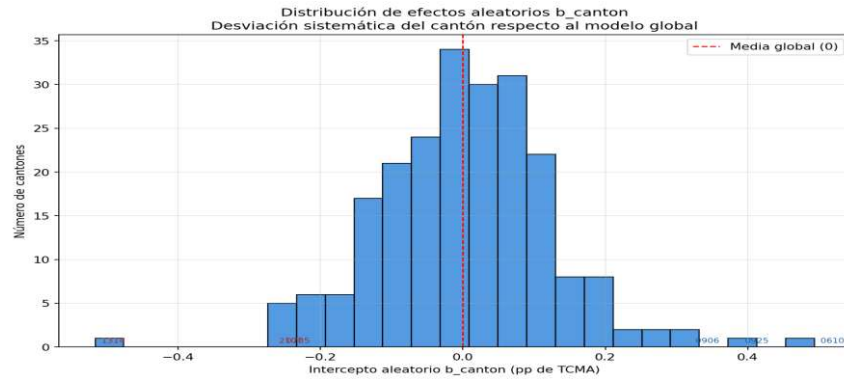
Figura 15

Clústeres espaciales LISA del crecimiento poblacional real, 2010-2022 (Script 25)



Con la tipología ya definida, se probaron dos extensiones del LightGBM orientadas a capturar heterogeneidad geográfica. El Script 27b incorporó efectos aleatorios por cantón bajo un esquema tipo BLUP, con contracción hacia la media para los cantones con menos datos (Hajjem et al., 2014; Bandara et al., 2022); obtuvo un MAPE de 9.55% y un RMSE algo menor que el LightGBM puro (13,769 frente a 14,354), aunque sin mejora en MAPE. El Script 27, en cambio, entrenó un LightGBM independiente por tipología sin corregir aún el problema de fuga de datos, y el resultado fue malo: 30.83% en evaluación directa y entre 158% y 364% en validación cruzada, peor que el propio benchmark de TCMA. La causa fue la misma que se corrigió después en el Script 28: la variable `target_tcma` seguía dentro del conjunto de entrenamiento.

La Figura 16 muestra la distribución de esos efectos aleatorios estimados por el Script 27b: la gran mayoría se concentra cerca de cero, coherente con que 136 de 221 cantones presenten una desviación menor a 0.1 puntos porcentuales respecto al promedio global.

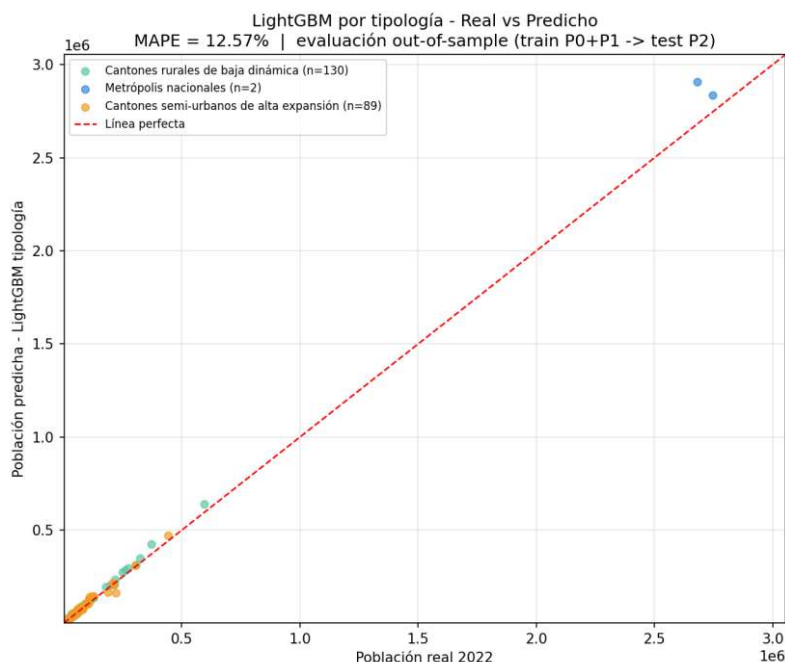
Figura 16*Distribución de efectos aleatorios por cantón (Script 27b)*

La versión corregida de ese modelo por tipología (Script 28) sí forma parte de la tabla final. Su primera ejecución había reportado un MAPE de 7.14%, evaluado sobre los mismos datos de entrenamiento y con `target_tcma` todavía presente como variable, dos problemas que en conjunto explican un resultado artificialmente bajo. Al aplicar partición temporal estricta (entrenamiento 1990-2010, prueba exclusiva sobre 2010-2022) y retirar esa variable, el MAPE real subió a 12.57%, con un sesgo nacional de +7.54%. Su promedio queda por debajo del benchmark de TCMA (14.29%), pero una prueba de Wilcoxon pareada sobre los 221 cantones no encuentra una diferencia estadísticamente significativa entre ambos ($p=0.63$): es el modelo más débil de los seis retenidos en la comparación final, y el único cuya ventaja sobre el benchmark no se sostiene con rigor estadístico.

La Figura 17 presenta el ajuste real de esta versión corregida.

Figura 17

Población real frente a predicha, LightGBM por tipología corregido (Script 28)



4.1.5. Método Hamilton-Perry amortiguado y búsqueda del modelo óptimo

El Script 29 aplicó el método demográfico de cohortes Hamilton-Perry (Hamilton & Perry, 1962), que proyecta la población futura extrapolando la razón de cambio entre los dos últimos censos: $P(t+n) = P(t) \times [P(t)/P(t-k)]^{(n/k \times d)}$. Se incorporó un factor de amortiguación $d=0.5$ para moderar el crecimiento extremo que la razón censal cruda tiende a proyectar: con $d=1$ la fórmula equivale a la razón censal sin ajustar, y con $d=0.5$ se modera a la mitad el exponente de crecimiento. En retrodicción sobre 2010-2022, este método obtuvo un MAPE cantonal de 7.79%, con un sesgo nacional de -3.56%. Ningún otro modelo del pipeline, ni siquiera los de machine learning, lo superó a nivel cantonal, pese a que Hamilton-Perry no usa más que dos observaciones censales y ningún entrenamiento.

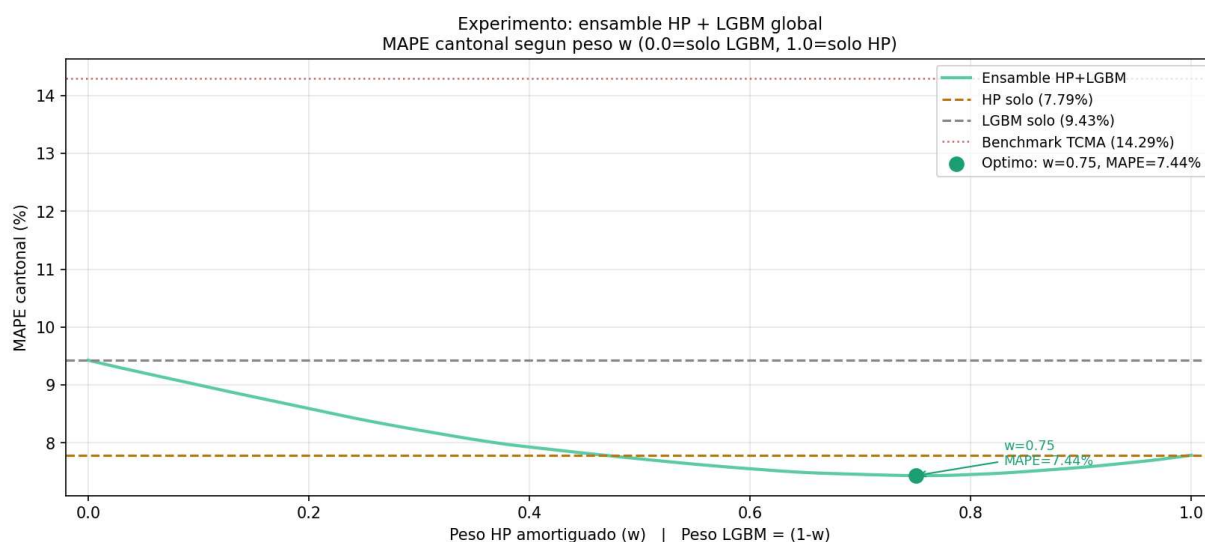
Para verificar si $d=0.5$ es un valor razonable o simplemente arbitrario, se realizó un barrido del factor de amortiguación entre $d=0.00$ y $d=1.00$ (incremento de 0.01), midiendo el MAPE cantonal agregado sobre el mismo periodo de retrodicción (proyección de 2001 y 2010 hacia 2022).

Con Hamilton-Perry y LightGBM global como los dos mejores modelos individuales, se probó si combinarlos mejoraba el resultado. El Script 30 hizo una búsqueda en grilla sobre

el peso de cada uno en un ensemble ponderado (siguiendo a Grossman et al., 2022) y encontró un óptimo de 0.75 para Hamilton-Perry y 0.25 para LightGBM, con MAPE de 7.44%, una mejora aparente de 0.35 puntos sobre Hamilton-Perry solo. La Figura 18 muestra esa curva de búsqueda.

Figura 18

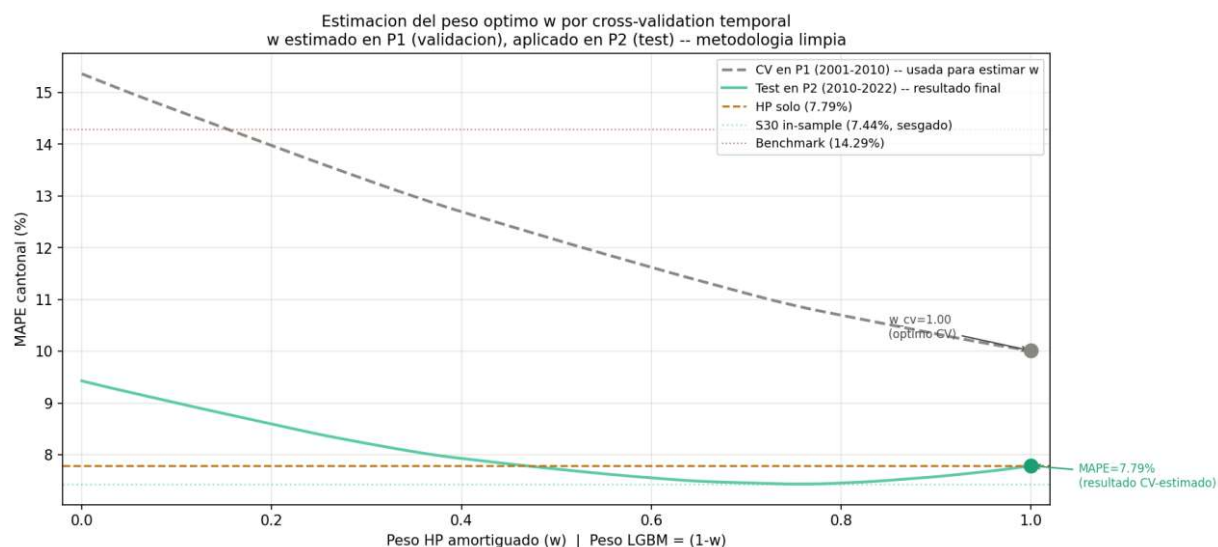
MAPE del ensemble según el peso asignado a Hamilton-Perry (Script 30)



El Script 31 sometió ese resultado a validación cruzada temporal, estimando el peso únicamente sobre el periodo 2001-2010, usado como validación, y aplicándolo sin reajustar sobre el periodo de prueba 2010-2022. Bajo ese esquema más exigente, el peso óptimo convergió a 1.00 para Hamilton-Perry, es decir, el ensemble colapsa de vuelta al modelo individual, con el mismo MAPE de 7.79%. La mejora del Script 30 resultó ser un artefacto de haber optimizado el peso directamente sobre el conjunto de prueba, no una ganancia real. La Figura 19 deja ver esta diferencia entre el MAPE estimado en validación cruzada y el obtenido al sobreajustar sobre el conjunto de prueba.

Figura 19

MAPE de validación cruzada frente a MAPE de prueba, por peso del ensemble (Script 31)

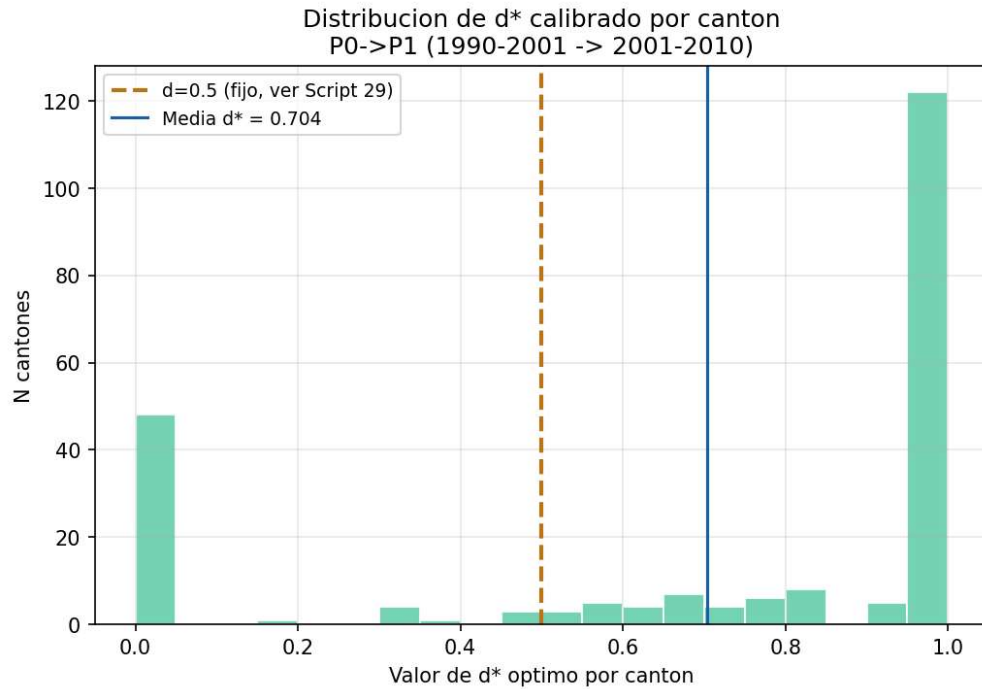


Como último método de experimentación (Script 32) en lugar de fijar $d=0.5$ para todos los cantones, se intentó con Ridge y LightGBM. La calibración histórica del d óptimo mostró una media de 0.70 y una desviación de 0.40, bastante lejos del 0.5 fijo, y el modelo sí encontró señal real en el problema (MAE de 0.15 frente a un baseline de 0.43 usando siempre $d=0.5$). Pero al aplicar esos valores aprendidos sobre el periodo de prueba, el resultado empeoró: 12.25% con Ridge y 10.04% con LightGBM, ambos peores que el 7.79% del d fijo.

La Figura 20 muestra esa distribución del d óptimo calibrado por cantón.

Figura 20

Distribución del factor de amortiguación d^ calibrado por cantón (Script 32)*



En lugar de un d por cantón, se calibró un único d global. Un primer barrido probó cada valor entre 0.00 y 1.00 midiendo el error directamente contra el periodo de prueba, y encontró un mínimo en $d=0.55$ (MAPE=7.71%). Este resultado representa un techo optimista, el mejor caso posible si se conociera de antemano el resultado a predecir, no una estimación confiable, ya que usa la misma información que después se reporta como resultado.

Para obtener una estimación honesta, se repitió el ejercicio usando el único periodo intercensal anterior disponible: proyectar 2010 a partir de los censos de 1990 y 2001, sin haber observado 2022 todavía. En ese periodo, el óptimo se ubica en $d=0.97$ (MAPE=7.96%), prácticamente sin amortiguar. La Tabla 17 y la Figura 21 muestran ambas curvas.

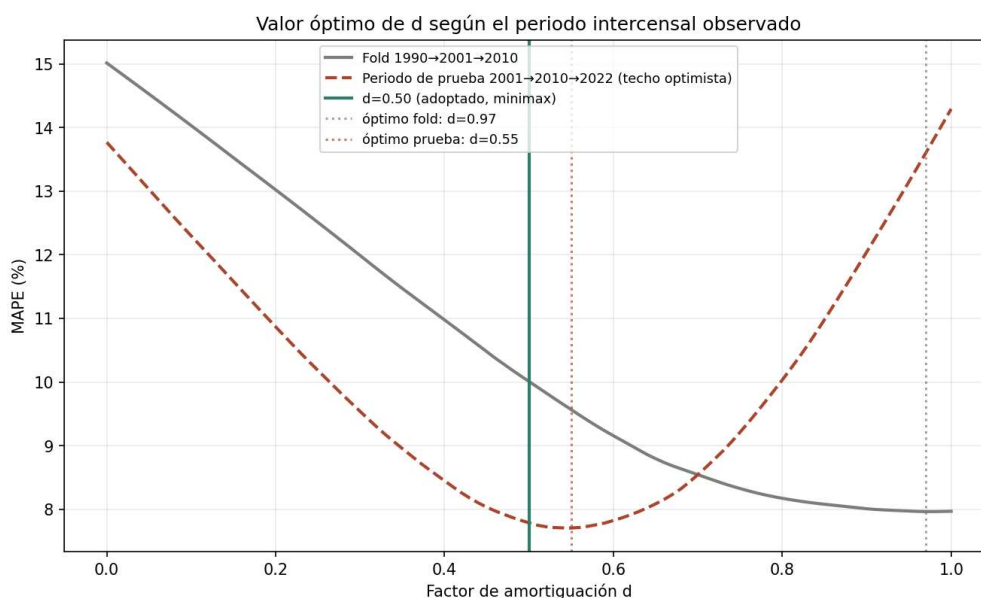
Tabla 17

Valor óptimo de d según el periodo intercensal observado

Criterio	d	MAPE
Barrido directo (techo optimista, periodo de prueba)	0.55	7.71%
Barrido en el fold anterior (1990→2001→2010)	0.97	7.96% (en el fold)
Ese mismo $d=0.97$, aplicado sin reajustar al periodo de prueba real	—	13.60%
Adoptado (criterio minimax, Wald 1950)	0.5	7.79%

Figura 21

Valor óptimo de d según el periodo intercensal observado (Script 29)



Al aplicar $d=0.97$ sobre el periodo de prueba real, sin haber visto el censo 2022 durante su calibración, el resultado es $\text{MAPE}=13.60\%$, cercano al benchmark de TCMA (14.29%). El valor que mejor funciona en un periodo no generaliza al siguiente: los dos únicos periodos intercensales disponibles en Ecuador arrojan óptimos muy distintos entre sí (0.97 y 0.55), y no hay un tercer periodo para saber cuál de los dos, si alguno, es representativo.

El ensamble ponderado, su corrección por validación cruzada, el aprendizaje de un d por cantón, y el barrido de un d global evaluado de dos formas: los cuatro intentos de calibrar d con los datos disponibles fracasan en generalizar del periodo donde se calibran al periodo de prueba. Por eso se adopta $d=0.5$ mediante un criterio minimax (Wald, 1950): sin información confiable sobre dónde cae el verdadero valor óptimo dentro de $[0,1]$, el punto

medio minimiza el error máximo posible, en vez de perseguir el óptimo de un periodo que no se sostiene en el siguiente.

4.2. Análisis de Resultados

Con el modelo cantonal ya definido en las secciones previas, esta sección reúne los resultados finales del pipeline: los Scripts 33 al 38, que cubren la validación retrospectiva completa, el comportamiento del error según el nivel de agregación geográfica y las proyecciones 2030-2035.

4.2.1. Marcos de evaluación

Los resultados de este capítulo se reportan bajo dos marcos distintos. El Marco A compara contra el censo real de 2022 y da un benchmark de TCMA de 8.65%. El Marco B, usado en los Scripts 33 a 38, compara contra las proyecciones oficiales EPP INEC 2024 y sube el benchmark a 14.29%. La diferencia entre ambos (unos 4.5 puntos) no es un error propio: el análisis del Script 33 encontró que el INEC sobreestimó la población nacional de 2022 en +4.58% (776,315 habitantes) frente al censo real, mientras que el LightGBM global de este trabajo estimó la misma magnitud con un error de apenas +0.49%. Mezclar los dos marcos al comparar modelos habría sido un error metodológico; por eso se mantienen separados a lo largo de todo el capítulo.

4.2.2. Tabla maestra de resultados

El Script 36 consolidó la validación retrospectiva 2010-2022 de los seis modelos propios, más la referencia externa del INEC y el benchmark de TCMA. La Tabla 18 y la Figura 22 resumen ese resultado.

Tabla 18

Validación retrospectiva 2010-2022 (Marco B)

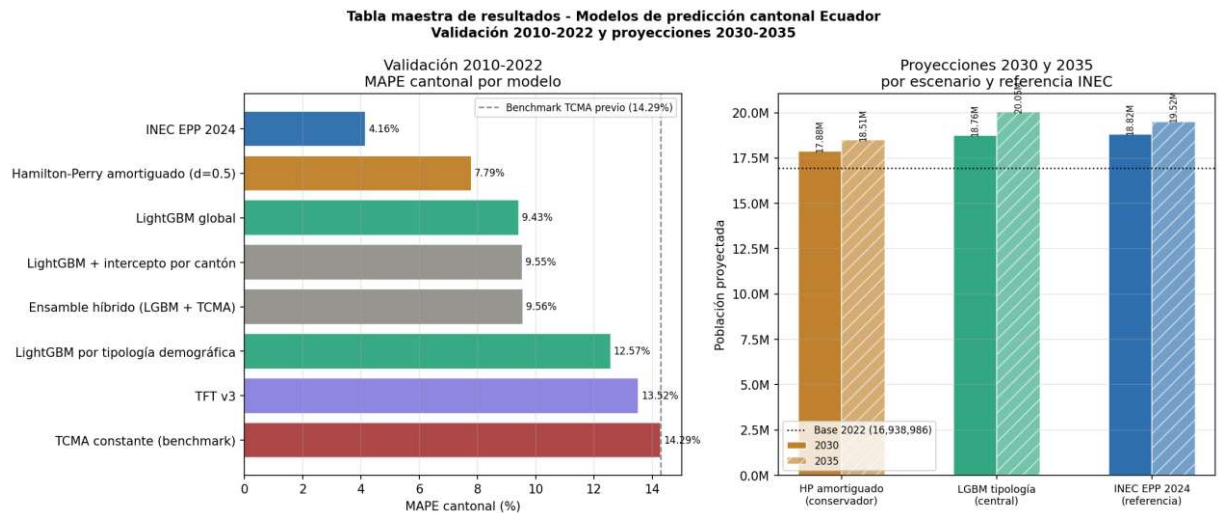
Modelo	MAPE cantón	MAPE provincia	Error nacional	¿Supera benchmark?
INEC EPP 2024 (referencia externa)	4.16%	3.81%	+4.58%	-
Hamilton-Perry amortiguado	7.79%	7.99%	-3.56%	Sí
LightGBM global	9.43%	5.45%	+0.49%	Sí
LightGBM + intercepto por cantón	9.55%	5.46%	+0.69%	Sí
Ensamble híbrido	9.56%	6.49%	+3.51%	Sí
LightGBM por tipología	12.57%	8.59%	+7.54%	Sí

TFT v3	13.52%	n/d	n/d	Sí
TCMA constante (benchmark)	14.29%	12.81%	+9.12%	Referencia

Nota. n/d: métrica no calculada para ese modelo. Benchmark: TCMA constante, MAPE=14.29% (Marco B).

Figura 22

Comparación de MAPE cantonal por modelo



Tres lecturas se desprenden de la tabla. Hamilton-Perry amortiguado es el más preciso a nivel cantonal (7.79%); LightGBM global lo es a nivel provincial (5.45%) y también el de menor sesgo a nivel nacional (+0.49%); y los seis modelos propios superan el benchmark de TCMA, aunque TFT v3 lo hace por el margen más estrecho de todos (13.52% frente a 14.29%).

4.2.3. Diagnóstico de errores por tamaño de cantón

El Script 35 desagregó el error por quintil de tamaño poblacional (Q1: cantones con menos de 10,000 habitantes; Q5: los más grandes) y por sesgo sistemático. Un patrón se repite en todos los modelos: los cantones pequeños son los que peor se predicen. En TCMA constante, Q1 llega a 13.9% frente a 12.7% en Q5; en LightGBM global, 9.9% frente a 8.2%; en LightGBM con intercepto, 10.0% frente a 8.1%. La escasez de datos censales afecta más a los cantones rurales pequeños que a los grandes, sin importar qué modelo se use.

También se confirma que Hamilton-Perry, con el mejor MAPE cantonal de todos, arrastra un sesgo negativo de -3.56% a nivel nacional: el factor de amortiguación modera el crecimiento proyectado, y eso subestima la población total en periodos de expansión. Las Figuras 23 y 24 ilustran ambos patrones.

Figura 23

Sesgo sistemático nacional por modelo

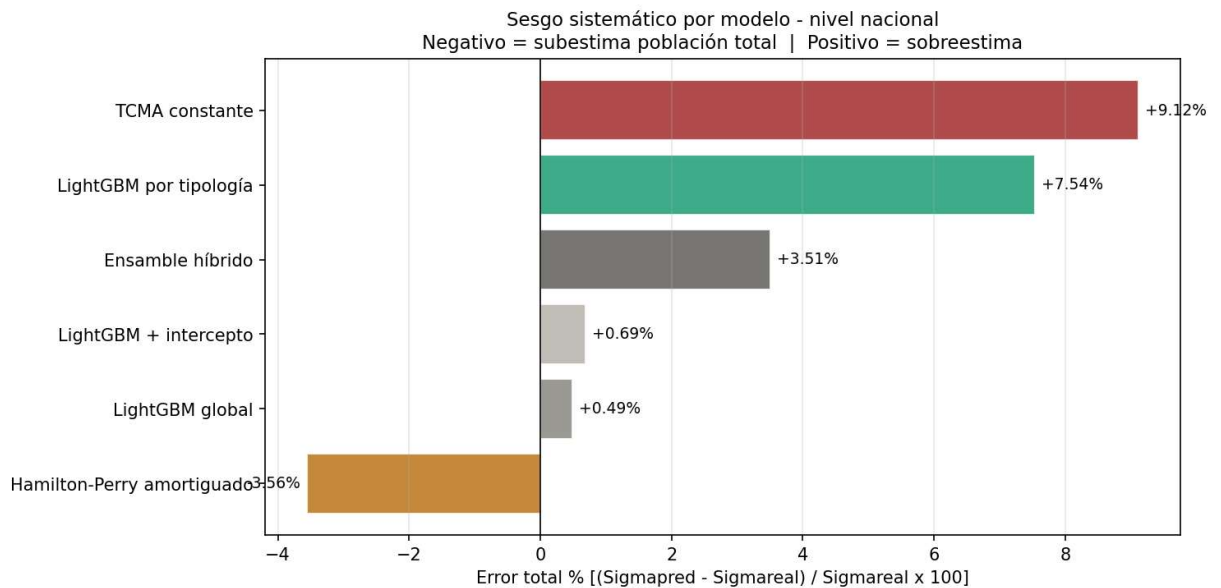
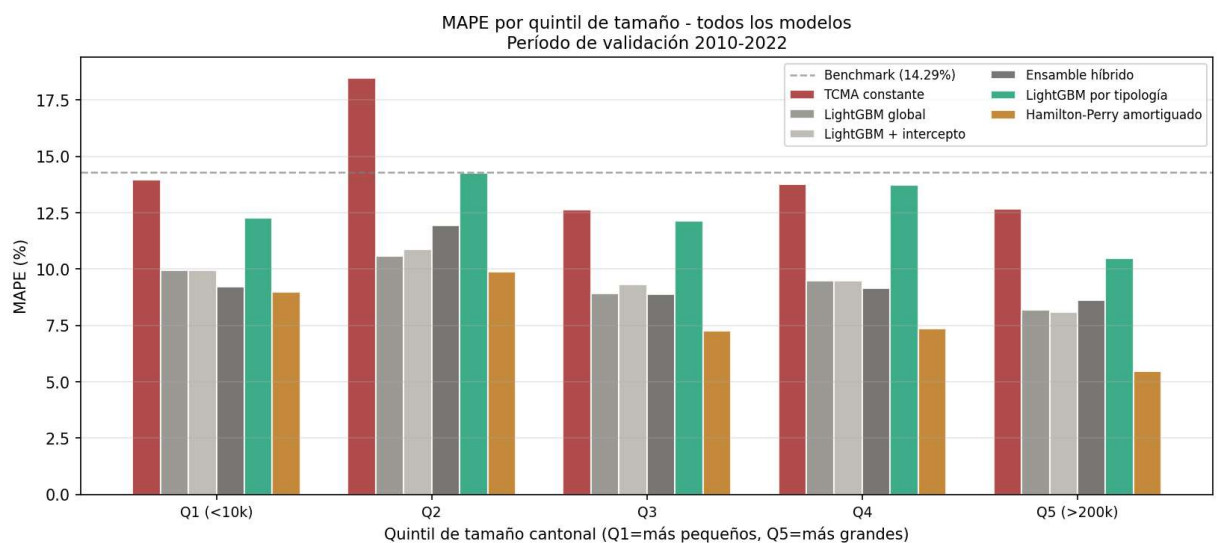


Figura 24

MAPE por quintil de tamaño poblacional



4.2.4. Análisis provincial

A nivel de las 24 provincias, el Script 33 encontró que la proyección oficial del INEC (3.81%) todavía supera al mejor modelo propio, LightGBM global (5.45%), por 1.64 puntos. El Script 37 comparó cantón por cantón: Hamilton-Perry gana en 76 de los 221 (34.4%). Sus mejores provincias son Santo Domingo, Los Ríos y Santa Elena; las peores, Sucumbíos, Chimborazo y Pichincha. Las Figuras 25 y 24 muestran ese conteo y su desagregación por quintil de tamaño.

Figura 25

Cantones ganados por Hamilton-Perry frente a los modelos de ML, por provincia

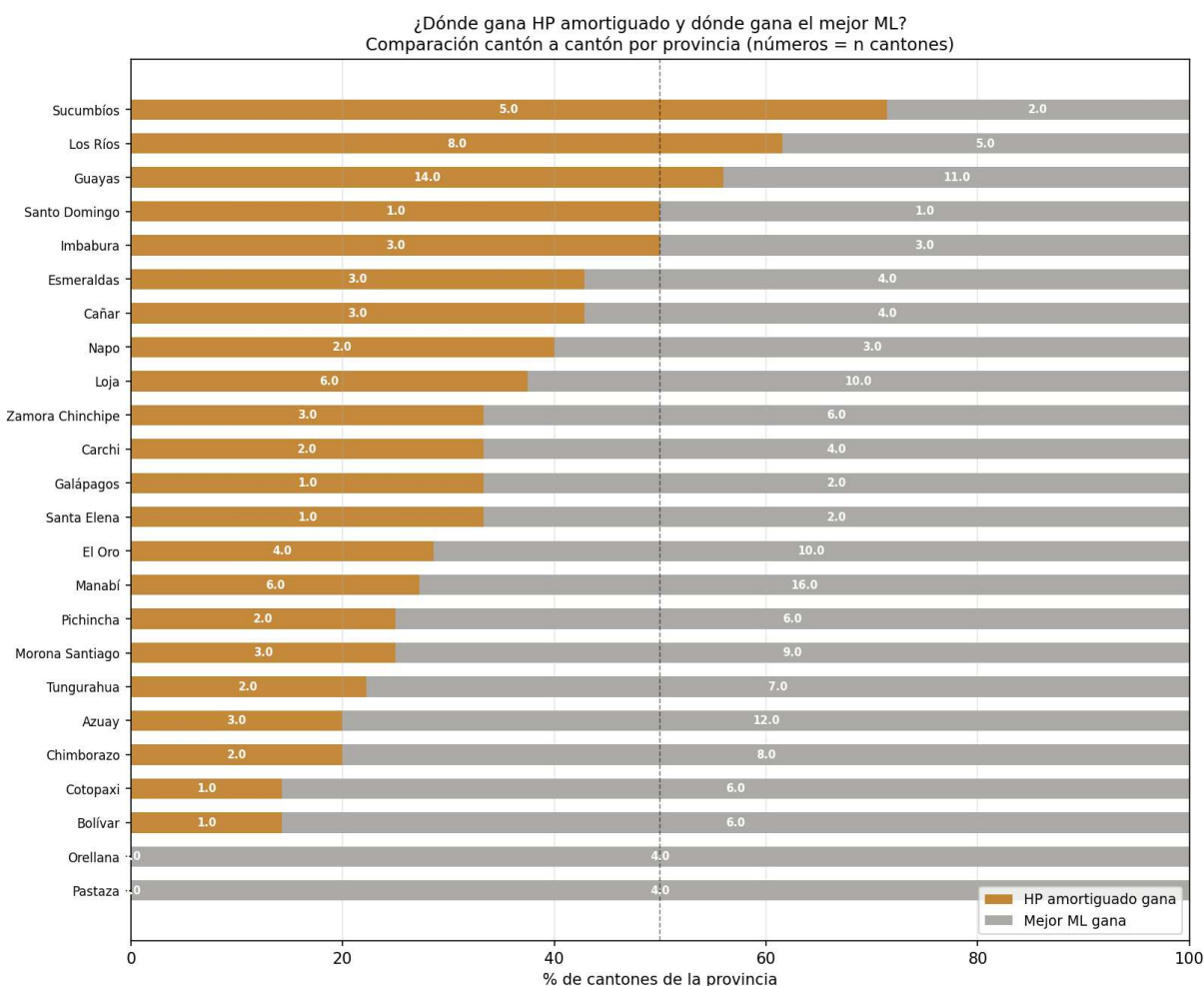
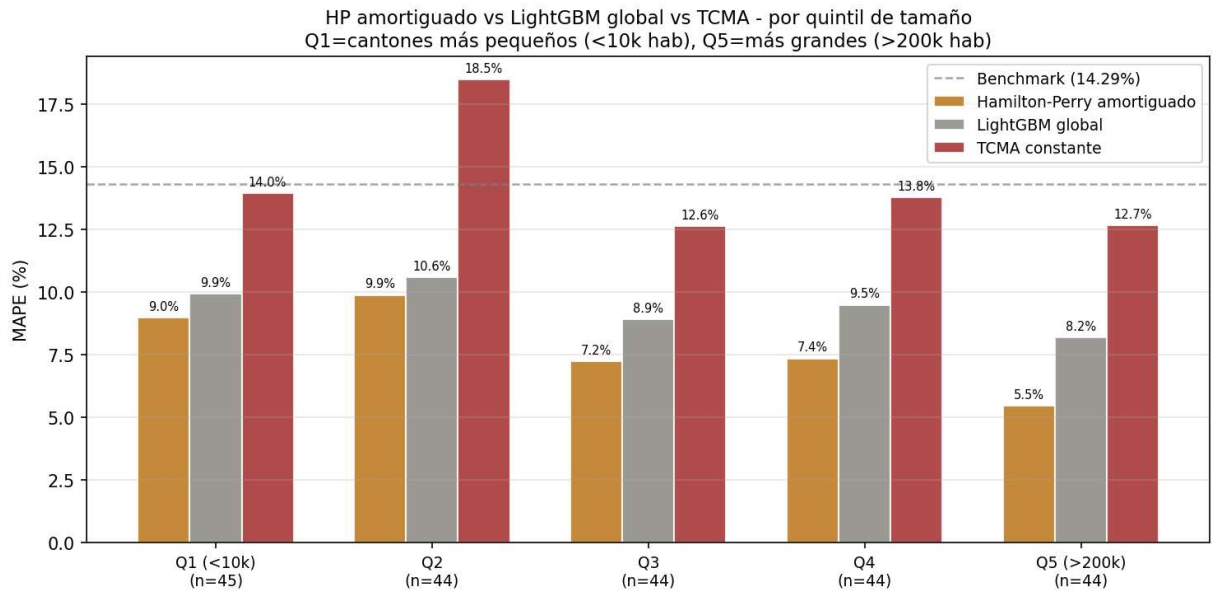


Figura 26

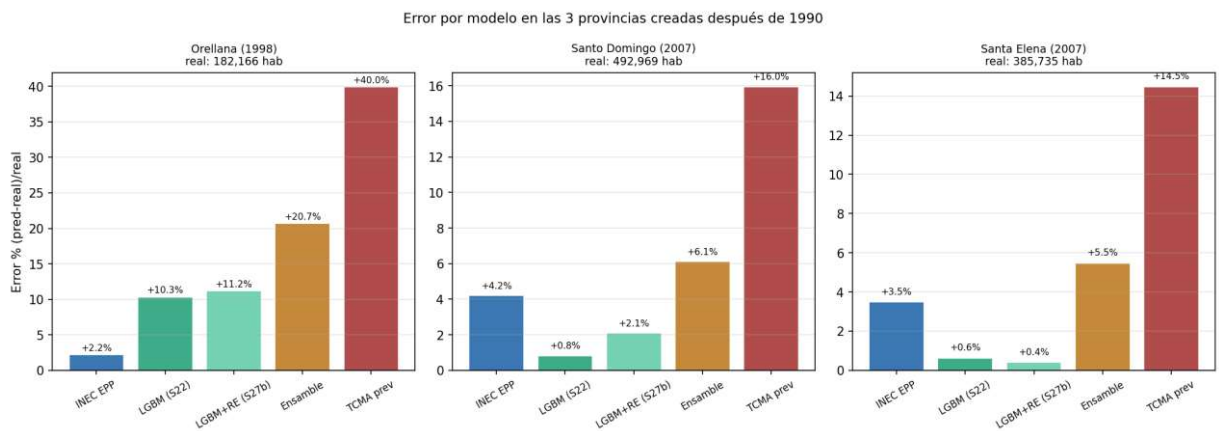
Hamilton-Perry frente a ML, por quintil de tamaño poblacional



Un caso particular son las provincias creadas después del censo base de 1990 (Orellana en 1998, Santo Domingo y Santa Elena en 2007). En Orellana, el INEC mantiene ventaja (+2.2% frente a +10.3% de LightGBM); en Santo Domingo y Santa Elena ocurre lo contrario, LightGBM supera claramente al INEC (0.8% y 0.6% frente a 4.2% y 3.5%). La Figura 27 detalla estos tres casos.

Figura 27

Desempeño en las provincias creadas después de 1990



Se identificaron además 42 cantones con contracción poblacional neta entre 2010 y 2022, un fenómeno que ningún modelo del pipeline fue diseñado para capturar de forma explícita, ni siquiera Hamilton-Perry, que asume implícitamente continuidad de la tendencia. La Tabla 19 lista los siete cantones con mayor error promedio; con una sola excepción, todos corresponden a contracciones superiores al 10%.

Tabla 19

Cantones con mayor error de predicción (outliers)

Código DPA	Provincia	Crecimiento real 2010-2022	APE promedio	Menor error
1707	Pichincha	-19.9%	68.1%	LightGBM global
1320	Manabí	-28.1%	63.7%	Hamilton-Perry
2103	Sucumbíos	-11.7%	52.0%	LightGBM global
0113	Azuay	-21.7%	37.9%	LightGBM global
0603	Chimborazo	-31.7%	35.2%	LightGBM + intercepto
0606	Chimborazo	-21.1%	34.0%	LightGBM global
2106	Sucumbíos	+4.6%	30.3%	Hamilton-Perry

Nota. Código DPA: codificación oficial del INEC. Fuente: script37_outliers_diagnostico.csv.

Las Figuras 28, 29 y 30 cierran el análisis provincial con el mapa de error, el heatmap por provincia y modelo, y el desglose por tipología demográfica.

Figura 28

Mapa coroplético de error de predicción por provincia

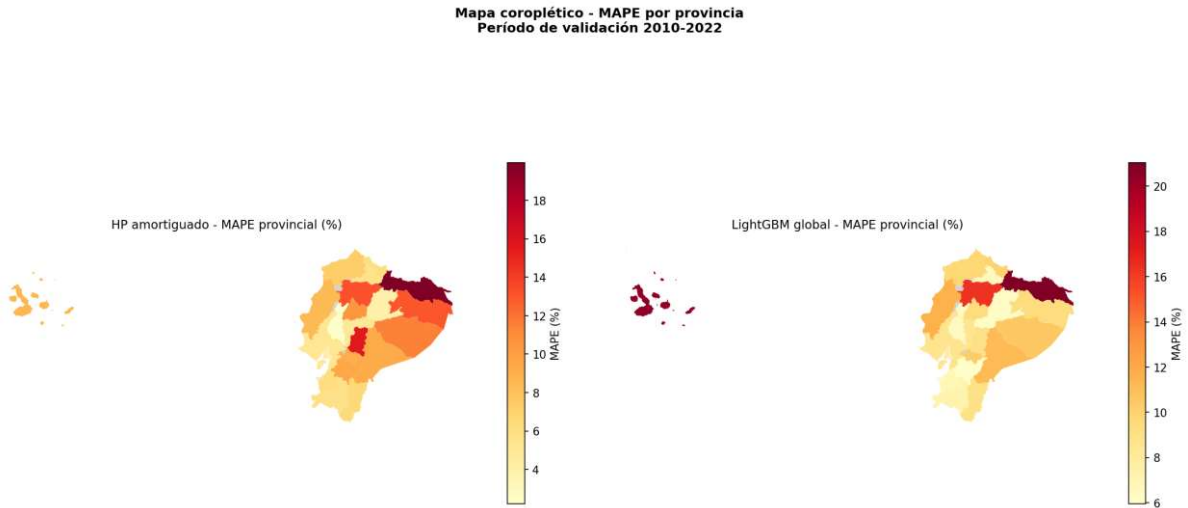


Figura 29

Heatmap de MAPE por provincia y modelo

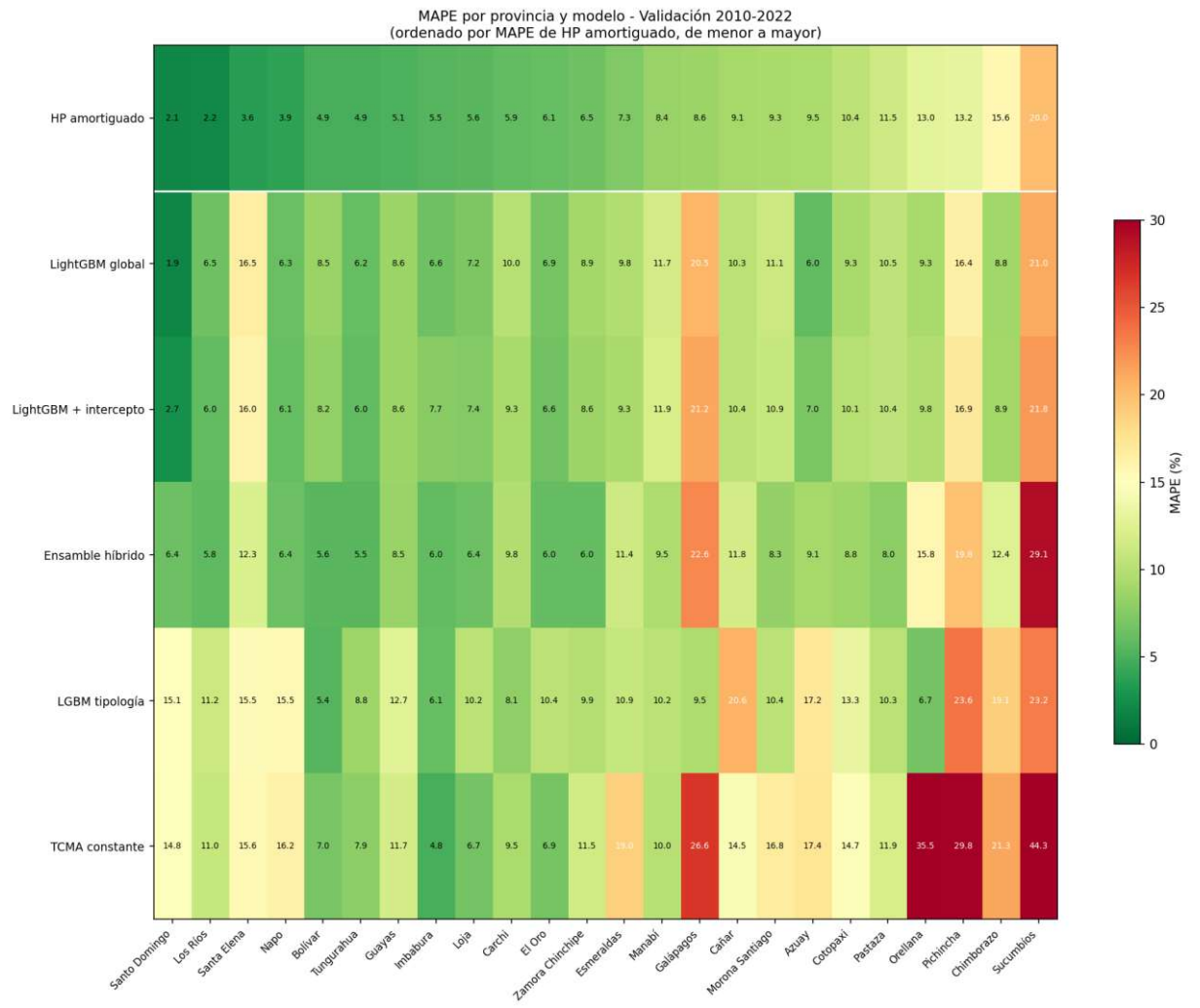
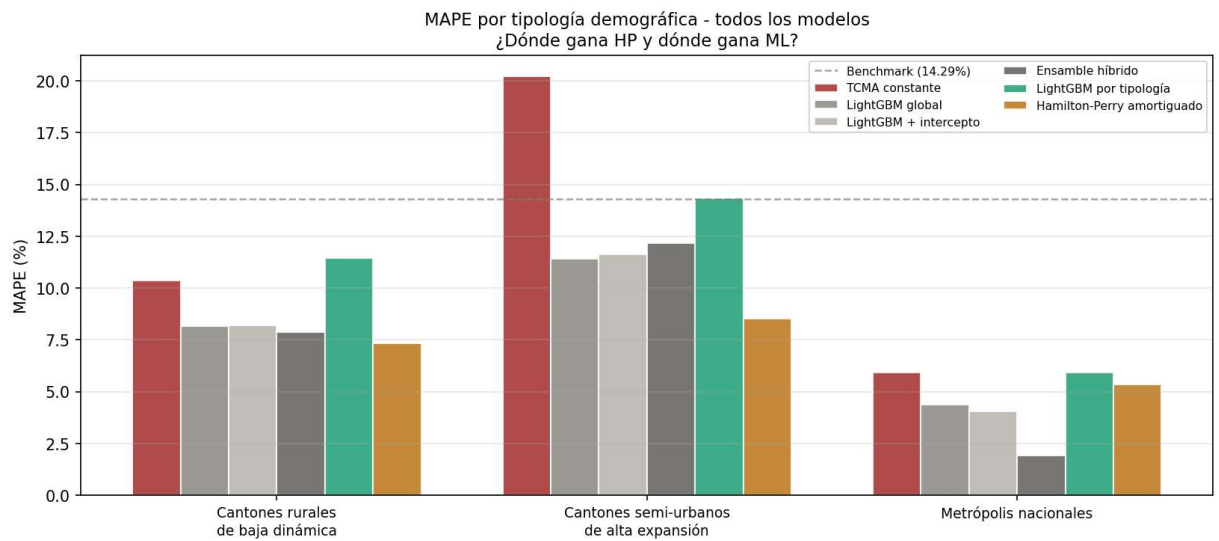


Figura 30

Hamilton-Perry frente a ML, por tipología demográfica



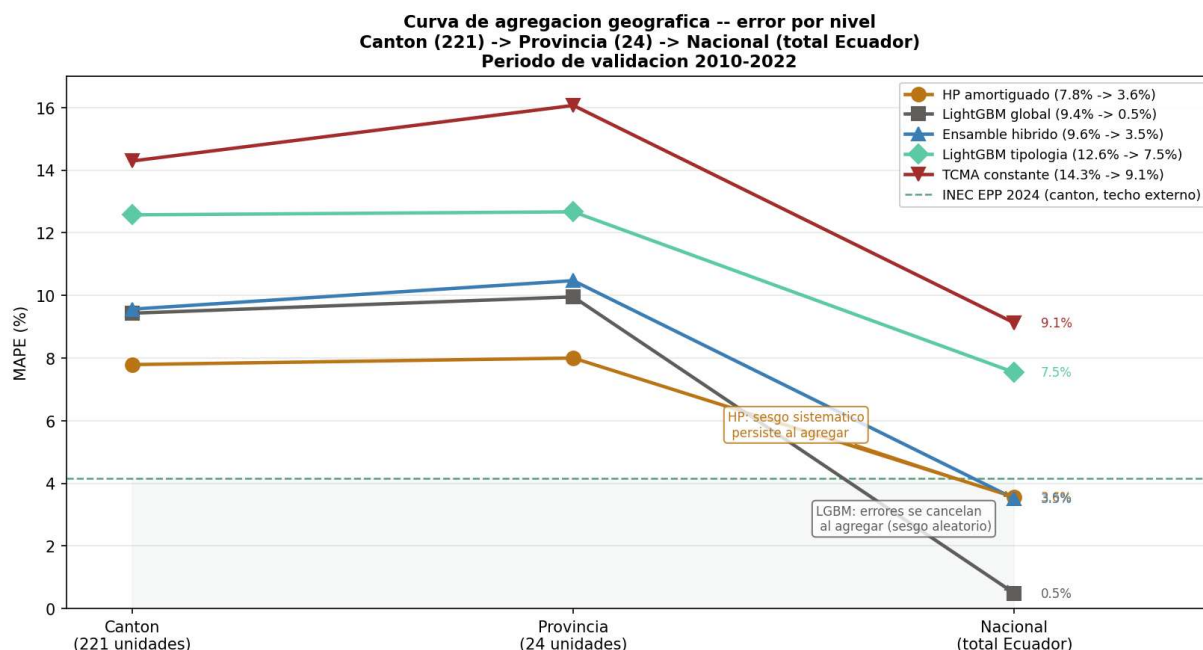
4.2.5. El nivel de agregación geográfica determina el modelo óptimo

El Script 38 cruza los tres niveles de agregación (cantón, provincia, nacional) para cada modelo. La Tabla 20 y la Figura 31 muestran el resultado, que es el hallazgo metodológico central de esta tesis.

Tabla 20

MAPE por modelo y nivel geográfico

Modelo	MAPE cantón	MAPE provincia	Error nacional
Hamilton-Perry amortiguado	7.79%	8.00%	-3.56%
LightGBM global	9.43%	9.95%	+0.49%
Ensamble híbrido	9.56%	10.47%	+3.51%
LightGBM por tipología	12.57%	12.67%	+7.54%
TCMA constante	14.29%	16.06%	+9.12%

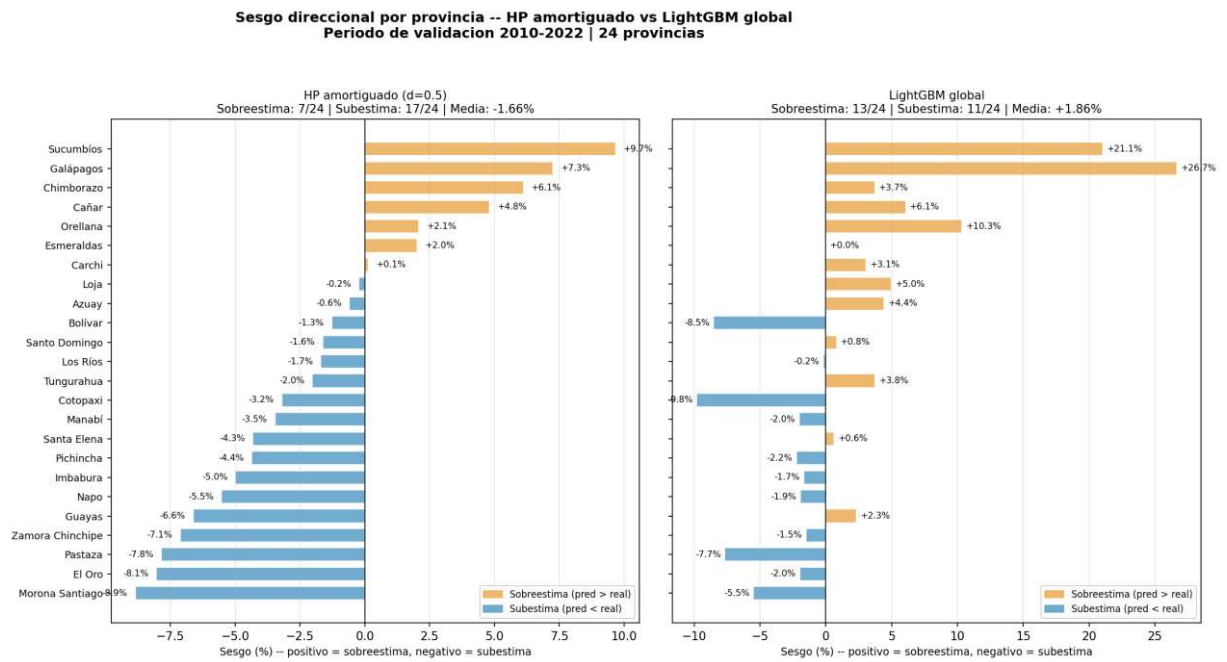
Figura 31*Curva de agregación geográfica del error*

El error de LightGBM se comporta como ruido: al agregar cantones en provincias y luego a nivel nacional, los errores individuales, de signo variable, tienden a cancelarse entre sí, y por eso el MAPE cantonal de 9.43% termina en un sesgo nacional de apenas 0.49%. El de Hamilton-Perry es distinto por naturaleza: el factor de amortiguación introduce una subestimación sistemática que no se cancela al agregar, y persiste como -3.56% incluso a escala nacional. No se encontró en la literatura revisada un patrón equivalente descrito de forma explícita, lo que convierte este resultado en el aporte distintivo del trabajo: no existe un modelo universalmente mejor para todos los niveles de agregación, la elección depende de la escala territorial que se necesite informar.

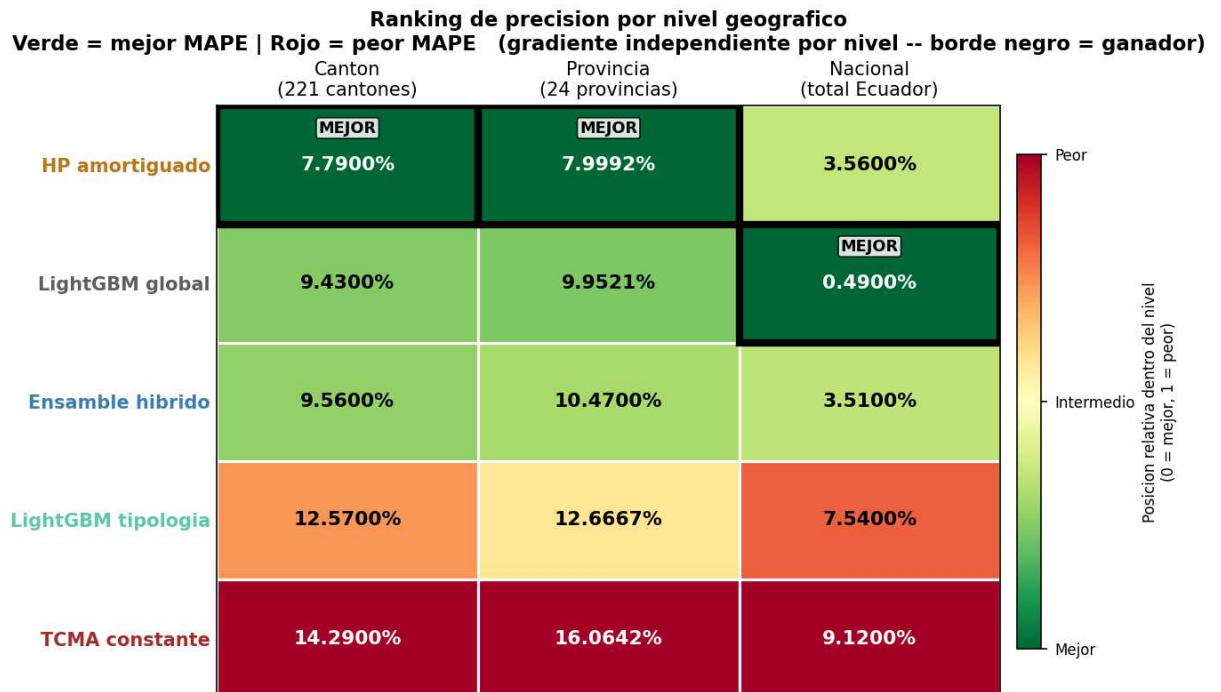
El sesgo direccional por provincia confirma la diferencia: Hamilton-Perry sobreestima en 7 de 24 provincias y subestima en 17, con una media de -1.66%; LightGBM sobreestima en 13 de 24 y subestima en 11, con una media de +1.86%, más cercana a cero y de signo mixto, coherente con un error que tiende a cancelarse. Las Figuras 32 y 33 muestran el sesgo direccional y el ranking de desempeño por nivel.

Figura 32

Sesgo direccional por provincia: Hamilton-Perry frente a LightGBM

**Figura 33**

Ranking de desempeño por modelo y nivel geográfico



En resumen: Hamilton-Perry gana a nivel cantonal y por margen estrecho a nivel provincial; LightGBM gana con claridad a nivel nacional; a nivel provincial la diferencia

entre ambos no es concluyente. Esto responde de forma directa a la hipótesis planteada en el Capítulo 1.

4.2.6. Proyecciones 2030 y 2035

El Script 34 proyectó dos escenarios a partir del censo real de 2022 (16,938,986 habitantes): uno conservador con Hamilton-Perry amortiguado, y uno central con LightGBM por tipología, que incorpora registros vitales. Las variables de estadísticas vitales se calcularon como el promedio de 2019 y 2022, excluyendo 2020-2021 para no introducir el sesgo de mortalidad de la pandemia. Para Quito y Guayaquil, por su comportamiento atípico, se usó como respaldo la TCMA 2010-2022. La Tabla 21 y la Figura 34 comparan ambos escenarios contra la proyección oficial.

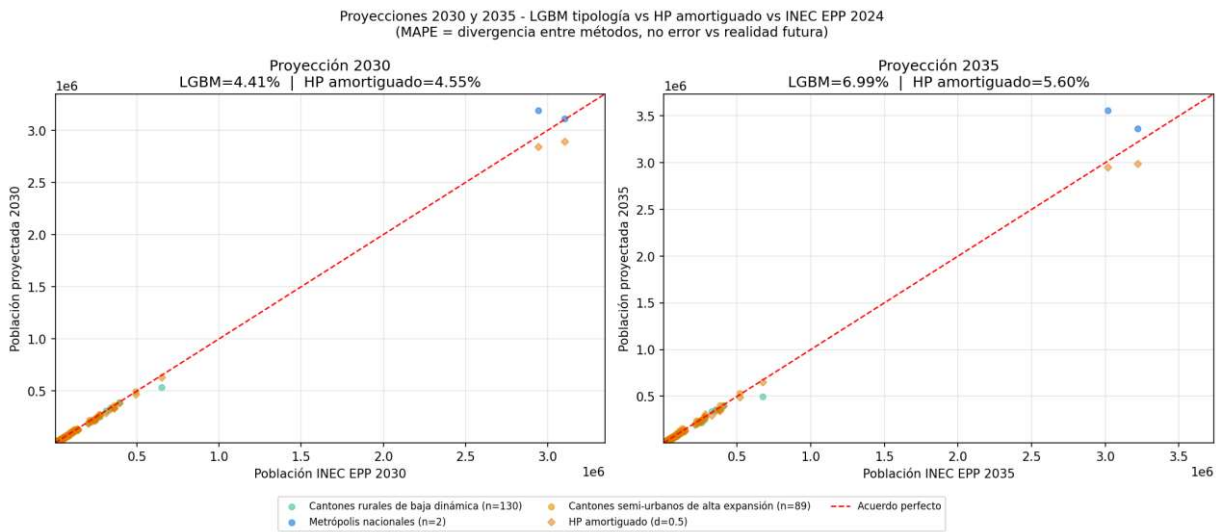
Tabla 21

Proyecciones 2030 y 2035, frente a EPP INEC 2024

Modelo	Población 2030	Dif. 2030	Población 2035	Dif. 2035
INEC EPP 2024 (referencia)	18,815,497	+0.00%	19,515,303	+0.00%
HP amortiguado (conservador)	17,883,046	-4.96%	18,509,315	-5.15%
LightGBM por tipología (central)	18,756,919	-0.31%	20,046,137	+2.72%

Figura 34

Proyecciones propias frente a EPP INEC 2024



Los dos escenarios propios enmarcan al INEC desde direcciones opuestas. El conservador subestima en ambos horizontes (-4.96% y -5.15%), porque asume una desaceleración demográfica mayor a la que prevé el INEC. El central converge casi exactamente con el INEC en 2030 (-0.31%) pero se aleja hacia arriba en 2035 (+2.72%), ya que proyecta la continuidad de las tendencias 1990-2022 sin incorporar de forma explícita los supuestos de transición demográfica que sí usa la metodología cohort-component del INEC. Como todavía no existe un censo posterior a 2022 contra el cual comparar, estas cifras no deben leerse como un error de predicción sino como una medida de qué tanto difieren dos metodologías distintas partiendo del mismo punto de arranque. Esa comparación honesta solo será posible con el próximo censo.

4.2.7. Síntesis

Los resultados de este capítulo respaldan la hipótesis planteada al inicio de la tesis: el límite de la predicción poblacional cantonal en Ecuador no está en la capacidad de los modelos, sino en la frecuencia censal disponible, cuatro observaciones en 32 años por cantón. Bajo ese régimen de datos escasos, el método demográfico de Hamilton-Perry supera de forma consistente a los enfoques de machine learning a nivel cantonal, incluyendo arquitecturas especializadas como el Temporal Fusion Transformer, y ningún esquema de ensamble ni de aprendizaje del factor de amortiguación logró mejorar ese resultado de forma genuina. El análisis SHAP de la sección 4.1.3 respalda además que ese desempeño limitado no se debe a variables mal elegidas: los factores que más pesan en el LightGBM (el intervalo entre censos, la mortalidad, la urbanización, la identidad provincial) son demográficamente razonables.

El segundo hallazgo, que el nivel de agregación geográfica determina qué modelo conviene usar, no aparece descrito de forma equivalente en la literatura revisada para este trabajo. De ahí se desprende la recomendación práctica de la tesis: la elección del modelo de proyección poblacional debe depender de la escala territorial de la decisión de planificación que se quiera informar, no de cuál modelo obtuvo el mejor promedio general.

CAPITULO 5:

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- En la revisión de la literatura realizada se resumen y clasifican modelos tradicionales (como el lineal, exponencial, logística, CCM, y Hamilton-Perry), y modelos que usan inteligencia artificial (como XGBoost, LSTM, LightGBM, y TFT). Donde se ha identificado que el uso de estos últimos modelos tienen amplia aplicación, sin embargo, se mantiene como una limitante la escasez de datos en países en desarrollo como es el caso de Ecuador.
- Las estimaciones y proyecciones actualmente deben ser probabilísticas en lugar de puramente deterministas con la finalidad de poder comunicar de manera estadística el grado de precisión y los riesgos involucrados.
- El crecimiento demográfico ha evolucionado de funciones deterministas generales a algoritmos de aprendizaje automático. La elección del método no debe ser arbitraria mientras la función logística es una herramienta poderosa ante la escasez de datos, los modelos de ML ofrecen objetividad.
- El método de Hamilton-Perry simplifica el proceso de proyección usando una relación de cambio de cohorte.
- La construcción del conjunto de datos demostró que es viable integrar en un único panel longitudinal los 221 cantones del Ecuador a partir de los cuatro censos disponibles (1990, 2001, 2010 y 2022) y de los registros de nacidos vivos y defunciones generales, aunque solo mediante tres decisiones de compromiso que quedaron documentadas y señalizadas: la imputación del indicador de necesidades básicas insatisfechas ausente en 2001, la exclusión de la escolaridad y la discapacidad por falta de comparabilidad entre censos, y el marcado del cambio en la definición de área urbana de 1990. El panel resultante conserva los 221 cantones, incluidos los 48 creados con posterioridad a 1990, y alcanza 884 registros, lo que bastó para entrenar y evaluar las cinco familias de modelos bajo un mismo protocolo. Su densidad temporal, en cambio, siguió siendo de cuatro observaciones por cantón en 32 años, y fue esa densidad, y no el número de variables disponibles, la restricción que condicionó el desempeño alcanzado.

- La evaluación mediante partición temporal estricta, con el censo de 2022 reservado como conjunto de prueba, no operó como una formalidad metodológica sino como la condición que hizo interpretables los resultados, y la evidencia procede del propio estudio. La primera versión del Temporal Fusion Transformer alcanzó un MAPE cercano a 0 % únicamente porque su conjunto de entrenamiento incluía, por un error de partición, el mismo censo que debía predecir; corregida la partición, su error real fue de 13,52 %. De forma análoga, la variante anual entrenada sobre la serie interpolada entre censos alcanzó un MAPE de 474 %, al haber aprendido el patrón de la interpolación en lugar de la dinámica demográfica. Ambos episodios muestran que, con solo tres períodos intercensales disponibles, una partición mal construida no produce un error ligeramente optimista sino una medición carente de significado.
- El nivel de agregación geográfica determina qué modelo conviene usar, y no existe un modelo universalmente mejor para todos los niveles: Hamilton-Perry amortiguado es el más preciso tanto a nivel cantonal (MAPE=7.79%) como a nivel provincial (MAPE=7.9992%), mientras que LightGBM global solo lo supera a nivel nacional, donde su error se reduce a apenas +0.49% frente al -3.56% de Hamilton-Perry. Es decir, la ventaja de Hamilton-Perry se sostiene en ambos niveles subnacionales, y solo se revierte cuando la agregación llega al total del país. No se encontró en la literatura revisada un patrón equivalente descrito de forma explícita, por lo que se plantea como el aporte distintivo de esta tesis.
- La limitación real de este trabajo no fue la capacidad de los modelos, sino la frecuencia censal disponible: cuatro observaciones en 32 años por cantón. Bajo ese régimen de datos escasos, ni el Temporal Fusion Transformer ni las variantes de LightGBM lograron superar de forma consistente a un método demográfico que solo necesita dos censos y ningún entrenamiento, lo que confirma que sumar complejidad computacional no compensa la ausencia de datos.
- Cuatro intentos independientes de mejorar sobre Hamilton-Perry con $d=0.5$ fijo, un ensamble ponderado, su corrección por validación cruzada, el aprendizaje de un factor de amortiguación por cantón, y el barrido de un valor global evaluado de dos formas, fracasaron en generalizar del periodo donde se calibran al periodo de prueba. Ante esa inestabilidad, $d=0.5$ se adopta por un criterio minimax (Wald, 1950): sin información confiable sobre dónde cae el verdadero óptimo, el punto medio del rango teórico minimiza

el error máximo posible. Que cuatro caminos distintos converjan en el mismo resultado le da a esta conclusión más peso del que tendría un solo experimento aislado.

- El análisis de interpretabilidad mediante SHAP mostró que el LightGBM global apoya sus predicciones en variables demográficamente razonables, como el intervalo entre censos, la mortalidad y la urbanización, y no en artefactos del dataset. Esto descarta que el desempeño limitado de los modelos de machine learning se deba a una mala selección de variables, y refuerza que la causa es la escasez estructural de datos.
- Los cantones en contracción poblacional concentraron los errores más altos de todo el análisis, sin que ningún modelo, incluido Hamilton-Perry, haya sido diseñado para capturar ese fenómeno de forma explícita. Queda entonces como una limitación compartida por todo el pipeline, no de un modelo en particular.
- El ejercicio quedó documentado como una secuencia ejecutable de scripts numerados que abarca desde la lectura de los microdatos censales hasta las métricas y proyecciones reportadas, con los productos intermedios persistidos en cada etapa y publicados en un repositorio de acceso abierto. Esto permite rastrear cualquier cifra del Capítulo 4 hasta la transformación que la originó y reejecutar el estudio completo a partir de fuentes públicas del INEC. La única condición que no queda cubierta es la disponibilidad local de los microdatos originales, cuyo volumen impide versionarlos junto con el código, por lo que su descarga desde los portales oficiales es el paso que el usuario debe realizar por su cuenta.

5.2.Recomendaciones

- La estimación de la migración es uno de los problemas más difíciles en demografía a causa principalmente de la mala calidad de los datos oficiales, por lo que es recomendable desarrollar modelos de recolección de datos con herramientas de Big Data para mejorar el cálculo de proyecciones.
- La elección del modelo de proyección poblacional debería depender del nivel de agregación que se necesite predecir: Hamilton-Perry amortiguado para decisiones a nivel cantonal o provincial, y LightGBM global únicamente cuando la decisión sea a escala nacional. Aplicar un solo modelo para todos los niveles, como suele hacerse en la práctica,

deja precisión sobre la mesa en al menos dos de los tres niveles de agregación.

- Se recomienda que investigaciones futuras trabajen en modelar de forma explícita la contracción poblacional, ya que ningún enfoque de los evaluados en esta tesis fue diseñado para ese caso y es justamente donde se concentran los mayores errores.
- Dado que la comparación entre marcos de evaluación (censo real frente a proyecciones oficiales) reveló una brecha sistemática de más de cuatro puntos porcentuales, se recomienda que cualquier estudio similar mantenga ambos marcos separados en vez de promediarlos, para no atribuir a los modelos un error que en realidad pertenece a la referencia usada.
- Incorporar técnicas de interpretabilidad como SHAP de manera rutinaria cuando se apliquen modelos de machine learning a estadísticas oficiales, ya que permite verificar que las predicciones respondan a variables demográficamente coherentes antes de comunicarlas como resultado válido para la toma de decisiones públicas.
- Que los conjuntos de datos contruidos a partir de censos sucesivos incorporen banderas que distingan los valores observados de los imputados y que señalicen los quiebres de definición entre rondas censales, como se hizo en este trabajo con el indicador de necesidades básicas insatisfechas y con la delimitación de área urbana. Sin esa señalización, decisiones de compromiso legítimas resultan indistinguibles del dato original y comprometen la interpretación posterior de los errores.
- Que los ejercicios de proyección poblacional que empleen aprendizaje automático adopten particiones temporales no solapadas en lugar de validación cruzada aleatoria, y que declaren de forma explícita el año reservado para prueba. La experiencia de este estudio sugiere, además, tratar un error inverosímilmente bajo como indicio de contaminación de la partición antes que como evidencia de un buen modelo.
- Poner el repositorio de código y su documentación a disposición de las instituciones que producen o utilizan estadística oficial, en particular el INEC y las direcciones de planificación de los gobiernos autónomos descentralizados, dado que la secuencia completa opera sobre fuentes públicas y está documentada paso a paso. Su utilidad no consiste en sustituir la metodología demográfica oficial, sino en permitir que las estimaciones intercensales a nivel cantonal puedan actualizarse y auditarse con un costo de

mantenimiento reducido.

Referencias Bibliográficas

1. Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2022). Methods for small area population forecasts: State-of-the-art and research needs. *Population Research and Policy Review*, 41(3), 865–898. <https://doi.org/10.1007/s11113-021-09671-6>
2. Raftery, A. E., & Ševčíková, H. (2023). Probabilistic population forecasting: Short to very long-term. *International Journal of Forecasting*, 39(1), 73–97. <https://doi.org/10.1016/j.ijforecast.2021.09.001>
3. Vanella, P., Deschermeier, P., & Wilke, C. B. (2020). An overview of population projections—Methodological concepts, international data availability, and use cases. *Forecasting*, 2(3), 346–363. <https://doi.org/10.3390/forecast2030019>
4. Hajirahimova, M. S., & Aliyeva, A. S. (2023). Development of a prediction model on demographic indicators based on machine learning methods: Azerbaijan example. *International Journal of Education and Management Engineering*, 13(2), 1–9. <https://doi.org/10.5815/ijeme.2023.02.01>
5. Wang, C.-Y., & Lee, S.-J. (2021). Regional population forecast and analysis based on machine learning strategy. *Entropy*, 23(6), Article 656. <https://doi.org/10.3390/e23060656>
6. Tanmoy, F. M., Hossain, Z., Tasfia, O., Hamim, M. A., Rahman, M., & Habib, M. (2025). Machine learning modeling for population forecasting. In *Innovations in electrical and electronics engineering* (pp. 213–228). Springer. https://doi.org/10.1007/978-981-97-9112-5_13
7. Cabello Blanco, J. J. (2024). Modelo de proyección poblacional con series de tiempo difusas. *Ciencia Latina Revista Científica Multidisciplinar*, 8(2), 1059–1071. https://doi.org/10.37811/cl_rcm.v8i2.10549
8. Cina, E., Elbasi, E., Elmazi, G., & AlArnaout, Z. (2025). The role of AI in predictive modelling for sustainable urban development: Challenges and opportunities. *Sustainability*, 17(11), Article 5148. <https://doi.org/10.3390/su17115148>
9. Alkema, L., Gerland, P., Raftery, A., & Wilmoth, J. (2015). The United Nations probabilistic population projections: An introduction to demographic forecasting with uncertainty. *Foresight*, 2015(37), 19–24.
10. Han, Y., Tian, Y., Yu, L., & Gao, Y. (2023). Economic system forecasting based on temporal fusion transformers: Multi-dimensional evaluation and cross-model comparative analysis. *Neurocomputing*, 552, Article 126500. <https://doi.org/10.1016/j.neucom.2023.126500>
11. Houssein, E. H., Mohamed, M., Younis, E. M. G., & Mohamed, W. M. (2025). Artificial intelligence and classical statistical models for time series forecasting: A comprehensive review. *Journal of Big Data*, 12, Article 271. <https://doi.org/10.1186/s40537-025-01318-z>
12. Lutz, W., & KC, S. (2010). Dimensions of global population projections: What do we know about future population trends and structures? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1554), 2779–2791. <https://doi.org/10.1098/rstb.2010.0133>
13. Villacís, B. (2019). *Proyecciones de población: Un estudio comparativo*. Banco Interamericano de Desarrollo. https://files.alapop.org/alap/Serie-E-Investigaciones/N2/Capitulos/Capitulo3_Estimaciones%26Proyecciones.pdf
14. Cabrera, G., & Resl, R. (2014). *Generación de una metodología para determinar la*

- tendencia del crecimiento poblacional en el cantón Mejía, aplicando técnicas de SIG y teledetección* [Tesis de grado, Universidad San Francisco de Quito]. Repositorio Institucional USFQ. <https://repositorio.usfq.edu.ec/bitstream/23000/3554/1/111812.pdf>
15. Mrkić, S. (2020). The 2020 round of population and housing censuses: An overview. *Statistical Journal of the IAOS*, 36(1), 35–42. <https://doi.org/10.3233/SJI-190574>
 16. Román-Sánchez, Y. G., Navarrete-López, E. L., & Barreto-Villanueva, A. (2023). Método para estimar proyecciones de población en la zona metropolitana del Valle de México. Años 2010–2020–2050. *Población y Desarrollo*, 29(56), 57–71. <https://doi.org/10.18004/pdfce/2076-054x/2023.029.56.057>
 17. Marquez, N., Bao, X., Kazura, E., Lapham, J., Sarma, P., Yu, C., Leibbrand, C., & Curran, S. (2024). An evaluation of projection methods for detailed small area projections: An application and validation to King County, Washington. *Population Research and Policy Review*, 43(2), Article 16. <https://doi.org/10.1007/s11113-023-09848-1>
 18. United Nations Population Fund. (2023). *State of world population 2023: 8 billion lives, infinite possibilities: The case for rights and choices*. <https://www.unfpa.org/sites/default/files/swop23/SWOP2023-SPANISH-230403-web.pdf>
 19. Comisión Económica para América Latina y el Caribe. (2024). *Breve panorama de los censos de población y vivienda de la ronda 2020 en América Latina y el Caribe y principales desafíos hacia la ronda 2030*. <https://www.cepal.org/sites/default/files/presentations/desafios-ronda-censal-2020-cepal.pdf>
 20. United Nations, Department of Economic and Social Affairs. (1974). *Methods for projections of urban and rural population (Manual VIII)*. https://www.un.org/development/desa/pd/files/files/documents/2020/Jan/un_1974_manual_viii_-_methods_for_projections_of_urban_and_rural_population_0.pdf
 21. Instituto Nacional de Estadística y Censos. (2025). *Estimaciones y proyecciones de la población de Ecuador: Revisión 2024* (3.^a ed.). https://www.ecuadorencifras.gob.ec/documentos/web-inec/Poblacion_y_Demografia/Proyecciones_Poblacionales/censo_2022/revision_2024_areas/Metodologia_Urbana_Rural_rev2025.pdf
 22. Bay, G., Acuña, M., & Villarroel, M. C. (2017). *Una propuesta metodológica para elaborar estimaciones y proyecciones de la población económicamente activa desde una visión demográfica*. Comisión Económica para América Latina y el Caribe. https://www.cepal.org/sites/default/files/static/files/pea_doc_metodologico_2.pdf
 23. Centro Latinoamericano de Demografía. (1984). *Métodos para proyecciones demográficas*. Comisión Económica para América Latina y el Caribe. <https://repositorio.cepal.org/server/api/core/bitstreams/3930982b-54e6-40ca-82dc-2c3bba71b8d7/content>
 24. Office for National Statistics. (2026, April 28). *National population projections: Quality and methods guide*. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/methodologies/nationalpopulationprojectionsqmi>
 25. Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6), 1313–1328. <https://doi.org/10.1080/00949655.2012.741599>

26. Hamilton, C. H., & Perry, J. (1962). A short method for projecting population by age from one decennial census to another. *Social Forces*, 41(2), 163–170. <https://doi.org/10.2307/2573607>
27. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>
28. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (Vol. 30, pp. 3146–3154). Curran Associates, Inc.
29. Rayer, S., & Smith, S. K. (2010). Factors affecting the accuracy of subcounty population forecasts. *Journal of Planning Education and Research*, 30(2), 147–161. <https://doi.org/10.1177/0739456X10380056>
30. Grossman, I., Bandara, K., Wilson, T., & Kirley, M. (2022). Can machine learning improve small area population forecasts? A forecast combination approach. *Computers, Environment and Urban Systems*, 95, Article 101806. <https://doi.org/10.1016/j.compenvurbsys.2022.101806>
31. Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>
32. Moore, A., & Bell, M. (2022). XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: A UK Biobank cohort study. *Clinical Medicine Insights: Cardiology*, 16, Article 11795468221133611. <https://doi.org/10.1177/11795468221133611>
33. Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, Investigación y Tecnología*, 21(3), Article 00002. <https://doi.org/10.22201/fi.25940732e.2020.21.3.022>
34. Zhou, T. (2023). Improved sales forecasting using trend and seasonality decomposition with LightGBM. In *2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 656–661). IEEE.
35. Babiker, A. (2025). Robust extreme gradient boosting model for predicting the behavior of RC slabs under impact loading: Key influencing factors and performance insights. *Revista IBRACON de Estruturas e Materiais*, 18(2). <https://doi.org/10.1590/S1983-41952025000200013>
36. Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., & Patil, K. R. (2023). *On leakage in machine learning pipelines* [Preprint]. arXiv. <https://arxiv.org/abs/2311.04179>
37. Gupta, M., Christopher, J., & Ramisetty, K. (2025). Mitigation strategies for data leakage in machine learning system design. In *2025 IEEE Region 10 Symposium (TENSYP)* (pp. 1–8). IEEE. <https://doi.org/10.1109/TENSYP63728.2025.11144995>
38. Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., et al. (2025). Overview of leakage scenarios in supervised machine learning. *Journal of Big Data*, 12, Article 135. <https://doi.org/10.1186/s40537-025-01193-8>
39. Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, 17(11), Article e70056. <https://doi.org/10.1111/cts.70056>

40. Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>
41. Bandara, K., Hewamalage, H., Godahewa, R., & Gamakumara, P. (2022). A fast and scalable ensemble of global models with long memory and data partitioning for the M5 forecasting competition. *International Journal of Forecasting*, 38(4), 1400–1404. <https://doi.org/10.1016/j.ijforecast.2021.11.004>
42. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
43. Durr, J.-M. (2021). Population and housing censuses: An overdue and old-fashioned instrument or still a modern, severely needed and steadfast tool? *Statistical Journal of the IAOS*, 37(2), 713–716. <https://doi.org/10.3233/SJI-210818>
44. Primicias. (2022). *Censo poblacional en aprietos y su extensión en Ecuador*. <https://www.primicias.ec/noticias/sociedad/censo-poblacional-aprietos-extension-ecuador/>
45. Wald, A. (1950). *Statistical decision functions*. John Wiley & Sons.

ANEXOS

Apéndices (si procede)

- **Documentación**

<https://github.com/byron-velasco/prediccion-poblacional-ecuador>

- **Tablas de análisis específicos**

- **Imágenes y gráficos complementarios**

- **Link en el que se encuentra compartidos los laboratorios prácticos desarrollados**

a) <https://github.com/byron-velasco/prediccion-poblacional-ecuador>