Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), enero-febrero 2025,
Volumen 9, Número 1.

https://doi.org/10.37811/cl_rcm.v9i1

EL ESTUDIO DE LA MOVILIDAD INDUCIDA POR CAUSAS AMBIENTALES. ENTRE LA INTELIGENCIA ARTIFICIAL Y LOS MODELOS EXPLORATORIOS

THE STUDY OF ENVIRONMENTALLY INDUCED MIGRATION. BETWEEN ARTIFICIAL INTELLIGENCE AND EXPLORATORY MODELS

Bernardo Bolaños-Guerra
Universidad Autónoma Metropolitana, México

Sazcha Marcelo Olivera-Villarroel
Universidad Autónoma Metropolitana, México

DOI: https://doi.org/10.37811/cl_rcm.v9i1.16563

El estudio de la Movilidad Inducida por Causas Ambientales. Entre la Inteligencia Artificial y los Modelos Exploratorios

Bernardo Bolaños-Guerra[1]
bbolanos@cua.uam.mx
https://orcid.org/0000-0002-8881-1638
Universidad Autónoma Metropolitana
Unidad Cuajimalpa.
Ciudad de México
México

Sazcha Marcelo Olivera-Villarroel
solivera@cua.uam.mx
http://orcid.org/0000-0003-1864-7374
Universidad Autónoma Metropolitana
Unidad Cuajimalpa
Ciudad de México
México

RESUMEN

El estudio cuantitativo de la movilidad impulsada por factores ambientales (tanto en forma de migración económica como de desplazamiento forzado) ha enfrentado desafíos importantes. Esto se debe no solo a la falta de bases de datos sólidas, sino también a la amplia libertad que tienen los investigadores para diseñar sus modelos exploratorios, lo que puede conducir a resultados sesgados o poco reproducibles. Al seleccionar diferentes variables ambientales, los investigadores pueden encontrar patrones aleatorios o ruido en los datos, en lugar de relaciones genuinas, siguiendo una lógica más exploratoria que el seguimiento de un proceso teórico basado en evidencia histórica. Afortunadamente, este problema se mitiga en los estudios computacionales que utilizan inteligencia artificial, concretamente técnicas de aprendizaje automático. Surge entonces la pregunta de si estas técnicas también pueden reducir el escepticismo sobre los desplazamientos climáticos. Sin embargo, una limitación importante de estos enfoques computacionales sigue siendo la escasez de bases de datos suficientemente grandes y confiables. Además, estos métodos enfrentan desafíos asociados con lo que Conway llama la "zona de peligro", donde la inteligencia artificial es tratada como una caja negra. Esto incluye problemas como la incomprensión por parte del usuario de los supuestos subyacentes, sesgos de confirmación no detectados, interpretación errónea de los datos y alucinaciones de los grandes modelos de lenguaje. Otras innovaciones que emplean también inteligencia artificial se están enfocando cada vez más en los datos de movilidad humana proporcionados por las redes sociales y otras aplicaciones web masivas. Estos datos representan una valiosa fuente de información que puede ser utilizada para comprender mejor los patrones y tendencias migratorias a nivel global.

Palabras clave: movilidad humana, cambio climático, inteligencia artificial, métodos cuantitativos, ciencias sociales computacionales

[1] Autor principal
Correspondencia: bbolanos@cua.uam.mx

The study of Environmentally Induced Migration. Between Artificial Intelligence and Exploratory Models

ABSTRACT

The quantitative study of environmentally driven mobility (both in the form of economic migration and forced displacement) has faced significant challenges, stemming not only from a lack of robust databases but also from the considerable freedom researchers have in designing exploratory models. This flexibility can lead to biased or poorly reproducible results. By selecting different environmental variables, researchers risk identifying random patterns or noise in the data instead of uncovering genuine relationships, often relying on exploratory logic rather than theoretical frameworks grounded in historical evidence. Fortunately, this issue can be mitigated by computational studies employing artificial intelligence, particularly machine learning techniques. This raises the question of whether these methods can also reduce skepticism surrounding climate-induced displacement. However, a key limitation of these computational approaches remains the scarcity of sufficiently large and reliable datasets. Additionally, these methods face challenges associated with what Conway terms the 'danger zone,' where AI is treated as a black box. This includes issues such as user misunderstanding of underlying assumptions, undetected confirmation biases, misinterpretation of data, and LLM hallucinations. Other AI-driven innovations are increasingly leveraging human mobility data from social networks and other large-scale web applications. These sources represent a valuable reservoir of information, offering promising potential for improving our understanding of global migration patterns and trends.

***Keywords*:** human mobility, climate change, artificial intelligence, quantitative methods, computational social science

Artículo recibido 20 enero 2025
Aceptado para publicación: 22 febreo 2025

INTRODUCCIÓN

La elección personal de variables o especificaciones climáticas basadas en los datos disponibles puede aumentar la flexibilidad de los investigadores, lo que a su vez puede dar lugar a resultados sesgados o no reproducibles. Esto es especialmente cierto cuando la construcción de los modelos estadísticos se basa en un proceso exploratorio en lugar de uno teórico (Ioannidis, 2005). Al elegir entre diferentes variables climáticas o especificaciones de los datos, los expertos pueden, sin darse cuenta, aprovechar patrones aleatorios o ruido en la información, en lugar de identificar relaciones genuinas. Este problema es descrito por Molina et al. (2023):

> "Los estudiosos de la migración a menudo recurren a sus datos para elegir entre variables climáticas alternativas o especificaciones alternativas (en lugar de probar hipótesis preregistradas, por ejemplo). Por supuesto, la selección de medidas o especificaciones del clima basada en datos no está exenta de problemas, ya que aumenta los "grados de libertad del investigador" (Simmons, Nelson y Simonsohn 2011), especialmente cuando se aplica a diferentes conjuntos de datos (Yarkoni y Westfall 2017). Cuantas más variables o especificaciones prueben los investigadores en sus datos, mayor será la probabilidad de obtener un resultado significativo puramente por casualidad y menor será la probabilidad de que los resultados se repliquen en un conjunto de datos diferente." (Molina et al., 2023, pp. 466-467).

El problema que hemos descrito se conoce comúnmente como "p-hacking" o "minado de datos" en el ámbito de la investigación científica, y es especialmente relevante en campos como la climatología, donde los conjuntos de datos son extensos y complejos.

Cuando los investigadores tienen demasiada flexibilidad en la selección de variables o en la especificación de modelos, sin una base teórica sólida, pueden (consciente o inconscientemente) elegir las variables o especificaciones que producen los resultados más "interesantes" o estadísticamente significativos (Ioannidis, 2005). Esto puede llevar a varios problemas:

1. Falsos positivos: Se pueden encontrar relaciones aparentemente significativas que en realidad son espurias.

2. Sobreajuste: Los modelos pueden ajustarse demasiado a los datos de la muestra, perdiendo capacidad de generalización.
3. Falta de reproducibilidad: Otros investigadores pueden no obtener los mismos resultados al intentar replicar el estudio.
4. Sesgo de confirmación: Los investigadores pueden seleccionar variables que confirmen sus hipótesis previas.

Las prácticas subjetivas descritas por Molina et al. pueden conducir a un sobreajuste, donde el modelo funciona bien con el conjunto de datos específico utilizado, pero falla al aplicarse a datos nuevos e invisibles. Además, cuantas más opciones tenga un investigador (variables, modelos o especificaciones), mayor será el riesgo de que su modelo final refleje esas elecciones subjetivas o personales en lugar de una verdadera relación subyacente. Esto es especialmente preocupante cuando se aplica a diferentes conjuntos de datos, ya que los hallazgos podrían no generalizarse más allá de los contextos específicos en los que se obtuvieron. Para abordar este problema, los investigadores pueden:

1. Establecer hipótesis y métodos de análisis de antemano, antes de examinar los datos.
2. Emplear técnicas de validación cruzada y conjuntos de datos de prueba independientes.
3. Ser transparente sobre todos los análisis realizados, incluyendo aquellos que no produjeron resultados significativos.
4. Basar la selección de variables en sólidos fundamentos teóricos, no solo en correlaciones empíricas.
5. Utilizar métodos estadísticos robustos que controlen el problema de las comparaciones múltiples.
6. Llevar a cabo análisis de sensibilidad para evaluar la solidez de los resultados.

Además, como lo señala Llopis (2007), entre otros autores, los datos utilizados en la mayoría de los estudios sobre migración sufren del "nacionalismo metodológico". Es decir, estos estudios suelen estar financiados por las administraciones públicas, lo que los limita a las necesidades político-administrativas y frena una reflexión científica más amplia. Esto lleva a que los investigadores adopten el marco territorial de la administración contratante como referencia, confundiendo el objeto social y el objeto de estudio. O bien, estos trabajos se inscriben en la perspectiva de políticas públicas, lo que con frecuencia se interpreta como que están pensados para administrar los flujos migratorios desde los supuestos intereses de un Estado o de su opinión pública.

Los electores nacionalistas pueden exigir controles migratorios y ello sesgar las investigaciones. Todo ello se traduce en una falta de investigación sobre las sociedades de origen de los inmigrantes, centrándose principalmente en las condiciones de la inmigración en el país receptor, sin analizar los procesos de emigración y sus efectos en los países de origen.

Los investigadores en ciencias sociales a menudo analizan la migración desde una perspectiva limitada al marco del estado-nación, lo cual Llopis (2007), Dumitru (2023) y otros consideran inadecuado. La movilidad humana, tanto en forma de migración económica como de desplazamiento forzado, es un fenómeno global que trasciende las fronteras estatales, por lo que se requiere una perspectiva más amplia que la conciba como un proceso inherente al sistema mundo, y no solo como algo que sucede entre estados separados. Es necesario superar las concepciones ideológicas que naturalizan al estado-nación como el ámbito "natural" desde el cual estudiar el fenómeno migratorio.

En este sentido los principales desafíos que enfrentan los estudios estadísticos sobre flujos migratorios son: 1. Problemas de categorización: - Confusión sobre si los datos se clasifican por lugar de origen, lugar de nacimiento o nacionalidad del inmigrante. - Falta de acuerdo en la definición de migraciones "internas" vs "externas". - Nuevas clasificaciones derivadas de la movilidad dentro de la Unión Europea. 2. Problemas conceptuales relacionados con diferencias idiomáticas y culturales: - Uso de instrumentos de investigación (cuestionarios, escalas) diseñados originalmente para la población autóctona. - Necesidad de garantizar el conocimiento suficiente del idioma de la investigación y de realizar adaptaciones lingüísticas. - Importancia de pruebas piloto, investigación exploratoria, traducciones paralelas y entrevistadores bilingües.

En este sentido se requiere una validación cruzada rigurosa, pruebas fuera de la muestra o controles de solidez para asegurar que los resultados no sean meros artefactos del proceso de selección de datos o que traigan sesgo de origen desde la recopilación de los datos o la selección de la muestra.

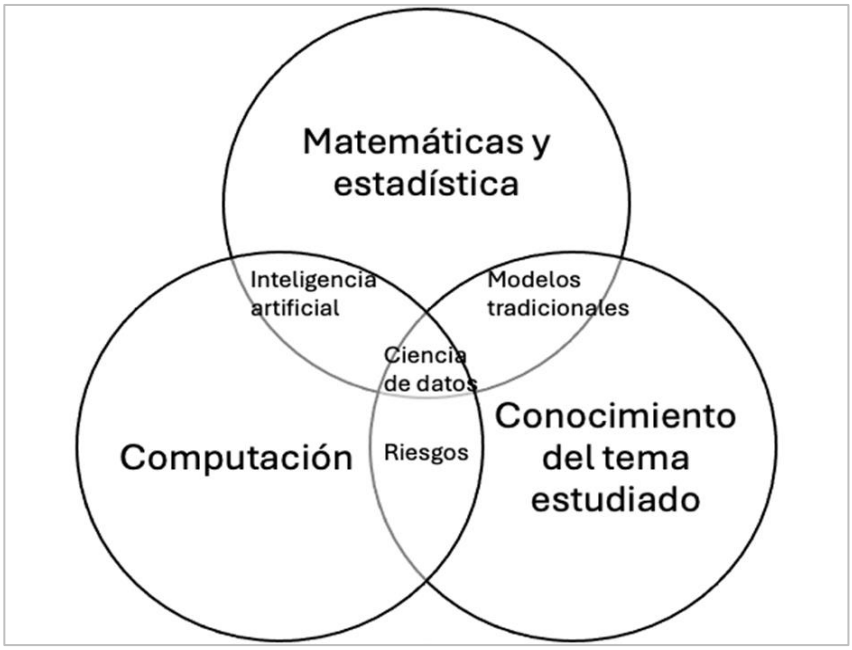
En el marco de esta discusión, el presente artículo presenta un análisis comparativo de las estrategias metodológicas utilizadas en tres estudios recientes y destacados sobre desplazamientos migratorios impulsados por factores ambientales, especialmente climáticos. El texto examina de manera clara y concisa los enfoques metodológicos empleados en estas investigaciones influyentes.

Dos de estos estudios (Xu et al., 2020; Hoffmann et al., 2020) emplearon modelos estadísticos tradicionales, mientras que el tercero (Molina et al., 2023) recurrió a técnicas de inteligencia artificial. La selección de estos tres trabajos se debe a su notable impacto en la actual investigación sobre migraciones.

El estudio de Xu et al. (2020) respalda posturas maximalistas sobre los futuros movimientos poblacionales provocados por la crisis ambiental. Por el contrario, la investigación de Hoffmann et al. representa una tendencia a relativizar el peligro de los desplazamientos ambientales. Este último es un metaanálisis que se presentó y discutió en la sesión plenaria de la Conferencia Mundial sobre Migración 2024, un importante evento académico anual dedicado al estudio multidisciplinario de la migración.

El estudio de Molina et al. (2023) es el resultado del trabajo de un grupo de investigación en ciencia de datos de la Universidad de Cornell, liderado por Filiz Garip. Garip, socióloga de la Universidad de Princeton y destacada experta en estudios migratorios, ha sido pionera en la aplicación de los estudios computacionales y la ciencia de datos a las investigaciones sociales. Como veremos, esta última tendencia interdisciplinaria está transformando no solo la metodología de la investigación social, sino también la enseñanza de la estadística en las ciencias sociales. La confluencia de estas disciplinas se puede visualizar mediante el siguiente diagrama de Venn: (Imagen 1. Elaboración propia a partir de Conway 2010).

Imagen 1. Análisis crítico de la ciencia de datos



El diagrama de Venn de Drew Conway (2010) ilustra las tres principales áreas de conocimiento que se intersectan para definir la ciencia de datos:

(a) Conocimientos técnicos en computación y programación. Esto abarca dominar lenguajes de programación como Python o R, trabajar con bases de datos y utilizar herramientas avanzadas de análisis de datos. Estas destrezas permiten a los científicos de datos investigar y explorar datos de maneras innovadoras.

(b) Sólidos conocimientos matemáticos y estadísticos. La capacidad de aplicar técnicas estadísticas y matemáticas como álgebra lineal, cálculo e inferencia estadística es fundamental para desarrollar modelos de datos precisos y útiles.

(c) Experticia en el campo de aplicación. Tener un profundo entendimiento del dominio específico donde se aplicarán los métodos de ciencia de datos (por ejemplo, medicina, migración o futbol) permite formular preguntas relevantes, interpretar los resultados de manera significativa y tomar decisiones informadas.

Intersecciones dentro del diagrama de Conway y riesgos de los enfoques computacionales que utilizan inteligencia artificial

La "ciencia de datos pura" (véase el centro del diagrama) surge de la combinación de habilidades de programación, conocimientos estadísticos y experiencia en un campo específico. Esta intersección de competencias permite a los nuevos científicos o equipos abordar problemas complejos de manera integral, aprovechando las tecnologías actuales.

De manera más acotada, técnicas de inteligencia artificial como el aprendizaje de máquina (*machine learning)* implican la confluencia de la computación más avanzada y de las disciplinas lógico-matemáticas. Aquí se ubican los profesionales capaces de programar y aplicar técnicas estadísticas para desarrollar algoritmos de aprendizaje automático. Estos expertos pueden crear modelos predictivos y automatizar análisis de datos, aunque sin necesariamente tener un conocimiento intuitivo del dominio específico en el que aplican sus modelos. Esto no es ideal, ya que un científico debería comprender los temas sobre los que trabaja.

Por su parte, los modelos tradicionales ("investigación tradicional"), anteriores a la inteligencia artificial, surgen de la intersección de los conocimientos matemáticos/estadísticos y de la experiencia en un campo determinado, sin recurso a la computación.

¿Y qué pasa cuando algunas personas tienen conocimiento de un campo y emplean la inteligencia artificial, pero sin tener suficiente capacidad de análisis estadístico? Drew Conway (2010) incluyó esta "zona de peligro" en su diagrama de Venn sobre la ciencia de datos para destacar los riesgos potenciales de combinar conocimientos sustantivos con habilidades de programación, pero sin una base sólida en conocimientos estadísticos o matemáticos. Nosotros hemos usado la palabra “riesgos”, aludiendo a los problemas de usar esta tecnología como caja negra (Carabantes, 2020). En esta zona, el riesgo radica en: (a) Interpretación errónea de los datos: sacar conclusiones erróneas debido a la falta de comprensión de la significación estadística, los sesgos o las variables de confusión. (b) Dependencia excesiva de las herramientas: utilizar modelos estadísticos o de aprendizaje automático sin comprender plenamente sus suposiciones subyacentes, lo que conduce a resultados incorrectos o engañosos. (c) Sesgo de confirmación: centrarse en los resultados que confirman nociones preconcebidas a partir de conocimientos sustantivos sin una validación estadística adecuada.

El objetivo de Conway (2010) era destacar la importancia de un conjunto equilibrado de habilidades en la ciencia de datos. Si bien la programación y el conocimiento del dominio son esenciales, una base sólida en matemáticas y estadísticas permite que los conocimientos extraídos de los datos sean válidos y confiables.

Este concepto de riesgo tiene una gran importancia relacionada con el fenómeno de las alucinaciones en los grandes modelos de lenguaje (LLM, por sus siglas en inglés). Las alucinaciones ocurren cuando una IA genera resultados que son factualmente incorrectos o inventados pero que parecen plausibles. Aunque los LLM imitan respuestas humanas, sus resultados se basan en patrones en los datos de entrenamiento, no necesariamente en una justificación lógica, matemática o estadística de sus conclusiones.

Por ejemplo, un LLM podría atribuir un avance matemático específico a una figura histórica equivocada (a Fermat en vez de a Pascal) o inventar una cita.

Los usuarios sin conocimientos sustantivos podrían aceptar esos resultados como factuales, lo que muestra los riesgos de la "zona de peligro". Obsérvese que aquí el problema no es el rigor estadístico *per se*, sino que el modelo arroje un dato falso. Por lo tanto, el diagrama de Conway (2010) debería mejorarse para reconocer dos "zonas de peligro" distintas: la zona de peligro originalmente marcada por Conway (errores clásicos de la ciencia de datos como son la incomprensión por parte del usuario de los supuestos subyacentes, los sesgos de confirmación no detectados y las interpretaciones erróneas de los datos); y una segunda zona de peligro sustantiva, que en el diagrama marcamos con las palabras “inteligencia artificial”, donde lo que faltan son bases de datos suficientemente grandes, lo que da como resultado datos falsos alucinados por los grandes modelos de lenguaje.

Dilema de los modelos exploratorios de estadística clásica: posible arbitrariedad o conclusiones debilitadas

El diseño de modelos exploratorios usando procesos estadísticos en estudios sobre migración y clima implica decisiones subjetivas por parte de los investigadores, lo que puede conducir a conclusiones aparentemente contradictorias.

Por ejemplo, Riosmena et al. (2018) y Nawrotzki el al. (2017) observaron que la reducción de las lluvias en México se asociaba con una mayor migración a Estados Unidos, pero solo en comunidades económicamente más favorecidas. Aparentemente, este hallazgo respalda la idea, previamente destacada en la literatura técnica como el Informe Foresight del gobierno del Reino Unido (2011), de que las personas en situación de pobreza extrema pueden verse "atrapadas", careciendo de los recursos necesarios para migrar en respuesta al estrés climático.

Pero el estudio de Riosmena et al. (2018) plantea una interesante pregunta: ¿sus hallazgos reflejan realmente lo que anticipaba el Informe Foresight de 2011 o fueron las expectativas incluidas en este último y otros similares los que orientaron el modelo hacia esa dirección? Esta cuestión cobra relevancia al considerar la metodología empleada.

En primer lugar, el estudio de Riosmena el al. (2018) integra diversas fuentes de datos, incluyendo los censos mexicanos del 2000 y 2010 armonizados en la Serie Integrada de Microdatos de Uso Público (IPUMS), sí como datos climáticos de la Unidad de Investigación Climática (CRU) de la Universidad de East Anglia, obtenidos vía el sistema de extracción Terra Populus.

Esta variedad de fuentes otorga a los investigadores una gran flexibilidad para seleccionar e integrar los datos, así como para elegir los marcos temporales, introduciendo un grado significativo de subjetividad. Además, Riosmena et al. (2018) utilizaron métodos estadísticos tradicionales, como modelos multinivel y análisis de subgrupos. Estas técnicas requieren que los investigadores tomen decisiones específicas, como la definición de comunidades "favorecidas" y "desfavorecidas", grandes y pequeñas, y la selección de variables o interacciones a incluir en los modelos. Estas elecciones metodológicas podrían influir en los hallazgos, posiblemente dando lugar a resultados mixtos o dependientes del contexto.

Como alternativa a estos estudios tradicionales, el uso de técnicas de aprendizaje automático ofrece un enfoque más objetivo, pero sometido a la “zonas de peligro” original del diagrama de Conway (2010). Por un lado, las especificaciones del modelo dependen más de los procesos de optimización del algoritmo que de las decisiones subjetivas de los investigadores, pero, por el otro, la inteligencia artificial puede funcionar en buena medida como una caja negra. Esto plantea una reflexión importante sobre cómo las decisiones metodológicas pueden influir en los resultados de la investigación.

El estudio de Riosmena et al. (2018), con mayor libertad en la toma de decisiones, podría estar más influenciado por intuiciones y preferencias personales, como la expectativa contenida en el Informe Foresight sobre las poblaciones inmovilizadas. Por el contrario, el enfoque más automatizado de Molina et al. (2023) puede proporcionar una perspectiva menos sesgada por expectativas previas, pero con la artificialidad propia a la inteligencia artificial.

Esta comparación resalta la importancia de examinar cuidadosamente las metodologías empleadas en la investigación sobre migración y clima, y de interpretar los resultados considerando las posibles influencias de las decisiones metodológicas en los hallazgos obtenidos.

Los problemas metodológicos en el estudio de la relación entre clima y migración han llevado a los investigadores a considerar técnicas más complejas para mejorar la robustez de sus conclusiones. Una de estas técnicas es el meta-análisis, que busca combinar los resultados de múltiples estudios para obtener una visión más amplia y potencialmente más precisa del fenómeno en cuestión.

El meta-análisis, al agregar los resultados de varios modelos, puede reducir el margen de acción o grados de libertad de los investigadores individuales. Esto, en teoría, debería disminuir la influencia de sesgos personales y decisiones metodológicas subjetivas en los resultados finales.

Sin embargo, este enfoque, aunque metodológicamente sofisticado, puede generar un nuevo desafío: el escepticismo o conservadurismo en las conclusiones. Al combinar múltiples estudios, que pueden tener resultados variados o incluso contradictorios, el meta-análisis tiende a producir conclusiones más matizadas y menos definitivas.

Existen muchos más enfoques para el desarrollo de modelos de migración por causas climáticas siendo las principales tendencias:

1. Análisis de series temporales avanzados (Aguado et al., 2016):

- Modelos ARIMA con variables exógenas (ARIMAX) para incorporar factores climáticos (Nawrotzki et al., 2013).
- Modelos VAR para examinar las interacciones dinámicas entre variables climáticas y migratorias (Feng et al., 2010).

2. Técnicas de análisis espacial (Navarro, 2015):

- Modelos de autocorrelación espacial para identificar clusters de migración relacionados con patrones climáticos (Leyk et al., 2012; Henry et al, 2004).
- Regresión geográficamente ponderada para capturar variaciones locales en la relación clima-migración (Dallmann & Millock, 2017).

3. Modelos de ecuaciones estructurales (SEM) (Gray & Wise, 2016):

- Para modelar relaciones complejas y multidireccionales entre variables climáticas, socioeconómicas y migratorias.

4. Análisis de redes (Sakdapolrak et al., 2016):

- Para mapear y analizar las rutas migratorias en relación con cambios climáticos a lo largo del tiempo.

5. Técnicas de aprendizaje profundo (Robinson & Dilkina, 2018):

- Redes neuronales recurrentes (RNN) o Long Short-Term Memory (LSTM) para capturar patrones temporales complejos en datos de clima y migración.

6. Modelos bayesianos jerárquicos (Porst & Sakdapolrak, 2018):

- Para incorporar incertidumbre y variabilidad a múltiples niveles (individual, comunitario, regional).

7. Análisis de eventos extremos (Bohra-Mishra et al., 2014):

- Modelos de valor extremo para examinar cómo los eventos climáticos extremos afectan los patrones migratorios.

8. Técnicas de emparejamiento (*matching*) (Mueller et al., 2014):

- Como Propensity Score Matching, para comparar áreas con características similares pero diferentes exposiciones al cambio climático.

9. Análisis de supervivencia (Warner & Afifi, 2014):

- Para modelar el tiempo hasta que ocurre la migración en relación con factores climáticos.

10. Modelos de agentes (Entwisle et al., 2016: Thober et al, 2018; Trinh et al, 2023):

- Para simular decisiones individuales de migración basadas en factores climáticos y socioeconómicos.

Si bien, como podemos ver, existen diferentes enfoques metodológicos para abordar el tema de la movilidad climática, en el resto de este artículo nos enfocaremos en la comparación de dos estudios recientes y multicitados (prestigiosos) que utilizan métodos estadísticos exploratorios clásicos en lugar de técnicas computacionales.

Por un lado, el estudio de Xu et al. (2020) sobre "El futuro del nicho climático humano", publicado en la prestigiosa revista *Proceedings of the National Academy of Sciences*, presenta escenarios preocupantes que proyectan miles de millones de desplazados climáticos en las próximas décadas. Denominaremos a este como el "enfoque maximalista".

Por otro lado, el metanálisis de Hoffmann et al. (2020) sobre migración inducida por causas ambientales adopta una postura más cautelosa. Este estudio concluye que existe una gran heterogeneidad en los impactos de los peligros ambientales sobre la migración, y que la fuerza y dirección de la relación desastres-desplazamientos depende de las condiciones locales y de la intensidad de los factores ambientales. Haremos referencia a este como el "enfoque cauteloso".

Un análisis comparativo de los grados de libertad de los investigadores en estos dos estudios implica examinar la naturaleza de sus metodologías, la flexibilidad en sus procesos analíticos y las decisiones inherentes que podrían influir en sus hallazgos.

El enfoque maximalista de Xu et al. (2020) les otorgó una mayor discrecionalidad en la elección de modelos climáticos, trayectorias de concentración representativas, marcos temporales, variables que definen el nicho climático humano, integración de diversas fuentes de datos y suposiciones sobre adaptabilidad humana y condiciones socioeconómicas futuras. Esto les permitió construir un escenario probable, pero también introdujo elementos subjetivos que reflejaban sus prioridades y enfoques personales.

Por el contrario, el metanálisis de Hoffmann et al. (2020) buscó reducir la arbitrariedad de las conclusiones, pero a cambio introdujo amplios márgenes de incertidumbre. Al seleccionar 30 estudios y extraer 1,803 estimaciones de efectos, los autores debieron tomar decisiones sobre los criterios de selección de los modelos y los métodos de estandarización, lo que redujo el margen para la interpretación subjetiva. Sin embargo, esta síntesis de una muestra heterogénea de estudios los llevó a concluir que existe gran variabilidad en los impactos de la migración frente a las condiciones locales y los factores ambientales, adoptando una postura escéptica o cautelosa.

El estudio de Xu et al. (2020) tuvo más flexibilidad en su modelado, mientras que el metanálisis de Hoffmann et al. (2020) buscó reducir la subjetividad, pero a costa de resaltar la complejidad y la falta de conclusiones definitivas de los resultados consolidados.

Posteriormente, veremos cómo los enfoques computacionales pueden aportar nuevas perspectivas y superar algunas de las limitaciones de estos dos estudios.

Ventajas metodológicas de los enfoques computacionales que utilizan inteligencia artificial

Los enfoques de aprendizaje automático supervisado pueden reducir los grados de libertad del investigador que se observaron en los estudios basados en métodos estadísticos clásicos. Esto se logra mediante la división de la muestra en conjuntos de entrenamiento y de prueba (Molina et al., 2023; Calonico et al., 2022; Engle, 2010).

La división de la muestra implica separar los datos en subconjuntos distintos antes de cualquier análisis.

Esto limita la capacidad de los investigadores para ajustar subjetivamente los modelos a los datos, lo que podría conducir a un sobreajuste y a evaluaciones demasiado optimistas del rendimiento del modelo (Hsiao, 1979; Imbens y Rubin, 2015).

Al comprometerse con esta división de la muestra, los investigadores que utilizan aprendizaje automático supervisado deben desarrollar y entrenar sus modelos en el conjunto de entrenamiento y luego evaluar su desempeño en el conjunto de prueba, que el modelo "nunca ha visto" (Hastie et al., 2009; Garson, 2016). Esto les permite evaluar mejor la capacidad de generalización de sus hallazgos a nuevas observaciones provenientes del mismo proceso de generación de datos (Molina et al., 2023; Friedman et al., 2001).

Además, los enfoques computacionales a menudo emplean técnicas de validación cruzada, donde los datos se dividen de diferentes maneras para minimizar aún más el sesgo (Kohavi, 1995; Arlot & Celisse, 2010). Esto ayuda a garantizar la solidez de los resultados y a evitar que se deban simplemente a idiosincrasias de los datos de entrenamiento (Breiman, 1996; James et al. 2013; Ashraf, Khudheir Salal y Abdullaev, 2021).

Sin embargo, es importante tener en cuenta que los enfoques computacionales no resuelven completamente los problemas de los modelos clásicos. Aún pueden existir sesgos relacionados con el pre-procesamiento de los datos, la determinación de la división de la muestra o la selección y ajuste de los modelos (Barocas & Selbst, 2016; Passonneau & Carpenter, 2014). Por lo tanto, los investigadores deben esforzarse por mantener la claridad, la coherencia y la transparencia metodológica en todo el proceso de investigación (Bem, 1987; American Psychological Association, 2020).

En este sentido los enfoques computacionales que utilizan aprendizaje automático supervisado y técnicas como la división de la muestra y la validación cruzada tienen el potencial de reducir los grados de libertad del investigador y mejorar la generalización de los hallazgos, en comparación con los métodos estadísticos clásicos (Molina et al., 2023; Calonico et al., 2022). Sin embargo, estos enfoques aún tienen limitaciones que deben ser abordadas cuidadosamente, en particular la “zona de peligro” de Conway (2010) por el uso de la inteligencia artificial como caja negra.

Otras aplicaciones de la inteligencia artificial para el estudio de la movilidad ambiental y climática

La información sobre usuarios de plataformas de Internet es una fuente potencial de datos sobre migración que está siendo explotada gracias a la inteligencia artificial (Leasure et al., 2023; Rampazzo et al., 2023). Por ejemplo, podemos revisar el número de usuarios de Facebook en México que se encontraban viviendo en Venezuela hace 5 años, como una estimación indirecta de la migración venezolana.

Grow et al. (2022) evalúan la precisión de los datos publicitarios de Facebook, que los científicos sociales utilizan cada vez más para realizar censos digitales y reclutar participantes para encuestas. Esos autores exploran si las estimaciones demográficas proporcionadas por la plataforma publicitaria de Facebook son lo suficientemente confiables como para ser utilizadas en la investigación en las ciencias sociales. Su investigación se basa en una encuesta en línea transnacional realizada en 14 países, en la que se compararon los datos de Facebook con fuentes de referencia externas, incluidas las oficinas nacionales de estadística y las Naciones Unidas (Grow et al., 2022).

En concreto, se centran en características demográficas como edad, género y nivel educativo, comparando estos indicadores con los datos oficiales para evaluar las discrepancias. Su estudio también considera la representatividad de los datos de los usuarios de Facebook de la población en general, lo que es crucial para los investigadores que utilizan estos datos para obtener hallazgos generalizables.

Grow et al., (2022) encuentran lo siguiente: los datos de Facebook generalmente sobreestiman la proporción de usuarios más jóvenes y subestiman la proporción de usuarios mayores, lo que refleja sesgos en la base de usuarios de la plataforma. La representación de hombres y mujeres es bastante precisa en general, pero existen variaciones entre países. Este estudio encontró discrepancias considerables en las estimaciones del nivel educativo de Facebook. En algunos países, los datos de Facebook sobrerrepresentan a individuos con niveles educativos más altos, lo que podría distorsionar los resultados de las investigaciones que se basan en estas métricas (Grow et al., 2022).

La precisión de los datos de Facebook varía significativamente entre países y algunas naciones muestran una alineación más estrecha con las fuentes de datos externas que otras.

Esta inconsistencia sugiere que los investigadores deben ser cautelosos al generalizar los hallazgos de los datos de Facebook a las poblaciones nacionales. Grow et al. (2022) destacan que los datos de Facebook son propensos a sesgos inherentes a la información autoinformada, como el hecho de que los usuarios tergiversen sus datos demográficos. Además, los datos de anuncios de Facebook reflejan sólo a los usuarios activos, no a toda la población, lo que limita la representatividad para estudios de población más amplios.

Dados los sesgos identificados, Grow et al. (2022) recomiendan que los investigadores que utilicen datos de Facebook para estudios de población apliquen técnicas de ponderación adecuadas y validen los datos frente a puntos de referencia confiables. También destacan la necesidad de transparencia en relación con las limitaciones de los datos de Facebook en las publicaciones de investigación (Grow et al., 2022).

Las técnicas de aprendizaje automático (*machine learning*) pueden servir para limpiar y mejorar la calidad de los datos demográficos de Facebook. Dados los sesgos e imprecisiones identificados en los datos publicitarios de Facebook, el aprendizaje automático puede ayudar a abordar algunos de estos desafíos al identificar patrones, corregir tergiversaciones y mejorar la confiabilidad general de la información.

Una de las formas en que se puede aplicar el aprendizaje de máquina para limpiar y mejorar los datos demográficos de Facebook es la detección de sesgos algorítmicos. Los algoritmos de inteligencia artificial se pueden entrenar para detectar sesgos demográficos en los datos de Facebook, comparándolos con datos de referencia de fuentes confiables, como las oficinas nacionales de estadística. Las técnicas como los modelos de aprendizaje automatizado que tienen en cuenta la imparcialidad pueden ayudar a ajustar los grupos sobrerrepresentados o subrepresentados.

En segundo lugar, los modelos de *machine learning* pueden ponderar de nuevo los datos para representar mejor a la población general. Por ejemplo, los modelos como la ponderación de probabilidad inversa pueden ajustar las distribuciones de muestra para que coincidan con las características demográficas conocidas, lo que mejora la representatividad de los datos (Kalton y Flores-Cervantes, 2003).

En tercer lugar, estas técnicas computacionales pueden imputar y corregir datos. Modelos de aprendizaje de máquina como k-Nearest Neighbors (k-NN) o los algoritmos de aprendizaje profundo pueden imputar información demográfica faltante en función de patrones en los datos disponibles (Fadlil, 2022; Jang et al., 2020).

Las técnicas de aprendizaje de máquina, como la agrupación en clústeres o la detección de anomalías, pueden identificar y corregir valores atípicos en los datos demográficos, como entradas de edad inverosímiles (Rousseeuw & Hubert, 2011; Leys et al., 2019). Los algoritmos de aprendizaje supervisado, como los árboles de decisión o las redes neuronales, pueden predecir y verificar las características demográficas en función de los patrones de actividad del usuario (LeCun et al., 2015; He at al. 2020; Kaparthi y Bumblauskas, 2020).

Los modelos de procesamiento del lenguaje natural (PLN) pueden analizar el contenido generado por el usuario (publicaciones, comentarios) para inferir atributos demográficos como el nivel de educación o la ocupación (Jurafsky & Martin, 2021). Al combinar datos de múltiples fuentes (por ejemplo, datos de Facebook con otras redes sociales o conjuntos de datos demográficos externos), los métodos de aprendizaje de conjunto pueden mejorar la precisión de los datos (Dietterich, 2000; Hastie et al., 2009). Las técnicas de aprendizaje automatizado se pueden utilizar para validar de forma cruzada los datos de Facebook con conjuntos de datos de referencia de alta calidad (Kohavi, 1995; Arlot & Celisse, 2010).

En el mediano plazo es previsible que las técnicas de aprendizaje de máquina mejoren la confiabilidad de los datos sobre migración inducida por causas ambientales provenientes de Facebook. Sin embargo, la transición no está libre de desafíos éticos y metodológicos que deben ser abordados cuidadosamente.

La tabla 1 busca ofrecer una visión detallada y matizada sobre cómo los enfoques tradicionales y los innovadores difieren en su tratamiento de los estudios migratorios. Resalta las limitaciones de los enfoques tradicionales, en contraste con la perspectiva más global y transnacional de los enfoques innovadores. Esta tabla proporciona una herramienta para comparar y contrastar los diferentes enfoques en el estudio de la migración. Como puede destacarse a partir de ella, existe una gran complejidad del tema y es necesaria una perspectiva amplia y adaptable en la investigación futura.

Tabla 1. Evolución de los Enfoques en Estudios Migratorios: Del Nacionalismo Metodológico a la Perspectiva Global

Aspecto	Enfoques Tradicionales	Enfoques Innovadores
Marco conceptual	"Nacionalismo metodológico" y "miopía estado centrista"	Perspectiva global y transnacional
Enfoque territorial	Limitado al marco del estado-nación	Considera la migración como proceso inherente al sistema global
Financiación y condicionamientos	Financiados por administraciones públicas, condicionados a necesidades político-administrativas	Búsqueda de fuentes de financiación más diversas y menos condicionadas
Objeto de estudio	Confusión entre objeto social y objeto de investigación.	Diferenciación entre objeto social y objeto de investigación
Alcance de la investigación	Centrado en condiciones de inmigración en el receptor, en particular en la Unión Europea y Estados Unidos.	Análisis integral de procesos de emigración e inmigración
Métodos principales	Modelos exploratorios diseñados por investigadores	Técnicas de aprendizaje automático y análisis de datos de redes sociales
Fuentes de datos	Bases de datos limitadas, censos nacionales	Datos de redes sociales, aplicaciones web masivas, fuentes transnacionales
Problemas de categorización	Confusión en clasificación de datos (origen, nacimiento, nacionalidad) - Dificultad para definir migración vs desplazamiento interno - Desafíos con nuevas clasificaciones (ej. movilidad en la UE y EU.)	Desarrollo de categorías más flexibles y adaptables a realidades transnacionales
Problemas conceptuales y metodológicos	Uso de instrumentos diseñados para población autóctona - Desafíos lingüísticos y culturales en la investigación.	Desarrollo de instrumentos culturalmente adaptados - Uso de IA para superar barreras lingüísticas.
Adaptaciones metodológicas	Necesidad de pruebas piloto - Investigación exploratoria - Traducciones paralelas - Entrevistadores bilingües.	Uso de tecnologías de traducción automática - Análisis de datos multilingües a gran escala.
Riesgos	Detección de patrones aleatorios en lugar de relaciones genuinas - Resultados poco reproducibles - Sesgo hacia perspectiva del país receptor. – Conclusiones debilitadas (en metanálisis).	Uso de la inteligencia artificial como caja negra - Posible sesgo en datos de redes sociales - Alucinaciones - Desafíos éticos en el uso de datos masivos.
Ventajas	Flexibilidad en la selección de variables - Profundidad en el análisis local	Mitigación de problemas en estudios computacionales - Potencial para análisis global y comparativo.
Reproducibilidad	Puede ser baja debido a la variabilidad en el diseño de modelos.	Potencialmente mayor debido a técnicas estandarizadas de aprendizaje automático.
Desafíos persistentes	Estrechez de miras impuesta por el nacionalismo metodológico - Falta de reflexión científica amplia.	Necesidad de superar concepciones ideológicas del estado-nación - Escasez persistente de bases de datos amplias y confiables.

Fuente: Elaboración Propia

CONCLUSIONES

Históricamente, el estudio cuantitativo de la movilidad impulsada por factores ambientales se ha enfrentado a la falta de bases de datos sólidas y a la excesiva libertad de los investigadores para diseñar sus modelos exploratorios, lo que ha conducido en ocasiones a resultados sesgados o poco reproducibles. Afortunadamente, la selección subjetiva de variables ambientales se mitiga gracias a los nuevos estudios computacionales que utilizan inteligencia artificial, concretamente técnicas de aprendizaje automático.

Mediante la división de la muestra en conjuntos de entrenamiento y prueba, los modelos de aprendizaje supervisado pueden reducir los grados de libertad del investigador y evaluar de manera efectiva la capacidad de generalización de los hallazgos. Esto permite a los investigadores tener un mayor control sobre la validez y robustez de sus análisis, lo que es crucial en un campo tan complejo y sensible como el estudio de los patrones migratorios impulsados por el cambio climático y otras causas ambientales.

Sin embargo, estos nuevos enfoques computacionales pueden caer en la llamada “zona de peligro” de Conway, cuando los investigadores carecen de suficientes conocimientos matemáticos o en las disciplinas de migración y estudios ambientales. En esos casos, se corre el riesgo de usar a la inteligencia artificial como una caja negra, sin la posibilidad de identificar sesgos de confirmación, la incomprensión por parte del usuario de los supuestos subyacentes, interpretaciones erróneas de los datos y alucinaciones por parte de los grandes modelos de lenguaje.

El otro desafío de los nuevos enfoques computacionales para el estudio de la migración y el desplazamiento ambientales lo comparten éstos con los estudios exploratorios: se trata de la falta de grandes bases de datos relativas al clima, la biodiversidad, el número de trayectos y otras variables propias del tema. Esta carencia se resiente especialmente en los países del Sur Global. No obstante, otras innovaciones que también emplean inteligencia artificial ya se están enfocando en procurarse sus propias bases de datos de movilidad gracias a las redes sociales y otras aplicaciones web masivas. Es decir, este reto común a los modelos exploratorios tradicionales y de los modelos computacionales podrá ser superado gracias a una segunda aplicación de las técnicas de inteligencia artificial (adicional a la identificación de las variables significativas).

Estas nuevas bases de datos no institucionales (distintas a censos y encuestas) representan una valiosa fuente de información que podrá ser utilizada para complementar las cifras oficiales y comprender mejor los patrones y tendencias migratorias a nivel global.

El aprendizaje de máquina ofrece soluciones innovadoras para detectar y corregir sesgos en los datos demográficos en general y de las redes sociales en particular. Técnicas como la detección de sesgos algorítmicos, la ponderación de los datos para mejorar la representatividad y la imputación y corrección de valores atípicos o faltantes, pueden ayudar a subsanar las limitaciones inherentes a este tipo de información recopilada a través de plataformas digitales. Esto es fundamental para garantizar que los hallazgos reflejen de manera lo más fiel posible la realidad de los patrones migratorios.

En resumen, las técnicas de aprendizaje automático son herramientas poderosas para mejorar la confiabilidad tanto de archivos históricos de migración como de la información, comercial en su origen, sobre desplazamientos físicos de usuarios, proveniente de plataformas como Facebook.

Sin embargo, la adopción de técnicas de inteligencia artificial conlleva desafíos éticos y metodológicos que deben ser abordados con sumo cuidado. Es crucial garantizar que los modelos de aprendizaje automático sean interpretables y transparentes, de modo que se puedan identificar posibles sesgos o alucinaciones. Es preciso que la minería de datos y los nuevos modelos no comprometan la privacidad de los usuarios. Asimismo, se requiere una validación continua de los modelos a medida que evolucionan los patrones de datos, a fin de mantener la fiabilidad y la relevancia de los resultados. Solo de esta manera podremos aprovechar plenamente el potencial de las tecnologías de aprendizaje automático en el estudio de la migración climática, sin perder de vista consideraciones éticas fundamentales.

Además de la “zona de riesgo” tempranamente destacada por Conway en su diagrama y que se refería a la falta de conocimientos matemáticos y estadísticos del investigador, otro de los riesgos a considerar será la confiabilidad de los datos. Los usuarios de las plataformas digitales pueden proporcionar una imagen parcial y sesgada de las poblaciones, más allá de la representatividad estadística, ya que no todas las personas tienen acceso a estas tecnologías o participan de la misma manera en ellas.

Es importante complementar estos datos con otras fuentes de información, como encuestas, registros oficiales y estudios de campo, siempre con la vision critica de los sesgos de tratamiento de datos de corte nacionalista (el nacionalismo metodológico típico de los estudios migratorios).

Asimismo, es crucial que los hallazgos obtenidos a partir de los enfoques computacionales sean contrastados con otras evidencias empíricas, menos tecnológicas y, por ello, menos occidentalocéntricas. Esto permitiría asegurar la comprensión de la movilidad ambiental en diferentes partes del mundo.

Mejorar la confiabilidad de los datos sobre migración climática y ambiental en general requiere un esfuerzo multidisciplinario. Los investigadores deben considerar cuidadosamente la vigencia de los viejos enfoques exploratorios, las limitaciones y sesgos inherentes a los nuevos modelos computacionales, tomar medidas para garantizar la privacidad y el uso ético de la información tomada de usuarios de Internet y contrastar todo lo anterior con estudios etnográficos. Solo de esta manera se podrá generar conocimiento sólido y confiable que contribuya a comprender, cuantificar y predecir la migración ambiental.

REFERENCIAS BIBLIOGRAFICAS

Aguado-Rodríguez, G. J., Quevedo-Nolasco, A., Castro-Popoca, M., Arteaga-Ramírez, R., Vázquez-Peña, M. A., & Zamora-Morales, B. P. (2016). Predicción de variables meteorológicas por medio de modelos ARIMA. *Agrociencia, 50*(1), 1-13.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Ashraf, M., Salal, Y. K., & Abdullaev, S. M. (2021). Educational data mining using base (individual) and ensemble learning approaches to predict the performance of students. In *Data Science: Theory, Algorithms, and Applications* (pp. 15-24).

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732.

Bem, D. J. (1987). Writing the empirical journal article. In M. P. Zanna & J. M. Darley (Eds.), *The compleat academic: A practical guide for the beginning social scientist* (pp. 171–201). New York: Random House.

Bohra-Mishra, P., Oppenheimer, M., & Hsiang, S. (2014). Nonlinear permanent migration response to climatic variations but minimal response to disasters. *Proceedings of the National Academy of Sciences, 111*(27), 9780–9785. https://doi.org/10.1073/pnas.1317166111

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/BF00058655

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2022). Regression discontinuity designs using covariates. *Review of Economics and Statistics, 104*(2), 229–247. https://doi.org/10.1162/rest_a_00971

Carabantes, M. (2020). Black-box artificial intelligence: An epistemological and critical analysis. *AI & Society, 35*(2), 309–317.

Conway, D. (2010). The data science Venn diagram. *Dataists*. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Dallmann, I., & Millock, K. (2017). Climate variability and inter-state migration in India. *CESifo Economic Studies, 63*(4), 560–594. https://doi.org/10.1093/cesifo/ifx014

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems* (Vol. 1857, pp. 1–15). Springer.

Drouhot, L. G., Deutschmann, E., Zuccotti, C. V., & Zagheni, E. (2023). Computational approaches to migration and integration research: Promises and challenges. *Journal of Ethnic and Migration Studies, 49*(2), 389–407.

Dumitru, S. (2023). The ethics of immigration: How biased is the field? *Migration Studies, 11*(1), 1-22.

Engle, R. F. (Ed.). (2020). *Handbook of financial econometrics: Tools and techniques (Vol. 1)*. Elsevier.

Entwisle, B., Williams, N. E., Verdery, A. M., Rindfuss, R. R., Walsh, S. J., Malanson, G. P., Mucha, P. J., et al. (2016). Climate shocks and migration: An agent-based modeling approach. *Population and Environment, 38*(1), 47–71. https://doi.org/10.1007/s11111-016-0254-y

Fadlil, A. (2022). K nearest neighbor imputation performance on missing value data graduate user satisfaction. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 6*(4), 570-576.

Feng, L., Shi, Y., & Chang, L.-T. (2021). Forecasting mortality with a hyperbolic spatial temporal VAR model. *International Journal of Forecasting, 37*(1), 255–273. https://doi.org/10.1016/j.ijforecast.2020.05.003

Feng, S., Krueger, A. B., & Oppenheimer, M. (2010). Linkages among climate change, crop yields and Mexico–US cross-border migration. *Proceedings of the National Academy of Sciences, 107*(32), 14257-14262.

Foresight, U. K. (2011). *Migration and global environmental change: Final project report*. The Government Office for Science, London. https://assets.publishing.service.gov.uk/media/5a74b18840f0b61df4777b6c/11-1116-migration-and-global-environmental-change.pdf

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning (Vol. 1)*. Springer Series in Statistics.

Garip, F. (2008). Social capital and migration: How do similar resources lead to divergent outcomes? *Demography, 45*(3), 591–617.

Garip, F. (2012). An integrated analysis of migration and remittances: Modeling migration as a mechanism for selection. *Population Research and Policy Review, 31*, 637–663.

Garip, F. (2012). Discovering diverse mechanisms of migration: The Mexico–US stream 1970–2000. *Population and Development Review, 38*(3), 393–433.

Garip, F. (2017). *On the move: Changing mechanisms of Mexico–US migration*. Princeton University Press.

Garip, F. (2023, May 19). Climate change, migration, and inequality. Ponencia en el *Data Science Institute*, University of Toronto.

Garson, G. D. (2016). *Partial least squares: Regression and structural equation models*. Statistical Associates Publishers.

Gould, R. (2021). Toward data-scientific thinking. *Teaching Statistics, 43*, S11–S22.

Gray, C., & Wise, E. (2016). Country-specific effects of climate variability on human migration. *Climatic Change, 135*, 555–568. https://doi.org/10.1007/s10584-015-1592-y

Grow, A., Flahault, A., & Devillanova, C. (2022). Are Facebook advertising data reliable proxies for population indicators? *Demography, 59*(1), 49–72. https://doi.org/10.4054/MPIDR-WP-2021-006

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., Zagheni, E., Flores, R. D., Ventura, I., & Weber, I. (2022). Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(2), S343–S363.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage, 206*, 116276.

Henry, S., Schoumaker, B., & Beauchemin, C. (2004). The impact of rainfall on the first out-migration: A multi-level event-history analysis in Burkina Faso. *Population and Environment, 25*(5), 423–460.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e1004085. https://doi.org/10.1371/journal.pmed.1004085

Jang, J.-H., Choi, J., Roh, H. W., Son, S. J., Hong, C. H., Kim, E. Y., Kim, T. Y., & Yoon, D. (2020). Deep learning approach for imputation of missing values in actigraphy data: Algorithm development study. *JMIR mHealth and uHealth, 8*(7), e16113.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing.* Pearson Education.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics, 19*(2), 81–97.

Kaparthi, S., & Bumblauskas, D. (2020). Designing predictive maintenance systems using decision tree-based machine learning techniques. *International Journal of Quality & Reliability Management, 37*(4), 659-686.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, *14*(2), 1137–1145.

Leasure, D. R., Kashyap, R., Rampazzo, F., Dooley, C. A., Elbers, B., Bondarenko, M., Verhagen, M., et al. (2023). Nowcasting daily population displacement in Ukraine through social media advertising data. *Population and Development Review*, *49*(2), 231-254. https://doi.org/10.1111/padr.12558

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leyk, S., Maclaurin, G. J., Hunter, L. M., Nawrotzki, R. J., Twine, W., Collinson, M., & Erasmus, B. (2012). Spatially and temporally varying associations between temporary outmigration and natural resource availability in resource-dependent rural communities in South Africa: A modeling framework. *Applied Geography*, *34*, 559–568. https://doi.org/10.1016/j.apgeog.2012.02.009

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1). https://doi.org/10.5334/irsp.289

Llopis Goig, R. (2007). El nacionalismo metodológico como obstáculo en la investigación sociológica sobre migraciones internacionales. *Empiria. Revista de Metodología de Ciencias Sociales*, *13*, 101–120. https://doi.org/10.5944/empiria.13.2007.1161

Molina, M., Pastukhov, A., & Kastellec, J. (2023). Causal inference in applied microeconomics: An introduction to machine learning methods. *Journal of Economic Literature*, *61*(2), 445–502. https://doi.org/10.1257/jel.20211503

Molina, M. D., Chau, N., Rodewald, A. D., & Garip, F. (2023). How to model the weather-migration link: A machine-learning approach to variable selection in the Mexico-US context. *Journal of Ethnic and Migration Studies*, *49*(2), 465-491.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*, *45*(1), 27-45.

Mueller, V., Quisumbing, A., Lee, H. L., & Droppelmann, K. (2014). Resettlement for food security's sake: Insights from a Malawi land reform project. *Land Economics, 90*(2), 222–236. https://doi.org/10.3368/le.90.2.222

Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the U.S. labor market. *The Quarterly Journal of Economics, 118*(2), 549–599.

Navarro, C. D. (2015). Migración y desempleo: Un análisis espacial para el noroeste argentino. *Documentos de trabajo, 14.* Universidad Nacional de Salta. https://www.economicas.unsa.edu.ar/ielde/archivos/docTrabajo/items_upload_items_upload-WPIelde_Nro_14.pdf

Nawrotzki, R. J., DeWaard, J., Bakhtsiyarava, M., et al. (2017). Climate shocks and rural-urban migration in Mexico: Exploring nonlinearities and thresholds. *Climatic Change, 140*, 243–258. https://doi.org/10.1007/s10584-016-1849-0

Nawrotzki, R. J., Riosmena, F., & Hunter, L. M. (2013). Do rainfall deficits predict U.S.-bound migration from rural Mexico? Evidence from the Mexican Census. *Population Research and Policy Review, 32*, 129–158. https://doi.org/10.1007/s11113-012-9251-8

Passonneau, R. J., & Carpenter, B. (2014). The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics, 2*, 311–326.

Porst, L., & Sakdapolrak, P. (2018). Advancing adaptation or producing precarity? The role of rural-urban migration and translocal embeddedness in navigating household resilience in Thailand. *Geoforum, 97*, 35–45. https://doi.org/10.1016/j.geoforum.2018.10.011

Rampazzo, F., Rango, M., & Weber, I. (2023). New migration data: Challenges and opportunities. In *Handbook of Computational Social Science for Policy* (p. 345).

Riosmena, F., Nawrotzki, R., & Hunter, L. (2018). Climate migration at the height and end of the great Mexican emigration era. *Population and Development Review, 44*(3), 455.

Robinson, C., & Dilkina, B. (2018). A machine learning approach to modeling human migration. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 1-8). https://doi.org/10.1145/3209811.3209868

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 73–79. https://doi.org/10.1002/widm.2

Sakdapolrak, P., Naruchaikusol, S., Ober, K., Peth, S., Porst, L., Rockenbauch, T., & Tolo, V. (2016). Migration in a changing climate: Towards a translocal social resilience approach. *DIE ERDE – Journal of the Geographical Society of Berlin*, *147*(2), 81–94. https://doi.org/10.12854/erde-147-6

Thober, J., Schwarz, N., & Hermans, K. (2018). Agent-based modeling of environment-migration linkages: A review. *Ecology and Society*, *23*(2). https://www.jstor.org/stable/26799102

Trinh, T. T., & Munro, A. (2023). Integrating a choice experiment into an agent-based model to simulate climate-change induced migration: The case of the Mekong River Delta, Vietnam. *Journal of Choice Modelling*, *48*, 100428. https://doi.org/10.1016/j.jocm.2023.100428

Warner, K., & Afifi, T. (2013). Where the rain falls: Evidence from 8 countries on how vulnerable households use migration to manage the risk of rainfall variability and food insecurity. *Climate and Development*, *6*(1), 1–17. https://doi.org/10.1080/17565529.2013.835707

Xu, C., Kohler, T. A., Lenton, T. M., Svenning, J.-C., & Scheffer, M. (2020). Future of the human climate niche. *Proceedings of the National Academy of Sciences*, *117*(21), 11350–11355.