# DISSERTATIO

## Estadística

COLEGIO DE GRADUADOS EN CIENCIAS ECONÓMICAS DE ROSARIO
**CONSEJO PROFESIONAL DE CIENCIAS ECONÓMICAS**
**DE LA PROVINCIA DE SANTA FE CÁMARA II**
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

# TRABAJOS FINALES

## RECENSIÓN DE TESINAS Y PRÁCTICAS PROFESIONALES DE LA CARRERA


CONSEJO PROFESIONAL DE CIENCIAS ECONOMICAS DE LA PROVINCIA DE SANTA FE
CAMARA II

UNR Universidad Nacional de Rosario

Colegio de Graduados en Ciencias Económicas de Rosario

Año VIII 2024, N.º 8, ISSN 2591-5215

# ÍNDICE


**CONFORMACIONES** 03

**UNA MIRADA HACIA EL FUTURO** 05

**ARTÍCULOS**

**IMPLEMENTACIÓN DE UN SISTEMA DE PRONÓSTICO DE ALTO RENDIMIENTO PARA EL DIMENSIONAMIENTO DEL TRÁFICO DE INTERNET EN NODOS DISTRIBUIDOS EN LA REPÚBLICA ARGENTINA** 06
LIC. CUESTA, VICTORIA

**CLASIFICACIÓN SUPERVISADA DE TEXTOS DE FICCIÓN SEGÚN GÉNERO UTILIZANDO BOSQUES ALEATORIOS** 17
LIC. GARCÍA SÁNCHEZ, SANTIAGO

**ESTRATIFICACIÓN DE POBLACIONES PARA ENCUESTAS CON PROPÓSITOS MÚLTIPLES UTILIZANDO ALGORITMO GENÉTICO** 29
LIC. GUASTELLA, MARINA GUADALUPE

**APLICACIÓN DE MODELOS DE CREDIT SCORING PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE DEFAULT UTILIZANDO INFORMACIÓN DE LA CENTRAL DE DEUDORES DEL BCRA** 42
LIC. ISAGUIRRE, MARÍA BELÉN

**CARACTERÍSTICAS DE PACIENTES CON ESQUIZOFRENIA: ANÁLISIS MEDIANTE MODELOS DE ECUACIONES ESTRUCTURALES** 54
LIC. LUST RIMOLDI, GRECIA

**META-ANÁLISIS PARA LA DETERMINACIÓN DE UN MODELO GENÉTICO QUE DESCRIBA LA ASOCIACIÓN ENTRE EL POLIMORFISMO FOK1 Y DIABETES TIPO 2** 66
LIC. MARTINEZ, JOSEFINA

**OPTIMIZACIÓN DEL ESPACIO EN GÓNDOLA PARA PRODUCTOS DE PRIMERA NECESIDAD A TRAVÉS DE MODELOS LINEALES DE PREDICCIÓN** 76
LIC. MELGRATTI, MATÍAS

**ESTUDIO DEL COMPORTAMIENTO DE LA VARIABLE ALTURA ELIPSOIDAL, UTILIZANDO HERRAMIENTAS DE LA ESTADÍSTICA ESPACIAL. CIUDAD DE ROSARIO, AÑO 2010** 85
LIC. SUAREZ, MARINA ALDANA

# CONFORMACIONES

**Año VIII 2024, N.º 8, ISSN 2591-5215**



Esta revista se pone a disposición de los profesionales matriculados al Consejo Profesional de Ciencias Económicas de la Provincia de Santa Fe Cámara II (CPCE), asociados del Colegio de Graduados en Ciencias Económicas de Rosario (CGCE), estudiantes y docentes de la Facultad de Ciencias Económicas y Estadística (FCEyE) de la Universidad Nacional de Rosario (UNR) y otras Instituciones vinculadas al quehacer profesional y académico.



El contenido de los trabajos finales no necesariamente refleja la opinión de los Comités responsables de esta publicación digital.

Las Instituciones no son responsables por el contenido de las informaciones y opiniones que viertan en esta revista quienes son identificados como autores de dichos trabajos finales, en todos los casos deberán ser cotejadas por los Profesionales y/o las fuentes.

DISEÑO: AXYOMA GLOBAL
Guadalupe Rocio Velazquez
Mateo Skalko

**Córdoba 2035**
S2000DFB, Rosario, Argentina
+54 (341) 5899241 / +54 (341) 2295698

# EDITORIAL

# UNA MIRADA HACIA EL FUTURO

En la presente edición de Dissertatio Estadística (UNR-CGCE-CPCE) se publica la recensión de siete tesinas y de un informe de práctica profesional. En todos los casos se trata de trabajos realizados por sus autores para cumplimentar con el último requisito obligatorio y poder alcanzar el título de Licenciado/a en Estadística, bajo la dirección y asesoramiento de docentes de la Escuela de Estadística, de la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario, o con la tutoría de profesionales estadísticos de instituciones públicas o privadas.

Esta octava edición de la Revista Dissertatio es la última de una serie que se publica de manera ininterrumpida desde su primera aparición en el año 2017. En un presente donde las políticas estatales atentan contra la Universidad Pública, el trabajo conjunto de docentes, estudiantes, directivos y profesionales es bastión de resistencia para sostener estos espacios de creación y difusión del conocimiento. Es de destacar la relevancia que tiene este tipo de revista digital al propiciar la difusión de una selección de trabajos finales de los/as egresados/as, teniendo además como objetivo incentivar la investigación y favorecer el tránsito de la vida académica a la profesional.

En los trabajos presentados en esta oportunidad se puede encontrar una gran variedad de abordajes de problemáticas y metodologías. Esto es una muestra de la diversidad de campos en los que se aplica la Estadística. Por un lado, se incluyen aplicaciones de métodos estadísticos en datos relacionados a la salud: un trabajo sobre las características de pacientes con esquizofrenia (Lic. Grecia Lust Rimoldi) y un meta-análisis para describir la asociación entre el polimorfismo fok1 y la diabetes tipo 2 (Lic. Josefina Martínez). Por otro lado, se presenta una aplicación en el área de scoring crediticio (Lic. María Belén Isaguirre) y un trabajo que aborda un problema de optimización del espacio en góndola para productos de primera necesidad (Lic. Matías Melgratti). Asimismo, se muestra cómo puede abordarse el estudio del relieve de la superficie terrestre en la ciudad de Rosario (Lic. Marina Suarez), factor fundamental a la hora de diseñar soluciones de planeamiento urbano y/o de gestión de emergencias. Por su parte, la Lic. Marina Guastella ofrece el análisis de métodos muestrales estratificados para encuestas. Otros de los trabajos aquí incluidos recogen aplicaciones de aprendizaje supervisado: el trabajo de la Lic. Victoria Cuesta se enfoca en el pronóstico con series temporales para grandes bases de datos en el área de tecnologías mientras que el Lic. Santiago García Sánchez aborda una aplicación de clasificación de textos literarios.

Desde el Comité Editorial aprovechamos la oportunidad para agradecer y felicitar a todas aquellas personas que hacen posible cada edición, en especial a quienes colaboraron en el presente año en un contexto tan particular donde resulta menester defender la ciencia y la educación superior para todos/as. Las instituciones participantes, Escuela de Estadística de la FCEyE, el Colegio de Graduados en Ciencias Económicas de Rosario y el Consejo Profesional en Ciencias Económicas de la Provincia de Santa Fe -Cámara II-, demuestran año a año un fuerte interés y compromiso en seguir fortaleciendo este valioso espacio de difusión técnico- estadístico, de acceso libre y gratuito al público en general. Una vez más, apostamos a seguir creciendo. La salida es colectiva.

# IMPLEMENTACIÓN DE UN SISTEMA DE PRONÓSTICO DE ALTO RENDIMIENTO PARA EL DIMENSIONAMIENTO DEL TRÁFICO DE INTERNET EN NODOS DISTRIBUIDOS EN LA REPÚBLICA ARGENTINA


## Lic. Cuesta, Victoria

**Directora: Mg. Méndez, Fernanda**



La necesidad de procesar grandes volúmenes de datos está creciendo rápidamente, y cada vez es más frecuente que las empresas basen sus decisiones en la inmensidad de información recolectada. En este contexto, el análisis de series temporales gana importancia progresivamente, ya que permite utilizar datos históricos para estimar valores futuros y prever eventos. Este trabajo se basa en automatizar el proceso de pronóstico del tráfico de internet en miles de nodos ópticos distribuidos por la República Argentina, esencial para optimizar inversiones y mantenimiento. Para lograrlo se desarrolló un sistema de pronóstico de alto rendimiento, utilizando modelos como ARIMA con errores XGBoost, ETS, Prophet con errores XGBoost y una Red Neuronal Autorregresiva. Todo el sistema se ha desarrollado íntegramente en el lenguaje de programación R. El resultado final es un pronóstico para cada nodo y tipo de señal (downstream o upstream), basado en su historial y en el modelo seleccionado como el mejor entre los ajustados. Además, se ha desarrollado una aplicación interactiva que permite al usuario visualizar la ubicación de los nodos caracterizada por su nivel de utilización y las gráficas de su historial y pronóstico, como así también monitorear el rendimiento del sistema en tiempo real. Aunque requiere gran capacidad computacional, beneficia optimizando inversiones, mejorando la calidad del servicio y liberando tiempo de los analistas. El sistema desarrollado, aunque fue diseñado para un caso específico, es una innovación práctica y replicable para pronosticar múltiples series de tiempo, de distinta naturaleza, de forma automática en diversas aplicaciones.


## INTRODUCCIÓN

Anticipar y prever escenarios futuros es clave para la toma de decisiones acertadas. Aunque existen desarrollos matemáticos y estadísticos para ajustar modelos de pronóstico, muchos requieren intervención humana para seleccionar el mejor. Entonces, surge el interrogante sobre qué sucede cuando la cantidad de datos es tanta que supera la capacidad humana de análisis.

El estudio de grandes volúmenes de series temporales de forma automática se vuelve crucial con el crecimiento del *Big Data*. Se generan enormes cantidades de datos de alta calidad en áreas como, por ejemplo, *IoT* (*Internet of Things*), atención médica digital, ciudades inteligentes, entre otras. Los datos generados por cualquier mecanismo de seguimiento pueden ser pronosticados para anticipar eventos y optimizar o prever pérdidas significativas. Aquí es donde la necesidad de procesar grandes volúmenes de series impulsa la creación de un sistema de pronóstico de alto rendimiento, basado en algoritmos informáticos que utilizan datos históricos y en tiempo real para predecir con precisión eventos futuros.

Para garantizar un alto rendimiento, es fundamental contar con una gran cantidad de datos precisos y actualizados, así como someter los modelos y algoritmos a una rigurosa validación. Esto asegura la confiabilidad y precisión de las predicciones realizadas.

Automatizar tareas en empresas que manejan grandes volúmenes de datos puede optimizar la eficiencia y el capital invertido. Este trabajo se enfoca en describir un sistema de pronóstico de alto rendimiento que pronostique múltiples series de tiempo, detallando los pasos desde la recolección hasta el pronóstico.

Se implementa este sistema en una empresa de telecomunicaciones con el fin de automatizar los cálculos de tráfico y pronósticos de internet en la red HFC (*Hybrid Fiber-Coaxial*), mejorando así su dimensionamiento. La predicción del tráfico en nodos es crucial para la gestión de redes de comunicaciones, optimizando la capacidad y mejorando la calidad del servicio. Se busca mejorar los procesos de pronóstico, que hasta el momento se basan en tendencias lineales sin base teórica.

## OBJETIVO

El objetivo principal de la tesina es mostrar la implementación de un sistema de pronóstico de alto rendimiento, aplicando modelos tanto de aprendizaje automático

como de estadística clásica para predecir el tráfico de internet de más de 9800 nodos de red HFC repartidos en toda la República Argentina.

## METODOLOGÍA

El trabajo describe detalladamente los pasos necesarios para la implementación de un sistema de pronóstico de alto rendimiento (Figura 1).

Figura 1. Etapas fundamentales de un sistema de pronóstico de alto rendimiento


- **Recolección de datos:** el sistema comienza con la identificación y recopilación de información proveniente de distintas fuentes relevantes para el cálculo de las series de tiempo que se desean pronosticar. Para automatizar este proceso, se puede utilizar un algoritmo que permita la integración de diferentes fuentes, tales como bases de datos, registros, información de proveedores, entre otros y que realice los cálculos necesarios. La complejidad de este algoritmo dependerá del nivel de sofisticación de las fuentes y de los cálculos requeridos.
- **Limpieza y preparación:** la calidad de los datos influye significativamente en la precisión de los modelos predictivos. Es importante garantizar que los datos estén ordenados cronológicamente y en el formato adecuado, así como no presentar valores faltantes ni anomalías que podrían sesgar los resultados. Para el tratamiento de valores faltantes es necesario examinar las posibles razones detrás de su aparición, ya sea debido a causas específicas o simplemente como errores aleatorios. Se presentan estrategias para abordar este problema: en casos donde la ausencia de datos esté justificada (por ejemplo, días festivos en un estudio de ventas), se pueden utilizar variables ficticias para identificar estos eventos; mientras que, para fallos o errores aleatorios, la interpolación lineal en series no estacionales o la

descomposición robusta de Loess de tendencia estacional (STL) en series estacionales pueden ser buenas alternativas. Por otro lado, para detectar y corregir anomalías se aplica un novedoso método, el cual utiliza la descomposición STL para detectar los *outliers* en la componente residual; este enfoque ha demostrado ser efectivo y robusto en series de distintas naturalezas. Por último, se tiene en cuenta que en las etapas subsiguientes se pretende implementar diversos modelos, incluyendo modelos de aprendizaje automático, algunos de los cuales requieren que los datos se ingresen de manera específica para poder procesar las características de las series, como por ejemplo desglosar la fecha en mes y año o normalizar los valores.

- **Ajuste de modelos:** en esta fase, es crucial considerar cómo se mide la precisión de cada modelo, dado que se busca su automatización. Existe un método, dividiendo los datos en un conjunto de datos de entrenamiento y otro de prueba, que es ejecutable sin intervención humana. Enfocándose únicamente en los datos de entrenamiento, se procede con la estimación de los parámetros de cada modelo. Se pueden elegir tantos modelos como se deseen, para cada uno se realiza el pronóstico con un horizonte temporal equivalente al número de observaciones que se separaron en los datos de prueba. De esta manera, al comparar los pronósticos generados con las observaciones reales del conjunto separado para prueba, podemos calcular los errores de pronóstico y evaluar la precisión de cada uno de ellos calculando la cantidad y tipo de métricas requeridas.
- **Selección del mejor modelo:** una vez que se han ajustado todos los modelos deseados, se selecciona aquel que sea señalado como el mejor por la mayor cantidad de métricas. En algunos casos, puede ocurrir un empate cuando la mitad de las métricas favorecen a un modelo en particular mientras que la otra mitad favorece a otro modelo diferente. En tal caso se combinan los pronósticos de todos los modelos empatados.
- **Re-ajuste, predicción y ensamble:** una vez seleccionado el o los modelos que mejor se ajustan a cada serie, se realiza una re-estimación de parámetros con el conjunto total de datos (entrenamiento + prueba). Este paso es fundamental, ya que los parámetros utilizados para la selección del modelo fueron estimados solo con una parte de la serie. Teniendo los

parámetros re-ajustados se procede a realizar el pronóstico con el horizonte deseado. En los casos que se encontraron empates de modelos se realiza el re-ajuste y la predicción con todos ellos; con los pronósticos resultantes se lleva a cabo un ensamble, que consiste en un promedio de las predicciones correspondientes al mismo momento en el tiempo.

## Modelos

Existen numerosos modelos de series de tiempo que pueden emplearse para hacer pronósticos. En este caso y teniendo en cuenta que el objetivo es pronosticar el tráfico de internet de nodos ubicados en toda la Argentina, se utilizan tanto modelos de estadística clásica como de *machine learning* o la combinación de ambos, lo que permite abordar la heterogeneidad inherente de estas series de manera efectiva. Los modelos utilizados son:

- ARIMA con errores XGBoost
- ETS (Error, Trend, Seasonality)
- Prophet con Errores XGBoost
- NNAR (Nonlinear neural network autoregressive)

## Implementación

Siguiendo los pasos enunciados anteriormente se puede obtener una predicción precisa sobre datos temporales teniendo en cuenta diversos modelos. Cada uno de los pasos fueron diseñados específicamente para que puedan ejecutarse de forma automática mediante algoritmos, sin la necesidad de intervención humana, lo cual permite escalar el proceso a múltiples series. No obstante, hay que tener en cuenta que el proceso puede resultar costoso en términos computacionales y requerir una mayor capacidad de procesamiento dependiendo del número de series que se desee pronosticar.

Para llevar a cabo la implementación, se utiliza el lenguaje de programación R. Este compila una serie de librerías que facilita las tareas de automatización necesarias para ejecutar el sistema. Si bien este trabajo se desarrolla en dicho lenguaje, es posible replicar la lógica con cualquier otro lenguaje que se desee.

## Monitoreo y optimización

Las series temporales tienen la particularidad que, con el paso del tiempo, se conoce el valor real de lo pronosticado, lo que permite observar el error cometido y su distribución. El último paso consiste en monitorear las predicciones realizadas, tarea que, si bien no puede realizarse de forma automática (el usuario debe observar los resultados y a su criterio tomar decisiones), debe efectuarse para controlar el desempeño del sistema.

## APLICACIÓN

Una empresa de telecomunicaciones se enfrenta a la problemática de tener que pronosticar el tráfico de internet en nodos ópticos de la red HFC. Con alrededor de 9800 nodos distribuidos en toda Argentina, la tarea de dimensionar el crecimiento de tráfico en ellos es crucial para prever posibles colapsos de equipos y determinar qué áreas son más rentables para invertir, entre otros beneficios.

La red HFC es un sistema de comunicaciones de banda ancha que combina fibra óptica y cables coaxiales para llevar señales de alta velocidad a los hogares y empresas. En este tipo de red, la señal de internet viaja desde el proveedor de servicios hasta un nodo óptico, que se encuentra en una ubicación central en el vecindario, a través de fibra óptica (FO). Desde allí, la señal se transmite a los hogares por medio de cables coaxiales.

Los nodos ópticos son los puntos de distribución de la señal en el vecindario y pueden cubrir aproximadamente mil hogares. Cada nodo es responsable de enviar y recibir señales a través de los cables coaxiales que se conectan a cada hogar o empresa. Cada nodo tiene una capacidad de tráfico limitada, lo que significa que, si se sobrecarga con demasiada demanda, la calidad del servicio puede disminuir significativamente. Por esta razón, es importante dimensionar adecuadamente el crecimiento de tráfico en cada uno de ellos para prevenir posibles colapsos de equipos y asegurar que la señal llegue a su destino sin pérdida de calidad.

Por último, es importante comprender la definición de tráfico de internet, el cual se refiere a la cantidad de datos que se emiten y reciben a través de la red. Cada vez que un usuario accede a una página web, descarga archivos o incluso envía correos electrónicos, genera tráfico. Éste puede viajar en dos direcciones, *upstream* o *downstream*. El primero (*upstream*) se refiere a la información que viaja desde el usuario hacia el proveedor, por ejemplo, cuando se envía un correo electrónico, y el

segundo (*downstream*) hace referencia a la información que viaja desde el proveedor hacia el usuario, por ejemplo, cuando se descarga un archivo. Además, todos los elementos de la red pueden medir y registrar la cantidad de tráfico que pasa a través de ellos en ambas direcciones, lo que resulta fundamental para su dimensionamiento.

## Materiales

Se cuenta con información del pico máximo de tráfico diario registrado por cada nodo óptico, tanto para *upstream* como para *downstream*. Existen más de 9800 nodos distribuidos en toda la Argentina, considerando que se registran valores para ambos sentidos de las señales se cuenta con más de 19000 registros diarios.

Esta información se ha estado recolectando desde abril de 2020 o desde el encendido del equipo (si esto ocurriese después de la fecha mencionada) hasta abril de 2023. Además, los nodos se encuentran georreferenciados y se dispone de su capacidad al momento de la medición para ambos sentidos de tráfico.

Para llevar a cabo el procesamiento de datos, se dispone de un servidor modelo Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz, con 20 procesadores y una memoria de 310 Gb. Además, opera con sistema operativo Linux.

Para alcanzar el objetivo planteado, es necesario realizar pronósticos de valores mensuales que indiquen el tráfico máximo esperado tanto en la dirección *downstream* como en la dirección *upstream* para cada nodo. El horizonte de predicción abarca un período de dos años, lo que permite disponer de tiempo suficiente para llevar a cabo las tareas de mantenimiento necesarias, como la construcción de nuevos nodos u otras actividades que requieran una planificación anticipada. Es crucial que estos pronósticos se actualicen de forma mensual y automática. El propósito final de este trabajo es proporcionar a la compañía una herramienta interactiva que contenga la información necesaria para llevar a cabo una planificación eficiente de sus redes.

## Resultados

La ejecución completa del sistema, desde la recolección de datos hasta la generación de pronósticos de las series, requiere un tiempo aproximado de 7 a 8 horas, utilizando 18 procesadores del servidor mencionado.

Un total de 18102 series, lo que equivale a 9051 nodos, han obtenido pronósticos exitosos. Los 830 nodos restantes corresponden a aquellos que se han excluido debido a su poca historia o se encuentran apagados en abril de 2023. Para los 522 nodos que tienen menos de 12 meses de historia, se ha optado por asignarles la tendencia de crecimiento que se observa en la suma de los nodos pertenecientes al mismo *hub* y al mismo sentido de la señal.

Utilizando las medidas de precisión *Mean Absolute Percentage Error* (MAPE), Mean Absolute Scaled Error (MASE), *Mean Absolute Error* (MAE) y *Root Mean Squared Error* (RMSE) para comparar el ajuste de los distintos modelos, en la Tabla 1 se expone cuántas veces cada modelo ha realizado el mejor ajuste.

Tabla 1. Cantidad de mejores ajustes por modelo

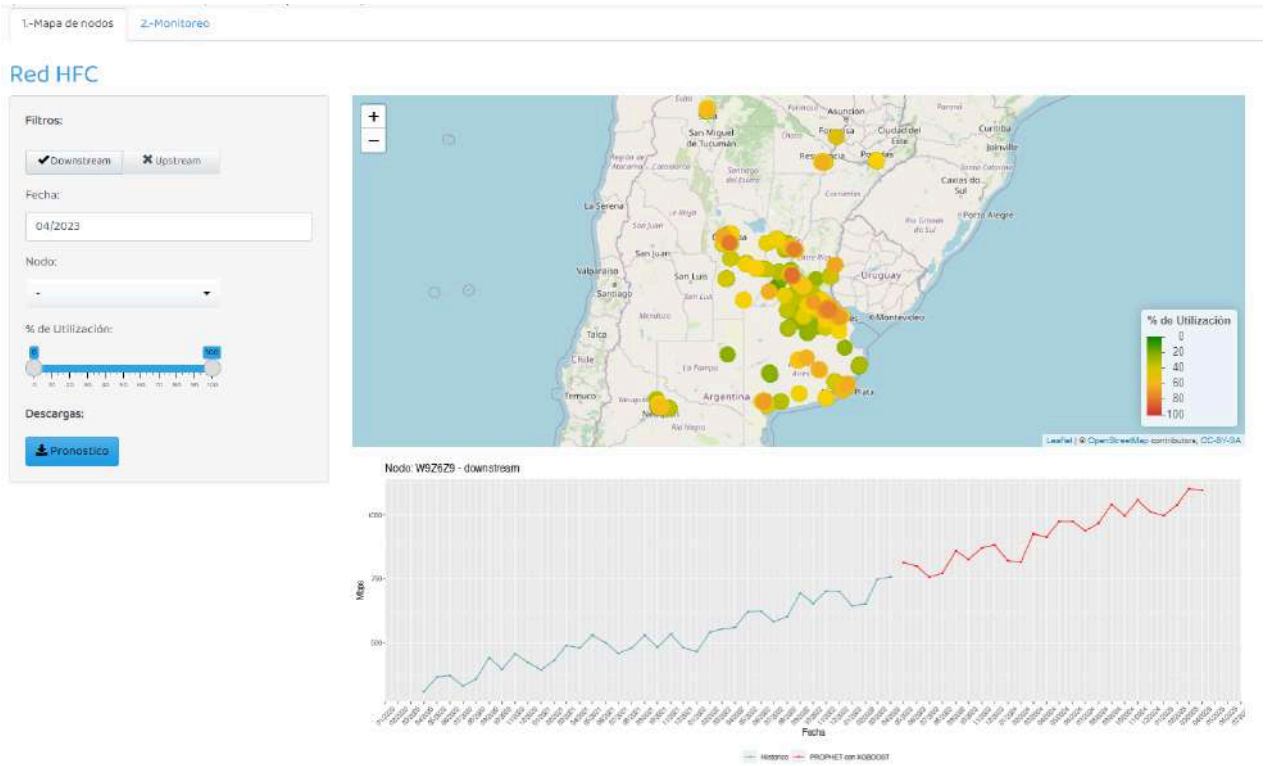
| Modelo | Frecuencia |
|---|---|
| ARIMA con errores XGBoost | 4204 |
| ETS | 5545 |
| Prophet con errores XGBoost | 2219 |
| Red Neuronal | 5882 |
| Ensamble | 250 |

Los modelos ETS y Red Neuronal han demostrado ajustarse con menor error sobre el total de series, y en 250 ocasiones se ha producido un empate en la elección del mejor modelo, lo que ha llevado a utilizar un ensamble de los modelos seleccionados. Sin embargo, es importante tener en cuenta que los modelos se han ejecutado utilizando los valores predeterminados de sus hiperparámetros, esto indica que existe un amplio margen para lograr mejores resultados en términos de precisión.

También se ha observado que 159 series ajustadas presentan valores menores o iguales a cero en algún momento del pronóstico, lo cual es incorrecto ya que no existe tráfico negativo y tampoco se espera que los nodos se apaguen. Estos casos, aunque representan solo el 1% de las series pronosticadas, pueden mejorarse mediante la optimización de hiperparámetros.

Para concluir, se desarrolló una aplicación interactiva que proporciona la capacidad de visualizar y descargar los pronósticos de todos los nodos, además de mostrar su estado de utilización, ubicación en el mapa y monitoreo del rendimiento en tiempo real (Figura 2).

Figura 2. Visualización de la interfaz gráfica de la aplicación interactiva




## CONSIDERACIONES FINALES

Este trabajo se enfoca en desarrollar un sistema que realice pronósticos confiables de forma automática, para poder escalar el mismo a un gran número de series de tiempo. Este sistema, si bien ha sido aplicado al tráfico de internet, tiene la capacidad de replicarse con cualquier conjunto de series.

Entre las ventajas del sistema desarrollado, se destacan las siguientes:

- No requiere volver a entrenar los modelos en intervalos regulares; a medida que se actualice la información y se ejecute el sistema, éste selecciona automáticamente el mejor modelo y lo entrena con los datos actualizados.
- El sistema puede funcionar de manera automática al ser programado a través de un orquestador o al ejecutar secuencialmente cada uno de sus pasos. Esto tiene la ventaja de eliminar la necesidad de intervención activa por parte de un analista especializado para obtener los pronósticos deseados.
- La automatización del proceso permite una mayor eficiencia y reproducibilidad en la generación de pronósticos, agilizando la obtención de resultados precisos y confiables.

- El sistema utiliza tanto técnicas clásicas de estadística como innovadoras de aprendizaje automático, así como diversas medidas de precisión, lo que brinda un amplio abanico de posibilidades para modelar cualquier tipo de serie.

Sin embargo, se deben tener en cuenta algunas desventajas:

- La selección del modelo mediante la división de la serie en dos partes requiere disponer de una cantidad considerable de información histórica y aun así no garantiza que el modelo seleccionado sea el más adecuado para esos datos.
- Es necesario que un humano supervise los resultados para detectar posibles errores.
- El sistema implica un alto costo computacional, que puede traducirse en tiempo o costo monetario.

Existen alternativas para mitigar estas desventajas. El problema de selección de modelos puede mejorarse significativamente utilizando validación cruzada, aunque recurrir a esta técnica requeriría mayor poder computacional. Con respecto a esto último, se ha comprobado que es posible ejecutar el sistema por partes, obteniendo resultados satisfactorios con una computadora hogareña de 32 GB de memoria RAM y 8 procesadores, finalizando todo el proceso en 24 horas. Este resultado es aceptable considerando la magnitud de los cálculos involucrados. Otra opción para ejecutar el sistema de forma más rápida, si no se cuenta con un servidor de gran potencia de procesamiento, es utilizar una plataforma de servicios en la nube. Esta alternativa, si bien no es económica, es especialmente adecuada cuando se necesita que el sistema se ejecute en poco tiempo y no se dispone de un servidor potente.

En resumen, el sistema de pronóstico desarrollado ha demostrado ser altamente escalable y adaptable a las necesidades del usuario, siendo una solución versátil y valiosa para aquellos que necesitan pronósticos precisos y eficientes en un entorno de gran escala. Su capacidad para generar pronósticos confiables para múltiples series de tiempo lo convierte en una herramienta de gran utilidad para diversas áreas de aplicación. La implementación exitosa de este sistema permite obtener resultados precisos y confiables en el pronóstico de datos, lo que puede beneficiar significativamente a diferentes sectores y campos de estudio que requieren de análisis predictivos.

## BIBLIOGRAFÍA


Bajaj, A. (2021). Neptune. Time Series Prediction vs. Machine Learning. Recuperado en abril de 2023 https://neptune.ai/blog/time-series-prediction-vs-machine-learning

Benrhmach, G., Namir, K., Namir, A. & Bouyaghroumni, J. (2020). Nonlinear Autoregressive Neural Network and Extended Kalman Filters for Prediction of Financial Time Series. Computational Intelligence and Neuroscience, 2020, 1-14. https://doi.org/10.1155/2020/4674261

Cleveland, R. B., Cleveland, W. S., McRae, J. E. & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1), 3-73.

Dancho, M. (2021). High Performance Time Series. Business Science. Recuperado en marzo 2023 de https://university.business-science.io/p/high-performance-time-series

Hyndman, R. J. & Athanasopoulos, G. (2018). Forecasting: principles and practice (2nd ed.). OTexts.

Hyndman, R. J. & Athanasopoulos, G. (2021). Forecasting: principles and practice (3rd ed.). OTexts.

Nielsen, A. (2019). Practical Time Series Analysis: Prediction with Statistics and Machine Learning. O'Reilly Media.

Taylor, S. J. & Letham, B. (2017). Forecasting at scale. PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2

Vallis, O., Hochenbaum, J. & Kejariwal, A. (2014). A novel technique for long-term anomaly detection in the cloud.

Zambrano, R. & Bartolomé, K. (2021). Rafael Zambrano: Múltiples modelos sobre múltiples series de tiempo: Un enfoque Tidy. Recuperado en marzo 2023 https://rafael-zambrano-blog-ds.netlify.app/posts/seriestemporales/

# CLASIFICACIÓN SUPERVISADA DE TEXTOS DE FICCIÓN SEGÚN GÉNERO UTILIZANDO BOSQUES ALEATORIOS


## Lic. García Sánchez, Santiago

**Director: Mg. Marfetán Molina, Diego**
**Codirector: Mg. Prunello, Marcos Miguel**



El aprendizaje automático es la disciplina que estudia el desarrollo y aplicación de algoritmos que descubren patrones presentes en datos provistos como entrada, generalmente con el propósito de realizar predicciones sobre nuevos datos. Entre estos algoritmos se encuentran los árboles de decisión, metodología que consiste en encontrar una serie de reglas para clasificar las observaciones dentro de determinadas categorías. Una mejora para esta técnica se presenta con el algoritmo de bosques aleatorios, que combina las conclusiones de múltiples árboles independientes para obtener resultados más precisos.

Este trabajo tiene como objetivo evaluar el rendimiento de bosques aleatorios en la tarea de clasificar textos de ficción en español dentro de siete géneros literarios predefinidos. Se plantean ocho escenarios de análisis para sintetizar la información de dichos textos, utilizando distintas formas de segmentarlos y construir variables a partir de ellos.

En general, los mejores resultados se obtienen al utilizar como variables a las frecuencias absolutas de las palabras. El algoritmo de bosques aleatorios permite obtener una precisión global aceptable, aunque la performance de la clasificación varía notablemente según el género literario.


## INTRODUCCIÓN

Se denomina aprendizaje automático o *machine learning* a la disciplina que estudia el desarrollo y aplicación de algoritmos que descubren patrones presentes en datos provistos como entrada. Estos patrones pueden ser usados posteriormente para realizar predicciones sobre nuevos datos. Las técnicas de aprendizaje automático son a menudo agrupadas en dos categorías:

- Aprendizaje supervisado: algoritmos que se aplican cuando se cuenta con la presencia de valores conocidos de una variable respuesta ($Y$) para guiar el proceso de aprendizaje. Frecuentemente tienen como objetivo predecir futuros valores de $Y$.
- Aprendizaje no supervisado: algoritmos que se utilizan cuando no se dispone de una variable respuesta. En estos casos, se tiene únicamente un conjunto de variables explicativas $X_1, X_2, ..., X_p$ y el objetivo suele ser hallar relaciones entre dichas variables o entre las observaciones, para describir cómo se organizan o agrupan los datos.

Una de las áreas de aplicación de los algoritmos de aprendizaje automático es la clasificación de textos, el proceso por el cual una serie de documentos son asignados a categorías o clases. En este contexto, un documento se puede definir como una unidad de datos textuales dentro de una colección, generalmente relacionado con algún documento del mundo real como un artículo, noticia o correo electrónico. A esta colección o conjunto de documentos se la denomina corpus.

El objetivo principal de este trabajo es investigar el uso de técnicas de aprendizaje supervisado para clasificar automáticamente obras de ficción, tales como novelas o cuentos, dentro de sus respectivos géneros literarios. Se planteó además explorar distintas metodologías para resumir la información de los documentos (es decir, cada obra de ficción), creando distintos escenarios de análisis y determinar cuál de ellos produce una mejor precisión en la clasificación.

## METODOLOGÍA

### Recolección de datos y etiquetado

Para la construcción del conjunto de datos se recurrió a diversos sitios *web* que ofrecen libros gratuitos y sin derechos de autor. Se incluyeron textos de ficción

escritos en español o traducidos a dicho idioma, excluyendo documentos que no sean de ficción (por ejemplo, manuales o textos periodísticos).

Seguidamente, se le asignó a cada documento una etiqueta, correspondiente a su género literario. Estas son las categorías que posteriormente se predijeron con las técnicas de aprendizaje supervisado. Para evitar ambigüedades al momento de adjudicarle un género a cada libro, se recurrió al sitio *web* Goodreads (www.goodreads.com). Dicho sitio permite a los usuarios asignarle una o más etiquetas a cada libro, por lo que el criterio utilizado fue tomar la etiqueta más votada como verdadera.

### Preprocesamiento

Antes de construir las variables que se utilizan para la clasificación de documentos, se debe realizar una serie de procedimientos colectivamente conocidos como preprocesamiento de los datos. Entre estas tareas se encuentran:

- Importar los textos al software, en este caso R (versión 4.1.3).
- Limpiar los textos, removiendo elementos innecesarios tales como signos de puntuación o números de página.
- Convertir todos los textos a minúscula homogeneizando formatos de escritura.
- Eliminar palabras vacías (también llamadas *stopwords*). Estas son palabras muy comunes en los documentos que ofrecen poca o nula información relevante. Por ejemplo, se remueven artículos, preposiciones y conjunciones.
- Lematizar las palabras de los textos. El proceso de lematización consiste en transformar cada término en su lema correspondiente, es decir, la forma en la que aparece en un diccionario. Por ejemplo, los sustantivos y adjetivos pasan a su forma masculina y singular y los verbos al infinitivo. Este proceso permite que la computadora reconozca las diferentes variantes de una palabra como distintas instancias de la misma.
- Segmentar los documentos. La segmentación (también llamada *tokenización*) es el proceso por el cual un documento es dividido en unidades significativas de texto conocidos como *tókenes*. Un *token* puede ser una letra, palabra o conjuntos de palabras. En particular, en el análisis de textos se suele trabajar con $n$-gramas, secuencias de "$n$" palabras consecutivas. En el caso particular en el que $n = 2$, esta se denomina bigrama, mientras que si $n = 1$ (es decir,

cuando se trabaja con palabras individuales) esta secuencia se denomina unigrama.

Para este trabajo, se realiza una segmentación en unigramas y otra en bigramas, generando así dos conjuntos distintos de variables a partir del mismo corpus.

Debido a que se trabaja con libros de ficción, es de esperar que entre las palabras con mayor frecuencia aparezcan nombres propios, correspondientes a personajes o lugares de las obras literarias. Se decidió entonces estudiar el efecto de estos términos en la clasificación de textos, evaluando si su presencia produce un peor desempeño debido al sobreajuste (fenómeno por el cual el algoritmo adopta criterios que arrojan buenos resultados en el conjunto de datos provisto, pero que resultan ineficaces en nuevos datos). Para ello, se crearon nuevas variantes para ambos conjuntos de datos mencionados previamente, en las cuales se filtraron las palabras reteniendo aquellas que figuran en el Diccionario de la lengua española de la RAE y descartando las que no. De este modo, se puede determinar si la presencia de nombres propios afecta negativamente o no a la clasificación de documentos de ficción.

## Análisis de textos

A diferencia de lo que ocurre en otras disciplinas, en el análisis estadístico de textos, las variables no se observan o miden de forma directa, sino que deben ser construidas por el investigador. Existen distintas formas de resumir la información provista por un documento.

La estrategia más sencilla es utilizar la frecuencia de términos (en inglés, *term frequency*). Esta se trata simplemente de la frecuencia absoluta de un determinado término en un documento. La frecuencia absoluta del término *t* en un documento *d* se representa como $tf_{td}$. A este enfoque, que consiste en considerar a cada documento como un vector numérico conteniendo las distintas palabras y sus respectivas frecuencias, se lo conoce como bolsa de palabras.

Otra estrategia posible consiste en calcular la frecuencia inversa del documento (*inverse document frequency* o *idf*) de los términos. La *idf* de un término *t* se calcula como:

$$idf_t = ln\left(\frac{Nº\ total\ de\ documentos\ en\ el\ corpus}{Nº\ total\ de\ documentos\ que\ contienen\ a\ t}\right).$$

Esta medida le asigna un valor bajo a términos muy comunes, mientras que un término infrecuente recibirá un valor alto. El menor valor posible se da en el caso en el que un término aparezca en todos los documentos del corpus.

Al combinar la frecuencia de términos con la frecuencia inversa del documento se obtiene la estadística “frecuencia de término - frecuencia inversa de documento” (más conocida como TF-IDF por sus siglas en inglés). Se utiliza para medir qué tan importante es una palabra respecto a un documento dentro de un corpus. El peso que se le asigna al término *t* en el documento *d* es:

$$tf - idf_{t,d} = tf_{t.d} \times idf_{t}.$$

Este enfoque le asigna una mayor ponderación a aquellos términos que presenten altas frecuencias absolutas en unos pocos documentos. Las palabras muy poco frecuentes, o que aparecen en una gran proporción de documentos del corpus, reciben bajas ponderaciones.

Todas las medidas mencionadas anteriormente pueden ser generalizadas para n-gramas con $n > 1$, para lo cual simplemente se trata a la secuencia de n palabras como si fuera un único término y se procede análogamente. Para este trabajo, se decidió calcular la frecuencia de términos y la estadística TF-IDF tanto para unigramas como para bigramas. Al contar además con dos versiones para estos conjuntos de datos (una que mantiene los nombres propios y una donde se eliminaron), esto dejó un total de 8 escenarios de análisis, presentados en la Tabla 1.

Para cada uno de estos escenarios, se construye su matriz de documentos-términos correspondiente. Esta estructura consiste en una matriz de dimensiones $m \times n$, siendo $m$ el número de documentos en el corpus y $n$ el número de total de tókenes (unigramas o bigramas según el caso). En cada celda $(i, j)$ se coloca entonces el valor de la frecuencia absoluta (o TF-IDF) del *token* $j$ en el documento $i$.

Tabla 1. Escenarios de análisis

| Escenario | Segmentación realizada | Medida calculada | Filtro de nombres propios |
|---|---|---|---|
| A | Unigramas | Frecuencia de términos | No |
| B | Bigramas | Frecuencia de términos | No |
| C | Unigramas | TF-IDF | No |
| D | Bigramas | TF-IDF | No |
| E | Unigramas | Frecuencia de términos | Sí |
| F | Bigramas | Frecuencia de términos | Sí |
| G | Unigramas | TF-IDF | Sí |
| H | Bigramas | TF-IDF | Sí |

## Aprendizaje supervisado

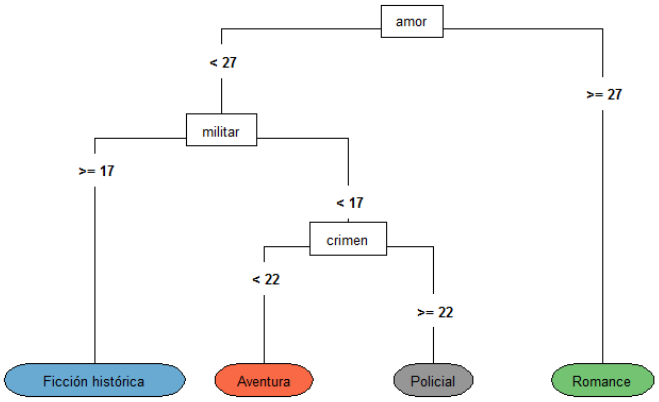
Los métodos de aprendizaje supervisado tienen como objetivo ajustar un modelo que permita predecir futuros valores desconocidos de la variable respuesta, en función de valores conocidos de las variables explicativas. En este trabajo, se decidió contrastar el desempeño de dos técnicas de aprendizaje supervisado: árboles de decisión y bosques aleatorios.

Los modelos basados en árboles son una clase de algoritmos no paramétricos que funcionan particionando el espacio de las variables predictoras en subregiones no superpuestas, agrupando individuos con valores similares de la variable respuesta según una serie de reglas de partición. Para hacer una predicción para una observación determinada, generalmente se usa la media o moda de la variable respuesta calculada entre las observaciones de entrenamiento ubicadas en la misma región. El conjunto de reglas puede resumirse gráficamente en forma de árbol, razón por la cual estos métodos son conocidos como árboles de decisión.

La Figura 1 muestra un ejemplo de un árbol de decisión, aplicado para identificar a qué género pertenece un determinado documento considerando el número de ocurrencias de los términos “amor”, “militar” y “crimen” dentro del mismo. Las regiones en las que se divide el espacio de predictoras se denominan nodos terminales o, por analogía con un árbol real, hojas. En el ejemplo, éstas corresponden a los distintos géneros literarios y se las representa con óvalos de

colores. Los puntos en el árbol en donde las variables se bifurcan se conocen como nodos internos y en el gráfico están representados como rectángulos blancos. Cuanto más arriba aparezca el nodo, mayor es la importancia de la variable para la clasificación. Finalmente, los segmentos que conectan los nodos entre ellos son las ramas del árbol.

Figura 1. Esquema de un árbol de decisión




Existen múltiples algoritmos de árbol de clasificación, que utilizan diversos criterios para realizar las particiones. Para el presente trabajo, se recurrió a uno de los algoritmos más utilizados en la actualidad: CART (del inglés *Classification and Regression Trees*, árboles de decisión y regresión). Este algoritmo intenta minimizar una medida conocida como índice de Gini, calculada de la siguiente manera:

$$G = \sum_{k}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

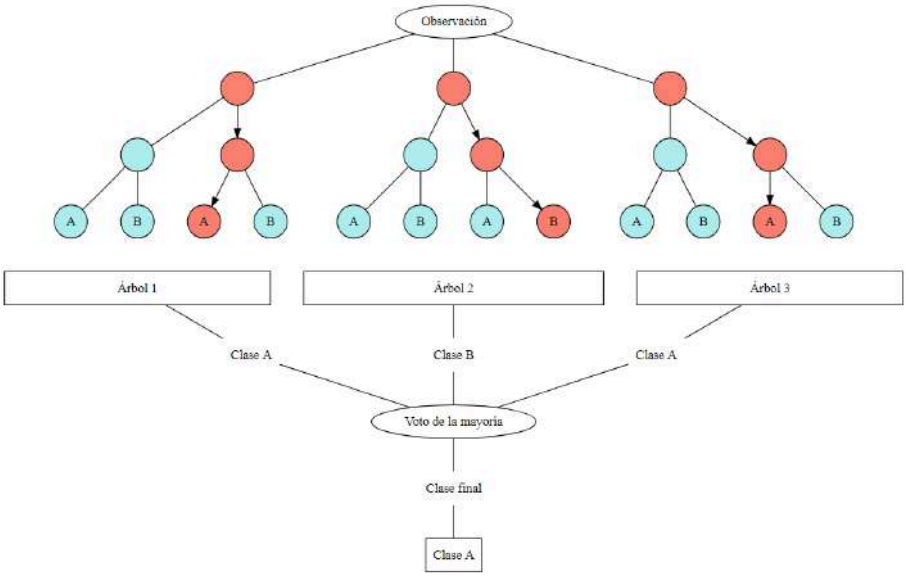
donde $\hat{p}_{mk}$ es la proporción de observaciones del conjunto de entrenamiento en la *m*-ésima región que pertenecen a la *k*-ésima clase. El índice de Gini toma entonces menores valores cuando todas las $\hat{p}_{mk}$ tengan valores cercanos a 0 o a 1, por lo que se la considera una medida de la "pureza" del nodo.

Los árboles de decisión tienen numerosas ventajas, incluyendo su flexibilidad, rapidez, bajo costo computacional y facilidad para graficar e interpretar. No obstante, suelen presentar una precisión predictiva muy baja en comparación a otras técnicas de aprendizaje supervisado. Además, los árboles de decisión suelen ser inestables, es decir que introducir pequeñas variaciones en el conjunto de observaciones puede

resultar en criterios de partición muy diferentes, generando árboles muy variables entre sí y haciendo las interpretaciones poco confiables. Para solucionar estos inconvenientes, surgió la técnica de bosques aleatorios (en inglés, *random forests*). Este método intenta obtener una mayor precisión en la clasificación al combinar los resultados de múltiples árboles de decisión.

El algoritmo comienza tomando una muestra simple al azar con reemplazo del conjunto de observaciones original y ajustando un árbol de decisión con éstas. A diferencia de lo que sucedía antes, no se consideran todas las variables para la construcción del árbol sino a un subconjunto elegido al azar en cada paso. Seguidamente, se toman nuevas muestras y se construyen nuevos árboles independientes con cada una de ellas. Esto quiere decir que cada árbol es ajustado con un conjunto distinto de observaciones y de variables, lo cual añade aleatoriedad al modelo y mejora su precisión respecto a los algoritmos de árboles de decisión. Una vez construido el bosque, se recurre al "voto de la mayoría" para clasificar las observaciones: para cada unidad, se registra a qué categoría fue asignada por cada árbol y se toma la más frecuentemente observada como clasificación final (Figura 2).

Figura 2. Esquema de un algoritmo de bosques aleatorios




Los bosques aleatorios suelen tener un mejor desempeño en tareas de clasificación que los árboles de decisión, siendo además más robustos que estos últimos respecto a variaciones en los datos. Sin embargo, su ajuste apropiado es más complejo y los tiempos de cómputo pueden llegar a ser elevados. Se pierde también

la posibilidad de graficar el algoritmo (pues ahora se cuenta con un bosque y no con un único árbol), lo cual hace a los resultados de esta técnica más difíciles de interpretar. Como solución a este último aspecto, usualmente se recurre al índice de Gini: para cada árbol del bosque, se calcula cuánto contribuye cada variable a disminuir el valor de *G*. Los valores calculados son sumados y promediados por el total de árboles, obteniéndose así una medida de la importancia de cada una de las variables en el modelo de clasificación. Estas medidas pueden luego ser graficadas, permitiendo obtener una interpretación más sencilla.

## RESULTADOS

Se hallaron nueve páginas *web* con libros sin derechos de autor. Al visitarlas, se logró adquirir 746 documentos, que luego fueron etiquetados. El conjunto de datos presentó siete géneros distintos, siendo ficción histórica el más frecuente (139 textos, 19% del corpus) y ciencia ficción el menos frecuente (72 textos o 9% del corpus). Ficción histórica fue también la categoría con textos más largos en promedio (con una media de 95.322 palabras por documento), mientras que terror fue la categoría con textos más cortos (25.510 palabras en promedio por documento).

La Tabla 2 presenta los resultados obtenidos con árboles de decisión y bosques aleatorios por escenario. Se calcularon la precisión global y el coeficiente Kappa como medidas del desempeño de los modelos.

Tabla 2. Precisión global y coeficiente Kappa observados según escenario y método

| Escenario | Árboles de decisión | | Bosques aleatorios | |
|---|---|---|---|---|
| | Precisión | Kappa | Precisión | Kappa |
| A | 43% | 0,33 | 64% | 0,57 |
| B | 41% | 0,30 | 51% | 0,41 |
| C | 30% | 0,16 | 59% | 0,51 |
| D | 37% | 0,24 | 50% | 0,39 |
| E | 36% | 0,23 | 64% | 0,58 |
| F | 35% | 0,21 | 48% | 0,38 |
| G | 43% | 0,33 | 52% | 0,42 |
| H | 37% | 0,24 | 47% | 0,36 |

Se observa que emplear bosques aleatorios superó, en todos los casos, el desempeño de los árboles de decisión individuales. La precisión promedio aumentó de 38% con árboles de decisión a 54% con bosques aleatorios.

Los mejores resultados se obtuvieron en general en los escenarios con unigramas, frecuencia de términos y sin filtro de nombres propios. Aun así, la mayor precisión se observó en el escenario E (unigramas, frecuencia de términos y con filtro). La Tabla 3 resume la precisión obtenida con bosques aleatorios según escenario.

La Figura 3 muestra la matriz de confusión para el escenario E, donde las filas corresponden a las clases observadas en el conjunto de prueba y las columnas a las clases predichas por el algoritmo. Los porcentajes en cada celda corresponden a las filas, y en particular los porcentajes de los elementos diagonales indican la precisión alcanzada por clase. Se aprecia que la precisión fue elevada para los géneros aventura, ficción histórica, policial y romance, superando el 68% en todos los casos. La categoría con la menor precisión fue ciencia ficción, dentro de la cual solo 6 documentos (26% del género) fueron correctamente clasificados. Las categorías más frecuentemente confundidas por el modelo fueron ciencia ficción con aventura (39% de los libros de ciencia ficción) y fantástico con romance (35% de los libros del género fantástico).

La Figura 4 muestra los diez términos más influyentes en el algoritmo. La longitud de las barras indica la reducción en el índice de Gini para el término respectivo.

Tabla 3. Precisión global según escenario usando bosques aleatorios

| | | Unigramas | Bigramas |
|---|---|---|---|
| Sin filtro | Frecuencia de términos | 64% | 51% |
| | TF-IDF | 59% | 50% |
| Con filtro | Frecuencia de términos | 64% | 48% |
| | TF-IDF | 52% | 47% |

Figura 3. Matriz de confusión correspondiente al escenario E

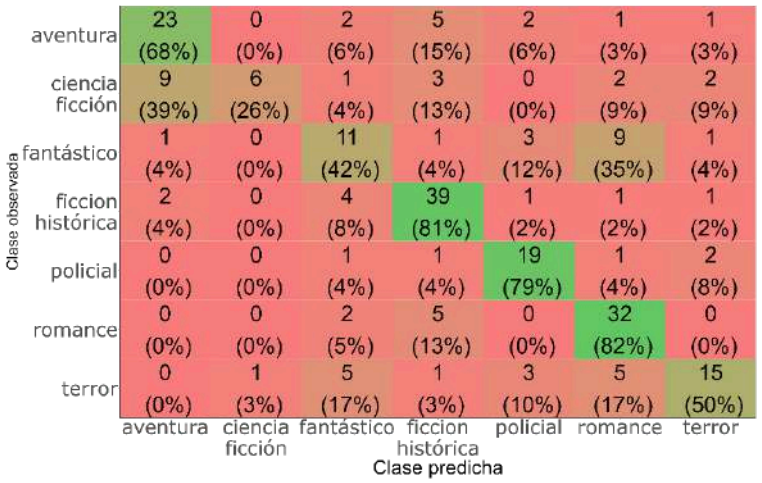
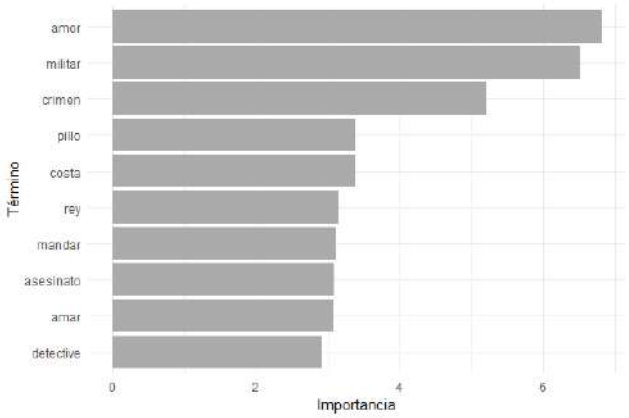



Figura 4. Términos más importantes para el modelo en el escenario E




## CONCLUSIONES

El objetivo de este trabajo fue plantear un modelo para clasificar automáticamente textos de ficción. Con ese fin, se construyó un corpus conformado por 746 documentos pertenecientes a 7 géneros distintos. Se plantearon ocho escenarios de análisis, al combinar distintas maneras de segmentar los textos (unigramas o bigramas), resumir su información (frecuencia de términos o TF-IDF) y filtrar términos potencialmente perjudiciales (eliminar o retener los nombres propios). Los mejores resultados se observaron en general para unigramas, frecuencia de términos y sin aplicar el filtro. En futuras investigaciones, se podrían construir nuevos escenarios utilizando el mismo corpus. Por ejemplo, se podrían combinar distintos

tipos de $n$-gramas o aplicar nuevas metodologías tales como análisis de sentimientos.

Los métodos que se contrastaron fueron árboles de decisión y bosques aleatorios, donde la segunda técnica presenta resultados superiores. En futuras investigaciones, se podrían aplicar nuevos algoritmos tales como *XGBoost*, el cual busca mejorar los resultados de bosques aleatorios al permitir que los nuevos árboles que se van construyendo "aprendan" de los anteriores, en lugar de ser independientes.

## BIBLIOGRAFÍA


Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Chen, T., y Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Feldman, R., y Sanger, J. (2013). *The text Mining Handbook: Advanced Approaches in analyzing unstructured data*. Cambridge University Press.

Hastie, T., Friedman, J., y Tisbshirani, R. (2017). *The elements of Statistical Learning: Data Mining, Inference, and prediction*. Springer.

Hvitfeldt, E., y Silge, J. (2022). *Supervised machine learning for text analysis in R*. CRC Press, Taylor & Francis Group.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R*. Springer.

Jena, M., y Dehuri, S. (2020). Decision tree for classification and regression: A state-of-the art review. *Informatica, 44*(4). https://doi.org/10.31449/inf.v44i4.3023

Kuhn, M., y Johnson, K. (2016). *Applied predictive modeling*. Springer.

Manning, C. D., Raghavan, P., y Schutze H. (2019) *Introduction to information retrieval*. Cambridge University Press.

R Core Team (2022). *R: A language and environment for statistical computing*. The Comprehensive R Archive Network. https://www.R-project.org/

Real Academia Española. (2014). Diccionario de la lengua española (23a ed.).

Silge, J., y Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly.

# ESTRATIFICACIÓN DE POBLACIONES PARA ENCUESTAS CON PROPÓSITOS MÚLTIPLES UTILIZANDO ALGORITMO GENÉTICO


## Lic. Guastella, Marina Guadalupe

**Director: Dr. Pagura, José Alberto**


La aplicación del muestreo estratificado requiere la elección de las variables de estratificación adecuadas, la decisión del número de estratos a utilizar, su conformación, y la definición del criterio de adjudicación (la forma en la que se reparte el tamaño total de la muestra en cada estrato). Desde el trabajo pionero de Neyman (1934) y hasta la actualidad se pueden encontrar diferentes aportes orientados a la obtención de una estratificación óptima en cuanto a lograr estimaciones más precisas. Barcaroli y Ballin (2014) presentaron una propuesta utilizando varias variables auxiliares categóricas y para muestras con las que se desean estimar varios parámetros poblacionales. Estos autores plantean conformar los estratos combinando las categorías de las variables de estratificación de modo que se logre un requisito de precisión al menor costo posible, para lo cual recurren al uso de Algoritmo Genético (AG).
En este trabajo se presenta un análisis de la propuesta basada en AG, se replica una de las aplicaciones realizada por los mencionados autores, pero comparando los resultados con los obtenidos por el método de Dalenius-Hodges, donde se verifican mejoras en los resultados. El estudio se completa con la aplicación del procedimiento dado por Barcaroli y Ballin al caso de una población con distribuciones asimétricas, comparando los resultados con procedimientos univariados específicos para esas situaciones.

## INTRODUCCIÓN

Entre los aspectos que deben tenerse en cuenta al momento de plantear un diseño muestral, es de gran importancia la estructura de la población en cuanto a la heterogeneidad que presentan las variables de interés. Conocer aproximadamente la variancia de estas variables aportará a la decisión del método de selección de la muestra a utilizar y al cálculo del tamaño de la misma para lograr una precisión aceptable de las estimaciones.

A modo de ejemplo, suele ser usual en estudios del ámbito económico o de la agricultura, encontrarse con poblaciones que presentan gran variabilidad y asimetría positiva de la distribución de la variable en estudio, es decir, la presencia de un gran número de observaciones con valores bajos de la variable en análisis y de algunas unidades con valores inusualmente grandes. Ante esta situación, utilizar el método de muestreo simple al azar conduciría a estimaciones poco precisas a no ser que el tamaño muestral sea exageradamente grande. Frente a poblaciones con alta variabilidad, el muestreo estratificado es una alternativa a considerar, que permite reducir la variabilidad del estimador mejorando así su precisión. Este método de selección de muestras consiste en dividir la población en grupos internamente homogéneos para luego extraer muestras aleatorias independientes de cada uno de ellos. El problema en la aplicación de esta metodología recae en:

- elegir adecuadamente la variable de estratificación,
- decidir el número de estratos a construir,
- determinar los límites de los estratos cuando la variable de estratificación es cuantitativa o los agrupamientos si la variable es categórica,
- decidir la forma de reparto de la muestra en los diferentes estratos.

A lo largo de los años se han publicado muchos estudios acerca del diseño óptimo del muestreo estratificado, cuyo fin es la obtención de estimadores más precisos. Una clasificación de estos diseños es la siguiente:

1) se establece la adjudicación o reparto de la muestra total en cada estrato, cuando se considera dada de antemano la estratificación -Neyman (1934)-,

2) se optimiza la estratificación o construcción de los estratos habiendo establecido una forma de reparto de la muestra - Dalenius y Hodges (1959), Lavallée e Hidiroglou (1988) y Rivest (2002)-,

3) se determina la estratificación y adjudicación de forma conjunta -Barcaroli y Ballin (2014)-.

En este trabajo se realizan breves cometarios de los aspectos fundamentales de tres de los métodos de estratificación tradicionales más divulgados, y se presentan en forma sintética, aspectos relevantes de la más reciente propuesta enunciada por Barcaroli y Ballin (2014).

Luego, se observa el comportamiento de los métodos mencionados en dos conjuntos de datos. El primero de ellos, conocido como Flores de Iris, el cual se eligió teniendo en cuenta que lo utilizan Barcaroli y Ballin (2014) como ejemplo de aplicación de su propuesta. El segundo, al que se hará referencia como “Edificios”, constituye una modificación de un trabajo anterior, realizado con la finalidad de mostrar el desempeño de los otros métodos de estratificación que aquí se mencionan (Borri, 2008). Como parte final, se realizan las comparaciones necesarias para derivar un conjunto de conclusiones.

## METODOLOGÍA

El método más divulgado para la construcción óptima de estratos es el que desarrollaron Dalenius y Hodges (1959). Este procedimiento se basa en una propuesta de Dalenius (1950) en la que se divide la población en L grupos o estratos definidos a partir de la variable en estudio "$Y$" de tipo cuantitativa encontrando límites $Y_{min} < Y_1 < Y_2 < \ldots < Y_{L-1} < Y_{max}$ de modo que minimizan la variancia del estimador empleando la adjudicación de Neyman. En dicho trabajo se llega a una solución matemática difícil de implementar, siendo el aporte de Dalenius y Hodge la deducción de una regla práctica aproximada conocida como “acumulada de la raíz cuadrada de la frecuencia”. Las fórmulas para su aplicación pueden encontrarse en la mayoría de los libros clásicos de muestreo. Cabe destacar que el uso de la variable en estudio como variable de estratificación es imposible en la práctica; para hacer posible la aplicación de este método a la práctica, se emplea una variable auxiliar altamente correlacionada con la variable en estudio obteniendo un resultado casi-óptimo.

Un método específico univariado para estratificar poblaciones asimétricas es el proporcionado por Lavallée e Hidiroglou (1988). La estratificación se lleva a cabo definiendo un estrato de inclusión forzosa que contiene las mayores unidades para

luego seleccionar una muestra aleatoria estratificada del resto de la población. Los límites de los estratos serán aquellos que conduzcan a minimizar el tamaño de la muestra para obtener un determinado coeficiente de variación del estimador, empleando adjudicación de potencia y se obtienen recurriendo a un proceso iterativo no siempre convergente. Si bien el desarrollo matemático del método se realiza considerando la variable en estudio como variable de estratificación, en la práctica, los límites de los estratos se obtienen en base al uso de una variable auxiliar que se encuentra correlacionada con la variable en estudio.

Rivest (2002) presenta una generalización del método de Lavallée-Hidiroglou considerando que la estratificación se realiza a partir de una variable auxiliar $X$ y asumiendo algún modelo de relación entre la variable de interés y la variable de estratificación. Propone un método iterativo que permite encontrar los límites de los estratos que minimizan el tamaño de la muestra para una precisión dada utilizando dos formas de adjudicar la muestra: adjudicación de potencia y adjudicación de Neyman.

### Estratificación multivariada para encuestas con propósitos múltiples (Barcaroli y Ballin, 2014)

En la práctica habitual se realizan estudios por muestreo donde se tiene más de una variable de interés -encuestas de propósitos múltiples- y donde se cuenta con múltiples variables auxiliares. Por este motivo, resulta relevante el aporte de Barcaroli y Ballin (2014) ya que provee un enfoque que es multivariado respecto a las variables auxiliares y respecto a las variables en estudio u objetivo.

Estos autores presentan un método para hallar la estratificación óptima a partir de variables auxiliares categóricas, con un criterio de optimalidad que consiste en minimizar una función de costo de la muestra para satisfacer una determinada precisión de cada uno de los estimadores calculados a partir de las variables objetivo. Las variables auxiliares que sean continuas podrán transformarse en categóricas según algún método apropiado, el cual dependerá del problema y procurando no producir una pérdida importante de información.

Para obtener la mejor agrupación de las unidades en estratos, se debería explorar la totalidad de las combinaciones de todas las categorías de las variables auxiliares disponibles, es decir el universo de las posibles estratificaciones que se podrían construir, tarea que requiere una excesiva cantidad de recursos computacionales. El

empleo de la metodología conocida como “Algoritmo Genético”, hace posible el tratamiento del problema ya que permite reducir el número de estratificaciones a examinar hasta encontrar la óptima.

## Formalización del problema de optimización

Se dispone del marco muestral de una población de N unidades, donde para cada una de ellas se encuentran registrados los valores de *M* variables auxiliares $X_m (m = 1, \ldots, M)$.

Cada variable auxiliar $X_m$ lleva asociada $k_m$ categorías, las que pueden identificarse con números enteros consecutivos $x_1, \ldots, x_{k_m}$.

El máximo número de estratos será $K = \prod_{m=1}^{M} k_m - I^*$, donde $I^*$ es el número de combinaciones de valores que no tienen unidades asociadas a ellos.

Se denomina ***estrato atómico*** $l_k (k = 1, \ldots, K)$, a cada uno de los estratos de la estratificación más detallada. Cada estrato atómico se caracteriza por una única combinación de categorías de las *M* variables auxiliares.

Si se considera el conjunto de estratos atómicos $L^* = \{l_1, \ldots, l_K\}$, se puede definir al conjunto de todas las *B* particiones posibles del conjunto $L^*$, $\{P_1, \ldots, P_B\}$, denominado ***espacio de estratificaciones***.

Dada una partición $P_i$ de $L^*$ caracterizada por *L* estratos, sean $N_h$ y $S^2_{h,g}$, $h = 1, \ldots, L$, $g = 1, \ldots, G$ el número de unidades y la variancia en el estrato *h* para cada una de las *G* distintas variables objetivos $Y_1, Y_2, \ldots, Y_G$, respectivamente.

Tomando una muestra aleatoria simple de tamaño $n_h$ de cada estrato y utilizando el estimador de Horvitz-Thompson del total, se tendrá como variancia del estimador correspondiente a la $g$-ésima variable objetivo:

$$Var\left(\hat{Y}_g\right) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S^2_{h,g}}{n_h} \quad g = 1, 2, \ldots, G.$$

Sea la siguiente función de costo: $C(n_1, \ldots, n_H) = C_0 + \sum_{h=1}^{H} C_h n_h$, donde $C_0$ indica el costo fijo (no depende del tamaño de la muestra) y $C_h$ representa el costo de obtener los valores de las G variables en una unidad del estrato *h*. El problema a resolver será el de encontrar aquella estratificación que permita obtener estimadores con una precisión establecida, al mínimo costo. La precisión se establece en términos de la variancia deseada y se asume diferente para cada uno de los G estimadores que se desean obtener.

Teniendo en cuenta las condiciones antes expresadas, los valores de $n_h$ se calculan por medio del algoritmo de adjudicación óptima multivariada propuesto por Bethel (1989), hallando los valores que minimizan la función C asumiendo que $Var(\hat{Y}_g) \leq V_g$, donde $V_g$ es la variancia deseada para cada estimador.

La mejor estratificación se puede lograr de la siguiente forma:

- Se genera la estratificación más detallada, es decir el conjunto $L^*$, de K estratos atómicos.
- Se enumeran todas las particiones $P_i$ de $L^*$.
- Para cada partición $P_i$ se determina la mejor adjudicación, lo que equivale a determinar el vector $(n_1, \ldots, n_L)$ por medio del algoritmo de Bethel y calcular el valor de la función de costo para dicha adjudicación $C_i(n_1, \ldots, n_L)$.
- Se selecciona la partición $P_i$ que logre minimizar $C_i(n_1, \ldots, n_L)$.

Es evidente que este procedimiento sólo puede aplicarse cuando *K* es pequeño.

Por medio de la utilización del algoritmo genético es factible identificar una solución óptima sin estudiar todas las estratificaciones posibles.

## Aplicación de un Algoritmo Genético para la estratificación óptima

Un algoritmo genético es una técnica de búsqueda computacional utilizada para hallar de forma aproximada o exacta soluciones óptimas. Se implementa mediante un proceso iterativo en el que la solución inicial es un conjunto de individuos, cada uno de ellos caracterizado por un conjunto de variables que se dispondrán en un vector que se denomina genoma. Ese conjunto de individuos evoluciona en cada

iteración, como resultado de operaciones de selección, mutación y cruza. En cada nueva iteración se obtiene un nuevo grupo de individuos que se denomina *generación*. En cada generación se evalúa la aptitud de cada individuo, es decir una característica del mismo definida de acuerdo al problema específico que se aborda, algunos de estos individuos son elegidos aleatoriamente y modificados: re-combinados y a veces mutados al azar, para conformar una nueva generación. Dado que los individuos que presenten una mejor aptitud tienen mayor probabilidad de ser seleccionados para integrar la próxima generación, el algoritmo genético incrementa el nivel de aptitud promedio a medida que va evolucionando.

Para implementar un algoritmo genético deben definirse algunos parámetros como la *velocidad de mutación* que expresa la cantidad de elementos del genoma que pueden mutar para cada individuo en el momento de generar individuos para la próxima generación. Una alta velocidad de mutación contribuye a que el algoritmo genético evite un óptimo local, teniendo como consecuencia una convergencia más lenta.

Por lo general, el algoritmo finaliza cuando se alcanza un número máximo de iteraciones que se establece de antemano, o bien cuando se van obteniendo nuevas generaciones sin lograr una mejora importante en la aptitud media.

La aplicación de este método para la obtención de una estratificación óptima se lleva a cabo de la siguiente forma:

- Una determinada estratificación se considera como un individuo, el cual se expresa en un vector de dimensión K, donde $v_i$ es un valor entero $\left(1 < v_i \leq U\right),\ con\ U \leq K.$
- Una estratificación $P(v)$ puede ser identificada por el vector $v = \left[v_1,\ ...,\ v_K\right]$ donde cada posible valor de i corresponde a un estrato atómico y $v_i$ tiene un valor entero en el intervalo [1,$U$].
- La función de aptitud de una determinada $P(v)$ es el valor de la función de costo $C\left(n_1,\ ...,\ n_L\right) = \sum_{h=1}^{L} n_h$, y los $n_1,\ ...,\ n_L$ se calculan aplicando el algoritmo de Bethel para la estratificación multivariada, bajo la condición de una precisión dada para el conjunto de estimadores a obtener.

Los pasos a ejecutar en su implementación son:

***Paso 0**: Creación de la generación inicial de individuos*

Basándose en un parámetro establecido *p,* tamaño de la generación, esa cantidad de individuos distintos son generados. Esto significa que, para cada individuo se generan aleatoriamente *K* enteros a través de una distribución uniforme en el intervalo $[1, U]$**.**

***Paso 1**: Evaluación de la aptitud de cada individuo*

Se calcula el costo total requerido para satisfacer la precisión deseada en los *G* distintos $\hat{Y}_g$, aplicando el algoritmo de Bethel que necesita conocer para cada estrato medias y desvío estándar de la variable objetivo y número de unidades en cada estrato.

***Paso 2**: Creación de una nueva generación*

La próxima generación se compone de un número de individuos de la generación anterior, aquellos con mejor aptitud (parámetro de elitismo), más un número de nuevos individuos obtenidos al seleccionar, cruzar y mutar individuos de la presente generación. La presencia de este segundo grupo ayuda a mantener una diversidad en la generación que sea suficiente para prevenir una convergencia prematura en soluciones alejadas del óptimo.

***Paso 3**: Iteración y punto de corte*

Se repite el procedimiento hasta cumplir el número máximo de iteraciones o bien hasta que no se obtengan mejoras.

## RESULTADOS

El interés principal en el presente trabajo fue el estudio de la nueva propuesta dada por Barcaroli y Ballin (2014), en adelante B-B, y la comparación con los resultados que se obtendrían utilizando técnicas para estratificación univariada. Con esa finalidad se eligieron dos poblaciones con características diferentes:

- El conjunto de datos popularmente conocido como “Flores de Iris” y que es utilizado por Barcarolli y Ballin en su artículo para ilustrar la aplicación de su propuesta multivariada pero en la que no se presentan comparaciones con métodos tradicionales de estratificación.
- La población utilizada por Borri (2008), que corresponde a un listado de obras en construcción en la ciudad de Rosario, con el agregado de una variable creada artificialmente. Esta población presenta la particularidad de que la

variable en estudio y las variables auxiliares presentan un comportamiento asimétrico. Fue elegida para la comparación con procedimientos univariados específicos para esta clase de poblaciones.

En ambos casos se construyen estratos por los métodos B-B y la conocida como regla de la acumulada de la raíz cuadrada de Dalenius-Hodges (1959). En la segunda población se agregan resultados obtenidos mediante los procedimientos de Lavallée-Hidiroglou (1988) y de Rivest (2002) específicos para poblaciones asimétricas. Las aplicaciones se realizaron utilizando los paquetes *stratification* y *SamplingStrata* del software estadístico R. Las comparaciones entre los procedimientos se hacen cotejando los coeficientes de variación poblacionales de las estimaciones. El requisito de precisión se estableció en términos del coeficiente de variación deseado y fue del 5%. Se agrega que, en el caso de Flores de Iris se comparan también los tamaños de muestra que deben utilizarse para obtener la precisión relativa deseada que se propone en la estratificación por B-B.

### Estratificación de la población "Flores de Iris"

Este conjunto de datos se encuentra en el software R y cuenta con 150 observaciones de ejemplares de tres especies de flores de iris en idénticas proporciones: setosa, virginica y versicolor. Las variables en estudio son "Largo del Pétalo" y "Ancho del Pétalo" las que se indicarán con $Y_1$ e $Y_2$ respectivamente y los valores poblacionales de interés son sus medias. Las variables auxiliares que se emplean para construir los estratos son "Largo del Sépalo" y "Especie", $X_1$ y $X_2$ en ese orden. La aplicación del algoritmo genético llevó a decidir la construcción de 4 estratos y un tamaño de muestra de 10. Los mismos números de estratos y tamaño de muestra se utilizaron con la regla de Dalenius y Hodges (D-H). La Tabla 1 contiene los coeficientes de variación $CV_1$ y $CV_2$ de los estimadores aplicando los métodos indicados en la tabla.

Tabla 1. Coeficientes de variación del estimador de las medias de Largo del Pétalo y Ancho del Pétalo con n=10 según método de estratificación y para muestreo aleatorio simple (MAS)

| Método de Estratificación | Largo del pétalo $CV_1$ | Ancho del pétalo $CV_2$ |
|---|---|---|
| **B-B** | 3,11% | 5,57% |
| **D-H** | 7,98% | 11,90% |
| **MAS** | 14,85% | 20,10% |

Tal como puede apreciarse en la Tabla 1, de las metodologías evaluadas el único método de estratificación en el que el nivel de precisión deseado en promedio para $Y_1$ es satisfecho, y para el caso de $Y_2$ es muy próximo al objetivo, es cuando se estratifica utilizando el método de B-B. De acuerdo a lo esperado, el muestreo aleatorio simple es mucho menos preciso para el mismo tamaño de muestra.

La Tabla 2 contiene los tamaños de muestra necesarios para lograr un coeficiente de variación de 5% para los estimadores de acuerdo al método de estratificación empleado o si se utiliza muestreo aleatorio simple. Puede observarse que el tamaño muestral al aplicar el diseño de estratificación para estimar el promedio poblacional de ambas variables de interés, es considerablemente menor cuando se emplea la técnica de B-B. Además, como se podía esperar, el muestreo aleatorio simple requiere un tamaño de muestra ampliamente mayor para la misma precisión.

Tabla 2. Tamaños de muestra necesarios para estimar las medias de Largo del Pétalo y Ancho del Pétalo con un Coeficiente de Variación igual a 5%

| Método de Estratificación | Largo del pétalo $n_1$ | Ancho del pétalo $n_2$ |
|---|---|---|
| **B-B** | 10 | 10 |
| **D-H** | 23 | 55 |
| **MAS** | 77 | 128 |

## Construcción de estratos en la población "Edificios"

El conjunto de datos Edificios contiene 627 observaciones y cinco mediciones para cada una de ellas: monto de aportes profesionales expresados en pesos $(Y_1)$, superficie a construir expresada en metros cuadrados $(X_1)$, número de pisos del

edificio $(X_2)$ , distrito en donde se realiza la obra $(X_3)$, valor de venta expresado en millones de pesos $(Y_2)$.

Las variables en estudio son “Monto de Aportes” y “Valor de Venta” y los valores poblacionales de interés son sus medias. Las variables auxiliares a utilizar en la estratificación son “Superficie a construir”, “Número de pisos” y “Distrito”. El método de estratificación utilizando algoritmo genético conduce a la construcción de 6 estratos y un tamaño de muestra de 92. Para las comparaciones, se construyeron estratos por los métodos de referencia y se mantuvo el tamaño de muestra y el mismo número de estratos.

Los resultados encontrados y presentados en la Tabla 3, muestran la superioridad del muestreo estratificado sobre el muestreo aleatorio simple. En cuanto a las comparaciones entre las diferentes estratificaciones, tanto los métodos específicos para poblaciones asimétricas como la propuesta de D-H aportaron mejores resultados que el método de B-B.

Tabla 3. Coeficientes de variación del estimador de las medias de Monto de aportes y Valor de venta que se obtienen con n=92 según métodos de estratificación y muestreo aleatorio simple

| Método de Estratificación | Monto de aportes $CV_1$ | Valor de venta $CV_2$ |
|---|---|---|
| **B-B** | 4,57% | 4,96% |
| **D-H** | 3,69% | 3,40% |
| **L-H** | 4,09% | 3,51% |
| **RIVEST** | 3,48% | 3,28% |
| **MAS** | 18,34% | 23,81% |

## COMENTARIOS FINALES

El interés principal del presente trabajo ha sido el estudio de la metodología propuesta por Barcaroli y Ballin (2014) y su comparación con métodos tradicionales. El método se aplicó en dos poblaciones que constituyen escenarios diferentes: Flores de Iris, utilizado por los autores como ejemplo en su presentación y Edificios, conjunto de datos en el que las variables en estudio tienen distribuciones asimétricas.

En la población de Flores de Iris, se obtuvieron resultados más precisos con el método de B-B. Esto puede deberse a que este método utiliza la información de la variable de estratificación categórica, la que tiene una gran influencia en la construcción de estratos.

La aplicación a la población de Edificios mostró que los métodos específicos para poblaciones asimétricas tienen mejor desempeño. La utilización de un estrato de inclusión forzosa, el de las unidades más grandes, proporciona una reducción importante en la variancia del estimador. También se observa una buena performance de D-H que podría explicarse por la alta correlación entre la variable continua de estratificación y ambas variables de interés, información que sufre pérdida al categorizar la variable para aplicar algoritmo genético. Debe tenerse en cuenta que el hecho de convertir en categórica una variable continua presupone una pérdida de información, por lo que debe ponerse especial cuidado en la definición de sus categorías de modo que se conserve lo más relevante de su comportamiento.

Cabe mencionar que las comparaciones realizadas en este trabajo, univariadas en su mayoría, deberán complementarse con contrastaciones de resultados de otras propuestas de estratificación multivariada, logrando así cotejar métodos en condiciones parecidas.

El método de B-B presenta un importante avance en el tratamiento del problema de la construcción de estratos. Por ese motivo se considera que merecen estudiarse otras alternativas que permitan mejorar su comportamiento de forma más general. Entre ellas, poner especial atención en la población que se estudia y en el procedimiento para construir las categorías de las variables auxiliares continuas. Por ejemplo, en el caso de contar con una población asimétrica, se podría considerar que una manera de mejorar los resultados sería construir un estrato de inclusión forzosa para determinar dichas categorías.

Por último, más detalles de la forma en la que fue implementado el método basado en algoritmo genético pueden encontrarse en la tesina de referencia.

## BIBLIOGRAFÍA


BAILLARGEON, S. & RIVEST, L. (2017). *"Univariate Stratification of Survey Populations"*, R Package "stratification", Version 2.2-6.

BARCAROLI, G. & BALLIN, M. (2014). *"Joint determination of optimal stratification and simple allocation using genetic algorithm"*, Survey Methodology, Vol. 39 No.2, pp. 369-393.

BARCAROLI, G., BALLIN, M., ONDENDAAL, H., PAGLIUCA, D., WILLIGHAGEN, E. & ZARDETTO, D. (2020). *"Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys"*, R Package "SamplingStrata", Version 1.5-1.

BETHEL, J.(1989). *"Sample allocation in multivariate surveys"*, Survey Methodology, Vol. 15 No.1, pp. 47-57, Statistics Canada.

BORRI, I. (2008). *"Estratificación en poblaciones asimétricas. Aplicación a la estimación del monto total de aportes de una asociación profesional.".* Tesina de Licenciatura en Estadística, Facultad de Ciencias Económicas y Estadística, UNR.

COCHRAN, W. (1977). *"Técnicas de Muestreo"*, 2º edición en español, Compañía Editorial Continental, México.

DALENIUS, T. (1950). *"The problem of optimum stratification"*, Skandinavisk Aktuarietidskrift, Vol. 33, pp. 203-2013.

DALENIUS, T. & HODGES, J. (1959). *"Minimum Variance Stratification",* Journal of the American Statistical Association, Vol. 54, pp. 88-101.

HIDIROGLOU, M. A. (1986). *"The construction of a Self-Representing Stratum of Large Units in Survey Design"*, The American Statistician, Vol 40, Nº 1.

LAVALLÉE, P & HIDIROGLOU, M. A. (1988). *"On the stratification of Skewed Population"*, Survey Methodology, Vol 14, Nº1, pp.33-43.

NEYMAN, J. (1934). *"On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection"*, Jornal of the Royal Statistics Society, 97, pp. 558-606.

RIVEST, L. (2002). *"A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys"*, Survey Methodology, Vol. 28 No.2, pp. 191-198.

SETHI, V. (1963). *"A note on optimum stratification of populations for estimating the population means"*, Australian Journal of Statistics, Vol. 5, pp. 20-33.

# APLICACIÓN DE MODELOS DE CREDIT SCORING PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE DEFAULT UTILIZANDO INFORMACIÓN DE LA CENTRAL DE DEUDORES DEL BCRA


## Lic. Isaguirre, María Belén

**Responsable de la Facultad de Ciencias Económicas y Estadística: Lic. Wibly, Adrián**

**Responsable de la entidad: Mg. Kovalevski, Leandro**



Para las empresas que ofrecen préstamos, resulta crucial mitigar las pérdidas asociadas a posibles incumplimientos de pago por parte de los deudores. En muchos casos, dicha mitigación se logra a través de evaluaciones basadas en el historial del solicitante, que a menudo conlleva costos elevados. A partir de las evaluaciones se determina si se aprueba o se rechaza la solicitud de préstamo.
En este trabajo se propone obtener una predicción de la probabilidad de default, es decir de un incumplimiento en el pago, minimizando los costos asociados y evaluando si la inclusión de variables relacionadas con el tipo de entidad (financiera, no financiera) mejora el rendimiento del modelo. Con este objetivo, se ajustaron múltiples modelos de regresión logística utilizando exclusivamente datos del Banco Central de la República Argentina (BCRA).
En cuanto a los resultados, se ha desarrollado un modelo que incorpora seis variables relacionadas con el historial crediticio del solicitante y el tipo de entidad, así como una aproximación de su edad. Este modelo proporciona una predicción de la probabilidad de default, la cual puede ser utilizada para generar un puntaje crediticio con el cual evaluar al solicitante.


## INTRODUCCIÓN

Este informe es el resultado de una Práctica Profesional realizada en la empresa San Cristóbal. Dentro de la compañía existe una unidad de negocio denominada Servicios Financieros, creada con el objetivo principal de proporcionar préstamos tanto a personas físicas como jurídicas. Dentro de Servicios Financieros, cada cliente que solicita un préstamo es evaluado en base a información actual y a información previa. Esta evaluación se lleva a cabo principalmente con el fin de evitar pérdidas por riesgo de crédito, es decir, por el incumplimiento del deudor de sus obligaciones.

La falta en el pago de un préstamo o cualquier otro tipo de obligación, en determinado horizonte temporal, se define formalmente como *default*. En este trabajo, se tomó un atraso mayor a 90 días, en un horizonte temporal de 12 meses, para señalar a un crédito como *default,* pero puede llegar a ser diferente en algunas jurisdicciones y dependiendo del tipo de crédito (Gutiérrez Girault *et al*., 2006).

El comportamiento de pago que posee cada deudor es registrado únicamente por la entidad que ha aprobado dicho préstamo y mantiene el seguimiento del mismo. Por esta razón, están obligadas a informar al BCRA, de manera mensual, la situación de morosidad en la que se encuentra cada uno de los clientes que posean al menos una deuda, junto a otros detalles referentes a información crediticia. A su vez, el BCRA ha clasificado a todas las entidades dentro de dos grandes grupos: entidades financieras y entidades no financieras.

Cada entidad, principalmente las financieras, informa los atrasos de todos los clientes que correspondan, categorizados en 6 posibles situaciones dependiendo de los días de atraso (siendo 1: Normal - Atraso de hasta 30 días, y 6: Irrecuperable).

Es importante destacar que, a raíz del aislamiento obligatorio y preventivo dictado por el gobierno de la República Argentina por causa de la pandemia en marzo de 2020, el BCRA tomó la decisión de realizar cambios en la medición de los atrasos de los clientes. En base a esta decisión, los días de atraso mencionados previamente se vieron afectados y, por ende, las diferentes categorías de situaciones también.

La información presentada por las entidades se encuentra disponible en la base de datos de la Central de Deudores del Sistema Financiero, dentro de la página de AFIP, la cual es actualizada periódicamente por el BCRA. Actualmente existen diferentes burós de crédito, como “Nosis” o “Veraz”, que utilizan dicha información de los individuos para obtener un puntaje crediticio, también denominado como *score*

crediticio. Este puntaje es un factor relevante en la decisión de aprobar o rechazar un préstamo, pero su uso conlleva un costo elevado. Actualmente San Cristóbal Servicios Financieros lo utiliza como herramienta clave en la evaluación de potenciales clientes, incurriendo en dicho costo.

## OBJETIVOS

Se ha planteado como objetivo obtener una predicción de la probabilidad de *default* de los deudores y evaluar a su vez si el desempeño del modelo mejora al incorporar alguna variable relacionada al tipo de entidad, utilizando únicamente la información disponible en el BCRA.

## MATERIALES

Los datos utilizados en este trabajo provienen de la base de la Central de Deudores del Sistema Financiero, que contiene los registros de todas las personas que tienen o han tenido al menos una deuda en el sistema financiero, agrupadas según el CUIT correspondiente a la persona y la entidad a la cual pertenece. Es decir que una persona tiene tantos registros como deudas en entidades diferentes. Específicamente, se encuentran aquellos casos cuyo importe de deuda sea de al menos $1000 (mil pesos) y en caso de tener más de una deuda en una misma entidad, las mismas se agrupan y se suman los montos (Banco Central de la República Argentina, 2021). Usualmente los préstamos otorgados por las entidades tienen carácter de pago mensual, a excepción de ciertos casos puntuales. Debido a esto, la base es actualizada mensualmente y en algunos casos hasta quincenalmente, para que se vean reflejados aquellos cambios que pudieron haber ocurrido en ese lapso de tiempo.

Cada registro se compone de 171 caracteres, y es posible estructurarlos definiendo cada una de las variables a través de un comunicado oficial del BCRA. En el mismo se indican los nombres de los campos, la longitud y el tipo de variable, y un comentario aclarando qué indica cada una. Una vez estructurada, la base de datos queda constituida por un total de 24 variables.

## MÉTODOS

### Regresión Logística

Se utilizaron modelos de regresión logística ya que la variable respuesta es de naturaleza binaria, indicando si la persona tuvo un atraso mayor a 90 días en al menos una deuda en el lapso de 12 meses, es decir, si tuvo *default*. Estos modelos son un caso puntual dentro de los Modelos Lineales Generalizados, en los cuales se utiliza el enlace *logit* para describir la relación entre la probabilidad de *default* y las variables explicativas.

### Random Forest

El bosque aleatorio (en inglés *Random Forest*) es una técnica de aprendizaje supervisado no paramétrico y, como su nombre indica, está formado por un gran número de árboles. Estos árboles son de decisión individual, con baja correlación entre ellos y funcionan como un conjunto. Dado un conjunto de variables **S** = $(X, Y)$, donde $X = (X_1, X_2, \ldots, X_p)$ es un vector de $p$ variables explicativas e $Y$ es la variable respuesta, el objetivo de los mismos es predecir los valores de $Y$ a través de los de $X$. Las variables, tanto respuesta como explicativas, pueden ser continuas o categóricas y la diferencia radica en el tipo de árbol de decisión utilizado. Cuando la variable respuesta es continua se utiliza un árbol de regresión y cuando es categórica se utiliza uno de clasificación.

Los árboles particionan recursivamente el espacio conformado por las $p$ variables, de manera tal que las regiones generadas contienen observaciones cuyos valores respecto a la variable respuesta $Y$ son lo más homogéneos posibles (Hastie *et al*., 2008). Para la construcción de los árboles se utiliza la técnica *CART*, de su nombre en inglés *Classification And Regression Trees*, la cual genera particiones homogéneas de forma automática. Para evaluar el nivel de homogeneidad en los grupos resultantes usa el índice Gini, el cual se calcula de la siguiente manera:

$$Gini = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

donde $\hat{p}_{mk}$ es la proporción de observaciones correspondientes a la clase $k$ en el nodo $m$. Cada árbol del bosque se entrena con diferentes muestras aleatorias con reposición obtenidas del conjunto de datos. Los bosques aleatorios, además de

obtener predicciones, permiten generar implícitamente medidas de importancia para las variables del modelo, las cuales pueden construirse utilizando el índice Gini. Con estas medidas lo que se busca es cuantificar el impacto que tiene cada variable predictora sobre la respuesta, es decir cuánto reducen el índice de impureza, y por ende determinar su relevancia o poder predictivo. Este procedimiento se ejecuta sobre cada uno de los árboles que conforman el bosque y posteriormente se promedian (o suman) todas las medidas de importancia que se generan para la variable $X_i$ con el fin de obtener una medida global del índice Gini (Liaw y Wiener, 2002). En este trabajo se consideraron los bosques aleatorios con la finalidad de obtener un análisis de importancia para seleccionar las variables más influyentes sobre la respuesta.

### Indicadores de desempeño

Para medir el rendimiento y eficiencia de los modelos es conveniente utilizar índices de desempeño o *KPIs* (por su nombre en inglés *Key Performance Indicators*). En este trabajo se consideraron los siguientes:

1. Raíz del error cuadrático medio: es una forma cuantificable de comparar los valores predichos de un modelo con los valores reales observados.
2. Falsos positivos y negativos: un falso positivo se refiere a un resultado en el que el modelo clasifica de manera incorrecta una observación como perteneciente a la clase positiva, cuando en realidad pertenece a la clase negativa. Un falso negativo se refiere a un resultado en el que el modelo clasifica de manera incorrecta una observación como perteneciente a la clase negativa, cuando en realidad pertenece a la clase positiva.
3. Sensibilidad y especificidad: la sensibilidad se refiere a la probabilidad de que el modelo prediga correctamente una clase positiva. La especificidad se refiere a la probabilidad de predecir correctamente una clase negativa.
4. Área bajo la curva ROC: el área bajo la curva ROC o AUC-ROC (de su nombre en inglés *Area Under the Receiver Operating Characteristic Curve*) representa la tasa de verdaderos positivos (sensibilidad) frente a la tasa de falsos negativos (1-especificidad) para distintos umbrales de clasificación (Kuhn y Johnson, 2018).

5. *Score F1*: es un indicador muy utilizado en modelos de clasificación binaria dado que combina medidas de precisión y sensibilidad. La precisión es una medida de la proporción de predicciones positivas verdaderas entre todas las predicciones positivas y la sensibilidad fue definida anteriormente.

## RESULTADOS

Con la finalidad de obtener una muestra y seleccionar variables relevantes para el trabajo, se definió un mes determinado como período de referencia. Teniendo en cuenta el impacto que tuvo el aislamiento preventivo dictado por el Gobierno Nacional a raíz de la pandemia a partir de marzo de 2020 y, siendo que se tenía información disponible desde junio de 2018, se definió el mes de junio de 2019. Del total de CUITs en dicho período se consideraron únicamente las personas físicas que, en ese momento, no contaban con deudas en situación 3 o mayor, es decir deudas con atrasos mayores a 90 días, y de éstas se obtuvo una muestra aleatoria de 160000 CUITs únicos. Una vez obtenida la muestra se realizó una selección de variables de las 24 disponibles en la base de datos de la Central de Deudores del Sistema Financiero, descartando aquellas irrelevantes para el trabajo.

De los CUITs seleccionados, se realizó un seguimiento de los 6 períodos previos, de los cuales se obtuvo la información de todas las variables seleccionadas previamente. Además, se realizó un seguimiento de los 12 períodos posteriores al período de referencia, de los cuales se registró la situación crediticia.

En base a la información recabada los seis meses previos y del período de referencia, se construyeron un total de 30 variables explicativas nuevas. Por otro lado, en base a la información de los 12 períodos posteriores se evaluó si la persona tuvo un atraso mayor a 90 días en al menos una deuda en dicho lapso de tiempo y se construyó la variable respuesta de tipo binaria indicando si tuvo *default* o no. Cabe destacar que de los 160000 CUITs hubo 951 (0.59%) en los que no fue posible evaluar la variable respuesta y no se consideraron en el análisis, quedando un total de 159049 CUITs.

Considerando que uno de los objetivos del trabajo radicaba en evaluar si el desempeño del modelo mejoraba al incorporar alguna variable relacionada al tipo de entidad, se clasificó a cada una de las entidades tomando dos criterios posibles y luego se construyeron diferentes predictores teniendo en cuenta dicha clasificación. Como primer criterio se utilizó una clasificación publicada por el BCRA a través de

un boletín oficial, basada en un indicador que toma en cuenta el promedio de los activos de cada entidad financiera en relación al total de activos en el sistema financiero. Como segundo criterio, se solicitó al Jefe de Riesgos de San Cristóbal Servicios Financieros su opinión profesional al respecto. Finalmente, se identificó a cada código de entidad de dos maneras:

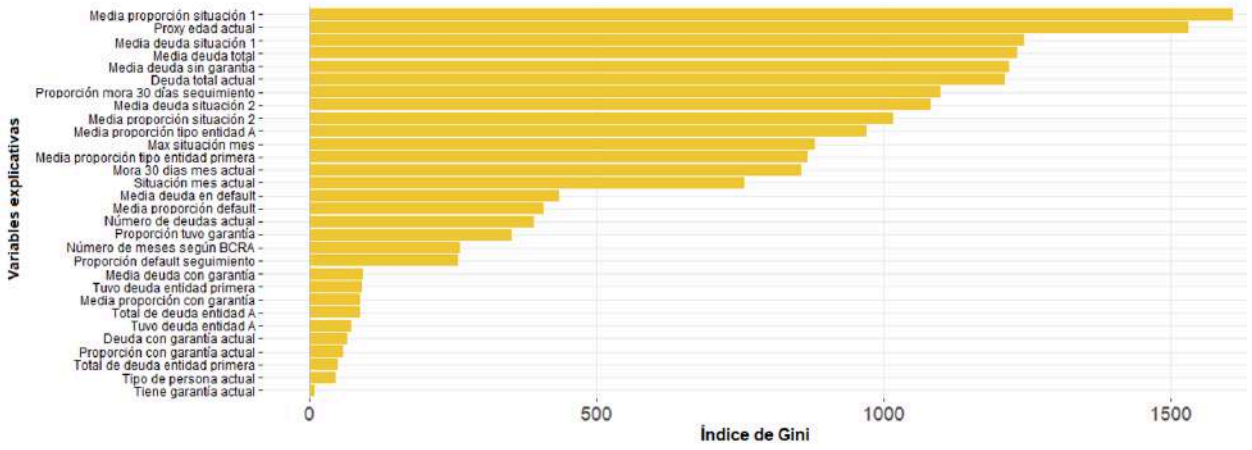
- La entidad corresponde, o no, al Grupo A según la clasificación del BCRA.
- La entidad corresponde, o no, a una de primera línea según la clasificación del negocio.

A modo de ejemplo, se definen dos de los predictores construidos a partir de la clasificación:

- “Tuvo deuda entidad A”: tuvo al menos una deuda en una entidad del tipo A en los 6 períodos de seguimiento previo.
- “Tuvo deuda entidad primera”: tuvo al menos una deuda en una entidad del tipo primera línea en los 6 períodos de seguimiento previo.

Inicialmente, se realizó un análisis descriptivo exploratorio marginal para evaluar la relación entre cada variable y el porcentaje de *default* observado. A partir de los resultados obtenidos, se identificaron las variables potencialmente influyentes. Posteriormente, se llevó a cabo un análisis para determinar la importancia de las variables explicativas en la probabilidad de *default*. Utilizando técnicas de *Random Forest*, se calculó el índice de Gini, el cual refleja la relevancia de cada variable; cuanto mayor sea este índice, mayor será la importancia de la variable.

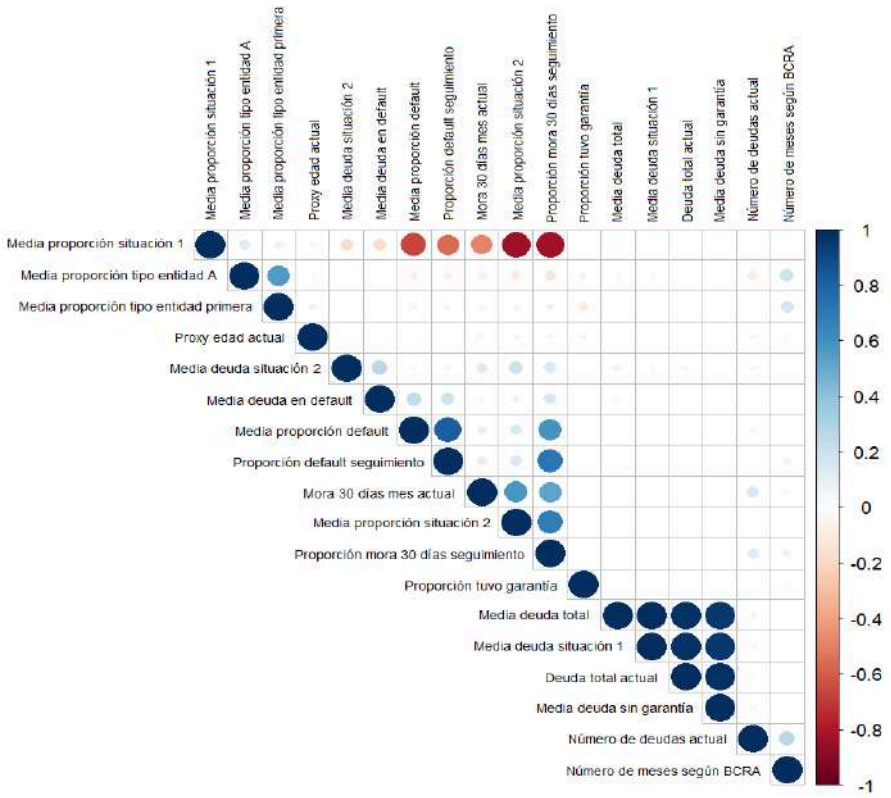
Gráfico 1. Valores del índice Gini para cada una de las variables explicativas




En el Gráfico 1 se muestran los valores del índice de $Gini$ correspondientes a cada variable explicativa. Se destaca una disminución significativa del índice en las

últimas diez variables en comparación con las demás, por lo que se tomó la decisión de descartarlas. Posteriormente, se llevó a cabo un análisis de correlación sobre 18 de las 20 variables restantes para eliminar aquellas con una alta correlación entre sí (dejando de lado dos variables ordinales).

Gráfico 2. Valores de correlación para cada combinación de variables explicativas




En el Gráfico 2 se muestra el nivel de correlación entre cada combinación de variables. Mientras más fuerte el color y más grande el círculo, mayor es el grado de correlación, tanto positivo como negativo.

Luego de los análisis realizados, se conservaron un total de 16 variables explicativas, dos de las cuales contenían información acerca del tipo de entidad.

Se realizó una partición de la muestra en tres subconjuntos de datos: entrenamiento, evaluación y validación (65%, 20% y 15% respectivamente). Utilizando el 65% y el 20% se construyeron y evaluaron distintos modelos de regresión logística con 14 de las 16 variables conservadas, sin considerar por el momento las dos variables relacionadas con el tipo de entidad, y el conjunto de validación fue utilizado para ratificar el punto de corte elegido para calcular los indicadores de desempeño. A

partir de las evaluaciones obtenidas, se seleccionaron los cinco predictores que mejor explicaban la probabilidad de *default*:

1. “Número de deudas actual”: número de deudas en el período de referencia.
2. “Situación mes actual”: situación en la que se encuentra el deudor en el período de referencia.
3. “*Proxy* edad actual”: aproximación de la edad del deudor en el período de referencia.
4. “Media proporción situación 1”: media de la proporción del total de deudas en situación 1 en los 6 meses de seguimiento previo.
5. “Proporción tuvo garantía”: proporción de meses en los que el deudor tuvo al menos una deuda con garantía.

A los cuales se sumaron las variables relacionadas con el tipo de entidad, es decir:

1. “Media de la proporción del tipo entidad A”: media de la proporción del total de las deudas en entidades del tipo A (según clasificación del BCRA) en los 6 períodos de seguimiento previo.
2. “Media de la proporción del tipo entidad primera”: media de la proporción del total de las deudas en entidades del tipo primera línea (según clasificación del negocio) en los 6 períodos de seguimiento previo.

En estas últimas dos se tomó la cantidad de deudas que la persona tenía en entidades de primera (o entidades A, dependiendo de la variable a calcular) en determinado período y se las dividió por el total de deudas en ese período, por ejemplo: 3 deudas en entidades del tipo primera de las 12 en total, en el período de mayo 2019. Este proceso se repitió en los 6 periodos correspondientes y se calculó el promedio.

Teniendo en cuenta los predictores seleccionados, se ajustaron tres nuevos modelos de regresión logística incorporando las relacionadas con el tipo de entidad una por vez.

- Modelo (1): “Número de deudas actual” + “Situación mes actual” + “*Proxy* edad actual” + “Media proporción situación 1” + “Proporción tuvo garantía”
- Modelo (2): “Número de deudas actual” + “Situación mes actual” + “*Proxy* edad actual” + “Media proporción situación 1” + “Proporción tuvo garantía” + “Media de la proporción del tipo entidad A”

- Modelo (3): “Número de deudas actual” + “Situación mes actual” + “*Proxy* edad actual” + “Media proporción situación 1” + “Proporción tuvo garantía” + “Media de la proporción del tipo entidad primera”

Luego se comparó el desempeño de cada uno, utilizando los indicadores correspondientes.

Tabla 1. Indicadores de desempeño de los diferentes modelos ajustados

| Modelo | n | RECM | AUC | FP (%) | FN (%) | S | E | F1 |
|---|---|---|---|---|---|---|---|---|
| (1) | 32446 | 0.264 | 0.781 | 3769 (11.6%) | 1370 (4.2%) | 55% | 87% | 0.393 |
| (2) | 32446 | 0.264 | 0.788 | 3976 (12.3%) | 1325 (4.1%) | 56% | 86% | 0.392 |
| (3) | 32446 | 0.264 | 0.787 | 3617 (11.1%) | 1368 (4.2%) | 55% | 88% | 0.400 |

Notas. n: Número de observaciones en la evaluación, RECM: Raíz del Error Cuadrático Medio, AUC: Área Bajo la Curva, FP: Falsos Positivos, FN: Falsos Negativos, S: Sensibilidad, E: Especificidad, F1: Score F1.

Al analizar los resultados arrojados en la Tabla 1 se pudo concluir que los tres modelos tuvieron un gran desempeño a la hora de predecir la probabilidad de *default* de un cliente. Incluso aquel que no consideraba las variables relacionadas con el tipo de entidad (1). Si bien esto fue un hallazgo interesante, era necesario elegir uno de los tres mediante la comparación de los diferentes indicadores.

Considerando la cantidad de falsos negativos, se eligió utilizar el modelo (2). Dado que, predecir que una persona no va a entrar en *default*, otorgarle el préstamo y que ocurra lo contrario implica una pérdida de dinero mayor.

## COMENTARIOS FINALES

En este trabajo se describió la tarea realizada en San Cristóbal como parte de una Práctica Profesional correspondiente a la carrera de Licenciatura en Estadística de la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario. Dentro de las tareas desarrolladas se incluyen el armado, procesamiento y análisis de grandes volúmenes de datos. En primer lugar se realizó la lectura y procesamiento de las bases, de las que se extrajo una muestra aleatoria de 160000 CUITs únicos. Luego, se descartaron ciertas variables del conjunto de datos original

ya que no se consideraron relevantes para el estudio. Sobre los CUITs incluidos en la muestra se recabó la información correspondiente a cada una de las variables en los seis períodos previos y en los 12 posteriores al de referencia. Una vez obtenida la información necesaria se construyeron 31 variables nuevas, 30 explicativas y una respuesta. Considerando que la cantidad de variables explicativas era muy elevada, se realizaron análisis descriptivos, se aplicaron técnicas de *Random Forest* y ajustes de modelos logísticos para lograr reducir la cantidad. Luego de haber seleccionado las variables se construyeron 3 modelos de regresión logística y se compararon los índices de desempeño. De los valores obtenidos se determinó que el modelo que tiene mejor desempeño a la hora de estimar la probabilidad de *default* es aquel que incluye las variables “Número de deudas actual”, “Situación mes actual”, “*Proxy* edad actual”, “Media proporción situación 1”, “Proporción tuvo garantía” y “Media de la proporción del tipo de entidad A”.

La importancia de poder clasificar a los clientes en base a su historial crediticio y utilizando diferentes técnicas estadísticas, fue creciendo a lo largo del tiempo. Sin dejar de lado el rol de los analistas de riesgo, cuyo juicio es clave para la toma de decisión final, el *Credit Scoring* es una herramienta que tiene un gran potencial y complementa al resto de información que se pueda tener del solicitante.

A través de los datos obtenidos se ha logrado construir un modelo para predecir la probabilidad de *default*, sin incurrir en costos monetarios adicionales, y que ha demostrado tener un gran desempeño. El análisis realizado y los resultados de este trabajo proporcionan una base de la cual se puede partir para continuar desarrollando y mejorando. Además de obtener una predicción sobre si el solicitante incumplirá o no, la probabilidad se puede incluir en la construcción de un puntaje, el cual podría utilizarse como medida interna complementaria en la toma de decisiones de la empresa en cuanto a la concesión de créditos y/o en relación a los mismos. Es por ello que se recomienda llevar a cabo una exploración y análisis más a fondo en estudios futuros.

## REFERENCIAS


Banco Central de la República Argentina (2012). Carta orgánica del B.C.R.A. ley N°24144. https://www.bcra.gob.ar/pdfs/bcra/cartaorganica2012.pdf.

Banco Central de la República Argentina (2018). Texto ordenado comunicación “a” 6439 sección 4. https://www.bcra.gob.ar/pdfs/texord/texord_viejos/v-ceninf_21-09-05.pdf.

Banco Central de la República Argentina (2020). Texto ordenado comunicación “a” 7169 sección 4. https://www.bcra.gob.ar/pdfs/comytexord/A7169.pdf.

Banco Central de la República Argentina (2021). Texto ordenado comunicación “a” 7257 sección 62. https://www.bcra.gob.ar/Pdfs/comytexord/A7257.pdf.

Gutiérrez Girault, M., Lippi, C., Canella, J., Díaz, M., Guión, A. y Nicolini, C. (2006). Sistemas de información para la administración del riesgo de crédito. Technical report, Banco Central de la República Argentina.

Hastie, T., Tibshirani, R. y Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference, and Predictions*. Springer New York Inc., 2nd ed. edition.

Hosmer, D., Lemeshow, S. y Sturdivant, R. (2013). *Applied Logistic Regression*. John Wiley and Sons, 3a ed. edition.

Kuhn, M. y Johnson, K. (2018). *Applied predictive modeling*. Springer New York Inc.

Lantz, B. (2015). *Machine Learning with R*. Packt Publishing, 2nd edition.

Liaw, A. y Wiener, M. (2002). *Classification and regression by randomforest*. Forest, 23.

R Core Team (2020). R: A language and environment for statistical computing. https://www.R-project.org/.

# CARACTERÍSTICAS DE PACIENTES CON ESQUIZOFRENIA: ANÁLISIS MEDIANTE MODELOS DE ECUACIONES ESTRUCTURALES


## Lic. Lust Rimoldi, Grecia

**Directora: Dra. Chiapella, Luciana**

**Codirectora: Mg. Arnesi, Nora**



La esquizofrenia es una enfermedad psiquiátrica crónica que afecta múltiples dominios de la personalidad, causando deterioro cognitivo, social y funcional progresivamente. Diversos estudios analizaron la relación de la conciencia de la enfermedad con características clínicas y de los pacientes. Sin embargo, emplearon abordajes estadísticos que no permiten estudiar de manera concurrente las relaciones entre variables. Desde la práctica clínica de médicos e investigadores, es posible plantear modelos teóricos de interrelación de las variables como los Modelos de Ecuaciones Estructurales (SEM), que permiten establecer relaciones de dependencia y considerar factores que no pueden ser medidos en forma directa.

Este trabajo propone utilizar SEM para estudiar las relaciones entre la conciencia de enfermedad en pacientes con esquizofrenia, los síntomas clínicos, la cantidad de psicofármacos utilizados y sus efectos adversos, la duración de la enfermedad y el consumo de sustancias, como droga y/o alcohol.

Se concluyó que los efectos adversos de la medicación y la conciencia de la enfermedad aumentan en función del tiempo de progreso de la misma. Además, los efectos adversos se ven potenciados cuando los pacientes consumen sustancias y repercuten sobre la sintomatología, la cual se hace mayor cuando los efectos adversos aumentan y se correlaciona con el grado de conciencia de la enfermedad.


## INTRODUCCIÓN

La esquizofrenia es una enfermedad psiquiátrica crónica que afecta a múltiples dominios de la personalidad causando deterioro cognitivo, social y funcional en forma progresiva (Naber *et al.*, 2015). Los pacientes requieren tratamiento psicofarmacológico a largo plazo y de manera continua a fin de evitar recaídas que puedan llevar a poner en riesgo sus vidas y la de quienes los rodean. Estos psicofármacos pueden causar efectos adversos sobre otros aspectos de la salud, modificando la sintomatología clínica de la enfermedad, por lo que en muchos casos se produce un abandono total o parcial de la medicación. Además, los pacientes esquizofrénicos pueden presentar menor grado de conciencia de la enfermedad (*insight*) en relación a la enfermedad que padecen y otras alteraciones cognitivas que afectan la percepción de necesidad de tratamiento.

Dada la relevancia de este padecimiento, diversos estudios han analizado la posible relación de la conciencia de la enfermedad con distintas características clínicas y de los pacientes. Sin embargo, se han utilizado abordajes estadísticos que no permiten estudiar de manera concurrente las relaciones existentes entre las variables, fundamentalmente en situaciones donde una característica puede ser dependiente de ciertas variables explicativas y, a la vez, ser explicativa del comportamiento de otras variables dependientes (Torio Palmero, 2019).

Desde el punto de vista estadístico, el abordaje de estudios en el campo de la psiquiatría presenta un desafío particular: muchas de las características de interés son de naturaleza compleja, resultado de varias variables que interactúan. Por lo tanto, suele ser relevante entender el complejo entramado entre variables, tanto latentes como observables. Desde la práctica clínica de médicos e investigadores, es posible plantear modelos teóricos de interrelación de las variables, entre los cuales se encuentran los modelos de ecuaciones estructurales (*Structural Equation Models*, SEM) que, para el análisis estadístico de estos casos, resultan una herramienta específica de gran utilidad, dado que permiten establecer la relación de dependencia entre las variables y considerar factores que no pueden ser medidos en forma directa.

Por todo lo mencionado, este trabajo se propone utilizar los SEM para estudiar las relaciones entre la conciencia de enfermedad en pacientes con esquizofrenia, los síntomas clínicos, la cantidad de psicofármacos utilizados, los efectos adversos de

estos medicamentos, la duración de la enfermedad y el consumo de sustancias (drogas y/o alcohol).

## MATERIALES Y MÉTODOS

Los datos utilizados en este trabajo corresponden a un estudio multicéntrico en el cual participaron investigadores pertenecientes al Instituto de Farmacología de la Facultad de Medicina de la Universidad de Buenos Aires, cuyo objetivo fundamental fue evaluar la adherencia al tratamiento psicofarmacológico y otras características clínicas en pacientes esquizofrénicos en hospitales públicos de la ciudad de Buenos Aires.

La información se recolectó en los servicios de emergencia del Hospital Neuropsiquiátrico Dr. Braulio A. Moyano, Hospital Interdisciplinario Psicoasistencial José Tiburcio Borda y Hospital General de Agudos Parmenio Piñero de la ciudad de Buenos Aires, Argentina, en el período comprendido entre mayo de 2015 y diciembre de 2017.

Se incluyeron en el estudio a todos los pacientes con edades comprendidas entre 18 y 65 años que ingresaron al servicio de internación de corto plazo (menos de un mes), que presentaron diagnóstico de esquizofrenia según la Entrevista Neuropsiquiátrica Internacional estructurada MINI (Ferrando *et al.*, 2000) y que no tenían lesiones en el sistema nervioso central ni utilizaban tratamiento antipsicótico de depósito. Así, se incluyó a 89 individuos evaluados al momento de la hospitalización, luego de la explicación de los objetivos de la investigación y de la aceptación del consentimiento informado.

Al momento del ingreso al estudio, se recopiló información acerca de las características socio-demográficas de los pacientes (edad, sexo, escolaridad, trabajo, hijos y estado civil) y sus características clínicas referidas a la patología (edad al inicio de la enfermedad, número de internaciones previas a la enfermedad, número de hospitalizaciones desde el inicio de la enfermedad clínica, consumo de sustancias y duración de la enfermedad). Conjuntamente, se midieron variables relacionadas específicamente con el tratamiento psicofarmacológico (sintomatología, conciencia de la enfermedad, efectos adversos de la medicación y cantidad de psicofármacos consumidos).

En este trabajo, el análisis se enfoca en evaluar la relación entre el grado de conciencia de la enfermedad, la sintomatología clínica, la duración de la

enfermedad, el consumo de sustancias, la cantidad de psicofármacos utilizados y los efectos adversos producidos por los mismos. Para alcanzar este objetivo, se realizó el análisis de los datos siguiendo los pasos para la modelización de ecuaciones estructurales propuestos por Kline (2005) y Kaplan (2012): especificación del modelo, identificación del modelo, evaluación de la calidad de la base de datos, estimación de parámetros, evaluación de la bondad de ajuste e interpretación de los indicadores del modelo y la re-especificación del modelo e interpretación de resultados.

Dado que se tenía una teoría previa analizada por el grupo de investigadores del estudio sobre las variables latentes y sus relaciones con los indicadores observados, se planteó en primera instancia un modelo de análisis factorial confirmatorio (AFC) para corroborar si los factores latentes mostraban un buen ajuste de los datos, luego un modelo SEM estableciendo las re-especificaciones necesarias en relación al criterio del índice de modificación elegido y tomando decisiones basadas en los diferentes índices sugeridos en la literatura (CFI, TCI, RMSEA y SRMR).

En el análisis de los datos, se utilizó el *software* estadístico R (versión 4.0.2) y sus paquetes *sem*, *lavaan*, *psych*, *MVN* y *semPlot*.

## RESULTADOS

### Especificación del modelo

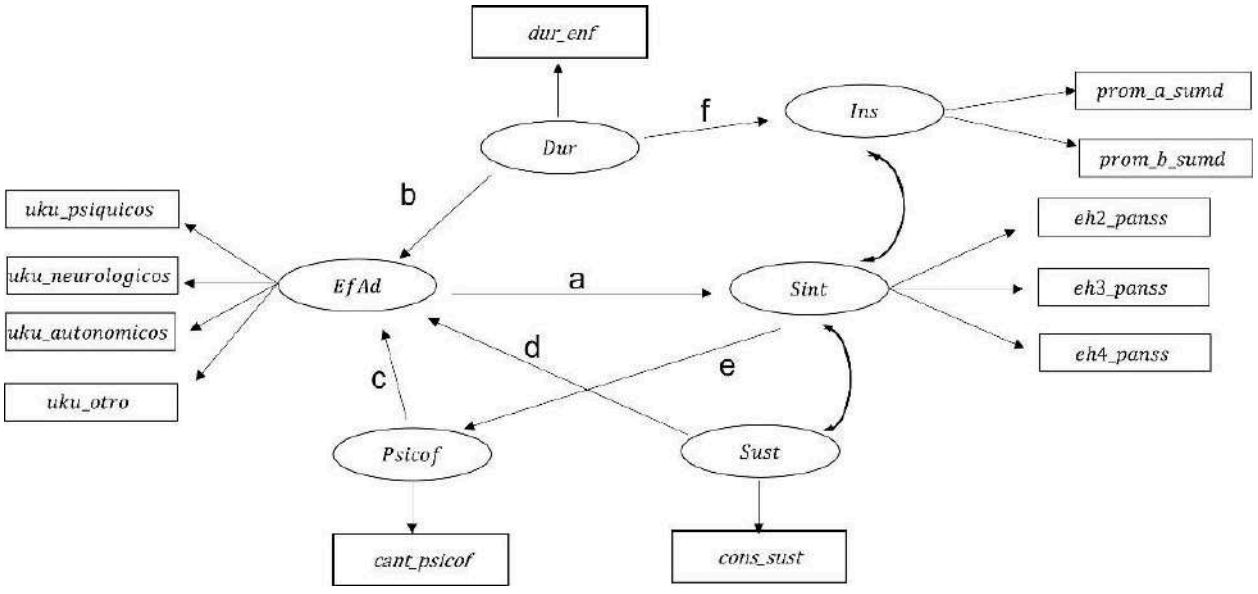
Se procede a plantear el modelo de la Figura 1, en el cual se relaciona a las variables latentes (representadas con elipses) endógenas: efectos adversos (*Ef Ad*), conciencia de la enfermedad (*Ins*), sintomatología (*Sint*) y cantidad de psicofármacos (*Psicof*), y variables latentes exógenas: duración de la enfermedad (*Dur*) y consumo de sustancias (*Sust*) con sus correspondientes variables observadas (representadas por rectángulos).

Cada variable latente es medida a través de variables indicadoras como lo son: efectos psíquicos (*uku_psiquicos*), efectos neurológicos (*uku_neurológicos*), efectos autonómicos (*uku_autonómicos*), otros efectos (*uku_otro*), conciencia de la enfermedad (*prom_a_sumd*), atribución de los síntomas (*prom_b_sumd*), sintomatología positiva (*eh2_panss*), sintomatología negativa (*eh3_panss*), sintomatología general (*eh4_panss*), duración de la enfermedad (*dur_enf*), cantidad de psicofármacos (*cant_psicof*) y consumo de sustancias (*cons_sust*).

Es necesario tener en cuenta que las variables *Dur*, *Psicof* y *Sust*, dado que tienen una única variable observada, son variables que pueden ser medidas en forma directa mediante las variables *dur_enf*, *cant_psicof* y *cons_sust*, respectivamente. Sin embargo, debido a las propiedades de los SEM, para poder considerar las relaciones entre estas variables y el resto de las variables del modelo, resulta útil generar las latentes ficticias *Dur*, *Psicof* y *Sust* teniendo en cuenta al momento de la modelización que se observan en forma directa, tal como se verá más adelante.

En este modelo, los efectos adversos tienen un impacto directo sobre la sintomatología a través del coeficiente *a* y la sintomatología, a su vez, tiene un efecto indirecto sobre los efectos adversos a través de la cantidad de psicofármacos con coeficientes *e* y *c*. Además, la duración de la enfermedad, la cantidad de psicofármacos y el consumo de sustancias tienen un impacto directo sobre los efectos adversos mediante los coeficientes *b*, *c* y *d* respectivamente. La duración de la enfermedad tiene un impacto directo sobre *insight* con coeficiente *f*. Las flechas curvas bidireccionales representan la correlación entre las variables *insight*, sintomatología y consumo de sustancias.

Figura 1. Modelo de ecuaciones estructurales inicial



## Identificación del modelo

Una vez propuesto el modelo por parte de los especialistas en el tema, en función a sus conocimientos y evidencias científicas sobre la relación entre las variables, se procedió a la identificación del mismo a través del análisis de los grados de libertad (*gl*).

Para la identificación del modelo se espera que los *gl* sean mayores o iguales que 0. El modelo inicial planteado en la Figura 1 es un modelo sobreidentificado con *gl* = 49.

## Evaluación de la calidad de la base de datos

Como primera etapa se realizó un análisis de correlación entre las variables y dado que los tipos de variables no son homogéneos, se deben calcular distintos coeficientes de correlación.

Para cuantificar la correlación entre las variables cuantitativas: duración de la enfermedad, sintomatología positiva, sintomatología negativa, sintomatología general, conciencia de la enfermedad, atribución de los síntomas, efectos psíquicos, efectos neurológicos, efectos autonómicos y otros efectos, se empleó la correlación de Pearson. En el caso de la correlación entre consumo de sustancias, con las variables cuantitativas mencionadas antes, se utilizó la correlación Biserial. Para conocer la correlación entre cantidad de psicofármacos y consumo de sustancias, se empleó la correlación de Kendal Tau C y, por último, la correlación de Spearman para conocer la relación entre cantidad de psicofármacos con el resto de las variables.

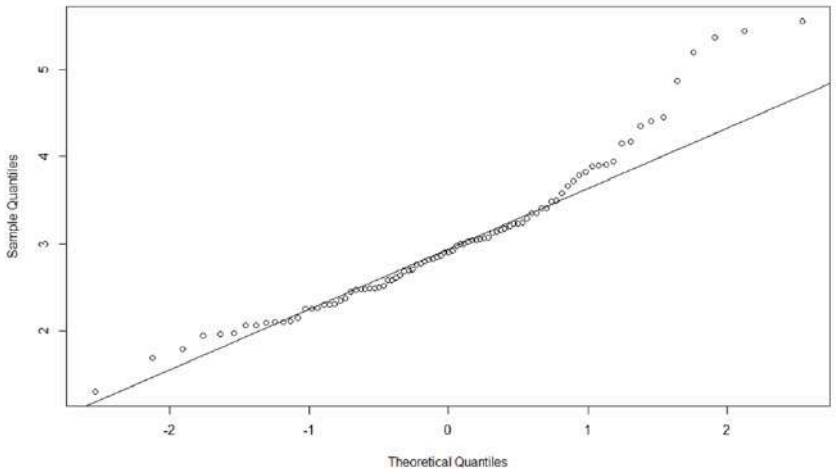
Se observó que todos los valores son inferiores a 0.85, por lo que se consideró que no hay variables altamente correlacionadas y, por lo tanto, no resultó necesario eliminar ninguna de ellas del análisis.

Antes de ajustar el modelo AFC a los datos, se debe determinar si la suposición de normalidad multivariada es sostenible para los datos disponibles. Para esto se usaron las pruebas de Mardia y Henze-Zinkler, donde la hipótesis nula propone que los datos siguen una distribución normal multivariada. A través de la observación en conjunto de la Figura 2 y las pruebas, se asumió que los datos no siguen una distribución normal multivariada. En el extremo superior del gráfico la línea de puntos se aleja de la recta de normalidad, lo cual indica que la distribución supuesta no es adecuada para estos datos. Esto además se comprobó con ambas pruebas, en las

cuales se obtiene un *p-value* menor a 0.05, lo que confirma lo mencionado anteriormente.

En relación a esta conclusión, al plantear el modelo SEM se utilizaron estimaciones robustas, (*weighted least squares mean and variance*, WLSMV) en lugar de máxima verosimilitud (ML).

Figura 2. Normal Q-Q plot para evaluar la normalidad multivariada

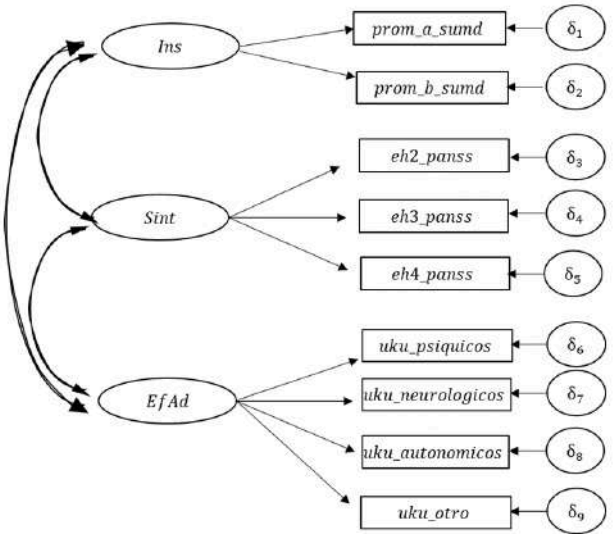



## Estimación de parámetros

El ajuste de los SEM se realiza en 2 etapas. Primero, los modelos AFC para los factores latentes deben ser estimados y mostrar un buen ajuste de los datos. Si el modelo no ajusta bien a los datos no es posible continuar con el análisis ni estimar los componentes del mismo porque los factores no son confiables.

Dado que se tenía una teoría definida sobre las variables latentes y sus relaciones con los indicadores observados, se comenzó el estudio planteando un modelo AFC en el cual se vinculó explícitamente a los indicadores conciencia de la enfermedad, atribución de los síntomas, sintomatología positiva, sintomatología negativa y sintomatología general, efectos psíquicos, efectos neurológicos, efectos autonómicos y otros efectos con los factores a los que teóricamente pertenecen, *insight*, sintomatología y efectos adversos (Figura 3).

Figura 3. Modelo de Análisis Factorial Confirmatorio




Para el ajuste del modelo AFC se obtuvo un $\chi^2$= 29.487, con 24 grados de libertad y un *p-value* mayor que 0.05, lo que indica que el modelo confirmatorio propuesto se ajusta a los datos. Además, CFI = 0.962 (≥ 0.95 buen ajuste), TLI = 0.943 (≥ 0.95 buen ajuste), SRMR = 0.065 (≤ 0.08 buen ajuste) y RMSEA = 0.051 (≤ 0.05 buen ajuste).

No se cumplió la normalidad multivariada de los indicadores y el modelo AFC ajusta a los datos, por lo que el estimador robusto proporciona un ajuste aceptable (el único índice que no genera un buen ajuste es el TLI, pero el valor es cercano a 0.95, por lo que no sería lógico rechazar el modelo por este valor). Por lo tanto, se continuó con el modelo estructural presentado en la Figura 1, estimando sus parámetros y determinando la bondad de su ajuste.

## Evaluación de la bondad de ajuste e interpretación de los indicadores del modelo

Al ajustar el SEM inicial, se observó que las relaciones que se vincularon con la variable cantidad de psicofármacos (efectos adversos y sintomatología), no resultaron ser significativas. Ambos coeficientes (*c* y *e* respectivamente) tienen valores de 0.443 y -0.002 con un *p-value* de 0.112 y 0.917 respectivamente. Tampoco resultó significativa la correlación entre sintomatología y consumo de sustancia con un *p-value* igual a 0.614.

Se concluyó que no existe una relación directa entre cantidad de psicofármacos y efectos adversos ni entre cantidad de psicofármacos y sintomatología, sin embargo,

al examinar el resto de los coeficientes de regresión, sí existen vínculos estadísticamente significativos entre las relaciones planteadas. Esto llevó a pensar que la cantidad de psicofármacos que consumen los pacientes no es importante en el modelo.

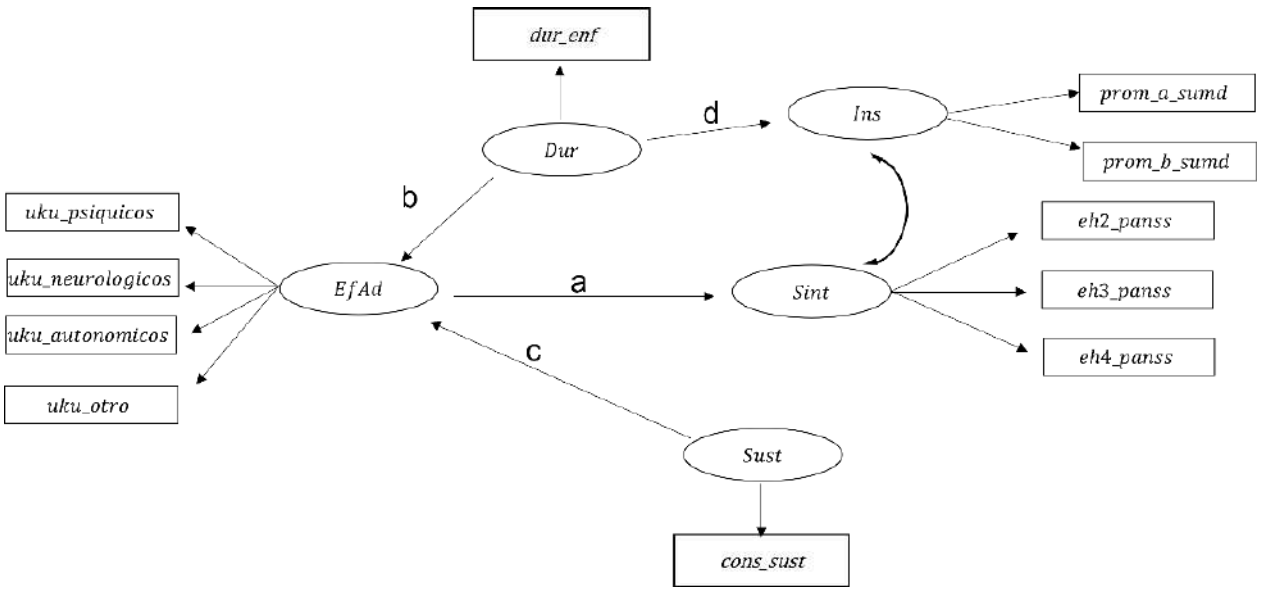
La estadística de la prueba de bondad de ajuste Chi cuadrado resultó estadísticamente significativa, el CFI (0.867 < 0.95), TLI (0.821 < 0.95), RMSEA (0.072 > 0.05) y SRMR (0.084 > 0.08) lo que indica que el modelo no se ajusta a los datos en forma aceptable, por lo que se procedió a realizar una re-especificación del modelo.

## Re-especificación del modelo e interpretación de los resultados

Se planteó el Modelo 2 (Figura 4) eliminando la variable cantidad de psicofármacos como también la relación indirecta entre sintomatología y efectos adversos y la correlación entre sintomatología y consumo de sustancias.

En términos de ajuste, el Modelo 2, al igual que el Modelo inicial, no se ajustó a los datos. La estadística de la prueba de bondad de ajuste Chi cuadrado sigue siendo significativa, y los coeficientes CFI=0.882, TLI=0.843, RMSEA=0.074 y SRMR=0.082 continúan sin indicar un buen ajuste a los datos para la estimación robusta. Al tener en cuenta el modelo ajustado se observó que todas las relaciones fueron estadísticamente significativas.

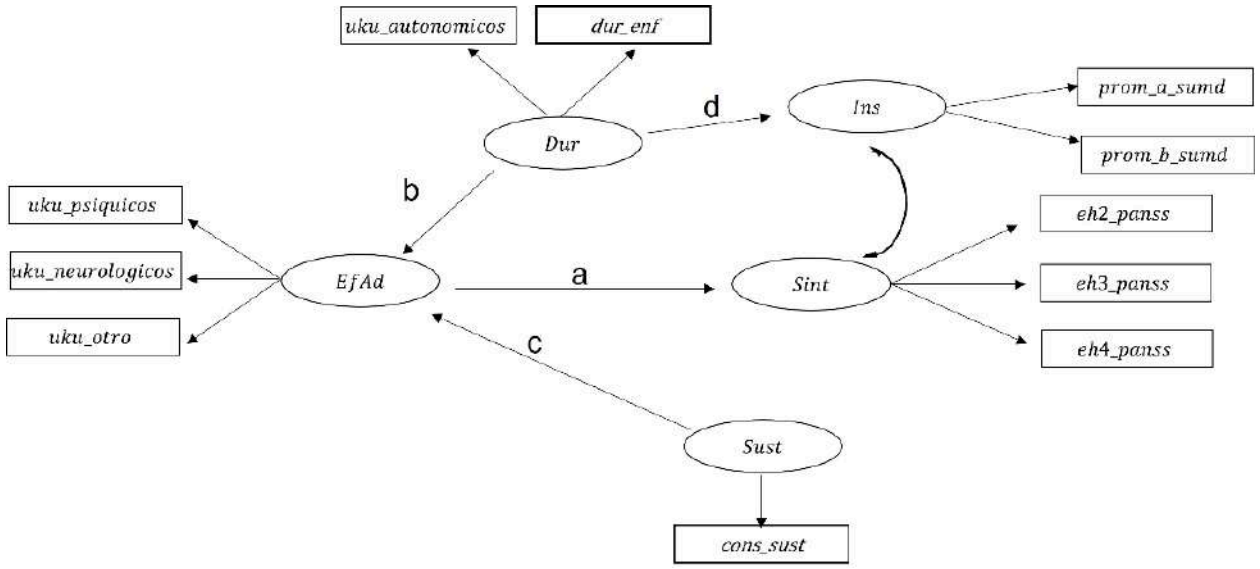
Figura 4. Modelo 2, luego de la re-especificación del Modelo inicial



Una vez conocida esta información, se continuó el análisis con la re-especificación del modelo examinando los valores de los índices de modificación (*mi*). Se obtuvo que la relación *insight - uku autonómicos* y *duración de la enfermedad - uku autonómicos* arrojaron valores de *mi* superiores a 13. Los valores de los índices de modificación son similares para ambas relaciones obtenidas a través de la re-especificación, pero, observando el resto de las estadísticas obtenidas se decidió seguir trabajando con la relación *duración de la enfermedad - uku autonómicos* que presenta valores superiores.

En función de los resultados obtenidos se planteó el Modelo 3 (Figura 5) modificando ahora el Modelo 2, adicionando la relación *duración de la enfermedad - uku autonómicos*.

Figura 5. Modelo 3, luego de la re-especificación del Modelo 2




A través del ajuste del Modelo 3, se observó que al eliminar la relación de *uku autonómicos* con efectos adversos y relacionarla ahora con la variable latente duración de la enfermedad, el modelo ajustó con un $\chi^2$ = 41.964 y un *p-value*=0.386, CFI=0.988, TLI=0.984, RMSEA=0.024 y SRMR=0.064; todas las relaciones fueron estadísticamente significativas (*p-value* < 0.05).

A partir de los resultados, hay una relación directa estadísticamente significativa entre sintomatología y efectos adversos, esta última con duración de la enfermedad y con consumo de sustancias e *insight* con duración de la enfermedad.

Todas estas relaciones son directas, lo cual confirma la importancia de cada una de ellas por sobre la otra. A medida que aumentan los efectos adversos (*a*=1.596), aumentan los síntomas que presentan los pacientes. El consumo de sustancias y el tiempo de duración de la enfermedad afectan de manera directa a los efectos adversos (*c*=2.073 y *b*=0.258 respectivamente), es decir, a medida que aumentan éstas, también lo hacen los efectos secundarios de la medicación. En lo que respecta a la conciencia de la enfermedad, se observa una relación positiva pero débil con duración de la enfermedad (*d*=0.074). La sintomatología está positivamente correlacionada con la conciencia de la enfermedad.

## CONSIDERACIONES FINALES

El análisis conjunto de las relaciones entre la conciencia de la enfermedad en pacientes con esquizofrenia, los síntomas clínicos, la cantidad de psicofármacos utilizados, los efectos adversos de la medicación, la duración de la enfermedad y el consumo de sustancias se basó en el estudio y utilización de modelos de ecuaciones estructurales.

En relación al problema planteado, los resultados obtenidos al ajustar el último modelo propuesto permitieron concluir que los efectos adversos de la medicación y la conciencia de la enfermedad aumentan en función del tiempo de progreso de la misma. A su vez, los efectos adversos se ven potenciados cuando los pacientes consumen drogas y/o alcohol y además repercuten sobre la sintomatología, la cual se hace mayor cuando los efectos adversos aumentan y se correlaciona con el grado de conciencia de la enfermedad.

Se debe tener en cuenta que el buen ajuste no significa que éste sea el modelo óptimo para los datos, ni que sea el único modelo que se ajuste bien a los datos. Más bien, se puede concluir que el modelo propuesto es capaz de reproducir con precisión las covarianzas entre los indicadores observados.

## BIBLIOGRAFÍA

Ferrando, L., Bobes, J., Gibert, J., Soto, M., & Soto, O. (2000). MINI Entrevista Neuropsiquiátrica Internacional (MINI International Neuropsychiatric Interview, MINI). *Instrum. Detección Orientación Diagnóstica*, 1–25.

Kaplan, D. W. (2012). *Structural equation modeling: Foundations and extensions* (2a ed.). SAGE Publications.

Kline, R. B. (Ed.). (2005). *Principles and practice of structural equation modeling. Methodology in the social sciences*. Guilford Publications.

Naber, D., Hansen, K., Forray, C., Baker, R. A., Sapin, C., Beillat, M., Peters-Strickland, T., Nylander, A. G., Hertel, P., Andersen, H. S., Eramo, A., Loze, J.-Y., & Potkin, S. G. (2015). *Qualify: a randomized head-to-head study of aripiprazole once-monthly and paliperidone palmitate in the treatment of schizophrenia*. Schizophrenia Research, *168*(1–2), 498–504. https://doi.org/10.1016/j.schres.2015.07.007

Torio Palmero, I. (2019). *Conciencia de enfermedad en esquizofrenia: relación con sintomatología y cognición*. Universidad Complutense de Madrid.

# META-ANÁLISIS PARA LA DETERMINACIÓN DE UN MODELO GENÉTICO QUE DESCRIBA LA ASOCIACIÓN ENTRE EL POLIMORFISMO FOK1 Y DIABETES TIPO 2


## Lic. Martinez, Josefina

**Directora: Mg. Cuesta, Cristina**



En los últimos años, se ha investigado cómo una adecuada reserva de Vitamina D podría mejorar la salud cardiovascular y prevenir la Diabetes Mellitus tipo 2 (DMT2). El gen receptor de Vitamina D (VDR) es crucial para que esta vitamina funcione en el cuerpo, y si presenta algún polimorfismo, su función puede verse afectada. Este estudio analiza la relación entre el polimorfismo Fok1 del gen VDR y la DMT2 mediante un meta-análisis. A diferencia de la mayoría de los estudios, los estudios de asociación genética requieren un mínimo de tres grupos, lo que dificulta la combinación de datos. Se utiliza la metodología de Takkistian (2004), que incluye verificar el Equilibrio de Hardy-Weinberg (H-W), explorar la heterogeneidad, usar análisis de regresión para agrupar datos y determinar el efecto genético, y analizar las diferencias de grupo por pares. Este método no asume un modelo genético a priori. El meta-análisis incluye 19 estudios, de los cuales 18 cumplen con el Equilibrio de H-W, sumando 5617 casos y 6196 controles. Se concluye que un modelo recesivo es adecuado para describir la relación entre el polimorfismo Fok1 y la DMT2.


## INTRODUCCIÓN

La DMT2 es una enfermedad endocrinometabólica crónica caracterizada por niveles elevados de glucosa en sangre, que con el tiempo conduce a daños graves en diferentes órganos. Ocurre cuando el cuerpo no produce suficiente insulina o se vuelve resistente a ella. Esta enfermedad está determinada por múltiples factores genéticos y ambientales.

Por su parte, la Vitamina D activa es una hormona esteroidea cuyo precursor (7-Dehidrocolesterol) puede ser incorporado con la dieta o sintetizado en la piel por la exposición a rayos UVB, el cual sufre modificaciones en el hígado y el riñón para convertirse en su forma activa. La Vitamina D activa cumple su función a través de su receptor: el gen VDR (Arias *et al.*, 2021-2022).

La Vitamina D tiene una amplia variedad de beneficios en nuestro cuerpo. Su déficit puede ocasionar raquitismo en los niños y osteomalacia en los adultos. Pero además, según investigaciones realizadas (Doheny, 2010; Harinarayan, 2014), un adecuado aporte de Vitamina D a nuestro organismo podría traducirse en un efecto protector en la salud cardiovascular de las personas y evitar la aparición de DMT2.

Varios estudios han demostrado un vínculo entre distintos polimorfismos del gen VDR y la DMT2, aunque los hallazgos difieren de una población a otra y entre los diferentes estudios. Los polimorfismos genéticos son variaciones en la secuencia del ADN. Estas variaciones pueden ser desde cambios en un solo nucleótido hasta cambios en largos tramos del ADN (Thakkistian *et al.*, 2003; Arias *et al.*, 2021-2022). El VDR puede presentar diferentes polimorfismos, entre los cuales se encuentra Fok1. La presencia de este polimorfismo se relaciona con cambios en la estructura del receptor, generando una proteína más corta.

Aún existen controversias sobre la relación entre este tipo de polimorfismo y el padecimiento de DMT2. De existir esta relación, es necesario establecer el modelo genético que la representa, es decir, determinar si se trata de un modelo recesivo, dominante, sobredominante, codominante, entre otros.

Dada la disparidad de resultados en la bibliografía actual, la cátedra de Fisiología de la Facultad de Ciencias Médicas de la Universidad Nacional de Rosario llevó a cabo una revisión sistemática (RS) para situar el estado del arte. A partir de esta RS, se intentó combinar los resultados a través de un meta-análisis.

Los meta-análisis para la determinación de modelos genéticos tienen características propias y distintivas de los meta-análisis estándares. En este sentido, Thakkinstian y sus colaboradores (2004) describieron una serie de pasos para llevar a cabo esta tarea.

En este trabajo se utiliza la metodología propuesta por Thakkinstian *et al.* (2004) para intentar encontrar un modelo genético que describa la asociación entre Fok1 y DMT2, en el marco de una RS realizada con este fin. Es importante remarcar que no se incluye la

problemática que conllevan los sesgos de las RS sino que se hará énfasis en la presentación metodológica para la determinación del modelo genético.

## OBJETIVO

El objetivo de este análisis es buscar un modelo genético que permita describir la asociación entre el polimorfismo Fok1 del gen VDR y el padecimiento de DMT2 a través de un meta-análisis.

## METODOLOGÍA

Como ya se mencionó, en este trabajo se sigue la metodología propuesta por Thakkinstan *et al.* (2004) para meta-análisis compuesto por estudios de casos y controles. El procedimiento consta de los pasos desarrollados a continuación.

### Evaluación del cumplimiento del Equilibrio de Hardy-Weinberg

La evaluación del Equilibrio de H-W se realiza sólo sobre los controles, dado que los casos pueden no estar bajo este equilibrio si existe asociación entre el genotipo y la enfermedad o condición estudiada (Pierce, 2011).

Para la evaluación del equilibrio se utiliza el test de bondad de ajuste Chi-Cuadrado. Sin embargo, por limitaciones propias de la prueba, ésta no es válida para tamaños de muestra pequeños.

Esta evaluación debe realizarse en cada estudio que participa del meta-análisis. Aquellos estudios que no verifican el Equilibrio de H-W deben eliminarse del análisis posterior.

### Evaluación de homogeneidad

La evaluación de la homogeneidad de los estudios se realiza a través de la metodología estándar para meta-análisis, en particular se usa la estadística $Q$ de Cochran, mediante la comparación de los genotipos de a pares. Ésta utiliza para su cálculo la estimación del efecto del genotipo de cada estudio, así como la estimación del efecto global de cada genotipo. En este estudio, el efecto de los genotipos se estima a través de razones de odds (RO).

## Análisis de Regresión

Antes de determinar el modelo genético, es necesario determinar si hay diferencias entre los genotipos. En caso de no haberse detectado heterogeneidad, el efecto del genotipo se estima ajustando un modelo de regresión logística, considerando al estudio como efecto fijo. En caso de existir heterogeneidad, se incluye en el modelo al estudio como un factor aleatorio.

Para determinar el efecto global de los genotipos, se compara un modelo que incluye su efecto contra uno que no, usando un test de razón de verosimilitud. Si el efecto global es significativo, se procede a realizar comparaciones entre las RO de los distintos genotipos.

## Determinación del modelo genético

Para determinar el modelo genético más apropiado, en primer lugar, se realiza un meta-análisis tradicional, calculando las RO para comparar los genotipos de a pares. De esta manera, se compara, para un locus con dos alelos (A y B), AA vs BB, y se estima $RO_1$; AB vs BB, dando lugar a $RO_2$; y AA vs AB, y se estima $RO_3$. Una vez obtenidas RO con sus respectivos intervalos de confianza (IC), se escoge un modelo a partir de las siguientes sugerencias:

Si $RO_1 \neq 1$, $RO_3 \neq 1$ y $RO_2 = 1$, se sugiere adoptar un modelo recesivo.

Si $RO_1 \neq 1$, $RO_2 \neq 1$ y $RO_3 = 1$, se sugiere adoptar un modelo dominante.

Si $RO_2 = 1$ o $RO_3 \neq 1$ y $RO_1 = 1$, se sugiere adoptar un modelo sobredominante completo.

Si $RO_1 > RO_2 > 1$ y $RO_1 > RO_3 > 1$ (o $RO_1 < RO_2 < 1$ y $RO_1 < RO_3 < 1$), se sugiere adoptar un modelo codominante.

Nótese que en la simbología anterior, por ejemplo $RO_1 = 1$ significa que el IC asociado a $RO_1$ cubre el valor 1, en cambio $RO_1 \neq 1$ significa que el IC para esa medición no cubre el 1.

## Estimación del efecto global

Una vez elegido el modelo, se combinan los grupos para estimar el efecto global genotipo mediante la teoría de meta-análisis convencional. Así, por ejemplo, si los

resultados se corresponden con un modelo recesivo, se colapsan los datos de AB y BB y se lleva a cabo el meta-análisis comparando AA vs AB + BB. La estimación global de ese meta-análisis será la apropiada para determinar la asociación bajo el modelo recesivo subyacente. Si ninguna de las situaciones de I a IV es plausible, entonces el meta-análisis no concluye respecto al modelo genético.

## MATERIALES

Los investigadores Arias, Martinelli y Petroni (2021-2022), de la cátedra de Fisiología de la Facultad de Ciencias Médicas de la Universidad Nacional de Rosario, llevaron a cabo una RS para identificar estudios que permitan evaluar la presencia del polimorfismo Fok1 en el gen VDR en pacientes diabéticos y controles sanos.

En este caso, los alelos son llamados mediante las letras T y C. Los alelos *wild type*, es decir, que no presentan ningún polimorfismo, se identifican con la letra T, mientras que los alelos polimórficos se identifican con la letra C.

La estrategia de búsqueda de la RS contenía palabras claves referentes al polimorfismo de VDR Fok1, a la Vitamina D y a la DMT2. Los criterios de inclusión de artículos fueron: (1) que sea un estudio de casos y controles, (2) que se haya estudiado el polimorfismo Fok1, (3) pacientes adultos de ambos sexos diagnosticados con DMT2, y (4) que se haya publicado en inglés, español, italiano, portugués o francés.

Al final de la revisión se obtuvieron 19 estudios, publicados entre 2003 y 2021, con 5817 casos y 6396 controles.

## RESULTADOS

Se evalúan todos los pasos propuestos para determinar la posible relación entre el polimorfismo de VDR Fok1 y el padecimiento de DMT2 y, en función de ello, determinar el modelo genético que relaciona ambas condiciones.

1. Evaluación del cumplimiento del Equilibrio de H-W

Uno de los estudios incluidos en el meta-análisis no cumple con el equilibrio de H-W, por lo que se lo excluye del análisis (Tabla 1). De esta manera, se cuenta con 18 estudios que contienen un total de 5617 casos y 6196 controles.

Tabla 1. Resultados del Test de Hardy-Weinberg para cada estudio (n=19)

| Estudio | Casos CC | Casos CT | Casos TT | Controles CC | Controles CT | Controles TT | p-value Test H-W |
|---|---|---|---|---|---|---|---|
| Sattar, 2021 | 120 | 280 | 100 | 65 | 105 | 30 | 0,239 |
| Selvarajan, 2020 | 115 | 75 | 10 | 70 | 110 | 20 | 0,014* |
| Satar, 2020 | 46 | 91 | 13 | 53 | 36 | 11 | 0,208 |
| Ma, 2019 | 237 | 394 | 43 | 344 | 161 | 16 | 0,587 |
| Rodrigues, 2019 | 16 | 31 | 9 | 31 | 24 | 7 | 0,482 |
| Gendy, 2018 | 16 | 26 | 8 | 32 | 15 | 3 | 0,498 |
| Angel, 2018 | 24 | 86 | 28 | 38 | 81 | 53 | 0,504 |
| Xia, 2017 | 129 | 94 | 19 | 38 | 50 | 12 | 0,468 |
| Rasheed, 2017 | 106 | 57 | 17 | 92 | 47 | 11 | 0,158 |
| Bertoccini, 2017 | 395 | 379 | 109 | 378 | 359 | 93 | 0,578 |
| Safar, 2017 | 46 | 42 | 12 | 38 | 47 | 15 | 0,940 |
| Mahjoubi, 2016 | 231 | 180 | 28 | 168 | 117 | 17 | 0,565 |
| Maia, 2016 | 46 | 42 | 12 | 38 | 47 | 15 | 0,940 |
| Yu, 2016 | 112 | 205 | 80 | 223 | 405 | 147 | 0,124 |
| Angel, 2015 | 24 | 96 | 40 | 53 | 75 | 32 | 0,560 |
| Jia, 2015 | 120 | 336 | 212 | 408 | 973 | 579 | 0,983 |
| Zhong, 2015 | 46 | 114 | 44 | 40 | 58 | 18 | 0,688 |
| Al-Daghri, 2012 | 22 | 133 | 213 | 19 | 111 | 129 | 0,461 |
| Malecki, 2003 | 85 | 159 | 64 | 77 | 110 | 52 | 0,284 |

* Significativo al 5%

2. Evaluación de homogeneidad

Se realiza el test de homogeneidad inter-estudios entre los genotipos de a pares. Cuando se consideran sólo los genotipos CC y TT, el 59% de la variabilidad total se debe a la heterogeneidad entre estudios ($I^2 = 0,59$), siendo ésta significativa según el test de homogeneidad ($Q = 41,15,\ p - value < 0,001$).

Cuando se consideran sólo los genotipos CT y TT, la variabilidad entre estudios resulta despreciable ($I^2 \simeq 0,\ Q = 15,76,\ p - value = 0,54$).

Cuando se consideran sólo los genotipos CC y CT, el 86% de la variabilidad total se debe a la heterogeneidad entre estudios ($I^2 = 0,86$), siendo ésta significativa según el test de homogeneidad ($Q = 117,60,\ p - value < 0,001$).

La estadística $Q$ para probar homogeneidad entre estudios para las comparaciones CC vs TT y CC vs CT reflejan que existe heterogeneidad entre estudios. Es por ello que se plantea un modelo con efectos aleatorios.

3. Análisis de Regresión

Con el objetivo de evaluar si el genotipo está relacionado con el padecimiento de DMT2, se plantean dos modelos $logit$, uno que incluye la variable genotipo y otro sin ésta. En ambos modelos se incluyó la variable "Estudio" como factor aleatorio, dada la heterogeneidad presente en los resultados de los mismos.

A fin de conocer si el genotipo está relacionado con el padecimiento de DMT2, se realiza un test de razón de verosimilitud comparando los dos modelos, el cual arroja un valor observado de 59,79 ($p - value < 0,001$). Esto indica que el efecto del genotipo es significativo.

4. Determinación del modelo genético:

Luego de haber probado que el efecto del genotipo sobre la probabilidad de padecer DMT2 es significativo, se procede a elegir un modelo genético adecuado para describir esta relación.

Se realizan las comparaciones entre los genotipos de a pares, a través de RO:

i. CC vs TT:

$$RO_{CCvsTT} = RO_1 = 0,742 IC_{RO1;\ 95\%} = [0,596; 0,922] p - value = 0,007$$

ii. CT vs TT:

$$RO_{CTvsTT} = RO_2 = 0,956$$

$$IC_{RO2;\ 95\%} = [0,858;\ 1,064] p - value = 0,410$$

iii. CC vs CT:

$$RO_{CCvsCT} = RO_3 = 0,716 IC_{RO3;\ 95\%} = [0,558;\ 0,919] p - value = 0,009$$

Dado que $RO_1 \neq 1$, $RO_3 \neq 1$ y $RO_2 = 1$, se sugiere adoptar un modelo recesivo.

5. Estimación del efecto global

Se combinan los genotipos CT y TT para la estimación del efecto global, ya que se optó por un modelo recesivo, considerando a la variable estudio como factor aleatorio.

De esta manera, se obtiene que la RO común estimada es $0,716$ ( $IC = [0,560;\ 0,914]$), lo que significa que la chance de padecer DMT2 cuando la persona posee genotipo CC es 28% menor que si la persona posee genotipo CT o TT.

La adopción de un modelo recesivo implica que son necesarias 2 copias de C para modificar el riesgo de padecer DMT2; por tanto, heterocigotos CT y homocigotos TT tienen el mismo riesgo. Se compara la combinación de ellos respecto a los homocigotos del alelo variante CC.

## CONSIDERACIONES FINALES

La mayoría de los meta-análisis de estudios de asociación genética han optado por reducir los tres grupos a dos ignorando los heterogicotos, realizando comparaciones separadas por pares o asumiendo algún modelo a priori. Cuando no se asume el modelo genético, muchos investigadores ajustan múltiples modelos y/o realizan comparaciones de a pares.

En caso de realizarse todas las comparaciones, las probabilidades asociadas deben ajustarse por comparaciones múltiples, lo que hace que el nivel de significación aumente, generando así una disminución en los resultados estadísticamente significativos. Sin embargo, rara vez se realizan ajustes por comparaciones múltiples y las estimaciones de a pares de las RO se obtienen mediante la realización de dos meta-análisis separados, ignorando la correlación entre las RO.

A fin de determinar el modelo genético que describe esta relación, se utilizó la metodología propuesta por Takkinstian *et al.* (2004), la cual construye el modelo final a través de una serie de pasos. Esto permite prescindir de la necesidad de hacer todas las comparaciones posibles. La evidencia mostró que el riesgo de padecer DMT2 es igual tanto para personas con genotipo CT como con genotipo TT, mientras que la chance de padecer DMT2 cuando la persona posee genotipo CC es 28% menor que si la persona padeciera cualquiera de los otros genotipos ($IC_{95\%}$=[0,56;0,91]).

Es importante destacar que esta metodología no tiene en cuenta si se tiene más de un polimorfismo en el mismo gen, ni si estos polimorfismos están relacionados entre sí.

En este trabajo no se pudo considerar otras posibles causas de heterogeneidad (por ejemplo, etnia), lo que habría hecho posible hacer análisis de subgrupos y, en consecuencia, mejorar las estimaciones dentro de cada uno de ellos.

## BIBLIOGRAFÍA


Arias, P., Martinelli, R., & Petroni, C. (2021-2022). *Programa de Formación y Perfeccionamiento en Investigación: Informe final.* Rosario.

Attia, J., Ioannidis, J. P., Thakkinstian, A., McEvoy, M., Scott, R. J., Minelli, C., et al. (2009). How to Use an Article About Genetic Association: C:What Are the Results and Will They Help Me in Caring for My Patients? JAMA , 304-308.

Attia, J., Thakkinstian, A., & D'este, C. (2003). Meta-Analyses of molecular association studies: Methodologic lessons for genetic epidemiology. Journal of Clinical Epidemiology, 297-303.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2023). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. R package version 1.1-33.

Benedetti, A., & Chen, B. (2017). *Quantifying heterogeneity in individual participant data meta-analysis with binary outcomes*. Recuperado el Mayo de 2022, de Systematic Reviews: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-017-0630-4

Bolar, K. (2019). *STAT: Interactive Document for Working with Basic Statistical Analysis*. R package version 0.1.0.

Doheny, K. (2010). *Relacionan baja vitamina D con mal control de la diabetes.* Recuperado el Agosto de 2022, de WebMD: https://www.webmd.com/a-to-z-guides/news/20100625/low-vitamin-d-linked-to-poor-diabetes-control

Harinarayan, C. V. (2014). Vitamin D and diabetes mellitus. *Hormones 13,* 163-181

Iniesta, R., Guinó, E. & Moreno, V. (2005) *Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos.* Gaceta Sanitaria.

Kalmes, R., & Huret, J. (2001). *Modelo de Hardy-Weinberg*. Recuperado en Mayo de 2022, de Atlas of Genetics and Cytogenetics in Oncology and Hematology: https://atlasgeneticsoncology.org/teaching/30100/modelo-de-hardy-weinberg

Martorell-Marugan J, T.-D. D.-R.-S. (2017). *MetaGenyo: A web tool for meta-analysis of genetic association studies*. Recuperado el 2022, de BMC Bioinformatics. 18:563: https://metagenyo.genyo.es/

Molina Arias, M. (2018). Aspectos metodológicos del metaanálisis (2). *Pediatría Atención Primaria*, 401-405.

Pierce, B. (2011). *Genética. Un enfoque conceptual.* Buenos Aires, Argentina: Panamericana.

Sánchez Meca, J., Marín Martínez, F., & Huedo, T. (2006). Modelo de efectos fijos versus modelo de efectos aleatorios. *En J.L.R. Martín, A. Tobías y T. Seoane (Coords.), Revisiones sistemáticas en ciencias de la vida: El concepto salud a través de la síntesis de la evidencia científica*, 189-204.

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). Meta-Analysis with R. Springer.

Thakkinstian, A., D'este, C., Eisman, J., Nguyen, T., & Attia, J. (2003). Meta-Analysis of Molecular Association Studies: Vitamin D Receptor Gene Polymorphisms and BMD as a Case Study. *Journal of Bone and Mineral Research*, 419-428.

Thakkinstian, A., McEldu, P., D'Este, C., Duffy, D., & Attia, J. (2004). A method for meta-analysis of molecular association studies. *Statistics in Medicine*, 1291-1306.

# OPTIMIZACIÓN DEL ESPACIO EN GÓNDOLA PARA PRODUCTOS DE PRIMERA NECESIDAD A TRAVÉS DE MODELOS LINEALES DE PREDICCIÓN


## Lic. Melgratti, Matías

**Directora: Mg. Cuesta, Cristina**

**Codirectora: Lic. Mryglod, María Eugenia**



Los supermercados de cercanía cuentan con un salón de ventas que no supera los 500 m2, no disponen de gran espacio físico para exhibir su mercadería y, en consecuencia, se requiere una precisa organización de los productos en las góndolas para generar resultados óptimos (tanto para el cliente como para el comerciante). En este trabajo se propone estimar el mínimo espacio en góndola necesario para exhibir diariamente un artículo de la canasta básica que requiere de mucho espacio, “papel higiénico”. Se utilizan modelos lineales para la predicción de la cantidad de rollos de papel higiénico a vender diariamente en una tienda. Se consideran como variables predictoras, características de las tiendas (como superficie, cantidad de cajas registradoras, entre otras) y de sus ventas diarias de otros productos (rollos de cocina, unidades de azúcar, yerbas, arroz, entre otras). Se comparan diferentes métodos de validación cruzada para seleccionar el mejor modelo predictivo, es decir, el que tenga menor error de predicción. A partir del modelo elegido, se realizan las predicciones y recomendaciones a los comerciantes acerca de la superficie que se debe disponer en góndola para ese producto.


## INTRODUCCIÓN

En este trabajo se aborda la problemática de la optimización del espacio requerido para exhibir un producto en góndola. Se estudia la categoría “papel higiénico”, debido a que pertenece a la canasta básica y además es una de las categorías que más espacio en góndola ocupa.

A fin de optimizar el espacio en góndola, primeramente, se debe tener una estimación de la venta diaria. La sobrestimación de este valor generaría una falta de espacio para exhibir otras categorías y la subestimación generaría “fueras de *stock*”, impactando negativamente en la facturación del supermercado.

La predicción de la cantidad de rollos de papel higiénico a vender, se realiza a través de modelos lineales. Se comparan distintos métodos de validación cruzada para elegir el mejor modelo predictivo (Tibshirani *et al*., 2008). A partir del modelo resultante, se obtienen la predicción de las unidades de rollos de papel higiénico a exponer diariamente en la góndola. Luego, se transforma el valor predicho en unidades (rollos), a “volumen a ocupar en la góndola”. Estos resultados son de utilidad para hacer recomendaciones a los comerciantes según las características de su tienda.

## OBJETIVO

El objetivo de este trabajo es hallar el modelo lineal con la mejor capacidad predictiva, a través de métodos de validación cruzada, para predecir la cantidad de rollos de papel higiénico a vender diariamente en un supermercado de cercanía.

## METODOLOGÍA

### Modelos Predictivos

La elección de un modelo estadístico usualmente se lleva a cabo según el propósito para el que es pensado. Algunos modelos intentan explicar el fenómeno en estudio (modelos explicativos). Otros, como en este caso, el objetivo es predecir la variable respuesta en nuevas situaciones, lo más adecuadamente posible (modelos predictivos). En general, es deseable lograr un balance entre un buen ajuste y una buena predicción, además de intentar lograrlo con un modelo relativamente parsimonioso.

La forma del modelo lineal general es: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$

donde $i = \overline{1,n}$ siendo $n$ el número de observaciones y $p$ el número de variables explicativas. Los $\beta_j$ ($j$=1, …, $p$) son los parámetros a estimar y $\varepsilon_i$ el término de error aleatorio.

Para poder elegir el "mejor" entre un grupo de modelos candidatos, se define una serie de medidas o criterios que permiten realizar esta comparación y evaluación.

## Medidas para la evaluación de un modelo lineal

Una de las estadísticas más utilizadas para evaluar la bondad de ajuste es el coeficiente de determinación múltiple ($R^2$). El valor del $R^2$ aumenta a medida que se agregan variables al modelo y por ello no es posible comparar modelos con distinto número de variables. Como alternativa se considera el coeficiente de determinación ajustado $\left(R_a^2\right)$, ya que éste tiene en cuenta el número de parámetros. Otra alternativa es desde el punto de vista Bayesiano, el criterio BIC (*Bayesian Information Criterion*). El BIC suele penalizar a los modelos con más variables explicativas.

Por otro lado, la medida más usual para evaluar la capacidad predictiva de un modelo es el Error Cuadrático Medio (ECM). El ECM tiende a ser pequeño cuando las estimaciones están cerca del valor observado. Análogamente al ECM, también se puede considerar el Error Absoluto Medio (EAM).

## Métodos de Validación Cruzada

El proceso de validación cruzada consiste en dividir aleatoriamente las observaciones en dos partes, la muestra de entrenamiento (*training set)* y la muestra de validación (*testing set).* Con el *training set* se busca el modelo que mejor se ajusta a los datos y con el *testing set* se estudia su capacidad predictiva.

Una vez definido el mejor modelo predictivo, se estiman sus parámetros a partir del conjunto total de datos. Algunos de los métodos de validación cruzada más usuales son:

**Validación cruzada simple:** se dividen aleatoriamente los datos en dos partes. Las particiones más comunes son 70/30, 80/20 o 90/10. La partición más grande corresponde al *training set* y es donde se ajustan todos los modelos posibles con 1,2, …, $p$ parámetros. Los modelos con la misma cantidad de parámetros son

comparados con la estadística $R^2$y BIC y se selecciona el mejor de ellos (mayor $R^2$ y menor BIC). Es decir, se tendrá el mejor modelo con 1 variable explicativa, con 2 variables explicativas, etc. Luego, se predicen los datos del *testing set* a partir de dichos modelos y se compara su capacidad predictiva a través del ECM y EAM. Es decir, entre todos los modelos que son probados en el *testing set* se selecciona uno que tiene la mejor capacidad predictiva.

Este proceso podría repetirse $h$ veces, lo cual da origen a $h$ modelos que podrían o no ser coincidentes. El modelo final será aquel que se presente con mayor frecuencia en las $h$ repeticiones.

**Validación cruzada dejando una observación fuera:** se divide al conjunto de datos en dos particiones, con la particularidad de que el *testing set* está formado solamente por una observación y las restantes $n - 1$ observaciones del conjunto forman el *training set*.

Una desventaja importante es que, si se cuenta con un conjunto de datos muy grande, el método puede resultar muy demandante computacionalmente.

**Validación cruzada con *K* grupos:** se divide aleatoriamente al conjunto de datos en $K$ grupos de aproximadamente igual tamaño y no solapados. Un grupo es considerado como *testing set* y los restantes $K - 1$ grupos juntos se consideran como *training set* para ajustar el modelo. El trabajo de selección del modelo se lleva a cabo de forma similar a la validación cruzada simple.

## MATERIALES

Los datos utilizados para dar respuesta al problema, corresponden a 420 supermercados de cercanía de las provincias de Buenos Aires y Córdoba, recolectados de mayo a julio del 2022 por *Scentia*, consultora argentina con más de 10 años de trayectoria en análisis de mercado y consumo masivo.

La predicción de la cantidad de rollos que se venderían por día se lleva a cabo teniendo en cuenta posibles variables tales como las unidades vendidas de los demás productos del supermercado, la provincia donde se encuentra, los metros cuadrados, la cantidad de cajas registradoras, la cantidad de tickets diarios, entre otras. Algunas de las categorías consideradas son: Aceite, Café, Lavandina, Galletitas, Gaseosas, Yerba, entre más de 200 diferentes.

## RESULTADOS

La selección del modelo de predicción se realiza a través de distintos métodos de validación cruzada que son luego comparados para obtener el modelo final. A fines prácticos y por limitaciones computacionales, no se aplica el método de validación cruzada dejando una observación fuera.

Para el procesamiento de la información, las técnicas de validación cruzada, las estimaciones de los modelos y las estadísticas de evaluación se utiliza el programa libre *R* en el entorno de *R Studio*.

Se prueba la validación cruzada simple bajo los escenarios que resultan de la combinación de las proporciones consideradas para cada partición (70/30, 80/20 y 90/10) y el número de iteraciones (10, 100 y 1000).

Los ECM y EAM resultantes de los modelos luego de 100 iteraciones de la partición 90/10 de los modelos se promedian entre los valores de cada iteración para cada uno de los modelos con $p$ parámetros, resultando la mejor alternativa en comparación con $h = 10$ y $h = 1000$.

Una vez determinado que con 100 iteraciones se puede encontrar el mejor modelo para la partición 90/10, se comparan los resultados de las particiones 70/30 y 80/20 con 100 iteraciones cada una. No se puede concluir cuál partición resulta mejor, aunque, en general, los ECM son levemente menores para la partición 90/10. Por ello, se elige como el método más adecuado a la partición 90/10 con 100 iteraciones.

El método de validación cruzada con *K* grupos, se prueba particionando el conjunto de datos en 5 grupos, donde se seleccionan los mejores modelos con distinto número de parámetros y se estiman sus errores absolutos y cuadráticos medios con 10 iteraciones, luego se promedian los errores entre todas las muestras para cada modelo con $p$ parámetros. Además, se itera el proceso 100 y 1000 veces.

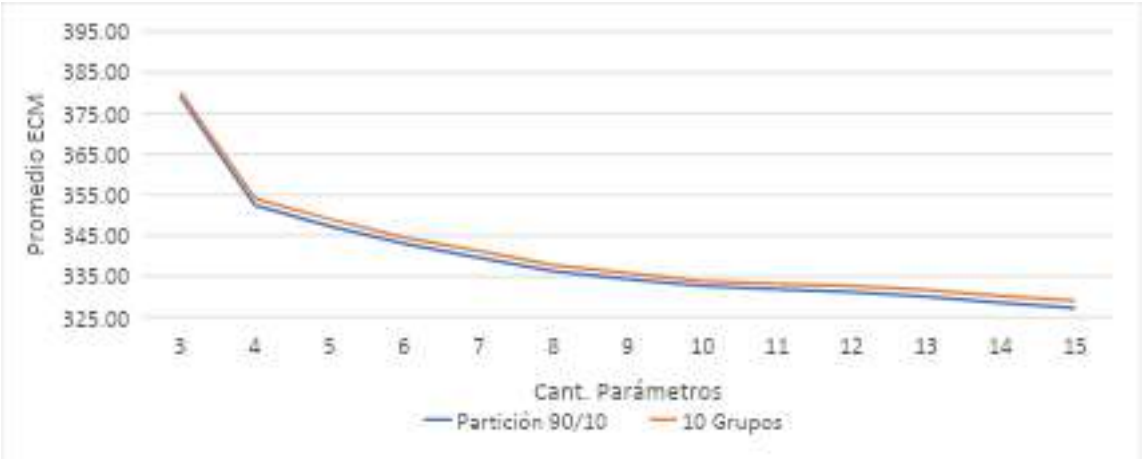
Para la validación cruzada con 5 grupos no se modifica demasiado la curva de los errores para los diferentes números de iteraciones. Sin embargo, para $h = 100$ la curva se encuentra sutilmente por debajo de las demás.

Análogamente, se compara la validación cruzada para 5 y 10 grupos con 100 iteraciones. En este caso, los ECM resultantes con $k = 10$ son levemente menores que para $k = 5$, por lo tanto, se puede concluir que el valor de $k = 10$ podría ser un

poco más efectivo para hallar un modelo con mayor predicción, aunque las diferencias son pequeñas.

Para completar la selección del método para hallar un modelo lineal con la mejor capacidad predictiva, se comparan las distintas técnicas de validación cruzada que se consideraron como “mejores” (validación cruzada simple con la partición 90/10 contra validación cruzada con 10 grupos, en 100 repeticiones cada uno).

Gráfico 1. Promedios de ECM de modelos con hasta 15 parámetros en 100 repeticiones de validación cruzada con partición 90/10 y de validación cruzada con 10 grupos




Aunque las diferencias sean pequeñas, el método que resulta más preciso y también el más eficiente computacionalmente para seleccionar un modelo lineal de predicción, es la validación cruzada simple con una partición de 90/10 en 100 iteraciones (Gráfico 1). Vale la pena mencionar que se consideran modelos con hasta 10 parámetros ya que se puede observar que, para cada uno de los métodos vistos, a partir de allí, la curva se vuelve aproximadamente plana, indicando que el modelo no mejora si se agregan más parámetros.

Luego de aplicar la validación cruzada simple con la partición 90/10 en 100 iteraciones, se busca cuáles son los modelos seleccionados en cada iteración ya que las variables seleccionadas pueden diferir en cada una de las iteraciones. Una vez identificados todos los modelos posibles con sus respectivas variables y la cantidad de veces que fueron seleccionados, se obtiene que el más frecuente resultó ser aquel que contiene la cantidad de tickets emitidos por día, las unidades promedio por ticket, la participación de la categoría “papel higiénico” y su porcentaje

de descuento. Además, contiene las unidades diarias de azúcar, dulce de leche, yerba, detergente, lavandina y rollos de cocina.

Tabla 1. Parámetros estimados del modelo seleccionado

| Parámetro | $\beta^{est}$ | Std.Err |
|---|---|---|
| **Tickets** | 0,01 | 0,01 |
| **Cant. prom. ticket** | -10,11 | 0,11 |
| **Participación** | 1718,67 | 8,77 |
| **Desc. papel higiénico** | 68,35 | 6,55 |
| **Dulce de Leche** | 0,89 | 0,03 |
| **Azúcar** | 0,05 | 0,01 |
| **Yerbas** | 0,62 | 0,02 |
| **Detergentes** | 1,43 | 0,04 |
| **Lavandinas** | 1,01 | 0,03 |
| **Rollos de Cocina** | 0,97 | 0,02 |

La predicción de la cantidad de rollos de papel higiénico a vender diariamente, a partir del modelo queda especificado de la siguiente manera (Tabla 1):

$$\hat{Cant.\ Rollos\ Papel\ Higiénico} = 0,01\ Tickets - 10,11\ Cant.\ Prom.\ Ticket + 1718,67\ Participaci$$

Para un supermercado de cercanía hipotético denominado **“Caso 1”:** emisión promedio de 360 tickets por día, con 4 artículos promedio por ticket, la categoría papel higiénico representa un 3 % de sus ventas y no tiene aplicado algún descuento, que venden 10 unidades de azúcar diarias y asignándole valores también al resto de las categorías de productos seleccionadas en el modelo, se espera que venda 55 ($IC_{95\ \%}$: 54,5 - 55,5) rollos diarios de papel higiénico (14 paquetes de 4 rollos).

Un supermercado de cercanía que emite en promedio 500 tickets diarios, con 5 unidades promedio cada uno, donde el 5 % de sus ventas están representadas por la categoría papel higiénico con un descuento del 5 %, que vende en promedio 15 unidades de dulces de leche y 10 de detergentes, se puede considerar como **“Caso 2”** y asignándole valores hipotéticos a las demás variables del modelo, se estima

que venda aproximadamente 111 ($IC_{95\ \%}$: 111,0 - 111,8) rollos de papel higiénico por día (28 paquetes de 4 rollos).

Para supermercados de cercanía hipotéticos, **"Caso 3"**, que emiten 600 tickets diarios con aproximadamente 6 unidades cada uno, que el 5 % de sus ventas correspondan a papel higiénico con un descuento aplicado del 10 % y venden en promedio 12 unidades diarias de lavandinas y unas 35 de yerbas, y asignando valores hipotéticos a las demás variables del modelo, se estima que vendan 150 ($IC_{95\ \%}$: 148,8 - 150,3) rollos de papel higiénicos diarios (38 paquetes de 4 rollos).

## DISCUSIÓN FINAL

En este trabajo se les dio respuesta a los comerciantes de los supermercados de cercanía sobre la cantidad de rollos a vender por día de la categoría "papel higiénico" mediante un modelo lineal. Para seleccionar el modelo lineal de predicción, se compararon los distintos métodos de validación cruzada pero no se encontraron diferencias sustanciales entre ellos.

El modelo seleccionado cumplió con los supuestos distribucionales. Sin embargo, la independencia podría ser discutible debido a que podría existir correlación entre las observaciones de una misma tienda. En un futuro trabajo podría considerarse esta situación incluyendo en el modelo el efecto aleatorio "supermercado".

Como sugerencia para futuros estudios resultaría de gran interés extender este análisis a otras categorías de productos como pueden ser Bebidas, Rollos de Cocina, Jabón y Detergentes para Ropa, Aceites o Yerbas, dado que son categorías que también ocupan un gran espacio en góndola y además tienen una alta rotación dentro de la canasta básica.

Otra alternativa a tener en cuenta es considerar la variable respuesta, en este caso, la cantidad de rollos a vender diariamente, como una variable Poisson y de este modo, ajustar un modelo lineal generalizado.

## BIBLIOGRAFÍA

Arlot, S. & Celisse, A. (2010). A Survey of Cross-Validation procedures for model selection. *Statistics Survey*, 40-79.

Curhan, R. (1972). The Relationship between Shelf Space and Unit Sales in Supermarkets. *American Marketing Association.*

James, G., Witten, D., Hastie, T. & Tibshinari, R. (2017). *An Introduction to Statistical Learning.* New York, Usa: Springer.

Kohavi, R. (2001). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *ResearchGate*.

Montgomey, D., Peck, E. & Vining, G. (2001). *Introduction to Linear Regression Analysis.* New York, Usa: John Wiley & Sons.

Searle, S. & Gruber, M. (2017). *Linear Models.* New Jersey, Usa: John Wiley & Sons.

Tibshirani, R., Hastie, T. & Friedman, J. (2008). *The Elements of Statistical Learning.* Stanford, California: Springer.

# ESTUDIO DEL COMPORTAMIENTO DE LA VARIABLE ALTURA ELIPSOIDAL, UTILIZANDO HERRAMIENTAS DE LA ESTADÍSTICA ESPACIAL. CIUDAD DE ROSARIO, AÑO 2010


## Lic. Suarez, Marina Aldana

**Directora: Mg. Balparda, Laura Rita**

**Codirectora: Mg. Borra, Virginia Laura**



La altura elipsoidal es la distancia vertical entre un punto de la superficie terrestre y un modelo matemático que representa la forma de la Tierra. El estudio de esta variable es necesario para conocer el relieve de la superficie terrestre, característica relevante al momento de planificar y diseñar políticas públicas.
Un Modelo Digital de Elevaciones es una matriz de datos en la cual cada elemento refiere a un valor de altura georreferenciado. Estos datos pueden ser estudiados usando el Análisis Exploratorio de Datos Espaciales, que permite caracterizar datos georreferenciados a través de técnicas que describen y visualizan distribuciones espaciales, identifican observaciones atípicas y descubren formas de asociación espacial.
El objetivo principal de este trabajo es estudiar el comportamiento de la altura elipsoidal en la ciudad de Rosario, utilizando herramientas de estadística espacial, para datos captados en el año 2010. Se encontró que dicha variable presenta autocorrelación espacial positiva y a partir del análisis de la correlación espacial local se encontraron zonas “altas” y “bajas” en la ciudad. El modelo de tendencia cuadrática entre las coordenadas (X, Y) y la variable en estudio es el que mejor ajusta y el que se debe incluir en la fase de modelización y predicción.


## INTRODUCCIÓN

La información geoespacial actualizada se utiliza para la planificación estratégica territorial y el diseño de políticas públicas, de modo tal que permita definir problemáticas a resolver, analizar variables que posibiliten relacionar diferentes elementos ubicados en la superficie de la tierra y, por último, postular una posible solución (IGN, s.f). Especialmente, la información geoespacial referida al relieve o superficie de la Tierra, donde se localizan diferentes objetos (edificaciones, rutas, calles, vegetación, etc.) es de suma importancia en trabajos de construcción, operaciones militares, navegación aérea, entre otras. Además, esta información se convierte en un insumo fundamental para usuarios cartógrafos, topógrafos, geólogos, hidrólogos, ingenieros, militares y para profesionales de diferentes disciplinas que hacen uso de los Sistemas de Información Geográfica (SIG).

Una de las variables estudiadas para resolver problemáticas tales como inundaciones, sequías y desertificación es la altura elipsoidal, que se define como la distancia perpendicular entre un punto de medición y la superficie del elipsoide. Conocer el comportamiento de dicha variable es el primer paso previo a la obtención de Modelos Digitales de Elevaciones (MDE) y así dar soluciones de planeamiento urbano y/o de gestión de emergencias, entre otros temas prioritarios vinculados con el territorio.

La información geoespacial está conformada por objetos y capas de información geográfica georreferenciadas, sus atributos y sus relaciones espaciales. Para su análisis, se utilizan métodos que permiten extraer las características de los datos georreferenciados y se conocen con el nombre de Análisis Exploratorio de Datos Espaciales (AEDE). Estos métodos se conciben como una disciplina dentro del análisis estadístico más general, diseñada para el tratamiento específico de los datos geográficos.

El objetivo principal de este trabajo es estudiar el comportamiento de la variable altura elipsoidal de la ciudad de Rosario, utilizando herramientas de la estadística espacial, para datos captados en el año 2010.

## MARCO TEÓRICO

Trabajar con información georreferenciada requiere conocer una serie de conceptos previos necesarios para poder realizar distintos tipos de operaciones. En los datos georreferenciados, la componente espacial refiere a una localización sobre la superficie de la Tierra. Por ello, se debe tener conocimiento de la forma de ésta y como se obtiene y procesa la información geográfica.

Es necesario comprender dos conceptos básicos: geoide y elipsoide (Olaya, 2014). El primero es la superficie teórica de la Tierra que une todos los puntos que tienen igual gravedad; esta superficie presenta irregularidades, por lo tanto, hay una

distancia distinta desde el centro de la Tierra a cada punto del geoide. Entonces para representar toda la superficie terrestre se emplea un modelo matemático, el cual es distinto para cada región, de forma que se adapte mejor a la forma de la Tierra. Este modelo de representación de la Tierra se denomina elipsoide.

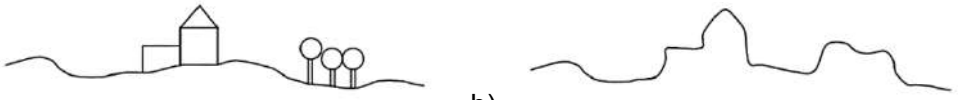
Una de las variables de interés en este trabajo es la altura elipsoidal que se define como la distancia vertical entre la superficie del elipsoide y el punto de medición; la magnitud depende principalmente del elipsoide empleado y de otros factores que intervienen en la medición.

## Modelos digitales de terreno

La elevación de la superficie terrestre es una de las variables más estudiadas en ciencias de la Tierra y medioambiente, ya que es utilizada en un amplio rango de aplicaciones científicas y civiles. Debido a limitaciones técnicas y económicas, no se puede medir y almacenar la elevación de cada punto sobre la superficie terrestre (Burgos & Salcedo, 2014). Por lo tanto, se hace inviable representar la superficie terrestre utilizando toda la muestra de elevación (puntos sobre el terreno) disponible y es necesario reducir el volumen de datos (Hernández *et al*., 2013).

Un MDE refiere, de manera genérica, a los datos topográficos digitales captados o relevados a través de distintos procedimientos, así como también, al método para interpretar a las elevaciones entre las observaciones (Maune *et al*., 2001). Con un MDE se pueden describir las características y elementos, tanto naturales como antrópicos, presentes sobre la superficie terrestre (Figura 1).

Figura 1. Perfil: a) esquema del mundo real y b) MDE



a) b)

Fuente: tomado de Fuentes *et al.,* 2012

A partir de un MDE, es posible calcular un Modelo Digital de Terreno (MDT). La principal diferencia radica en que este último modelo no incluye las características y elementos antrópicos y la vegetación (Figura 2), y sólo considera “los valores de elevación de los puntos más bajos de una superficie (terreno)” (Fuentes *et al.*, 2012, p. 259).

Figura 2. Perfil: a) esquema del mundo real y b) MDT

a)  b) 

Fuentes: tomado de Fuentes *et al*., 2012

## Área de estudio y datos

Se cuenta con un conjunto de datos que contiene información sobre la variable altura elipsoidal (metros) medida sobre distintas intersecciones de ejes de calles de la ciudad de Rosario. Estas mediciones corresponden a datos derivados del procesamiento de imágenes satelitales captadas en el año 2010. El conjunto de datos incluye tres variables, “Altura elipsoidal”, “Coordenada X” y “Coordenada Y” y un total de 10.403 observaciones.

# METODOLOGÍA

## Datos espaciales

Se entiende por dato espacial a todo aquel que tiene asociada una referencia geográfica, de modo tal que se puede localizar exactamente dónde sucede dentro de una región o área geográfica. Esta posición especificada en el espacio es relevante en la descripción y en el análisis de los datos (Olaya, 2014). El análisis de datos en todas las áreas del conocimiento hizo necesario recurrir a diferentes herramientas estadísticas. Más precisamente, para el análisis de los datos que presentan una componente espacial se puede recurrir a la estadística espacial, que se define como la reunión de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios (puntos del espacio o agregaciones espaciales) de una región. Es decir que la estadística espacial analiza las relaciones de un proceso estocástico $\{y_i : i \in \mathrm{D}\}$, en el que *i* representa una ubicación en el espacio euclidiano *d*-dimensional, $y_i$ es una variable aleatoria en la ubicación *i*, donde *i* varía sobre un conjunto de índices $\mathrm{D} \subset R^d$ (Giraldo, 2002).

La estadística espacial se divide en tres áreas (Giraldo, 2002), en función de las características del conjunto D de índices del proceso estocástico de interés. Éstas son: geoestadística, *lattice* y patrones espaciales.

## Autocorrelación espacial

A partir de la década de los setenta, la geografía ha experimentado un gran cambio. Entre los elementos que han aportado a este desarrollo se encuentra la formalización del concepto de autocorrelación espacial (o dependencia espacial). Tobler (1970) define que el efecto de dependencia espacial es la relación funcional existente entre los valores que asume una variable en una unidad del espacio y en unidades vecinas. Es decir, el valor que toma una variable en una unidad no sólo está explicado por condiciones internas, sino también por los valores que toma esa misma variable en otra unidad del espacio. Esta dependencia puede ser expresada según la primera ley de la geografía, en la cual todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes. Formalmente:

$$y_i = f\left(y_{i'}\right) \quad i = 1, ..., n \qquad i' \neq i,$$

donde $y_i$ e $y_{i'}$ son realizaciones (valores) de una variable aleatoria localizadas en las posiciones $i$ e $i'$ del espacio y $n$ es el número de unidades espaciales en la región D.
El efecto de la autocorrelación espacial puede ser de signo positivo, negativo o nulo. En particular, la autocorrelación espacial positiva se refiere a la asociación entre valores similares de una variable y locaciones cercanas, es decir, cuando la presencia de un fenómeno en una región hace que ese fenómeno se extienda hacia las regiones que lo rodean favoreciendo la concentración del mismo. Es llamado efecto contagio o desbordamiento.

## Índice *I* de Moran global y Diagrama de dispersión de Moran

La distribución no aleatoria de fenómenos en el espacio no puede estudiarse a partir de los métodos de la estadística clásica, sin embargo, tiene varias consecuencias para el análisis estadístico. Uno de ellos es que introduce “redundancia” entre los datos tal que cada unidad adicional aporta menos información nueva. Esto afecta, por ejemplo, al cálculo de los intervalos de confianza y a la estimación de parámetros. Por ende, es necesario evaluar el grado de autocorrelación en un

conjunto de datos espaciales antes de cualquier análisis estadístico (Acevedo Bohórquez & Gómez Álvarez, 2008).

La medida más utilizada para dicha evaluación es el índice I de Moran, que es una adaptación de una medida de correlación no-espacial a un contexto espacial y se define como (Moran, 1950):

$$I = \frac{n}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{i'=1}^{n} w_{ii'}(y_i-\bar{y})(y_{i'}-\bar{y})}{\sum_{i=1}^{n}\sum_{i'=1}^{n} w_{ii'}} \qquad \forall\, i \neq i',$$

donde:

$n$ es el número de unidades espaciales en la región D,

$y_i$ es la variable aleatoria en la ubicación $i$,

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ es la media de la variable,

$w_{ii'}$ es el elemento de la matriz de conectividad que recoge la relación de vecindad entre las unidades $i$ e $i'$ $(i = 1, 2, \ldots\ldots, n \; y \; i' = 1, 2, \ldots\ldots, n)$.

Una representación gráfica habitual de la autocorrelación espacial es el diagrama de dispersión de Moran (Anselin, 1996). Consiste en un gráfico de dispersión donde se representa la variable en estudio estandarizada y el retardo espacial de dicha variable estandarizada (promedio ponderado de los valores que toma la variable en el sub-conjunto de observaciones vecinas).

Este diagrama permite analizar los diferentes tipos de autocorrelación: positiva, negativa y nula. Si los valores se encuentran concentrados sobre la diagonal que cruza los cuadrantes I (derecha superior) y III (izquierda inferior) existe una elevada correlación espacial positiva de la variable, de forma que su pendiente es igual al valor obtenido para el contraste de la estadística *I* de Moran. La dependencia será negativa si los valores se concentran en los dos cuadrantes restantes. Si, por el contrario, la nube de puntos está dispersa en los cuatro cuadrantes es indicio de ausencia de correlación espacial (Serrano & Vayá, 2000).

## Dependencia espacial local y mapa LISA

Los indicadores globales de dependencia espacial no son capaces de detectar la inestabilidad de ciertas estructuras locales de asociación, que pueden estar presentes o no, en una estructura global de dependencia. De allí, la importancia de

considerar medidas con una perspectiva local, como la versión local del índice de Moran (LISA, por su sigla en inglés *Local Indicators of Spacial Association*). Este índice permite la identificación de patrones locales de asociación espacial. Descomponiendo el Índice Moran se puede evaluar la influencia de ubicaciones individuales en la estadística global. Además, se encarga de representar aquellas localizaciones con valores significativos en indicadores estadísticos de asociación espacial local, alertando así de la presencia de puntos calientes (*hot spots*) o atípicos espaciales, cuya intensidad depende de la significación asociada de los datos estadísticos analizados (Anselin, 1995).

Se define el índice de asociación espacial local de Moran para la unidad i como (Anselin, 1995):

$$I_i = (y_i - \bar{y}) \sum_{i'=1}^{n} w_{ii'}(y_{i'} - \bar{y}).$$

A partir de los valores LISA se pueden obtener representaciones gráficas llamadas Mapas LISA, que destacan las unidades espaciales con valores significativos de $I_i$, poniendo de manifiesto, a través de una gradación de colores, la presencia de cinco tipos de conglomerados espaciales: Alto-Alto, Bajo-Bajo, Alto-Bajo, Bajo-Alto, Relación no significativa (Yrigoyen, 2003).

## RESULTADOS

A partir del cálculo de medidas resúmenes, se conoce que la distancia vertical promedio entre el punto de medición y la superficie del elipsoide en la ciudad de Rosario (año 2010) es de 43,19 m. El 50 % de las observaciones presenta un valor de altura elipsoidal menor a 42,95 m. Se encuentra que el 25 % de las observaciones toma valores menores a 39,72 m y el 75 % menores a 46,83 m, siendo el rango de variación 38,46 m.

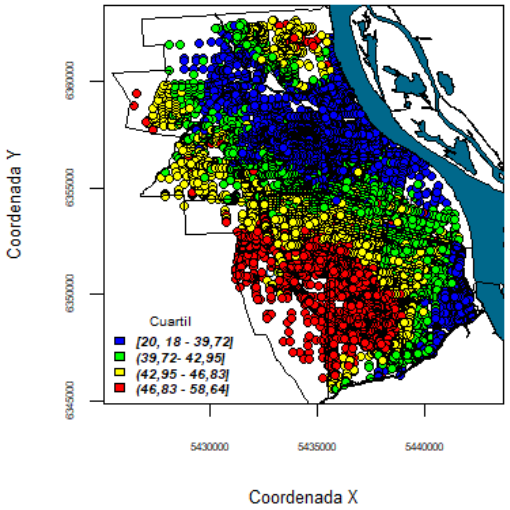
La ciudad de Rosario se encuentra dividida en 6 zonas territoriales denominadas “distritos” (Centro, Norte, Noroeste, Oeste, Sudoeste y Sur).

En la distribución espacial de la altura elipsoidal (Figura 3), se observa que los valores más altos (puntos rojos) se encuentran en la zona Sudoeste de la ciudad de Rosario y al norte del distrito Norte. En cambio, los valores más bajos (puntos azules) se encuentran en el distrito Noroeste y parte del distrito Norte, donde se ubica la desembocadura del arroyo Ludueña. Asimismo, se observan valores bajos

en el distrito Sur, donde desemboca el arroyo Saladillo y en el distrito Centro, en la zona del Monumento Nacional a la Bandera.

Los valores intermedios de la altura elipsoidal se ubican en el distrito Centro, al norte del distrito Sur y al Oeste del distrito Noroeste; también se observan algunos valores intermedios en el distrito Norte.

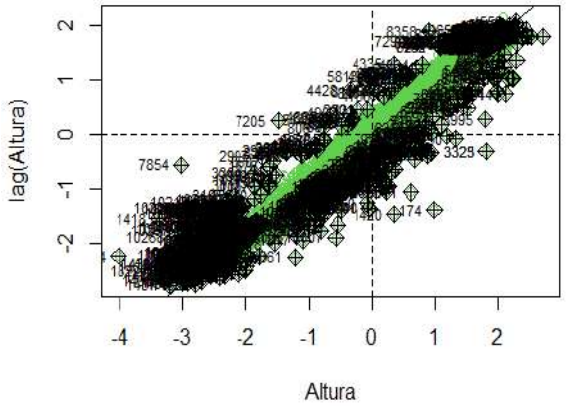
Figura 3. Distribución espacial de la variable altura elipsoidal según cuartil. Ciudad de Rosario, año 2010




El índice *I* de Moran es igual a 0,91 (p-value< 0,001) a partir del cual se concluye con un nivel de significación del 5 % que existe autocorrelación espacial positiva, es decir, puntos con valores altos de altura elipsoidal están rodeados de otros puntos con valores altos de altura elipsoidal, mientras que puntos con valores bajos están rodeados de otros similares.

A partir del diagrama de dispersión de Moran se analiza cómo las observaciones que corresponden a la altura elipsoidal se encuentran autocorrelacionadas espacialmente. Se detecta que la autocorrelación espacial es positiva, ya que la mayoría de las observaciones se encuentra concentrada en la diagonal que cruza los cuadrantes I y III (Figura 4).

Figura 4. Diagrama de dispersión de Moran

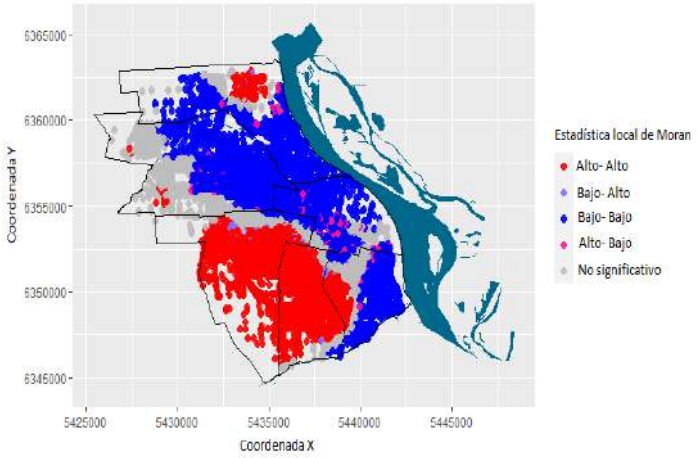



A continuación, se calculan índices de asociación espacial local de Moran para cada punto de la ciudad incluido en la muestra. En la Tabla 1 se cuantifica el número de observaciones según categorías del índice. En tanto que, en el mapa LISA (Figura 5) se representa la distribución espacial, donde se visualizan observaciones asociadas significativamente en un área local permitiendo identificar conglomerados y *outliers* espaciales.

Tabla 1. Cantidad de observaciones según conglomerado LISA

| Alto-Alto | Bajo-Alto | Bajo-Bajo | Alto-Bajo | No sig. | Total |
|---|---|---|---|---|---|
| 3.926 | 123 | 4.356 | 150 | 1.848 | 10.403 |

Figura 5. Mapa de conglomerado LISA




Con un nivel de significación del 10%, se detecta que 8.555 puntos resultaron significativos, mientras que 1.848 no resultaron significativos. En la Figura 5 donde se representa el mapa LISA, los puntos en rojo (3.926) son observaciones con

valores altos de altura elipsoidal rodeados de observaciones con valores también altos. Estos puntos se ubican principalmente en el distrito Sudoeste, Oeste y una zona del distrito Norte de la ciudad. Los puntos en color azul (4.356) son observaciones con valores bajos de altura elipsoidal rodeados de puntos con valores bajos, ubicados sobre todo en el distrito Centro, Sur, Norte y Noreste de la ciudad.

Por último, para poder conocer si el valor de las coordenadas X e Y explican la variación de la variable altura elipsoidal, se evalúan distintos modelos de regresión. Se obtiene que un modelo con tendencia cuadrática en X e Y es el que mejor explica la variabilidad total de la variable altura elipsoidal con $R^2_{ajus} = 0,56$. Se obtiene un valor similar de $R^2$ ajustado, cuando la tendencia es cuadrática en X y cúbica en Y. Los valores más pequeños de los criterios de bondad de ajuste AIC y BIC también se obtienen con dichos modelos.

El modelo de tendencia cuadrática en ambas coordenadas es el más apropiado para explicar la relación entre la altura elipsoidal y las coordenadas, ya que obtiene los mejores valores de las estadísticas de bondad de ajuste y es el más parsimonioso. Estos resultados deberán tenerse en cuenta si se desea realizar predicciones en sitios de la ciudad donde no se hayan hecho mediciones (*Kriging*).

## CONSIDERACIONES FINALES

Cuando se realizan planificaciones territoriales para resolver una o varias problemáticas como inundaciones, deslizamientos de tierra, sequías, proyectos de grandes obras de ingeniería, planificación de vuelos, análisis de riesgos ambientales, etc., es necesario estudiar la superficie de la Tierra con herramientas que describan y visualicen la distribución espacial y que luego permitan la modelación y predicción de la variable en estudio. Conocer el comportamiento de la altura elipsoidal de un territorio es de importancia para resolver este tipo de problemáticas.

En el estudio se describió la distribución espacial de la variable según su ubicación en la ciudad de Rosario y se encontró que la distancia vertical en los puntos de medición con la superficie del elipsoide es menor en zonas ubicadas en la desembocadura de los arroyos Ludueña y Saladillo, las cuales se corresponden aproximadamente a zonas con riesgo de inundaciones.

Por otro lado, a partir del índice *I* de Moran y el diagrama de dispersión se comprobó la existencia de autocorrelación espacial positiva de la altura elipsoidal, es decir, en términos generales, los valores bajos de altura elipsoidal en la ciudad de Rosario se encontraron rodeados por valores bajos, encontrándose el mismo patrón en ubicaciones donde se presentaron valores altos de altura elipsoidal.

En el análisis de regresión lineal que explica la relación entre la altura elipsoidal y las coordenadas X e Y, se concluyó que el modelo de tendencia cuadrática en ambas coordenadas es el que mejor ajusta y el que se debe incluir en la fase de modelización y predicción para la obtención de un modelo digital de terreno.

Agradecimiento a: Ingeniera Agrimensora y Licenciada en Física María Cecilia Torralba.

## BIBLIOGRAFÍA

Acevedo Bohórquez, I. & Gómez Álvarez, N. M. (2008). *Algunos elementos para el análisis de datos espaciales: teoría y aplicación* [Tesis de maestría no publicada]. Universidad EAFIT, Colombia.

Anselin, L. (1995). Local indicators of spatial association LISA. *Geographical analysis*, 27(2): 93-115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. *In Spatial analytical perspectives on GIS* (pp. 111-126). Routledge.

Burgos, V. H. & Salcedo, A. P. (2014). *Modelos digitales de elevación: Tendencias, correcciones hidrológicas y nuevas fuentes de información*. Encuentro de Investigadores en Formación en Recursos Hídricos (2, 2014, Ezeiza, Buenos Aires, Argentina). [Consulta: marzo de 2022]. Disponible en: http://www.ina.gov.ar/ifrh-2014/Eje1/1.11.pdf.

Fuentes, J. E., Bolaños, J. A. & Rozo, D. M. (2012). Modelo Digital de Superficie a partir de Imágenes de Satélite Ikonos para el análisis de Áreas de Inundación en Santa Marta, Colombia. Boletín de Investigaciones Marinas y Costeras - INVEMAR, 41 (2), 251-266. [Consulta: marzo de 2021]. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0122-97612012000200001&lng=en&tlng=es.

Giraldo, H. R. (2002). *Introducción a la geoestadística: Teoría y aplicación.* Bogotá: Universidad Nacional de Colombia.

Hernández, C. U. P., Castillo, W. E. S., Becerra, X. M. & Hernández, S. B. J. (2013). Evaluación y comparación de métodos de interpolación determinísticos y probabilísticos para la generación de modelos digitales de elevación. *Investigaciones Geográficas, Boletín del Instituto de Geografía*, *2013*(82), 118-130.

Instituto Geográfico Nacional. República Argentina. (s.f). [Consulta: marzo de 2021]. Disponible en: https://www.ign.gob.ar/.

Maune, D. F. & American Society for Photogrammetry and Remote Sensing. (2001). Digital elevation model technologies and applications: The DEM user's manual. Bethesda, Md: American Society for Photogrammetry and Remote Sensing.

Olaya, V. (2014). Sistemas de información geográfica. Libro libre de Víctor Olaya. [Consulta: agosto de 2021]. Disponible en: https://volaya.github.io/libro-sig/.

Serrano, R. M. & Vayá, E. V. (2000). *La Utilidad de la Econometría Espacial en el Ámbito de la Ciencia Regional* por Esther Vayá Valcarce. Documento de trabajo, 2000, 13. Universidad de Barcelona.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:34–240.

Yrigoyen, C. C. (2003). *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*. Dirección General de Economía y Planificación. Consejería de Economía e Innovación Tecnológica. Madrid, España.

CONSEJO PROFESIONAL
DE CIENCIAS ECONOMICAS
DE LA PROVINCIA DE SANTA FE
CAMARA II

Colegio de Graduados
en Ciencias Económicas
de Rosario

UNR
Universidad
Nacional
de Rosario