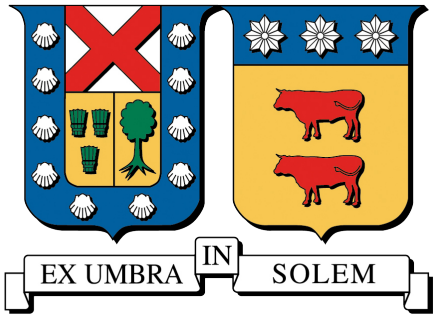

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INDUSTRIAS

ENFOQUE DATA-DRIVEN PARA MANTENIMIENTO PREDICTIVO: ROBUSTECIENDO EL MODELO DE RIESGOS PROPORCIONALES CON DATA SCIENCE PARA LOS DESAFÍOS ASOCIADOS A MÚLTIPLES COVARIABLES.

TESIS PARA OPTAR AL TÍTULO DE

MAGÍSTER EN CIENCIAS DE LA INGENIERÍA INDUSTRIAL

AUTOR

VÍCTOR DANIEL ÁLVAREZ BARRÍA

PROFESOR GUÍA	:	DAVID GODOY R.
PROFESOR INTERNO	:	VÍCTOR ALBORNOZ S.
PROFESOR EXTERNO	:	HANS MILLÁN R.
PROFESOR EXTERNO	:	RODRIGO BARRAZA V.

SANTIAGO, 27 de MARZO, 2024

Citas y Agradecimientos

A mi familia,

por su paciencia y sabiduría inconmensurable que guió mi camino durante estos años voraces.

A mi madre y hermana,

por ser los pilares fundamentales de mi vida. Sin su amor y apoyo jamás hubiera logrado todo esto.

A todas esas personas que,

de mil formas distintas y únicas, aportaron en mi formación profesional y personal.

"But remember this,... airplanes are not tools for war... airplanes are beautiful dreams. Engineers turn dreams into reality."

(H. Miyazaki, 2013)

"... Si puedes soñar sin que los sueños te dominen;
Si puedes pensar y no hacer de tus pensamientos tu único objetivo;
Si puedes encontrarte con el Triunfo y el Desastre, y tratar a esos dos impostores de la misma manera..."

(R. Kipling, 1895)

Data-driven Approach to Predictive Maintenance: Strengthening Proportional Hazards Models with Data Science for Multiple Covariate Challenges.

Abstract

In asset-intensive industries, the integration of Condition-Based Maintenance (CBM) and Proportional Hazards Model (PHM) is crucial for improving predictive maintenance strategies. This integration is employed to predict reliability, failure rates, and guide maintenance decisions. This analysis requires understanding the distribution parameters and covariate weights; the parameter estimation can have multiple feasible solutions. Therefore, strategies for bounding and seed values require special attention. Once these parameters are estimated, it is necessary to define the transition probabilities of the asset's conditions/states over time. Considering more than one covariate involves a composite measure whose physical magnitude is challenging for expert judgment to interpret, complicating the estimation of state bands to construct the transition probability matrix. Thus, this research addresses the challenges associated with orders of magnitude related to the use of multiple covariates in parameter weighting and estimation of the transition probability matrix of the Weibull PHM model using Data-Driven algorithms to enhance the estimation of the impact of conditions on equipment life. Innovative strategies for initial value estimation and bounds are proposed with Gradient Boosting, Random Forest, and IPCRidge, increasing the robustness of covariate weighting. GMM clustering is used as an innovation for band estimation, providing significant results as a complement to expert criteria. Therefore, this contributes to the advancement of predictive maintenance, using Data Science to improve the interpretation of asset health, so that maintenance decision-making occurs at the optimal moment, thereby increasing the remaining useful life of the assets.

Enfoque Data-driven para mantenimiento predictivo: Robusteciendo el modelo de riesgos proporcionales con Data Science para los desafíos asociados a múltiples covariables.

Resumen

En industrias intensivas en activos, la integración del Mantenimiento Basado en Condiciones (CBM) y el Modelo de Riesgos Proporcionales (PHM) es crucial para mejorar las estrategias de mantenimiento predictivo. Esta integración se emplea para predecir la confiabilidad, las tasas de falla y orientar las decisiones de mantenimiento. Este análisis requiere conocer los parámetros de distribución y los pesos de las covariables; la estimación de parámetros puede tener múltiples soluciones factibles. Por lo tanto, las estrategias para el acotamiento y valores semilla requieren de especial atención. Una vez que se estiman estos parámetros, es necesario definir las probabilidades de transición de las condiciones/estados del activo a lo largo del tiempo. Considerar más de una covariable implica una medida compuesta cuya magnitud física es difícil de interpretar para el juicio experto, lo cual dificulta la estimación de bandas de estados para construir la matriz de probabilidad de transición. Así, esta investigación aborda los desafíos de órdenes de magnitud asociadas al uso de múltiples covariables en la ponderación de parámetros y estimación de la matriz de probabilidad de transición del modelo Weibull PHM utilizando algoritmos Data-Driven con el fin de robustecer la estimación del impacto de las condiciones en la vida útil de los equipos. Se plantean estrategias innovadoras de estimación de valores iniciales y cotas con Gradient Boosting, Random Forest e IPCRidge, incrementando la robustez de la ponderación de covariables. Para la estimación de bandas se utiliza clustering GMM como innovación, proporcionando resultados significativos como complemento a los criterios expertos. Por lo tanto, se contribuye al avance del mantenimiento predictivo, usando Data Science para mejorar la interpretación de la salud de los activos, de modo que la toma de decisiones de mantenimiento sea en el instante óptimo y, de esta forma, aumentar la vida útil remanente de los mismos.

Índice

Introducción **1**

Problema De Investigación **3**

Objetivos **4**

Objetivo General 4

Objetivos Específicos 4

Revisión de Literatura **5**

Gestión de Activos Físicos 5

Modelo de riesgos proporcionales 6

Data Driven PHM 9

Metodología **16**

Preparación de Datos 17

Escalamiento de datos 17

Estimación de parámetros 18

Parámetros de Distribución Weibull 18

Pesos de covariables 19

Estrategias de optimización 21

Estimación de bandas de estado 23

Clustering 24

Cálculo de rangos de covariables 25

Matriz de Probabilidad de Transición 25

Cálculo de la confiabilidad 26

Resultados **28**

Estimación de parámetros 29

Cálculo de Bandas 40
Confiabilidad Condicional y RUL 47

Conclusiones y Recomendaciones **52**
Conclusiones Secundarias 54

Trabajos Futuros **55**

Introducción

Dentro de una empresa u organización intensiva en activos, uno de los principales objetivos a cumplir es la mejora de la productividad, es decir, mejorar el rendimiento de los sistemas. Este objetivo se puede alcanzar a través de la Gestión de Activos Físicos (PAM), el cual se enfoca en alcanzar un resultado sostenible de la productividad de los equipos, considerando, por ejemplo, el establecimiento de una política de mantenimiento óptima que permita alcanzar las metas deseadas para dichos equipos, ya sea maximizar la disponibilidad o minimizar los costos, entre otros.

Entre las políticas de mantenimiento, una de las más interesantes es el mantenimiento predictivo, la cual define el momento más adecuado de intervención sobre un equipo para minimizar la probabilidad de falla mediante diversas técnicas. El mantenimiento basado en condición (CBM), es una técnica predictiva que estima la tasa de falla y la confiabilidad de un equipo en función de las condiciones actuales y futuras. Esto se consigue mediante el monitoreo constante de estas condiciones y el uso de herramientas como el Modelo de Riesgos Proporcionales (PHM), el cual asigna un peso a cada condición para luego calcular el riesgo de falla del equipo en ese momento, estos pesos junto con los parámetros de la distribución utilizada pueden ser calculados a partir de la maximización de la función de verosimilitud (MLE), esta función presenta no convexidad considerando una distribución Weibull y múltiples covariables, por lo que pueden haber múltiples soluciones factibles. Definir un punto de partida (valor semilla) para los parámetros junto con cotas que ayuden en la resolución al algoritmo de optimización empleado se convierten entonces en decisiones importantes de estudiar. Sin embargo, esto no es suficiente para decidir sobre la intervención del activo; es necesario definir los distintos valores del riesgo de falla que precisan el momento oportuno para mantener el equipo en funcionamiento y para realizarle mantenimiento, es decir, establecer ciertos rangos que definan las distintas decisiones posibles.

Al combinar el mantenimiento basado en condición (CBM) con el modelo de riesgos proporcionales (PHM) es posible controlar las constantes vitales o covariables para predecir

la tasa de fallos, la función de confiabilidad y las decisiones de mantenimiento. Este análisis requiere definir las probabilidades de transición de las condiciones de los activos que se mueven entre estados a lo largo del tiempo. Cuando sólo se evalúa una covariable, los parámetros del modelo suelen derivarse de la opinión de expertos para proporcionar directamente bandas de estados. El reto reside en los problemas con múltiples covariables, en los que el juicio experto pierde utilidad, ya que la medida compuesta no representa ninguna magnitud física. Además, la selección de covariables requiere más procedimientos para priorizar las más relevantes.

El resto de la presente tesis está organizada de la siguiente forma. Se expone el **Problema De Investigación** que surge como el resultado de la necesidad de negocio a resolver, luego se definen los **Objetivos** que se esperan alcanzar para dar solución a la problemática, después se realiza una revisión de la literatura existente relacionada al tema de la investigación en la sección **Revisión de Literatura**, en la sección siguiente se explica la **Metodología** seguida para alcanzar los objetivos propuestos, después se muestran los **Resultados** y las **conclusiones** obtenidas, para finalizar con una serie de **Trabajos Futuros** y líneas de investigación propuestas a seguir a partir del trabajo desarrollado en esta tesis.

Cabe destacar que dos artículos de investigación (Godoy et al., 2023; Godoy et al., 2024), en los cuales participa el autor, sirvieron de base para la metodología expuesta en la presente Tesis.

Problema De Investigación

En años recientes, ha habido un creciente interés en utilizar Data Science (DS) y algoritmos avanzados para mejorar la precisión de los modelos PHM. La estimación de parámetros en modelos PHM con múltiples covariables es un desafío debido a la naturaleza no convexa de la función de verosimilitud y la alta dimensionalidad del espacio de parámetros. Para abordar este desafío, es necesaria una metodología para estimar los pesos de covariables y los parámetros Weibull en modelos PHM.

Actualmente, no existe un método predeterminado que calcule de forma precisa y sistemática los rangos o el estado de los equipos para diferentes contextos y condiciones. Por ello, es relevante desarrollar un modelo para determinar múltiples bandas de covariables para la matriz de probabilidades de transición para fortalecer el modelo PHM y complementar el conocimiento experto.

En resumen, este trabajo busca aportar una nueva contribución en la formulación, las propiedades analíticas y la práctica para los problemas de múltiples covariables, que da lugar un proceso de estimación de parámetros complejo y bandas con una unidad de medida que a menudo no representa ninguna magnitud física. El uso de técnicas para la estimación de parámetros de distribución Weibull y pesos para covariables junto con criterios no arbitrarios para bandas de múltiples covariables aún no se ha abordado en profundidad. Aquí es cuando entra en juego esta innovadora propuesta de evaluación del estado CBM-DS, que espera ser de utilidad para los gestores de activos.

Objetivos

Objetivo General

Resolver los desafíos de órdenes de magnitud asociados al uso de múltiples covariables en la ponderación de parámetros y estimación de la matriz de probabilidad de transición del modelo Weibull PHM utilizando algoritmos Data-Driven con el fin de robustecer la estimación del impacto de las condiciones en la vida útil de los equipos.

Objetivos Específicos

Evaluar la integración de un escalamiento de datos para el manejo de covariables con diferentes órdenes de magnitud en base a su impacto en la estimación de parámetros.

Evaluar estrategias Data-driven de acotamiento y valores semilla con respecto a los resultados alcanzados en MLE para la estimación de parámetros Weibull PHM con múltiples covariables.

Determinar múltiples bandas de covariables para la matriz de probabilidades de transición, utilizando los parámetros estimados, mediante clustering tomando en consideración diferentes índices de validación.

Determinar el impacto de las condiciones en la vida útil de los equipos, a partir de la obtención de las confiabilidades condicionales y RUL de los estados estimados usando técnicas DD.

Revisión de Literatura

Gestión de Activos Físicos

Recientemente, la gestión de activos físicos, o PAM (siglas en inglés) se ha convertido en un elemento clave para las industrias del tipo capital intensivo. La minería, aeronáutica o defensa, por nombrar algunas, han volteado su atención a métodos más eficientes y seguros para realizar sus operaciones y han apuntado a mecanismos para integrarse con las decisiones de mantenimiento (Godoy et al., 2014; Nehring et al., 2018; Safaei & Jardine, 2018). Además, la naturaleza compleja de los sistemas de ingeniería modernos ha potenciado el desarrollo de la PAM como una disciplina (Hastings, 2015; Syed & Lawryshyn, 2020). Esto requiere de esfuerzos coordinados y sistemáticos en procesos estratégicos, operacionales, técnicas y de mantenimiento para lograr entender el valor de un activo, o un conjunto de ellos, de una forma sustentable a través de su ciclo de vida, considerando riesgos (Hastings, 2015).

El desarrollo de la PAM se ha destacado por la necesidad de un enfoque integral, que optimice el valor de los equipos a lo largo de su ciclo de vida, utilizando e implementando explícitamente un sistema de gestión de activos (AMS). Es decir, reconocer valor para el negocio, reducir el riesgo y reducir el costo total de vida de los activos (Maletič et al., 2020). De hecho, el uso de AMS ayuda a las organizaciones a mejorar sus procesos de toma de decisiones, además de la gestión de riesgo y sus resultados financieros (Alsyouf et al., 2021).

Dado el reciente y transversal desarrollo tecnológico de las industrias, herramientas avanzadas pueden adoptarse para mejorar la gestión de activos y procesos, como el mantenimiento predictivo y técnicas Data-driven (DD) (Crespo et al., 2020; Galar & Kans, 2017). El Mantenimiento Basado en Condiciones (CBM) es una técnica de mantenimiento predictivo que permite tomar decisiones de mantenimiento basadas en información recopilada mediante el monitoreo de condiciones del activo, como temperatura o presión, para prever fallas, mejorando así la gestión de la salud de los activos y reduciendo el costo del ciclo de vida (Godoy et al., 2023; A. K. Jardine et al., 2006). El CBM es un factor clave para el PAM y es fundamental para el desarrollo de un ASM, según Maletič et al. (2020).

El mantenimiento basado en condiciones (CBM) permite tomar decisiones de mantenimiento basadas en información recolectada por medio del monitoreo de las condiciones de los activos (Godoy et al., 2014; A. K. Jardine et al., 2006). Esta técnica de mantenimiento preventivo usa la información proveniente del monitoreo de múltiples parámetros (covariables) como la vibración, nivel de contaminación o la temperatura para predecir la ocurrencia de fallas, mejorando la gestión de la salud de los activos y reduciendo el costo de sus ciclos de vida (Ahmad & Kamaruddin, 2012). La investigación de Maletič y col. (2020) propone que el monitoreo de condiciones es uno de los factores considerados como prácticas básicas de la PAM, además de ser importante para el desarrollo de un AMS con un enfoque integral. Jardine y Tsang (2005) sostienen que el CBM se origina del monitoreo de condiciones y que el desarrollo e integración de los AMS plantea un desafío para la PAM.

Modelo de riesgos proporcionales

El modelo de riesgos proporcionales (PHM) es un procedimiento estadístico para estimar la tasa de falla de un componente de un equipo basado en la información de monitoreo de condición (Cox, 1972; A. K. Jardine & Tsang, 2005). Para optimizar las decisiones de CBM y gestionar efectivamente el riesgo de falla de un activo bajo monitoreo de condición, se comenzaron a realizar estudios utilizando PHM como una solución satisfactoria (A. K. Jardine & Tsang, 2005). La función de riesgo de PHM se ha utilizado ampliamente como una distribución Weibull debido a su flexibilidad y su función de riesgo y confiabilidad de forma cerrada (A. Jardine et al., 1987; H. Liu & Makis, 1996; Zheng et al., 2022; Zheng et al., 2023). Banjevic y Jardine (2006) presentan un método para el cálculo de la confiabilidad y la vida útil remanente (RUL) en función de las condiciones actuales. La función de confiabilidad se asume bajo la premisa de un modelo de proceso de Markov y su solución es obtenida como una aproximación numérica al sistema de ecuaciones diferenciales de Kolmogorov. En esta misma investigación, Banjevic y Jardine aplican este método a una sola covariable, el nivel de hierro presente en el aceite de sistemas de transmisión de transportes mineros. Luego, los datos son discretizados en diferentes estados, de modo que se calculan varias funciones de

confiabilidad, representativas de cada estado. Estas funciones de confiabilidad se representan gráficamente, mostrando las diferencias en sus tasas de mortalidad. Finalmente, A. K. Jardine y Tsang (2005) desarrollaron un software destinado a optimizar CBM para el proceso de toma de decisiones de intervención de activos; El software comprende múltiples covariables, MLE para la estimación de parámetros, matrices de probabilidad de transición, PHM y los costos asociados con las intervenciones correctivas y preventivas. Específicamente, toma señales de monitoreo procesadas y las correlaciona con fallas pasadas y eventos de fallas potenciales. Luego, mediante el uso de modelos, proporciona estimaciones del riesgo de falla y de la vida útil restante en sintonía con las consideraciones económicas y los requisitos de disponibilidad de ese activo en su contexto operativo actual (A. K. Jardine & Tsang, 2005).

Estos últimos estudios sirven de base para el tema central de esta investigación, los desafíos asociados a la uso de múltiples covariables en el modelo Weibull PHM, considerándolos en la tasa de riesgo y confiabilidad. Existen muchas aproximaciones para este asunto en particular, algunas de las más relevantes se discutirán a continuación. Cox (1972) muestra una forma de obtener coeficientes/pesos de covariables con estimación de máxima verosimilitud (MLE); los parámetros de distribución no se estudian, pero Cox señala la posibilidad de asumir una distribución de dos parámetros, como Weibull o exponencial, para modelar la tasa de riesgo base y agregar estos parámetros dentro de MLE.

Las dos técnicas más ampliamente utilizadas para el análisis de riesgo de falla/supervivencia son el Modelo de Riesgos Proporcionales (PHM) de Cox y el modelo de tiempo de falla acelerado (AFTM) (Newby, 1988), debido a su flexibilidad y eficiencia. Una de las principales diferencias es que el PHM de Cox es un método semi-paramétrico en el que no se imponen suposiciones sobre la función de riesgo base (Wilson et al., 2021). H. Liu y Makis (1996) proponen un método que utiliza la Estimación de Máxima Verosimilitud (MLE) para obtener parámetros de distribución Weibull (escala y forma) y pesos/importancias de las covariables. Específicamente, consideraron el PHM como un AFTM para poder estimar los parámetros de forma, escala y pesos de covariables con una sola función de verosimilitud. El principal

problema es la naturaleza no lineal de estas funciones MLE, por lo que las soluciones analíticas no son simples de obtener (M. Xu et al., 2017). Vlok et al. (2002) muestran una forma no analítica de resolverlo, utilizando una tasa de riesgo base definida, específicamente, la tasa de riesgo Weibull, por lo que el modelo fue bautizado como Modelo Weibull PHM o modelo de regresión Weibull. Ellos maximizaron la función de verosimilitud asociada utilizando el método cuasi-Newton Broyden–Fletcher–Goldfarb–Shanno (BFGS) (Press et al., 1994), de esta forma se estimaron los parámetros de Weibull y 6 pesos de covariables. Una limitación del método propuesto es que BFGS es sensible a las condiciones iniciales. Si la suposición inicial está lejos del mínimo global, BFGS puede no converger o converger a un mínimo local. Una segunda limitación es que BFGS no garantiza la convergencia global, por lo que para funciones objetivo no convexas como la función MLE de Weibull PHM, BFGS puede fallar (Lawless, 2011; Mascarenhas, 2004). Además, la función MLE de Weibull PHM especialmente necesita valores iniciales razonables, según Smith (1991). Por lo tanto, los límites y la estrategia de valores iniciales son asuntos importantes a considerar para la resolución del problema.

Como se mencionó anteriormente, Banjevic y Jardine (2006) presentan un método para el cálculo de la confiabilidad y la vida útil remanente (RUL) en función de las condiciones actuales considerando una sola covariable, el nivel de hierro presente en el aceite de sistemas de transmisión de transportes mineros. Los datos son discretizados en diferentes estados, de modo que se calculan varias funciones de confiabilidad, representativas de cada estado. Dado que es una covariable, las bandas son determinadas por criterio experto. A. K. Jardine y Tsang (2005) se basan en la investigación anterior para realizar el cálculo de confiabilidad y RUL con múltiples covariables, dejando la estimación de bandas a criterio experto. Zheng y col. (2020) proponen una política de CBM con umbrales dinámicos y múltiples acciones de mantenimiento para un sistema sujeto a inspección periódica; Los umbrales son usados para decidir cuándo realizar una acción de mantenimiento en particular. Uno de los aspectos más destacables de esta investigación es la incorporación del PHM con un proceso de covariables

de estado continuo para describir la tasa de riesgo. Los límites dinámicos se determinan utilizando un marco de proceso de decisión semi-Markov. Sin embargo, estos umbrales son definidos en base a bandas de covariables pre-definidas para discretizar los estados de la cadena de Markov, con un rango igual entre ellos.

Data Driven PHM

Recientemente, algunos estudios han intentado incorporar enfoques no paramétricos para obtener parámetros de Weibull PHM. Los métodos Data-driven (DD) utilizan reglas meta-heurísticas para establecer relaciones entre las características internas de los datos disponibles y sus insights (W. Liu et al., 2023). Los modelos DD han sido ampliamente utilizados en diversos campos con resultados exitosos (Ferraz Júnior et al., 2023; Godoy et al., 2023; Hesabi et al., 2022; W. Liu et al., 2023; Mikhail et al., 2024; M. Xu et al., 2017).

Uno de los aspectos relevantes que esta investigación apunta a desarrollar es la integración de Machine Learning (ML) dentro de los modelos destinados a establecer políticas de intervención que lleven la gestión de activos a la Industria 4.0. El uso de ML y gestión de datos es uno de sus principales pilares (Mofolasayo et al., 2022).

De forma genérica, ML puede ser entendido como el uso de algoritmos que aprenden de experiencias como los seres humanos (Shinde & Shah, 2018). En el caso de estos algoritmos, la experiencia está directamente relacionada a los datos, esto es, mostrándoles, o no, qué buscar (aprendizaje supervisado y no supervisado, respectivamente), o castigándolos y recompensándolos durante el aprendizaje (aprendizaje por refuerzo) (Rajendra et al., 2022). El aprendizaje supervisado relaciona variables de entrada o explicativas con variables de salida, dependientes o objetivo conocidas. El aprendizaje no supervisado no tiene resultados predefinidos y, por lo tanto, intenta inferir patrones naturales a partir de un conjunto de datos (El Ouadi et al., 2022). Finalmente, el aprendizaje por refuerzo implica el aprendizaje de algoritmos a través de recompensas y castigos para guiarlos mientras aprenden, en un proceso de prueba y error (Rajendra et al., 2022).

El trabajo de Rao et al. (2019) aplica técnicas de Gradient Boosting (aprendizaje

supervisado) para evaluar la calidad de las entradas para un problema de optimización, el cual se resuelve utilizando el algoritmo de Artificial Bee Colony (ABC), un algoritmo evolutivo/data-driven para optimización cercano a la familia de aprendizaje por refuerzo, como Genetic Algorithm (GA) o Ant Colony Optimization (ACO). Así, Gradient Boosting minimiza el potencial de que ABC quede atrapado en un óptimo local identificando las características/variables más relevantes. Este estudio demuestra que la selección de características utilizando Gradient Boosting no sacrifica la precisión del modelo y supera los resultados de optimización con diferentes conjuntos de datos. Otros estudios han utilizado Random Forest para mejorar el rendimiento de algoritmos evolutivos (Paul et al., 2017; Wang et al., 2022), llegando a conclusiones similares a las de (Rao et al., 2019).

Los algoritmos evolutivos (EA) se caracterizan por tener parámetros simples, amplia aplicabilidad, generalidad, facilidad de implementación y han tenido éxito en diversas áreas (He et al., 2023). GA es el EA más comúnmente utilizado en el campo de los problemas de optimización (Katoch et al., 2021), debido a su simplicidad de implementación y su capacidad para converger hacia una solución cercana a un punto factible en un tiempo prudente, pero su precisión depende de la selección de parámetros (Cavallaro et al., 2024), como límites y valores iniciales. En Cavallaro et al. (2024), se muestra que el GA potenciado con algoritmos de Machine Learning obtiene un mejor rendimiento que simplemente usar GA, proporcionando evidencia de que utilizar técnicas de Machine Learning para ajustar los parámetros del algoritmo es un enfoque prometedor. De hecho, el Machine Learning desempeña un papel fundamental en navegar por el espacio de búsqueda, permitiendo que los algoritmos concentren sus esfuerzos computacionales en las direcciones más prometedoras para descubrir soluciones mejores.

Wilson et al. (2021) presentan un enfoque utilizando aprendizaje de Kernel múltiples (MKL), un algoritmo de aprendizaje supervisado, para modelos de predicción de supervivencia y una función conjugada convexa para la función de verosimilitud del modelo PHM de Cox. El estudio muestra un rendimiento sólido en la estimación de la función de verosimi-

litud y el MKL evita la complejidad del acotamiento. Sin embargo, el estudio no considera ninguna suposición sobre la función de riesgo base, por lo que los parámetros de distribución no se estiman en la MLE y utilizan una función convexa modificada para la MLE.

La ponderación de características/covariables es un tema importante que la MLE intenta abordar en el presente trabajo. C. Chen et al. (2020) proponen un enfoque de autoencoding de redes neuronales (NN) (aprendizaje no supervisado) para estimar las características más robustas dentro de las existentes. Luego, se utilizan el modelo PHM de Cox y redes neuronales de memoria a corto y largo plazo (LSTM) para predecir el tiempo entre fallas (TBF). De esta forma, pudieron seleccionar las características más importantes para la predicción y utilizaron el PHM de Cox para obtener una curva de confiabilidad. Se utiliza la MLE para determinar la función de riesgo base y la ponderación de covariables, pero no especificaron ningún método para resolver la MLE. Mikhail et al. (2024) utilizan aprendizaje por refuerzo (Q-learning), el algoritmo de Random Forest (RF) (aprendizaje supervisado) y un proceso de decisión de Markov (MDP) para la optimización del CBM. La matriz de probabilidad de transición no se utiliza porque el estudio se basa completamente en los algoritmos de meta-heurística para obtener curvas de confiabilidad (de niveles de deterioro) basadas en el método de confiabilidad Kaplan-Meier (no paramétrico) (Klein et al., 2003; Satten & Datta, 2001), por lo que la ponderación de covariables se realiza mediante el uso de este método. Además, no asumen ninguna distribución para el deterioro, RF se utiliza para predecir este fenómeno. MDP incorpora el espacio de estados (definido por RF), espacio de acciones (reemplazos preventivos y correctivos, no hacer nada y reparación preventiva multinivel, incluida la reparación 'as good as new'), recompensas (RUL) y castigos (costo de mantenimiento). Q-Learning se utiliza para simular un agente tomador de decisiones que decide la acción en función de las condiciones 'actuales' de los datos, los costos y el RUL esperado. Un problema preocupante de este estudio es el alto costo computacional relacionado con el método Q-learning.

En la búsqueda de mejorar la estimación de parámetros para la distribución q-Weibull,

M. Xu et al. (2017) introducen el algoritmo Adaptive Hybrid Artificial Bee Colony (AHABC). Este enfoque innovador integra sin problemas las capacidades de exploración global de ABC con las técnicas de explotación local de la búsqueda simplex de Nelder-Mead, mejorando así el rendimiento de la búsqueda local de ABC. Análisis comparativos con otros métodos de optimización, como el algoritmo genético (GA) y la optimización por enjambre de partículas (PSO), revelan que el algoritmo AHABC propuesto los supera en velocidad de convergencia y precisión. Aunque el estudio actual no profundiza en las complejidades de la ponderación de covariables, refleja los beneficios potenciales de amalgamar varios modelos estadísticos y heurísticos. Esto se alinea con uno de los objetivos del presente trabajo, que busca explorar e implementar una integración similar de varios modelos para abordar el acotamiento y la estrategia de valores iniciales para la MLE de Weibull PHM.

Entre los algoritmos de aprendizaje no supervisado, la familia de algoritmos de clustering (Sircar et al., 2021) busca agrupar un conjunto de datos según características similares compartidas, haciéndolos lo más diferentes posible de otros grupos o clusters (Sancho et al., 2022). Existen diferentes enfoques y formas de realizar clustering. En esta investigación se usó el clustering particionario (PC) y clustering basado en modelos (MC). El PC se caracteriza por trabajar con un número de clusters previamente establecidos a generar y cada observación pertenece a un único cluster (Ezugwu et al., 2022; Li et al., 2022). Los algoritmos MC primero crean un modelo primario para identificar los parámetros de la distribución estadística de cada cluster y, a partir de ellos, clasificar las observaciones en estos clusters (Ezugwu et al., 2022; Li et al., 2022).

Uno de los algoritmos PC más usados es K-means (Dhanachandra et al., 2015; Li et al., 2022). Este algoritmo, propuesto por Stuart Lloyd (1982) y Edward W. Forgy (1965), encuentra 'k' clusters minimizando la raíz cuadrada de la suma de las distancias entre cada punto y el centroide correspondiente a su cluster (Kanungo et al., 2002). El proceso consta de dos fases: primero, se identifican los centroides de cada conglomerado y, segundo, cada observación se clasifica según su distancia a cada centroide (Dhanachandra et al., 2015).

Debido a su proceso de optimización iterativo, la convergencia es rápida, además de trabajar bien con datasets no tan extensos. Estos aspectos lo convierten en un algoritmo popular en ML (L. Chen et al., 2021). K-means se ha utilizado para mejorar la zonificación urbana para la logística de carga (El Ouadi et al., 2022), la segmentación de imágenes (Dhanachandra et al., 2015) y el mapeo del riesgo de inundaciones urbanas (H. Xu et al., 2018), entre otros (Troccoli et al., 2022).

El modelo de mezcla gaussiana (GMM) es un algoritmo MC ampliamente usado en diversas áreas como la segmentación de imágenes, procesamiento de señales y ciencias biomédicas (Ezugwu et al., 2022; Xinmin et al., 2022). GMM tiene la capacidad de hacer aproximaciones leves de funciones generales de densidad de probabilidad a través de sumas ponderadas de múltiples funciones gaussianas (Huang et al., 2005). Además, GMM no sólo proporciona un ajuste de distribución generalizado uniforme, sino que sus componentes (o grupos) pueden describir claramente la densidad multimodal si es necesario (Reynolds, 2002). Por lo tanto, se puede utilizar para predecir y clasificar datos asociándolos con diferentes componentes basados en distribuciones de probabilidad (Hamdi et al., 2022). Por lo que tiene la capacidad de trabajar con variables que presentan diferentes comportamientos, periodicidad e identificar patrones complejos.

Un aspecto a considerar en K-means y GMM son sus dependencias a hiper-parámetros K que representan el número de clusters o componentes que los algoritmos deben generar. En consecuencia, el desafío radica en encontrar el valor óptimo de estos hiper-parámetros (H. Xu et al., 2018). En general, estos problemas se resuelven evaluando diferentes valores para k y comparando sus rendimientos con indicadores de validación (Fahim, 2021) como Silhouette (El Ouadi et al., 2022; Sancho et al., 2022; H. Xu et al., 2018), el método Elbow (Troccoli et al., 2022), criterio de información de Akaike (Steinley & Brusco, 2011; Xinmin et al., 2022) o el criterio de información Bayesiano (Ezugwu et al., 2022; Steinley & Brusco, 2011). También se pueden resolver utilizando métodos visuales (Fahim, 2021) como dendrogramas (Habib et al., 2022). Aunque estos métodos ayudan a determinar el número óptimo K de acuerdo

con la naturaleza de las observaciones, el investigador o dueño del proceso siempre debe tener la última palabra. El número K puede estar predefinido para desarrollar un cuadro de mando con diferentes KPIs. Por ejemplo, un indicador tipo semáforo tiene sólo tres estados por defecto. De forma similar, el objetivo del software EXAKT es ayudar en la toma de decisiones de intervención generando un gráfico con tres regiones para identificar la condición del equipo en base a PHM y CBM (A. K. Jardine & Tsang, 2005). En estos ejemplos, el número K de clusters a considerar es un valor predefinido por el juicio de expertos.

El párrafo anterior destaca un elemento relevante de este estudio. Después de calcular los centroides generados por K-means y las distribuciones de probabilidad extraídas de GMM, se espera que se determinen los estados (clusters) y sus bandas de covariables (rangos de valores de covariables que describen un determinado estado (Lam & Banjevic, 2015)). Ambos son necesarios para la creación de la matriz de probabilidad de transición que, a su vez, se requiere para el desarrollo del PHM. Esto dado que también se busca encontrar un número de bandas óptimo, estos dos algoritmos conversan muy bien con la necesidad a resolver. Desarrollan clusters a partir de un número predefinido de los mismos.

Con el objeto de mejorar las estrategias de CBM relacionadas a determinar probabilidades de transición, algunas investigaciones han usado diferentes técnicas de ML. Por ejemplo, Yang y col. (2022) proponen una estrategia de CBM considerando políticas de mantenimiento dinámicas. El modelo de optimización involucra un proceso de decisión de Markov mejorado con redes neuronales y aprendizaje-Q (aprendizaje por refuerzo), también conocido como aprendizaje Deep-Q (DQL). Este enfoque es adaptable a sistemas multicomponentes y puede aprender de forma efectiva la política de mantenimiento óptima de sistemas redundantes. Sin embargo, los umbrales de cada estado se definen arbitrariamente.

Nguyen et al. (2022) desarrollaron un enfoque de CBM basado en inteligencia artificial (IA), que primero construye un predictor basado en una red neuronal artificial para estimar el costo de mantenimiento del sistema y luego emplea un algoritmo DQL multiagente personalizado para optimizar las decisiones de mantenimiento. Aunque el proceso de estado

no se analiza en el artículo, en este se asume que el conjunto de datos está completo en términos de contener todos los estados de transición del sistema.

Azar et al. (2022) desarrollaron un marco semisupervisado basado en clustering para la toma de decisiones de mantenimiento basado en el modelo PHM de Cox. Para ello, utilizaron K-means para estimar los estados del sistema basados en la edad y los valores entregados del monitoreo de condiciones, el espacio de los estados con las probabilidad de transición asociadas se definen en base a la clusterizacion. Esta investigación contribuye desarrollando una metodología innovadora para estimar el estado del sistema sin juicio experto. Sin embargo, el parámetro K se establece aleatoriamente, además de que el algoritmo K-means presenta problemas de robustez frente a valores atípicos y casos extremos.

Metodología

Para una mejor explicación de la metodología propuesta, se creó un diagrama de flujo del proceso analítico, que se puede ver en la Figura 1, a continuación.

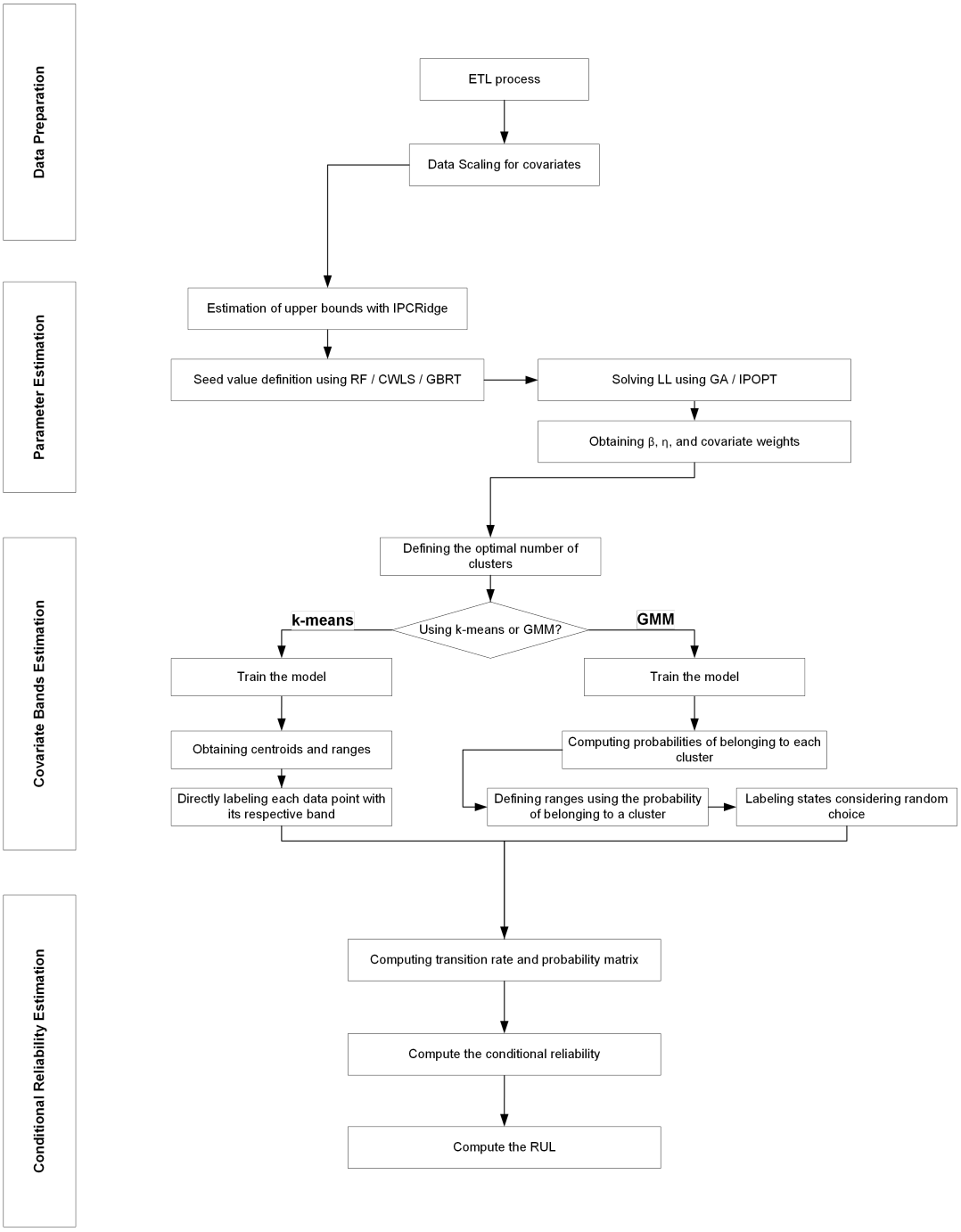

Figura 1

Diagrama de flujo de la metodología propuesta.

Preparación de Datos

En primer lugar, la preparación de datos es crucial para garantizar la integridad de la información antes de su utilización. Esto implica realizar comprobaciones exhaustivas para eliminar datos nulos (null), faltantes (Nan) o duplicados. Además, se lleva a cabo una integración de varias tablas mediante el proceso ETL (Extract, Transform, and Load) para consolidar toda la información disponible en un solo Dataset.

Simultáneamente, un aspecto clave implica la identificación y clasificación de los tipos de mantenimiento mediante un sistema binario. Este sistema binario asigna valores de 1 y 0 a los mantenimientos preventivos y correctivos, respectivamente. Para el tratamiento de valores NaN o null, hay varios caminos que se pueden tomar, algunos de estos son los siguientes:

1. Completar los valores NaN con el valor siguiente o anterior si los datos siguen una cierta tendencia.
2. Completar los valores NaN con el promedio de los datos.
3. Completar los valores NaN por medio de interpolaciones y/o regresiones lineales o polinómicas.
4. Eliminar el registro si representa un porcentaje ínfimo del total de datos.

La tabla resultante encapsula solo los datos pertinentes esenciales para el desarrollo del modelo. Esto incluye un identificador para los activos y sus componentes (si corresponde), el tipo de intervención representado en el sistema binario, los intervalos de tiempo entre intervenciones y columnas individuales para cada covariable.

Escalamiento de datos

Antes de trabajar en los datos de covariables, se recomienda transformar o estandarizar los datos. Una alternativa es escalar estos datos de manera que tengan el mismo orden de magnitud entre ellos, así como los mismos valores mínimo y máximo. Esto facilita el análisis de los pesos de cada covariable, permitiendo establecer su importancia relativa entre sí. En

este estudio, se analizan dos técnicas de transformación. El escalamiento Min-Max y una transformación particular que representa una alternativa de no escalamiento (NS). Ambas se mostrarán a continuación.

El escalamiento Min-Max es un método ampliamente utilizado en Ciencia de Datos para la estandarización que escala los datos entre dos valores, su fórmula se muestra a continuación.

$$x_{Scaled}(min, max) = \frac{(x - X_{min})}{(X_{max} - X_{min})} \cdot (max - min) + min \tag{1}$$

Así, el valor (x) se escala utilizando el valor mínimo (X_{min}) y el valor máximo (X_{max}) de los datos de la covariable, luego se pondera considerando un nuevo valor mínimo (min) y máximo (max).

La transformación NS se considera en este estudio únicamente con fines comparativos, por lo que la alternativa de no escalamiento puede representarse mediante este método. Su fórmula se muestra a continuación.

$$X_{Scaled} = \frac{(X + X_{min})}{X^*_{max}} \tag{2}$$

Donde X^*_{max} es el valor máximo de la columna de datos transformados $X + X_{min}$. Por otro lado, X_{min} es el valor mínimo de la columna de datos de la covariable X.

Estimación de parámetros

En este estudio, tanto los parámetros de la distribución Weibull como los pesos de las covariables son estimadas a partir de MLE, a continuación se detallan las estrategias para la estimación de valores semilla y acotamiento para la obtención de estos parámetros.

Parámetros de Distribución Weibull

Se pretende que la obtención de los valores iniciales (semilla) de estos parámetros se lleve a cabo linealizando el modelo de confiabilidad propuesto por Jardine (1987) A. Jardine et al., 1987, quien indicó que la confiabilidad puede modelarse como una distribución Weibull

de dos parámetros a través de:

$$R(t) = e^{-(t/\eta)^{\beta}} \tag{3}$$

Al linealizar la expresión mencionada anteriormente, se obtiene la siguiente ecuación:

$$ln(-ln(R(t)) = \beta \cdot ln(t) - \beta \cdot ln(\eta) \tag{4}$$

Donde β es el parámetro de forma (pendiente) y η es el parámetro de escala. La confiabilidad de los datos se obtiene mediante el método de Lewis:

$$R(t) = \left[\frac{n+1-i}{n+2-i}\right]^{(1-\delta)} \cdot R(t_{i-1}) \tag{5}$$

Donde δ es una variable binaria que toma el valor 1 cuando ocurre una censura (por ejemplo, intervenciones preventivas) y 0 cuando ocurre una falla.

$$\lambda = \frac{\beta}{e^{ln(\eta)\beta}} \tag{6}$$

Finalmente, los valores iniciales para los parámetros de la distribución Weibull se obtienen ajustando un polinomio de grado 1 utilizando la ecuación 4 y el método de mínimos cuadrados. El parámetro λ se calcula utilizando η y β, como se puede ver en la ecuación 6. Desde la MLE se obtienen λ y β factibles, luego se calcula η. Para el acotamiento de los parámetros de Weibull, se utiliza el conocimiento experto.

Pesos de covariables

Gradient Boosting. En varios problemas, combinar las predicciones de un conjunto de modelos a menudo resulta en un modelo con un rendimiento mejorado (Ridgeway, 1999). Gradient Boosting (GB) construye modelos de regresión aditivos utilizando funciones parametrizadas simples (base learners) de manera secuencial (Friedman, 2001, 2002). En este estudio se consideran dos base learners para los métodos de Gradient Boosting, árbol de regresión (GBRT) y mínimos cuadrados componente a componente (CWLS). Ambos métodos

se basan en los estudios de Ridgeway (1999) y Hothorn et al. (2006), respectivamente. Para este estudio, se utiliza la biblioteca de Python *Scikit-survival* para la implementación del algoritmo de ambas técnicas de Gradient Boosting.

Específicamente, GBRT y CWLS se utilizan como estrategia de estimación inicial de pesos de covariables, utilizando el concepto de clasificador por votación (Breiman, 1996) pero en el contexto de Gradient Boosting. A medida que los algoritmos aprenden y avanzan a través de las iteraciones (un nuevo base learner aprende de los anteriores), el base learner actual ha considerado alguna covariable como la mejor para explicar la variable dependiente (TBF). Por lo tanto, se podrían adoptar dos estrategias para la estimación de pesos iniciales. La primera es considerar como pesos el número de veces que un base learner elige alguna covariable como la mejor, esta estrategia se utilizó para RF. La segunda estrategia es capturar el índice (el número de iteraciones) en el cual se elige a cada covariable por primera vez, considerando que un base learner de las primeras iteraciones tiene necesariamente un rendimiento predictivo menor que uno relacionado con iteraciones posteriores, debido al como funciona la técnica de Gradient Boosting. Por lo cual, el rendimiento predictivo se ve reflejado en la magnitud del índice, es decir, otorgándole un proto-peso a cada covariable en base a su rendimiento predictivo. Esta última estrategia se utilizó para CWLS y GBRT.

Después de determinar los proto-pesos para cada covariable, se almacenan en un vector que se escala con el mismo escalador con el que se estandarizaron las covariables. Finalmente, este vector escalado ya contiene los pesos iniciales que se utilizarán en los algoritmos de optimización.

Random Forest. RF se utiliza como una tercera alternativa para obtener pesos iniciales de covariables. Para este algoritmo, se realiza un método más similar al clasificador por votación, cada árbol en el RF puede elegir una covariable como la mejor candidata. Luego, se realiza una votación, cada covariable se asocia a un proto-peso que representa el número de árboles que la eligieron como la mejor de las covariables para explicar la variable dependiente (TBF). Finalmente, el vector de proto-pesos se escala con el escalador que

estandarizó las covariables para obtener el vector de pesos iniciales. Para la implementación de la regresión de RF, se utiliza la biblioteca de Python *Scikit-learn*.

Para ambos algoritmos, GB y RF, se utiliza el índice de impureza Gini (Biró & Néda, 2020) para elegir al mejor candidato de covariable en cada base learner.

IPCRidge. El modelo de tiempo de falla acelerado con ponderación de probabilidad inversa de censura (IPCRidge), es un modelo propuesto por Stute (1993) que asume un modelo de regresión de la forma:

$$\log y = \beta_0^{ipc} + X\beta^{ipc} + \epsilon \tag{7}$$

Cada muestra de los datos se pondera por la inversa de la probabilidad de censura considerando censura a derecha (bajo el supuesto de que la censura es independiente de las covariables, es decir, censura aleatoria). ϵ es el término de error aleatorio del modelo de regresión. El término de penalización α^{ipc} asociado con la contracción L2 y aplicado a los coeficientes β^{ipc} debe ser un valor pequeño y positivo (entre 0 y 1); de esta manera, mejora la condición del problema y reduce la varianza de las estimaciones. Además, no se recomienda usar $\alpha^{ipc} = 0$ por razones numéricas. La biblioteca de Python *Scikit-survival* se utiliza para la implementación. Para esta investigación, se fuerza a que los coeficientes sean positivos.

Si el vector de pesos se transforma tomando la inversa (es decir, $1/\beta^{ipc}$), entonces este vector contiene los estimadores de Kaplan-Meier Satten y Datta, 2001 asociados con cada covariable. El valor máximo de los estimadores se utiliza como cota superior para los pesos de las covariables en los algoritmos de optimización. Por lo tanto, en el peor de los casos, el algoritmo podría considerar como máximo estas cotas como pesos factibles para las covariables.

Estrategias de optimización

Las estrategias de optimización implican la minimización del negativo del logaritmo parcial de verosimilitud, para este estudio se utiliza el modelo de logaritmo parcial de verosimilitud propuesto por H. Liu y Makis (1996), pero considerando todas las muestras como

fallas. La ecuación se presenta a continuación.

$$LL = \sum_{i=1}^{n}(ln(\lambda t_i^{(\beta-1)}) + \sum_{j=1}^{m}\gamma_j ln(z_{ij}) - \frac{\lambda t_i^{\beta}}{\beta}\prod_{j=1}^{m} z_{ij}^{\gamma_j}) \tag{8}$$

Donde n es el número de muestras, m es el número de covariables y γ_j es el peso asociado con la j-ésima covariable.

Una vez estimados los valores iniciales y las cotas, se utilizan como entrada para los algoritmos de optimización. Se consideran dos métodos de optimización, ambos se mostrarán a continuación.

Genetic Algorithm. Los algoritmos genéticos (GA) establecen una técnica de búsqueda combinando la idea de la supervivencia del más apto y una población intercruzada. Los GA operan representando soluciones en forma de cromosomas, sometiéndolos a fases de selección, apareamiento y mutación para generar nuevas soluciones que posteriormente se evalúan utilizando una función de aptitud (Holland, 1992; Peiravi et al., 2020). La función de aptitud en este estudio es la ecuación (8), las cotas y el valor inicial se obtienen como se explica en la sección anterior.

Los cromosomas son cadenas que representan una solución al problema de optimización. Esta investigación define un cromosoma como una matriz $(2 + m) \times 1$, donde m es el número de covariables. Así, las dos primeras filas contienen los parámetros de Weibull (λ y β) y las demás contienen los pesos γ de las covariables. Para obtener soluciones viables, se implementa una función de penalización si la solución no está dentro de las cotas pre-definidas (IPCRidge). Además, la iteración inicial se crea utilizando los valores iniciales obtenidos de las estrategias mencionadas anteriormente (RF, GBRT, CWLS).

Cuando se crea una población, se seleccionan los mejores cromosomas de la población para la siguiente población según los valores de aptitud. La función (8) intenta minimizarse, por lo que los cromosomas seleccionados (con el método del torneo) son aquellos con los valores de aptitud más bajos.

Para la recombinación de cromosomas, se utilizan apareamiento y mutación. El apa-

reamiento, también conocido como crossover, es el proceso en el que se mezclan los genes de dos padres seleccionados para generar nuevas descendencias, este proceso se realizará con una probabilidad preestablecida. La mutación, por otro lado, se aplica a los cromosomas descendientes generados por cruces para evitar la convergencia prematura a una solución local. Además, cada padre se selecciona aleatoriamente a una tasa de mutación predeterminada. Además, la mutación de cada valor dentro de la descendencia cambia aplicando una mutación aditiva gaussiana. Para la implementación de GA, se utiliza la biblioteca de Python *Deap*.

Interior Point OPTimizer. El optimizador de puntos interiores (IPOPT) implementa un método de filtro de búsqueda de línea de puntos interiores que tiene como objetivo encontrar una solución local de problemas no lineales (NLP) (Wächter & Biegler, 2006). Este algoritmo considera límites para las funciones de restricción y para las variables. La función objetivo puede ser lineal o no lineal y convexa o no convexa (pero debe ser dos veces continuamente diferenciable).

Dado que el enfoque es más parecido a la programación de un modelo de optimización, requiere menos toma de decisiones de hiperparámetros (tuning). De hecho, este enfoque solo necesita una función objetivo, variables, cotas y un valor inicial para comenzar. En este estudio, la función objetivo es la ecuación (8), las variables son los parámetros de Weibull y los pesos de las covariables. Las cotas y los valores iniciales se estiman como se menciona en las secciones anteriores (RF/GBRT/CWLS + IPCRidge). Para la implementación de IPOPT, se utiliza la biblioteca de Python *Pyomo*.

Estimación de bandas de estado

Una vez que se obtienen los pesos, se introduce un nuevo parámetro en el dataset que se utilizará para la clusterización, el cual corresponde a la suma de $\gamma \cdot Z(t)$ de cada covariable, es decir,

$$f(\gamma, z) = \sum_{i \in Z} \gamma_i Z_i(t) \tag{9}$$

Clustering

Para utilizar los algoritmos de clustering, se recomienda la simplificación de datos cuando los datasets son muy grandes, este caso implica que los datos $f(\gamma, z)$ sean subdivididas en intervalos de clases. A los intervalos se les asigna su marca de clase correspondiente, lo que reduce la cantidad de datos con la que tiene que trabajar el algoritmo. Para determinar el número de intervalos en los que se divide la base de datos, se puede emplear la 'Regla de Sturges', que se basa en el número de muestras para definir los intervalos.

Posteriormente, este trabajo propone dos algoritmos de clustering diferentes, K-means y GMM. El principal hiperparámetro que ambos métodos requieren es el número de clusters en los que se espera que se dividan los datos. Para determinar este número, se pueden iterar diferentes números y comparar los resultados para seleccionar uno de ellos por medio de diversos indicadores como el índice Silhouette o criterios de información como Akaike (AIC) o Bayesiano (BIC), que permiten comparar los resultados para decidir qué número 'K' de clusters se adapta mejor a la naturaleza de los datos. A su vez, se debe tener en cuenta que siempre se recomienda la selección de modelos más simples, es decir, si la diferencia en los valores del indicador entre dos números de clusters es baja, se prefiere el valor 'K' menor. También se debe considerar lo que representarán estos clusters en el análisis a realizar, la información del negocio que pueda aportar el tomador de decisión puede proporcionar algunas pistas sobre el número más realista de clusters para el caso de negocio a resolver. No se debe pasar por alto el número de intervalos de clase en los que se dividieron los datos al principio, ya que es probable que si se dividen en el mismo número de clusters, los resultados de dicha división puedan estar sesgados, generando estructuras artificiales y no necesariamente ser la mejor opción.

Finalmente, una vez seleccionado el número de clusters y dividido los registros a través de los algoritmos mencionados anteriormente mediante sus parámetros correspondientes, el último paso es clasificar/marcar cada registro en el cluster al que pertenece para facilitar la observación del resultado del algoritmo.

Cálculo de rangos de covariables

A continuación se detallan las diversas alternativas a estudiar para el cálculo de estos rangos.

1. ***Usando la distancia entre centroides:*** Basándose en el clustering por GMM y K-means, se calculan los centroides de cada cluster; luego, se calcula el promedio entre centroides sucesivos (es decir, el promedio entre 1 y 2, 2 y 3, etc.) para definir estos valores como rangos límite. De esta manera, el valor del rango indica en qué cluster se encontrará cada $f(\gamma, z)$.

2. ***Por medio de probabilidades:*** Utilizando información proveniente del modelo GMM y los clusteres encontrados en el mismo, se calcula la probabilidad de que cada observación pertenezca a ese cluster para clasificarlos. Luego, se identifican las probabilidades mínimas de pertenencia al modelo para todas las observaciones, y estos delimitan los rangos.

Metodología de clasificación para situaciones de borde: En casos en los que el dato esté cerca de uno de los límites de las cotas y no sea fácil dilucidar a qué cluster pertenece, se propone una solución utilizando las probabilidades proporcionadas por GMM. Si la diferencia entre probabilidades es menor al 5 %, se aplica una elección aleatoria utilizando las probabilidades de pertenencia a cada cluster, es decir, qué tan probable es que un cierto valor pertenezca a un cluster u otro.

Matriz de Probabilidad de Transición

Una vez calculados los rangos, el siguiente paso es generar una matriz de probabilidades de transición. Para ello, se debe identificar el estado al que pertenece cada registro del dataset. Luego, ordenando estas observaciones cronológicamente, se identifican las transiciones de estados, específicamente cuántas veces un estado i ha transitado a un estado j, para todas las combinaciones de estados posibles. Este valor se identificará con el parámetro n_{ij}. Posteriormente, el siguiente parámetro a calcular es A_i, que representa el tiempo que el

equipo permanece en el estado i. Con estos dos parámetros, se calculan las tasas de transición según la siguiente expresión:

$$\lambda_{ij} = \frac{n_{ij}}{A_i}, i \neq j \tag{10}$$

$$\lambda_{ii} = -\sum_{i \neq j} \lambda_{ij} \tag{11}$$

Finalmente, las probabilidades de transición se calculan de la siguiente manera:

$$\pi_{ij} = e^{\lambda_{ij}} \tag{12}$$

Cálculo de la confiabilidad

En este paso, primero se establece la cantidad de iteraciones a realizar, para lo cual se requiere un punto temporal de inicio y fin, y un valor para los intervalos entre el valor inicial y final de tiempo, ya que cuanto menor sea el valor, más preciso será el resultado. Luego, la cantidad de iteraciones k realizadas es:

$$k = (t_{\text{final}} - t_{\text{initial}})/\Delta = x/\Delta \tag{13}$$

Posteriormente, el valor del exponente de cada covariable x_j debería obtenerse mediante la ecuación (14):

$$e^{(x_j)} = e^{\left((\Delta/\eta)^{\beta}\ e^{(\gamma x_j)}\right)\left(k^{\beta}-(k+1)^{\beta}\right)} \tag{14}$$

Donde γ son los pesos de cada covariable y Δ es la longitud del intervalo de aproximación. Los valores obtenidos mediante la ecuación anterior se expresan en una matriz diagonal y luego se multiplican por la matriz de probabilidad de transición, que también debe expresarse en una matriz diagonal, obteniendo así la matriz $\mathbf{L}[i]$.

El siguiente paso es obtener el producto entre $\mathbf{L}[i]$ de la iteración anterior y la iteración actual. La confiabilidad condicional se obtiene mediante la suma de cada fila de la última matriz. Esto se realiza utilizando el método de propiedad de producto explicado en detalle

en el trabajo de Banjevic y Jardine (2006), para el cual primero se estima la matriz de tasa de falla mediante la ecuación:

$$\lambda(t, Z(t)) = (\beta/\eta)(t/\eta)^{\beta-1} e^{\sum_i \gamma_i * Z_i(t)} \tag{15}$$

Utilizando lo anterior junto con la matriz de probabilidad de transición, finalmente se resuelve la matriz $\mathbf{L}[i]$ y $\mathbf{L}[x, t]$.

Resultados

El caso de estudio de esta investigación intenta poner en práctica la metodología propuesta utilizando datos reales de una empresa de distribución eléctrica en Chile. El Dataset utilizado requirió pre-procesamiento como el que se detalla en la sección Metodología. Se muestra un extracto del Dataset pre-procesados a continuación.

Cuadro 1

Extracto de los datos del caso de estudio.

TBF [hora]	**Tipo de Intervención**	**ID Equipo**	**Demanda Eléctrica [MVA]**	**Temperatura Interna [°C]**	**C2H4 (%)**	**Rigidez Dieléctrica [kV]**
162770	0	1	18,1	64	2,1392	64,1948
163210	0	1	16,4	54	2,1082	64,0284
186900	0	1	16,6	59	3,3276	63,7482
156630	0	2	8,9	53	1,6896	66,2913
157060	0	2	12,8	50	1,7452	65,9824

La Tabla 1 muestra las diferentes columnas disponibles en el conjunto de datos. La primera columna es el tiempo entre fallas (TBF) de cada muestra en horas, la segunda indica el tipo de intervención (1 para preventiva y 0 para correctiva), la columna ÏD Equipo"trae el identificador del transformador eléctrico, y las últimas cuatro columnas son las covariables consideradas en este caso de estudio. La demanda eléctrica satisfecha por el transformador se mide en megaVolt-Ampere [MVA], la temperatura interna del transformador se registra en grados Celsius [°C], el etileno [C2H4] en el aceite del transformador se mide como porcentaje de presencia de gas. Finalmente, la rigidez dieléctrica del aceite del transformador se mide en kiloVolt [kV]. Además, la Tabla 2 muestra algunos indicadores estadísticos relevantes para los datos de las covariables, como la media, desviación estándar y cuartiles. La selección de estas cuatro covariables se basó en la disponibilidad de sus datos. Sin embargo, es importante tener en cuenta que el modelo tiene la capacidad de considerar un mayor número de covariables, con

la posibilidad de rechazar algunas durante el paso posterior de ponderación de covariables.

Cuadro 2

Análisis de datos de las covariables.

	Demanda Eléctrica [MVA]	**Temperatura Interna [°C]**	**C2H4 (%)**	**Rigidez Dieléctrica [kV]**
count	60,0000	60,0000	60,0000	60,0000
mean	11,6550	53,5833	5,1092	68,8628
std	5,5781	14,0728	7,8651	3,6886
min	0,0000	29,0000	0,0000	61,1533
25 %	8,9000	42,7500	1,7313	65,9329
50 %	12,4000	52,5000	3,0694	69,6306
75 %	16,0750	64,0000	5,4385	71,7698
max	20,4000	96,0000	45,7537	75,5950

Estimación de parámetros

Tal y como se detalla en la página 18, los parámetros de Weibull se estimaron para ser utilizados como valores iniciales para los métodos de optimización, estos valores se presentan a continuación.

Cuadro 3

Valores iniciales para los parámetros Weibull.

Weibull parameter	**Initial/Starting Value**
β	1,6013
η	209019,3596

Los próximos pasos requieren que se haya realizado el escalamiento de datos, las covariables tienen diferentes órdenes de magnitud (ver Tabla 2) por lo que, por las razones ya explicadas, es importante estandarizarlas. Por lo tanto, para los próximos pasos, se van a considerar los tipos de escalamiento como caminos separados en paralelo con el fin de

contrastar resultados para MLE. Para las 4 covariables consideradas en este estudio de caso, se presentan las tres alternativas de estimación de valores iniciales, como se muestra en la página 19. La elección de hiperparámetros de estos algoritmos se realizó utilizando *GridSearchCV* de la librería de Python *Scikit-learn*. Los proto-pesos estimados se presentan a continuación.

Cuadro 4

Proto-pesos obtenidos para covariables.

Escalador	**Método de Ponderación**	**Demanda Eléctrica (p1)**	**Temperatura Interna (p2)**	**C2H4 (p3)**	**Rigidez Dieléctrica (p4)**
NS	CWLS	9	65	1	578
MinMax(0,1)	CWLS	15	8	1	64
MinMax(0,2)	CWLS	15	8	1	64
NS	GBRT	19	49	4	1
MinMax(0,1)	GBRT	19	49	4	1
MinMax(0,2)	GBRT	19	49	4	1
NS	RF	181	189	157	173
MinMax(0,1)	RF	181	189	157	173
MinMax(0,2)	RF	181	189	157	173

Como se puede ver en la Tabla 4, estos proto-pesos tienen diferentes órdenes de magnitud en relación con los datos, por lo que es por eso que se necesita la escalabilidad. También es posible observar que, por ejemplo, CWLS elige primero la covariable número 3 y la cuarta por último, por lo que serían la menos y la más importante, respectivamente, para explicar el comportamiento del TBF. RF es más imparcial en la importancia que asigna a cada covariable basándose en su votación. Para el caso del método 'NS', los proto-pesos se escalan utilizando el método MinMax entre 0 y 3, esto es completamente arbitrario y debe considerarse como un caso con influencia de conocimiento experto. Por lo tanto, los pesos iniciales se hacen con un enfoque intuitivo y son fáciles de interpretar. Los pesos escalados,

que se utilizarán como valores iniciales, se presentan a continuación.

Cuadro 5

Pesos iniciales estimados para cada covariable.

Escalador	Ponderación	p1	p2	p3	p4
NS	CWLS	0,0416	0,3328	0,0000	3,0000
MinMax(0,1)	CWLS	0,2222	0,1111	0,0000	1,0000
MinMax(0,2)	CWLS	0,4444	0,2222	0,0000	2,0000
NS	GBRT	1,1250	3,0000	0,1875	0,0000
MinMax(0,1)	GBRT	0,3750	1,0000	0,0625	0,0000
MinMax(0,2)	GBRT	0,7500	2,0000	0,1250	0,0000
NS	RF	2,2500	3,0000	0,0000	1,5000
MinMax(0,1)	RF	0,7500	1,0000	0,0000	0,5000
MinMax(0,2)	RF	1,5000	2,0000	0,0000	1,0000

Dado que las cotas inferiores de los pesos de las covariables son cero, se consideraron dos formas para las cotas superiores: el método IPCRidge y no considerar cotas superiores. El propósito de esto es intentar descifrar qué parte es más importante para la optimización, si las cotas o el valor inicial. Las cotas calculadas se muestran en las Tablas 6 y 7. Las cotas superiores para el método NS son más relajados que los demás; este método no estandariza los órdenes de magnitud de todas las covariables como los otros métodos.

Cuadro 6

Cotas para los algoritmos de optimización considerando Cotas superiores (con IPCRidge).

Tipo Cota	Escalador	λ	β	p1	p2	p3	p4
Inferior	-	0	1,1	0	0	0	0
Superior	NS	10	11	16,8393	16,8393	16,8393	16,8393
Superior	MinMax(0,1)	10	11	1,3164	1,3164	1,3164	1,3164
Superior	MinMax(0,2)	10	11	2,4083	2,4083	2,4083	2,4083

Cuadro 7

Cotas para los algoritmos de optimización sin cotas superiores.

Tipo Cota	λ	β	**p1**	**p2**	**p3**	**p4**
Inferior	0	1,1	0	0	0	0
Superior	10	11	Inf	Inf	Inf	Inf

Ahora es posible poner en acción los algoritmos de optimización. Para el enfoque de IPOPT, se obtuvieron los siguientes resultados.

Cuadro 8

Resultados con IPOPT.

Cota Sup	**Pond.**	**Escalador**	**Valor LL**	λ	η	β	**T. Ejec. [s]**
No	CWLS	NS	748,3538	1,3063E-14	144571,8120	2,7766	5,7463
Yes	CWLS	NS	748,3538	1,3063E-14	144571,6497	2,7766	1,2766
No	GBRT	NS	748,3538	1,3063E-14	144571,6005	2,7766	15,3736
Yes	GBRT	NS	748,3538	1,3063E-14	144571,6497	2,7766	8,4927
No	RF	NS	748,3538	1,3063E-14	144571,6005	2,7766	13,0116
Yes	RF	NS	748,3538	1,3063E-14	144571,6497	2,7766	4,4905
No	CWLS	MinMax(0,1)	747,1369	5,0210E-15	145706,4476	2,8576	3,5969
Yes	CWLS	MinMax(0,1)	747,1369	5,0210E-15	145706,4502	2,8576	2,2430
No	GBRT	MinMax(0,1)					2,5358
Yes	GBRT	MinMax(0,1)	747,1369	5,0210E-15	145706,4399	2,8576	4,5415
No	RF	MinMax(0,1)					4,4798
Yes	RF	MinMax(0,1)	747,1369	5,0210E-15	145706,4399	2,8576	5,7032
No	CWLS	MinMax(0,2)					1,8389
Yes	CWLS	MinMax(0,2)	748,7728	1,2140E-17	1726525,4024	2,7834	2,1190
No	GBRT	MinMax(0,2)					4,0725
Yes	GBRT	MinMax(0,2)	747,1155	4,5026E-15	151169,8890	2,8579	4,2528
No	RF	MinMax(0,2)					5,2494
Yes	RF	MinMax(0,2)	748,7728	1,2140E-17	1726525,4024	2,7834	7,4086

Cuadro 9

Pesos estimados utilizando IPOPT.

Cota Sup	Ponderación	Escalador	p1	p2	p3	p4
No	CWLS	NS	1,3852E-07	1,2185E-05	7,3698E-02	1,9637E+00
Yes	CWLS	NS	1,9493E-07	1,6818E-05	7,3698E-02	1,9637E+00
No	GBRT	NS	1,9493E-07	1,6818E-05	7,3698E-02	1,9638E+00
Yes	GBRT	NS	1,9493E-07	1,6818E-05	7,3698E-02	1,9637E+00
No	RF	NS	1,9493E-07	1,6818E-05	7,3698E-02	1,9638E+00
Yes	RF	NS	1,9493E-07	1,6818E-05	7,3698E-02	1,9637E+00
No	CWLS	MinMax(0,1)	1,7519E-07	7,8057E-07	7,5726E-02	6,8442E-02
Yes	CWLS	MinMax(0,1)	1,7519E-07	7,8057E-07	7,5726E-02	6,8442E-02
No	GBRT	MinMax(0,1)				
Yes	GBRT	MinMax(0,1)	1,9281E-07	8,5577E-07	7,5726E-02	6,8442E-02
No	RF	MinMax(0,1)				
Yes	RF	MinMax(0,1)	1,9281E-07	8,5577E-07	7,5726E-02	6,8442E-02
No	CWLS	MinMax(0,2)				
Yes	CWLS	MinMax(0,2)	3,1597E-03	1,3733E-01	6,0173E-02	1,1878E+00
No	GBRT	MinMax(0,2)				
Yes	GBRT	MinMax(0,2)	1,3115E-07	5,9619E-07	7,2702E-02	6,6582E-02
No	RF	MinMax(0,2)				
Yes	RF	MinMax(0,2)	3,1597E-03	1,3733E-01	6,0173E-02	1,1878E+00

Para el caso de IPOPT (Tabla 8), el método de escala que logra los valores más bajos de LL en promedio es el método MinMax(0,1), pero el método NS logra obtener soluciones factibles en todos sus escenarios. El método MinMax(0,2) logra el valor más bajo de LL con GBRT y límites superiores, pero en el resto sus soluciones son peores o infactibles. En cuanto a esto último, los valores estimados de β y η muestran una etapa de desgaste en el ciclo de vida del activo, por eso $\beta > 1$, a su vez, eso explica el orden de magnitud de η. El método seguido disminuye el valor de η, por lo tanto, el valor estimado de η debe ser menor al valor que se muestra en la Tabla 3. Por lo tanto, el método MinMax(0,2) logra solo una solución

factible, utilizando GBRT con límites superiores. CWLS es el método de ponderación de semillas que obtiene más soluciones factibles en todos sus escenarios en relación con los otros métodos, por lo que es el más consistente de todos. El método MinMax(0,2) obtiene η con órdenes de magnitud que escapan al resto de las soluciones y logrando sólo una solución factible, es el menos consistente de todos para IPOPT. Por las mismas razones, el método NS (con conocimiento experto) es el más consistente en los resultados de IPOPT.

En la Tabla 9, se puede ver que no hay mucha diferencia en la elección del método para la estimación de los pesos iniciales y los valores estimados de los pesos de las covariables. De hecho, el método de escala tiene más importancia para el valor en la estimación de los pesos que el método de peso inicial. Para MinMax(0,2) hay más variación, pero podría explicarse debido a la escasa consistencia de este método de escala. Además, los pesos cambian cuando cambia la escala, el peso de la covariable con más variación es Resistencia Dieléctrica (p4) y Temperatura Interna (p2), que son las covariables con los órdenes de magnitud más altos de todas (ver Tabla 2). Sin embargo, no hay mucha diferencia entre los métodos de escala MinMax para estas dos covariables (teniendo en cuenta la consistencia del método MinMax(0,2)), la principal diferencia ocurre entre los métodos NS y MinMax.

Para el enfoque de Algoritmos Genéticos (GA), la probabilidad de apareamiento se establece en 1, la tasa de mutación es igual a 0,7, la mutación de cada valor dentro de la descendencia, como se mencionó antes, cambia aplicando una mutación aditiva gaussiana con media igual a 0, desviación estándar igual a 1 y probabilidad independiente para que cada valor mute igual a 0,3. Finalmente, el tamaño de la población es de 500 individuos y se simularon 1000 generaciones en este estudio para cada iteración. Esta elección de hiper-parámetros fue realizada mediante un proceso de sensibilización usando lenguaje *Python* y jupyter notebooks como entorno de ejecución, alcanzando un tiempo de ejecución cercano a dos días. Utilizando los hiper-parámetros resultantes de la sensibilización, se obtuvieron las siguientes estimaciones de la optimización con GA.

Cuadro 10

Resultados con GA.

Cota Sup	Pond.	Escalador	Valor LL	λ	η	β	T. Ejec. [s]
No	CWLS	NS	756,7873	4,8460E-09	127529,7255	1,6723	1175,9755
Yes	CWLS	NS	756,7873	4,8460E-09	127419,1777	1,6724	1174,8894
No	GBRT	NS	756,7873	4,8460E-09	127315,7402	1,6725	1201,4359
Yes	GBRT	NS	756,7873	4,8460E-09	127490,0126	1,6723	1182,3434
No	RF	NS	756,7873	4,8460E-09	127387,1861	1,6724	1184,8883
Yes	RF	NS	756,7873	4,8460E-09	127490,1241	1,6723	1177,2808
No	CWLS	MinMax(0,1)	756,3793	4,8460E-09	133799,8400	1,6651	1126,4223
Yes	CWLS	MinMax(0,1)	756,3793	4,8460E-09	133799,8405	1,6651	1056,5227
No	GBRT	MinMax(0,1)	756,3793	4,8460E-09	133799,8405	1,6651	1165,5899
Yes	GBRT	MinMax(0,1)	756,3793	4,8460E-09	133799,8412	1,6651	1046,7461
No	RF	MinMax(0,1)	756,3793	4,8460E-09	133799,8409	1,6651	1118,3745
Yes	RF	MinMax(0,1)	756,3793	4,8460E-09	133799,8408	1,6651	1072,9131
No	CWLS	MinMax(0,2)	757,1333	4,8460E-09	156459,0487	1,6422	1118,3847
Yes	CWLS	MinMax(0,2)	756,4737	4,8460E-09	139765,1829	1,6586	1112,1583
No	GBRT	MinMax(0,2)	757,1333	4,8460E-09	156459,0482	1,6422	1140,9477
Yes	GBRT	MinMax(0,2)	756,4737	4,8460E-09	139765,1844	1,6586	1103,4481
No	RF	MinMax(0,2)	757,1333	4,8460E-09	156459,0473	1,6422	1105,1039
Yes	RF	MinMax(0,2)	756,4737	4,8460E-09	139765,1829	1,6586	1114,9228

De la Tabla 10, es posible observar que hay una mayor consistencia en todos los valores obtenidos sin importar el método elegido. MinMax(0,1) obtiene los valores más bajos de LL. GA obtiene diferentes β que IPOPT, pero se compensa con valores más bajos de λ, por eso los órdenes de magnitud de η no cambian entre algoritmos de optimización. Además, β sigue siendo mayor que 1. Los resultados del método MinMax(0,2) son sensibles a la presencia o no de cotas superiores, reafirmando la inconsistencia del método al cambiar el algoritmo de optimización y la importancia de las cotas superiores para obtener valores de LL más bajos en menos tiempo. La elección del método para obtener los pesos iniciales no es lo suficientemente importante como para afectar los resultados obtenidos.

Cuadro 11

Pesos estimados utilizando GA.

Cota Sup	Ponderación	Escalador	p1	p2	p3	p4
No	CWLS	NS	9,7500E-16	1,8989E-01	5,0954E-02	1,0428E+00
Yes	CWLS	NS	5,8443E-16	1,9170E-01	5,0992E-02	1,0507E+00
No	GBRT	NS	1,1242E-13	1,9340E-01	5,1027E-02	1,0581E+00
Yes	GBRT	NS	2,3680E-15	1,9054E-01	5,0968E-02	1,0456E+00
No	RF	NS	1,2804E-15	1,9223E-01	5,1003E-02	1,0530E+00
Yes	RF	NS	1,2422E-15	1,9054E-01	5,0968E-02	1,0456E+00
No	CWLS	MinMax(0,1)	2,3341E-16	3,5369E-16	5,1027E-02	4,4779E-02
Yes	CWLS	MinMax(0,1)	4,7749E-16	4,0651E-16	5,1027E-02	4,4779E-02
No	GBRT	MinMax(0,1)	3,8399E-16	3,5341E-16	5,1027E-02	4,4779E-02
Yes	GBRT	MinMax(0,1)	4,7636E-16	4,0128E-16	5,1027E-02	4,4779E-02
No	RF	MinMax(0,1)	3,7140E-16	3,8111E-16	5,1027E-02	4,4779E-02
Yes	RF	MinMax(0,1)	2,9873E-16	1,7825E-16	5,1027E-02	4,4779E-02
No	CWLS	MinMax(0,2)	3,2435E-16	2,2551E-15	3,5347E-02	2,9013E-15
Yes	CWLS	MinMax(0,2)	1,4053E-16	5,8639E-16	4,8051E-02	4,0673E-02
No	GBRT	MinMax(0,2)	3,2011E-16	2,2414E-16	3,5347E-02	1,2945E-15
Yes	GBRT	MinMax(0,2)	1,7579E-16	3,4217E-16	4,8051E-02	4,0673E-02
No	RF	MinMax(0,2)	3,4526E-16	6,2181E-16	3,5347E-02	9,1053E-16
Yes	RF	MinMax(0,2)	1,1811E-16	4,1346E-16	4,8051E-02	4,0673E-02

En la Tabla 11, se observa un comportamiento similar al logrado por IPOPT (ver Tabla 9). Sin embargo, GA muestra una mejor consistencia en los pesos estimados para todas las covariables. El método MinMax(0,2) aún presenta una consistencia pobre a lo largo de las iteraciones, en las mismas covariables que en el caso de IPOPT, y muestra sensibilidad a la existencia (o no) de cotas superiores. El método NS aún tiene una diferencia de órdenes de magnitud mayores en los pesos de las covariables 2 y 4 en relación con los valores estimados utilizando los métodos MinMax. La Tabla 12 muestra un menor costo computacional promedio alcanzado por los métodos MinMax en comparación con la alternativa sin escala.

Por lo tanto, es más fácil para los algoritmos de optimización trabajar con datos escalados que manejar datos de covariables con una transformación creciente (NS).

Cuadro 12

Tiempos promedios de ejecución para cada método de escalamiento y optimización.

Escalador	**Tiempo promedio GA** [s]	**Tiempo promedio IPOPT** [s]
NS	1182,8022	8,0652
MinMax(0,1)	1097,7614	3,8501
MinMax(0,2)	1115,8276	4,1569

Además, con fines de bondad de ajuste, se calcularon los p-values para cada parámetro estimado mediante ambas estrategias de optimización, basándose en intervalos de confianza de verosimilitud del 95 % y el test de Wald utilizando la biblioteca de Python *VeMoMoTo*. Notar que $p < 0{,}05$ significa que los resultados son estadísticamente significativos, rechazando la hipótesis nula. Los p-values se muestran en las tablas 13 y 14.

Como se puede observar en las tablas anteriores (8 y 10), ambos algoritmos de optimización alcanzan soluciones factibles en la mayoría de las iteraciones simuladas. Las iteraciones que no alcanzaron una solución factible tienen como factor común la ausencia de una cota superior. Además, de la Tabla 12, se puede apreciar que GA, en promedio, requiere un costo computacional mucho mayor que IPOPT, pero logra llegar siempre a soluciones factibles, a diferencia de IPOPT. En casi todos los casos (con solución factible), los tiempos más cortos se logran cuando se consideran cotas superiores. Por lo tanto, las cotas superiores tienen un impacto a tener en cuenta en los métodos de optimización, y es más evidente en IPOPT.

Cuadro 13

P-values de los parámetros estimados por GA.

Cota Sup	Ponderador	Escalador	η	β	p1	p2	p3	p4
No	CWLS	NS	0,360	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	CWLS	NS	0,360	<0,001	<0,001	<0,001	<0,001	<0,001
No	GBRT	NS	0,360	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	GBRT	NS	0,360	<0,001	<0,001	<0,001	<0,001	<0,001
No	RF	NS	0,360	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	RF	NS	0,360	<0,001	<0,001	<0,001	<0,001	<0,001
No	CWLS	MinMax(0,1)	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	CWLS	MinMax(0,1)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	GBRT	MinMax(0,1)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	GBRT	MinMax(0,1)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	RF	MinMax(0,1)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	RF	MinMax(0,1)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	CWLS	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	CWLS	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	GBRT	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	GBRT	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	RF	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	RF	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001

El método para la estimación del valor inicial para los pesos que presenta la mayor robustez es CWLS. En el algoritmo de optimización menos consistente (IPOPT), logra obtener soluciones factibles sin cotas superiores en la mayoría de los casos. Esto podría indicar que CWLS entrega un valor semilla más cercano a una solución factible para las covariables que los otros métodos. Los resultados de IPOPT (Tabla 8) alcanzan valores más bajos de LL que GA (Tabla 10) en todos los casos con solución factible.

Cuadro 14

P-values de los parámetros estimados por IPOPT.

Cota Sup	Ponderador	Escalador	η	β	p1	p2	p3	p4
No	CWLS	NS	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	CWLS	NS	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	GBRT	NS	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	GBRT	NS	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	RF	NS	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	RF	NS	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	CWLS	MinMax(0,1)	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001
Yes	CWLS	MinMax(0,1)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	GBRT	MinMax(0,1)						
Yes	GBRT	MinMax(0,1)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	RF	MinMax(0,1)						
Yes	RF	MinMax(0,1)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	CWLS	MinMax(0,2)						
Yes	CWLS	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	GBRT	MinMax(0,2)						
Yes	GBRT	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001
No	RF	MinMax(0,2)						
Yes	RF	MinMax(0,2)	1,000	<0,001	<0,001	<0,001	<0,001	<0,001

De las tablas 13 y 14, es posible observar que en la mayoría de las iteraciones se alcanzan resultados estadísticamente significativos, siendo de especial interés el caso sin cota superior, CWLS y MinMax(0,1) que logra estimar valores estadísticamente significativos para todos los parámetros, sin importar el algoritmo de optimización.

Con el fin de seguir con el cálculo de bandas y pasos posteriores, se tomarán como valores de referencia a las estimaciones obtenidas en la iteración que consideró cotas superiores, CWLS para los valores semilla, IPOPT como algoritmo de optimización y NS para el escalamiento de los datos. Por lo tanto, los parámetros estimados son los siguientes.

Cuadro 15

Parámetros estimados escogidos.

η	β	**p1**	**p2**	**p3**	**p4**
144571,6497	2,7766	1,9493E-07	1,6818E-05	7,3698E-02	1,9637E+00

Cálculo de Bandas

Los datos de las covariables entonces cuentan con la transformación 'NS'. Dado que el Dataset es pequeño, sólo 60 registros, se decidió obviar el proceso de división en intervalos de clase. Luego, se utilizaron los dos métodos de clustering, K-means y GMM, para definir el número de clusters que mejor se adapta al Dataset. Se contrastaron los resultados aplicando el método Elbow, Silhouette, AIC y BIC. De esta forma se obtienen las figuras 2, 3 y 4.

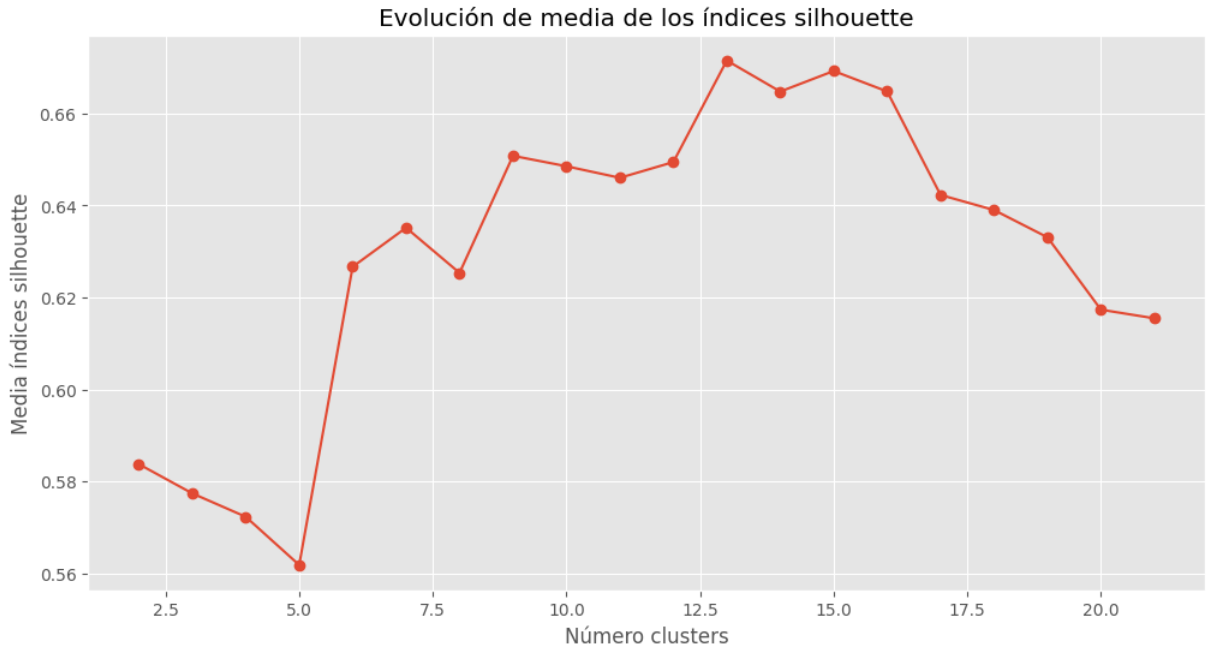

Figura 2

Evolución de las medias de los índices de Silhouette para diferentes números de clusters con el método K-means.

Como es posible observar en la figura 4, los criterios de información no aportan

resultados concluyentes, AIC presenta valores menores para un K igual a 2 y 6, mientras que BIC encuentra el menor valor con un K igual a 1. Cabe destacar que los valores de BIC son en general más robustos que AIC por su método de cálculo. Silhouette (Figura 2) presenta peaks cuando K es igual a 2 y 7, considerando los modelos menos complejos. Sin embargo, las iteraciones comienzan en 2, siendo imposible poder saber si para un K igual a 1 el índice cae o sube, para poder considerar $K = 2$ como la solución más adecuada.

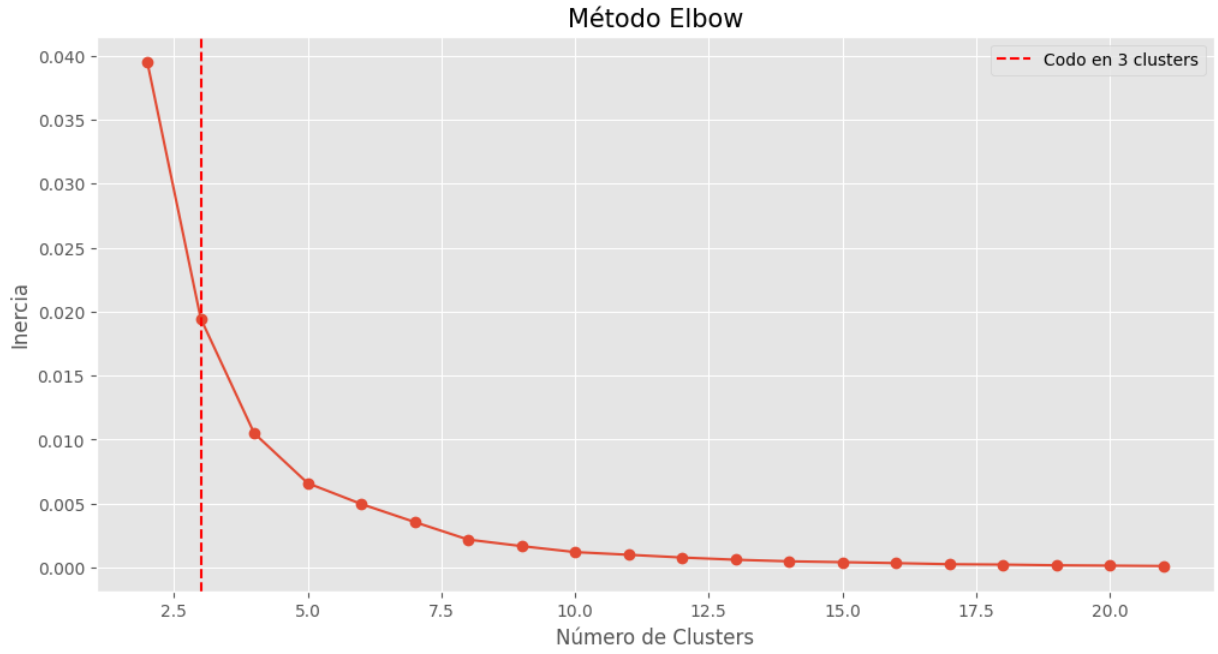

Figura 3

Método Elbow para diferentes números de clusters con el método K-means.

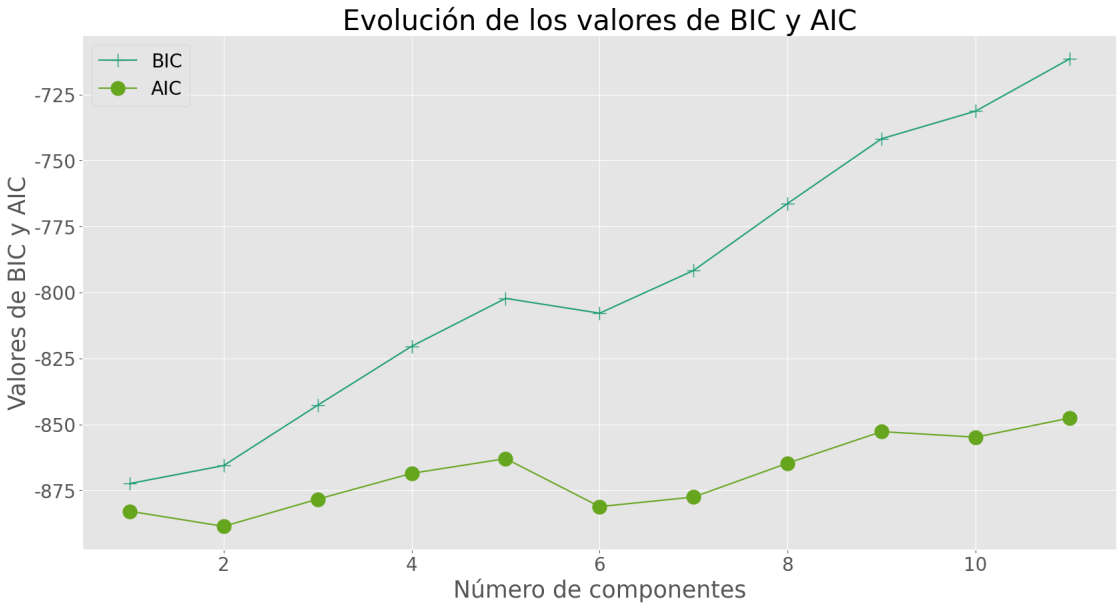

Figura 4

Valores de los criterios de información BIC y AIC para diferentes números de clusters (componentes) con el método GMM.

Por último, la Figura 3 presenta índices de inercia para diferentes valores de K, esto con el fin de poder aplicar el método Elbow o del codo, el cual entrega un K óptimo igual a 3, esto sin objeciones observables. Por lo tanto, se utilizará un $K = 3$ para los algoritmos de clustering.

Luego de establecer un K adecuado, es posible generar los límites que definirán los estados para la matriz de transición. Como se explicó anteriormente, diferentes métodos se proponen para su estimación.

La primera es mediante los centroides de cada cluster. Los límites entre dos clusters se definen como el punto medio entre ambos centroides, que se presentan en la Figura 5 utilizando K-means. Si bien es posible utilizar GMM, la elección natural para este método es K-means por la forma en como determina cada cluster (mínima distancia entre puntos y centroide).

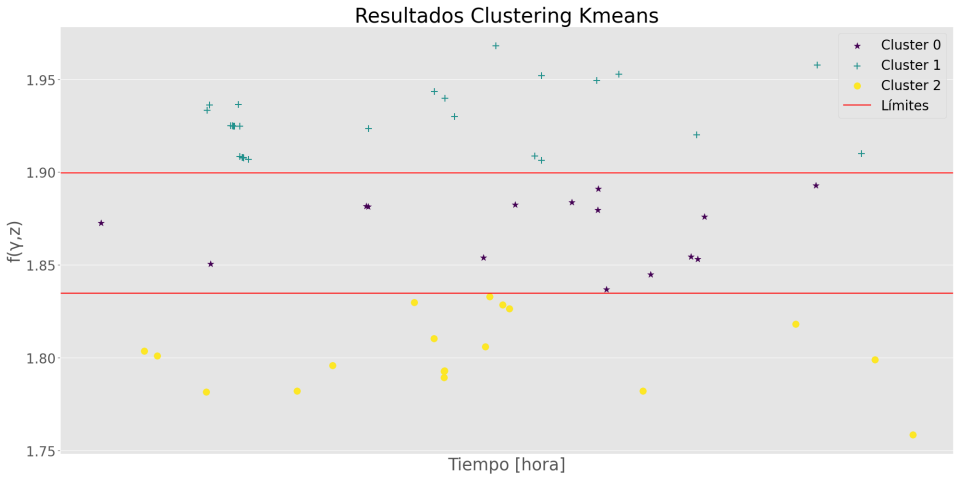

Figura 5

Clusters y límites obtenidos con K-means.

A continuación, se presentan los estados generados junto con sus límites numéricos para este primer método con K-means.

Cuadro 16

Estados/Clusters con K-means.

Estado	Límite Inferior	Límite Superior
1	0	1.8353
2	1.8353	1.8984
3	1.8984	Inf

El segundo método para calcular los rangos de cada estado se basa en las probabilidades proporcionadas únicamente por el método GMM. Esto último puede observarse en la Figura 6, la cual muestra la distribución de densidad de probabilidad relacionada al modelo GMM completo, junto el valor en escala de colores de la función de densidad de probabilidad, los clusters generados y, por último, el límite encontrado a partir de estos valores. Este límite se obtiene al encontrar el mínimo valor de la función de densidad de probabilidad (áreas más blancas). Al utilizar esta aproximación, es posible observar que sólo encuentra un límite de

los dos que debería estimar. Producto de lo anterior, es que se extraen las probabilidades discretas de pertenecer a algún cluster en la Figura 7.

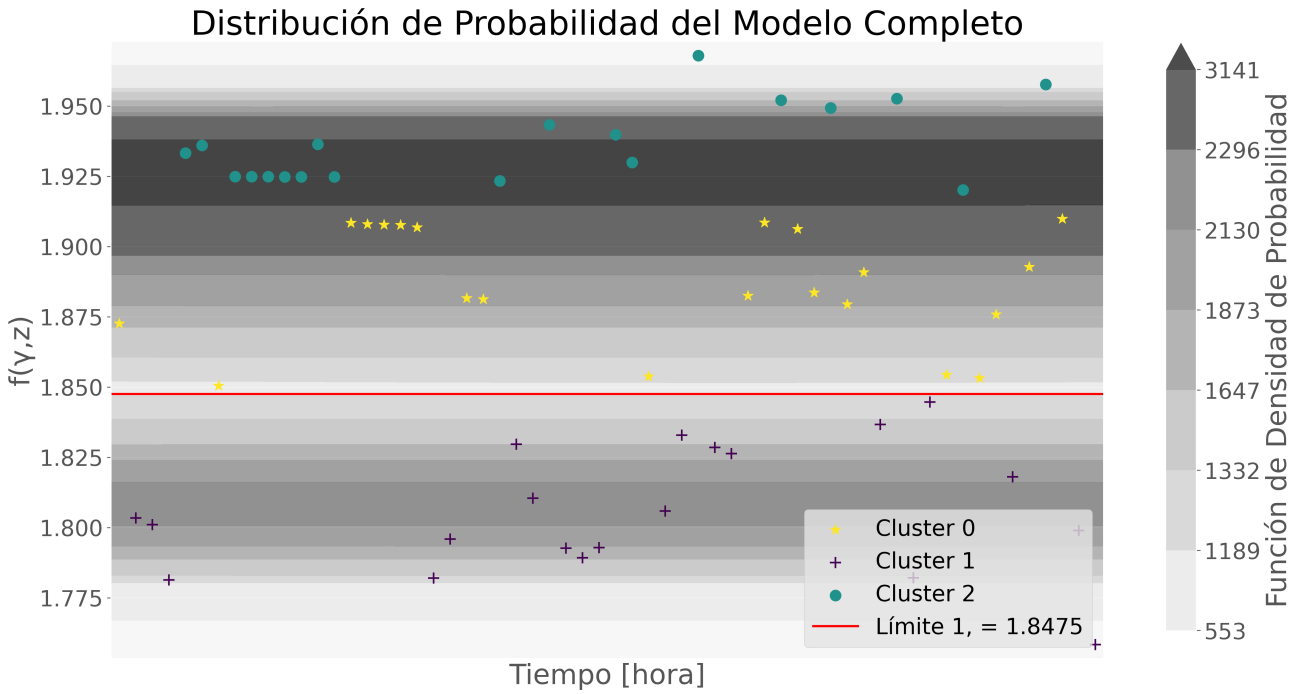

Figura 6

$f(\gamma, z)$ a lo largo del tiempo y su distribución de densidad de probabilidad del modelo completo.

La figura 7 muestra un mapa de colores de probabilidades discretas de pertenencia a algún cluster según el valor de $f(\gamma, z)$. Esta probabilidad es el valor máximo dentro del vector de probabilidades relacionadas a pertenecer a un cluster, para un punto en particular. De esta forma, se busca hacer un mapa que indique cuando un punto tiene una probabilidad baja de pertenecer al cluster al cual fue asignado por el modelo GMM. Estas regiones son las más oscuras e indicarían el valor de los límites para los estados. Una representación más simple de estos 'valles' se observa en la Figura 8.

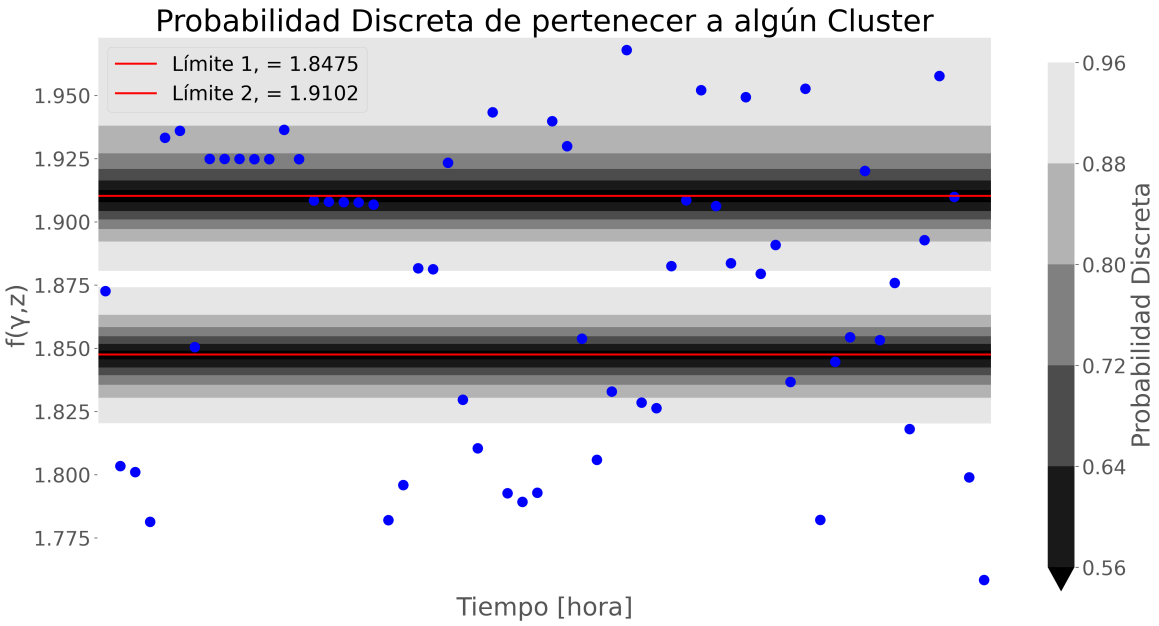

Figura 7

Distribución de $f(\gamma, z)$ a lo largo del tiempo y su probabilidad discreta de pertenecer a un cluster del modelo completo.

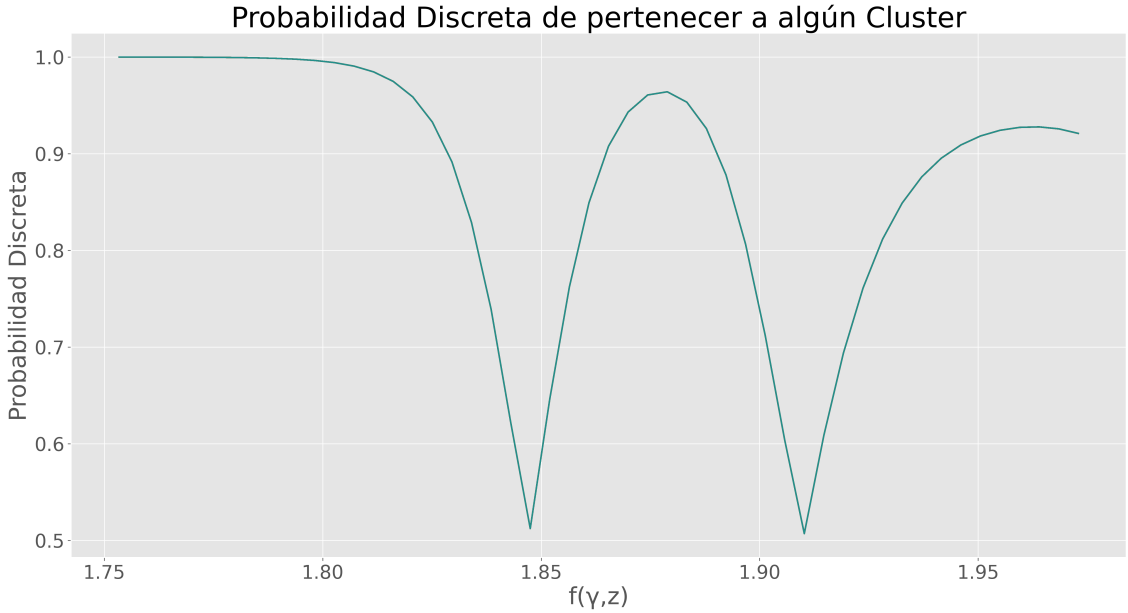

Figura 8

$f(\gamma, z)$ en el tiempo y su distribución de probabilidad discreta de pertenecer a un cluster del modelo completo.

De esta forma se observa que, al utilizar la función de densidad de probabilidad, se pierde información relacionada a los límites para cada cluster (ver Figura 6). Es por esto que se desarrolla el proceso de estimación a partir de las probabilidades de pertenencia por punto (ver Figura 7 y 8).

Cuadro 17

Estados/Clusters con GMM.

Estado	**Límite Inferior**	**Límite Superior**
1	0	1.8475
2	1.8475	1.9102
3	1.9102	Inf

Como se observa en la Figura 7 y 8, a partir de estas probabilidades discretas se encuentran los dos límites necesarios para definir los tres estados relacionados a cada cluster. Entonces se definen los estados, tal y como se observan en la Tabla 17. Al observar los resultados para el cálculo del rango (ver Tabla 17 y 16), ambos métodos alcanzan resultados similares pero no iguales. Al clasificar cada registro en su estado (con GMM) correspondiente se obtienen 21 puntos de datos que pertenecen al estado 2 (cluster central), 20 al estado 1 (cluster inferior) y 19 al estado 3 (cluster superior).

Para las situaciones de borde, se considera que las probabilidades de pertenencia a un cluster u otro son menos claras para los puntos cercanos al límite entre clusters; por lo tanto, se propone un rango de confiabilidad alrededor de este límite. Si la diferencia de probabilidades de pertenencia entre clusters es menor al 5 %, un punto dentro de este rango se clasifica aleatoriamente en otro cluster con probabilidades iguales de pertenecer a este límite. De esta manera, la distribución de los datos en cada estado/cluster es el siguiente:

- Estado 2 = 36 datos
- Estado 1 = 16 datos

- Estado 3 = 8 datos

Confiabilidad Condicional y RUL

En este punto ya es posible calcular las matrices de probabilidad de transición para ambos métodos, tal y como se explica en la Página 25. Los resultados son los siguientes.

Cuadro 18

Matriz de Probabilidad de Transición a partir de centroides (K-means).

Estado	**1**	**2**	**3**
1	99,9917 %	0,0050 %	0,0033 %
2	0,0049 %	99,9828 %	0,0123 %
3	0,0024 %	0,0048 %	99,9929 %

Cuadro 19

Matriz de Probabilidad de Transición a partir de probabilidades de pertenencia (GMM).

Estado	**1**	**2**	**3**
1	99,9981 %	0,0005 %	0,0014 %
2	0,0011 %	99,9966 %	0,0023 %
3	0,0006 %	0,0002 %	99,9992 %

Finalmente, la función de confiabilidad condicional se estima a través de los pasos descritos en la Página 26. En este caso, se establece que t_{inicial} es 0, t_{final} es 200000 y Δ es 100. Luego, se obtiene la confiabilidad en función del tiempo para ambos métodos (centroide y probabilidades de pertenencia). La Figura 9 y la Figura 10 presentan las funciones de confiabilidad condicional correspondientes utilizando las matrices de transición de probabilidad como entrada para el mencionado método de propiedad del producto.

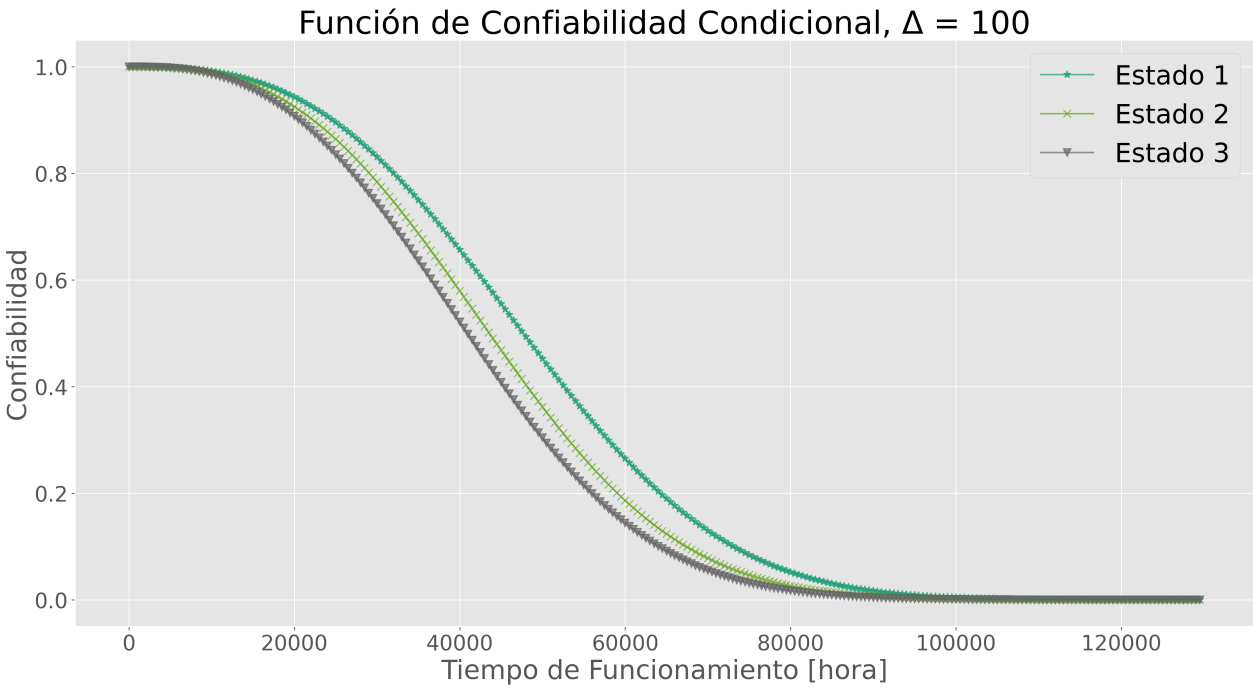

Figura 9

Función de confiabilidad condicional por método de centroides (K-means).

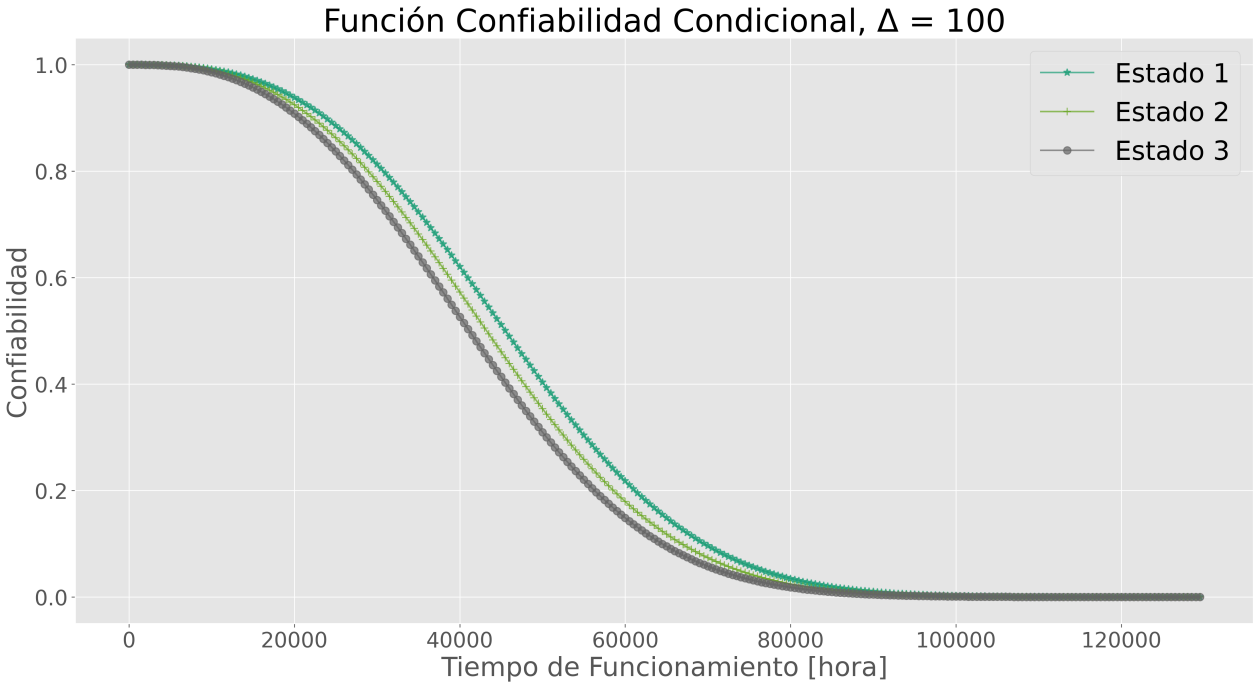

Figura 10

Función de confiabilidad condicional por método de probabilidades de pertenencia (GMM).

Estas figuras muestran que, para ambos casos, el comportamiento de confiabilidad para cada estado es similar. Específicamente, se puede determinar, basándose en lo mismo, que el estado 1 corresponde al estado en el cual el componente exhibe la mayor confiabilidad, seguido por el estado 2 y luego el estado 3, que presenta la menor confiabilidad con el tiempo. Esto concuerda con la cantidad de datos presentes en cada estado. Sin embargo, es posible notar que las curvas del método GMM son más cercanas entre sí que el método K-means. Además, GMM genera curvas de confiabilidad que caen levemente más rápido, lo cual podría indicar que este método castiga un poco más las curvas.

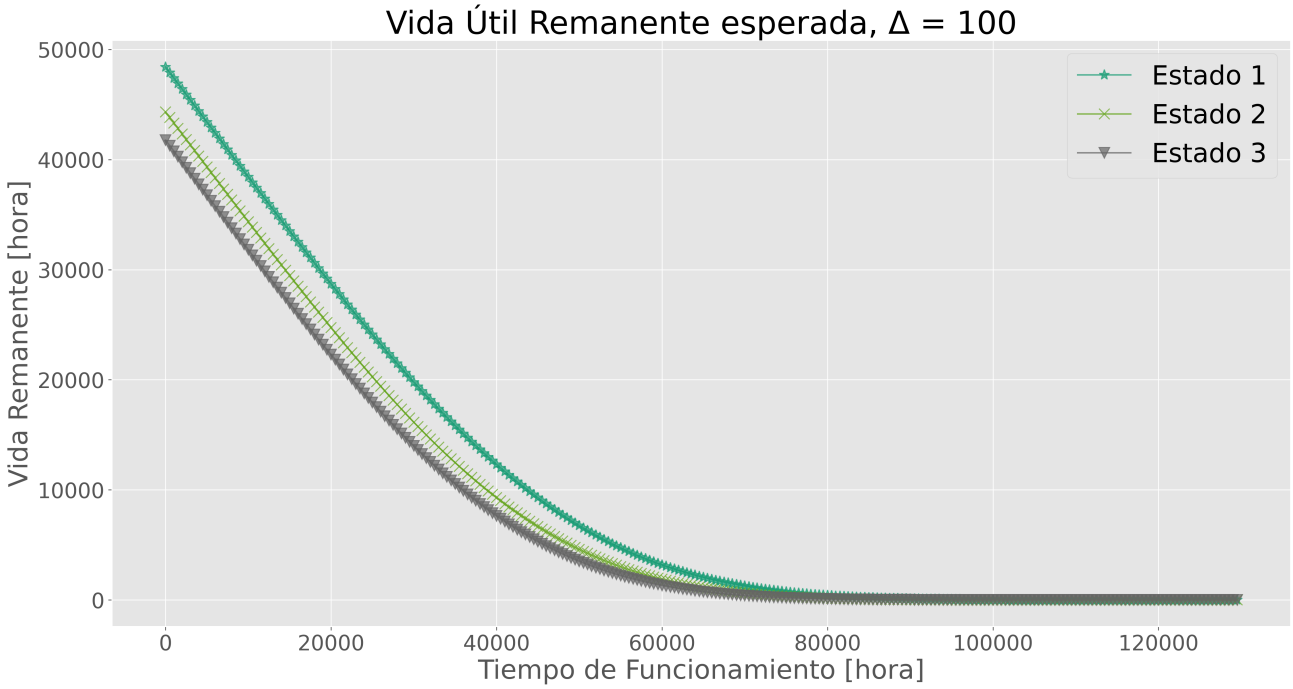

Figura 11

Vida Útil Remanente por método de centroides (K-means).

Para propósitos predictivos y utilizando las funciones de confiabilidad condicional estimadas como entradas, las Figuras 11 y 12 muestran la vida útil remanente (RUL) para ambos métodos, considerando la evolución de los datos de covariables clusterizados a lo largo de los estados. Se observa que el Cluster-Estado 1 puede considerarse como la mejor condición, ya que su declive desde la RUL máxima presenta un comportamiento suave.

Por otro lado, el Estado-3 es la peor condición, ya que su declive es el más agresivo a lo largo del tiempo de funcionamiento. Aunque este efecto podría ser esperado, la novedad de la presente propuesta radica en lo mencionado anteriormente. Es decir, la entrada de conocimiento experto podría ser directa al separar los límites de banda para una covariable, ya que tiene solo una unidad de medida. Sin embargo, es difícil para un experto proporcionar un valor límite de banda al tratar con diversas covariables, especialmente cuando la unidad de medida combinada no representa directamente una magnitud física para evaluar. Por lo tanto, el método de ML propuesto complementa el conocimiento experto bajo una novedosa evaluación de condiciones.

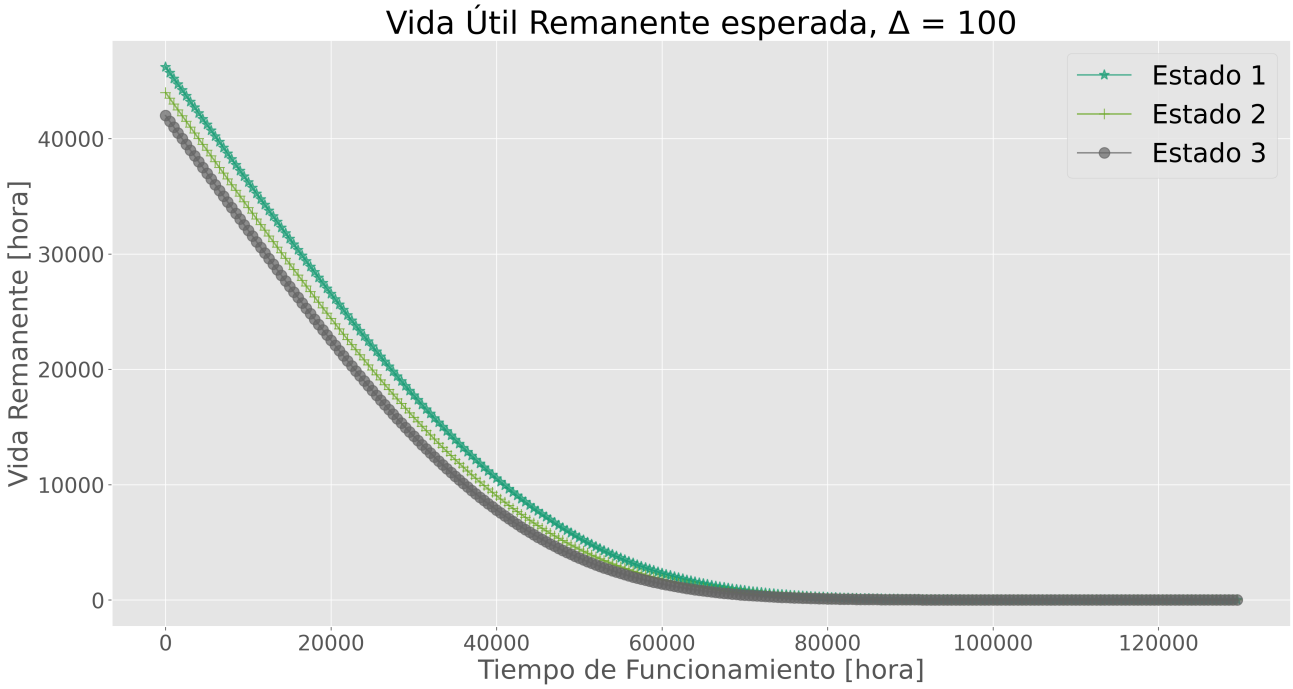

Figura 12

Vida Útil Remanente por método de probabilidades de pertenencia (GMM).

Por otro lado, lo observado con las curvas de confiabilidad condicional se traspasa a las curvas de la RUL de cada estado. El método GMM genera curvas de RUL que caen más rápidamente en comparación con K-means, además de tener una menor distancia entre curvas, restandole importancia a las condiciones.

Una vez explicado el efecto de la condición, la otra componente del Modelo predictivo PHM (Ecuación (15)) es la contribución de la edad a la tasa de riesgo. Los valores de β y η son consecuentes con su interpretación en la confiabilidad condicional y RUL de las Figuras 9 a 12. El valor de β en la Tabla 15 es superior a 1, lo que indica la etapa de desgaste en el ciclo de vida del activo. Esto significa que los equipos en estudio están envejeciendo a lo largo del período. La vida característica η concuerda también con los valores del tiempo de funcionamiento. Sin embargo, es interesante notar que en la etapa inicial del ciclo de vida, el impacto de la condición clusterizada es un poco más significativo que el efecto de la edad; es decir, la diferencia en la RUL entre los estados es más notoria. Por otro lado, consecuente con el modelo PHM, las diferencias en RUL se acortan en las etapas posteriores de la vida útil del activo. Esto significa que en tiempos de funcionamiento elevados, el efecto del tiempo prevalece porque la condición general del activo está muy degradada al alcanzar un tiempo de operación avanzado.

Finalmente, la propuesta de evaluación de condición mediante ML añade conocimiento innovador sobre el proceso de covariables clusterizadas. Es decir, maneja varias señales vitales con diversas fuentes y unidades de medida, permitiendo así un enfoque holístico. Al implementar una política predictiva, esta novedosa evaluación de condición permite una mejor comprensión de la salud operativa del activo, interviniendo en el momento exacto (en contraste con una política de reemplazo fijo por edad), maximizando así la RUL de los activos. Se demuestra que este enfoque de ML para abordar políticas de mantenimiento puede proporcionar resultados significativos como alternativa o complemento al criterio de expertos, ofreciendo ventajas, especialmente cuando se trata de más de una covariable, y aumentando en última instancia la expectativa y precisión de la vida útil restante para los activos críticos.

Conclusiones y Recomendaciones

La presente investigación introduce enfoques novedosos para los desafíos asociados a contemplar múltiples covariables en el contexto del Modelo Weibull PHM, aprovechando técnicas avanzadas Data-driven. En el mismo sentido, varios métodos de ML se han demostrado efectivos para estimar cotas y valores iniciales en el problema no convexo de la estimación de verosimilitud parcial de Cox. Dos algoritmos de optimización, un algoritmo genético (no paramétrico) y un optimizador de puntos interiores (semi-paramétrico), se utilizaron para resolver este problema no convexo, alcanzando resultados factibles en ambos casos. GA muestra una mayor robustez en el sentido de que logra encontrar soluciones factibles en todos los escenarios evaluados, a diferencia de IPOPT, aunque GA no logra obtener resultados tan buenos como IPOPT en cuanto a los valores de la verosimilitud (LL) y el costo computacional. Es importante señalar que la LL mide la bondad de ajuste de un modelo estadístico, pero adicionalmente se calcularon los p-values relacionados con intervalos de confianza del 95 % para los parámetros estimados, siendo estadísticamente significativos la mayoría de ellos.

La introducción del estimador de Kaplan-Meier como cota superior proporciona resultados superiores en comparación con escenarios sin cotas superiores, mejorando el rendimiento de los algoritmos de optimización en cuanto a valor de verosimilitud alcanzado y tiempo de ejecución (costo computacional), estableciéndolo como una técnica útil para la estimación de límites superiores. Además, Gradient Boosting con mínimos cuadrados (CWLS) como base learner demuestra robustez al proporcionar estimaciones de valores iniciales factibles en todas las iteraciones analizadas, y estadísticamente significativas para todos los parámetros en un caso particular. Por lo tanto, es evidente que abordar las cotas y emplear una estrategia para los valores iniciales es crucial para resolver eficazmente este problema. Además, la estrategia de estimación introducida en este estudio mejora la robustez de la Estimación de Máxima Verosimilitud (MLE) en el Modelo de Riesgo Proporcional con distribución Weibull.

Además, se ha analizado el escalado de datos y la inclusión de métodos MinMax muestra resultados de máxima verosimilitud consistentes cuando el valor máximo es 1, siendo

superior al método basado en conocimiento experto (NS). En cuanto al costo computacional, el escalado de datos simplifica el trabajo para los métodos de optimización y contribuye a obtener una estimación de máxima verosimilitud mejorada. Además, el enfoque de Gradient Boosting y el método de votación de Random Forest (RF) son fácilmente interpretables, gracias a su operación intuitiva. Cuentan con procesos de cálculo sencillos de seguir. Sin embargo, el método de mínimos cuadrados por componente (CWLS) supera en LL y tiempos de ejecución a otros métodos para los algoritmos de optimización considerados.

Continuando con los desafíos asociados a múltiples covariables, se presenta un método innovador y novedoso para definir bandas óptimas de covariables en la evaluación de condiciones dentro del marco de mantenimiento predictivo PHM-CBM. Las técnicas de clustering (K-means y GMM) contribuyen a establecer el número de bandas y rangos coherentes y confiables de clusters para todas las covariables, haciendo posible la construcción de las matrices de probabilidad de transición. Se utilizaron diversos indicadores para calcular el número de clusters que se ajusta mejor al dataset, finalmente considerando por validez de los resultados al método Elbow, el cual arrojo un número de clusters óptimo igual a 3. Para conocer los límites de las bandas, se utilizaron dos métodos. En primer lugar, los rangos se calcularon en función del punto medio entre centroides de dos clusters sucesivos, usando K-means. El segundo método define los rangos en base a las probabilidades generadas por GMM, definidas como el punto más bajo de probabilidad de pertenencia al cluster asignado. Para este último método, también se definió un procedimiento de clasificación para situaciones de borde entre bandas.

A partir de las matrices de probabilidad de transición calculadas, se obtuvieron las confiabilidades condicionales y RUL asociados a los estados establecidos a través de las técnicas de clustering. De esta forma, es posible estimar el impacto de las condiciones en la vida útil de los equipos. Se evidencia que en la etapa inicial del ciclo de vida, el impacto de la condición es más significativo que el efecto de la edad, como se puede observar en las curvas de RUL entre los estados. Por otro lado, consecuente con el modelo PHM, las diferencias en

RUL se acortan en las etapas posteriores de la vida útil del activo. Por lo tanto, en tiempos de funcionamiento elevados, el efecto del tiempo prevalece por sobre la condición.

Finalmente, a partir de estimaciones robustas y Data driven del impacto de las condiciones en la vida útil, esta tesis conduce a decisiones efectivas de mantenimiento predictivo para diversos escenarios de falla en diferentes parámetros operativos, contribuyendo en la mejora de la interpretación de la salud de los equipos considerando múltiples condiciones, lo cual permite que la toma de decisiones pueda ser en el instante óptimo, haciendo posible el potencial aumento de la vida útil de los activos.

Conclusiones Secundarias

Se proponen diferentes métodos para el cálculo de bandas, uno basado en centroides usando el algoritmo K-means, y otro basado en probabilidades de pertenencia usando GMM. Se destaca que GMM le resta importancia a las condiciones en el cálculo de la RUL y genera curvas que decaen más rápidamente. Al existir un castigo mayor de las curvas, GMM se asoma como un método más estricto en cuanto al impacto de las condiciones y la edad en la vida útil de los equipos. Esto se explica por la mayor información que considera al determinar los estados y clasificar cada punto, ofreciendo nuevas líneas de investigación al respecto. En términos prácticos, es más sencillo el uso de la alternativa K-means. Determinar puntos medios entre centroides y obviar situaciones de borde simplifica la estimación.

Más allá de las validaciones que se deben realizar utilizando otros datasets sobre los algoritmos utilizados en esta investigación, la metodología desarrollada se compone de etapas generales y aplicables en otros contextos. Por ejemplo, la preparación de datos es un desafío transversal entre industrias. Además, la metodología permite la integración de más covariables y utilizar datasets con un mayor número de registros. Sin embargo, se debe tener en consideración que, por criterio experto, no es aconsejable utilizar más de 10 covariables sobre estos modelos sin tener la suficiente capacidad computacional o usar métodos de paralelización del trabajo. Por otro lado, la metodología permite el uso de otras distribuciones, ya que sólo impactaría en principio a la función de verosimilitud a optimizar, la obtención

de estas funciones es un tema que se escapa de este estudio pero existe vasta literatura al respecto.

Dada las múltiples interacciones entre las covariables con diferentes comportamientos temporales, factores externos no considerados y la vida útil de los equipos, es necesario realizar periódicamente estimaciones de parámetros con el fin de tener pesos que representen la situación actual. En este sentido, contar con modelos de estimación de parámetros que sean eficientes en cuanto a tiempos de ejecución es clave. Si bien GA encuentra soluciones factibles en todos los escenarios estudiados, tiene tiempos de ejecución mucho mayores a IPOPT. Sin tener en consideración los tiempos destinados a la elección de hiper-parámetros que requiere GA. Para el caso de IPOPT, estas sensibilizaciones no son requeridas. El tiempo de ejecución impacta directamente en los costos computacionales en los que se incurre por el uso de estos modelos. Por lo tanto, tener en consideración estos tiempos de ejecución y de hiper-parametrización es importante.

Por último, la investigación integrada no solo fortalece la comprensión de las estrategias de mantenimiento predictivo, sino que también establece una conexión vital entre estas estrategias y el campo en evolución de Data Science, ofreciendo oportunidades enriquecedoras para la toma de decisiones en escenarios prácticos.

Trabajos Futuros

Todos los hallazgos presentados se derivaron de un dataset real de una empresa de distribución eléctrica, enfrentando desafíos comunes en el manejo de datos, como valores faltantes y comportamientos no ideales de las distribuciones subyacentes de los datos, una ocurrencia común en varios campos. Para abordar estos desafíos, se emplearon algoritmos mixtos semi/no paramétricos capaces de manejar múltiples variables con comportamientos y unidades diversos. Sin embargo, el enfoque de resolución de problemas multi-objetivo propuesto en este estudio puede necesitar una validación adicional con covariables adicionales cuando se presentan conjuntos de datos diversos de otros escenarios intensivos en capital.

Además, un estudio dedicado centrado en los desafíos relacionados con la estimación

de parámetros y el cálculo de bandas de covariables es esencial para desarrollar un Weibull PHM robusto e integrado Data driven y que considere múltiples covariables. Por lo tanto, el principal trabajo futuro deberá ser la completa integración de los dos modelos expuestos con el fin de desarrollar una estructura integrada para la definición de políticas de mantenimiento predictivo. Esto requerirá de sensibilización de ambos modelos en conjunto. Además, es necesario estudiar más las diferencias relacionadas a la elección de una técnica de clustering por sobre otra, es decir, K-means en contraste con GMM. Esto último considerando su impacto en la confiabilidad condicional y la RUL.

La función de densidad de probabilidad que entrega el método GMM ofrece oportunidades de investigación futura en la línea de estimar directamente confiabilidades condicionales a partir de estas funciones de densidad, e intentar llegar a un proceso estocástico cuyas realizaciones sean las posibles confiabilidades condicionales que puede tener el equipo en cuestión, por dar un ejemplo.

Mirando hacia el futuro, se sugiere la validación y prueba de la metodología propuesta en diversas aplicaciones industriales más allá de la industria eléctrica, como minería. Adicionalmente, investigaciones futuras podrían explorar la integración de otras técnicas avanzadas impulsadas por datos para mejorar la precisión y confiabilidad de la predicción del Tiempo de Vida Útil Remanente (RUL). Finalmente, explorar distribuciones de probabilidad alternativas, como log-normal o gamma, podrían mejorar aún más la precisión de las estimaciones de la RUL.

Referencias

Ahmad, R., & Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & Industrial Engineering*, *63*(1), 135-149.

Alsyouf, I., Alsuwaidi, M., Hamdan, S., & Shamsuzzaman, M. (2021). Impact of ISO 55000 on organisational performance: evidence from certified UAE firms. *Total Quality Management & Business Excellence*, *32*(1-2), 134-152.

Azar, K., Hajiakhondi-Meybodi, Z., & Naderkhani, F. (2022). Semi-supervised clustering-based method for fault diagnosis and prognosis: A case study. *Reliability Engineering & System Safety*, *222*, 108405.

Banjevic, D., & Jardine, A. (2006). Calculation of reliability function and remaining useful life for a Markov failure time process. *IMA Journal of Management Mathematics*, *17*(2), 115-130.

Biró, T. S., & Néda, Z. (2020). Gintropy: Gini index based generalization of entropy. *Entropy*, *22*(8), 879.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123-140.

Cavallaro, C., Cutello, V., Pavone, M., & Zito, F. (2024). Machine Learning and Genetic Algorithms: A case study on image reconstruction. *Knowledge-Based Systems*, *284*, 111194.

Chen, C., Liu, Y., Wang, S., Sun, X., Di Cairano-Gilfedder, C., Titmus, S., & Syntetos, A. A. (2020). Predictive maintenance using cox proportional hazard deep learning. *Advanced Engineering Informatics*, *44*, 101054.

Chen, L., Shan, W., & Liu, P. (2021). Identification of concrete aggregates using K-means clustering and level set method. *Structures*, *34*, 2069-2076. https://doi.org/10.1016/j.istruc.2021.08.048

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202.

Crespo, A., Gómez, J. F., Martınez-Galán, P., & Guillén, A. (2020). Maintenance management through intelligent asset management platforms (IAMP). Emerging factors, key impact areas and data models. *Energies*, *13*(15), 3762.

Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Computer Science*, *54*, 764-771. https://doi.org/10.1016/j.procs.2015.06.090

El Ouadi, J., Malhene, N., Benhadou, S., & Medromi, H. (2022). Towards a machine-learning based approach for splitting cities in freight logistics context: Benchmarks of clustering and prediction models. *Computers & Industrial Engineering*, *166*, 107975. https://doi.org/10.1016/j.cie.2022.107975

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/10.1016/j.engappai.2022.104743

Fahim, A. (2021). K and starting means for k-means algorithm. *Journal of Computational Science*, *55*, 101445. https://doi.org/10.1016/j.jocs.2021.101445

Ferraz Júnior, F., Romero, R. A. F., & Hsieh, S.-J. (2023). Machine Learning for the Detection and Diagnosis of Anomalies in Applications Driven by Electric Motors. *Sensors*, *23*(24). https://doi.org/10.3390/s23249725

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, *21*, 768-769.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367-378.

Galar, D., & Kans, M. (2017). The impact of maintenance 4.0 and big data analytics within strategic asset management. *Maintenance Performance and Measurement and Management 2016 (MPMM 2016). November 28, Luleå, Sweden*, 96-104.

Godoy, D. R., Álvarez, V., & López-Campos, M. (2023). Optimizing Predictive Maintenance Decisions: Use of Non-Arbitrary Multi-Covariate Bands in a Novel Condition Assessment under a Machine Learning Approach. *Machines*, *11*(4), 418.

Godoy, D. R., Álvarez, V., Mena, R., Viveros, P., & Kristjanpoller, F. (2024). Adopting new Machine Learning approaches on Cox's partial likelihood parameter estimation for predictive maintenance decisions. [Under Review]. *Machines*.

Godoy, D. R., Pascual, R., & Knights, P. (2014). A decision-making framework to integrate maintenance contract conditions with critical spares management. *Reliability Engineering & System Safety*, *131*, 102-108.

Habib, A., Akram, M., & Kahraman, C. (2022). Minimum spanning tree hierarchical clustering algorithm: A new Pythagorean fuzzy similarity measure for the analysis of functional brain networks. *Expert Systems with Applications*, *201*, 117016. https://doi.org/10.1016/j.eswa.2022.117016

Hamdi, M., Hilali-Jaghdam, I., Elnaim, B. M., & Elhag, A. A. (2022). Forecasting and Classification of New Cases of Covid 19 before Vaccination Using Decision Trees and Gaussian Mixture Model. *Alexandria Engineering Journal*.

Hastings, N. A. J. (2015). *Physical asset management: With an introduction to ISO55000*. Springer.

He, Y., Wang, J., Liu, X., Wang, X., & Ouyang, H. (2023). Modelling and solving of knapsack problem with setup based on evolutionary algorithm. *Mathematics and Computers in Simulation*.

Hesabi, H., Nourelfath, M., & Hajji, A. (2022). A deep learning predictive model for selective maintenance optimization. *Reliability Engineering & System Safety*, *219*, 108191.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355-373.

Huang, Y., Englehart, K. B., Hudgins, B., & Chan, A. D. (2005). A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses. *IEEE Transactions on Biomedical Engineering*, *52*(11), 1801-1811.

Jardine, A., Anderson, P., & Mann, D. (1987). Application of the Weibull proportional hazards model to aircraft and marine engine failure data. *Quality and reliability engineering international*, *3*(2), 77-82.

Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, *20*(7), 1483-1510.

Jardine, A. K., & Tsang, A. H. (2005). *Maintenance, replacement, and reliability: theory and applications*. CRC press.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 881-892.

Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia tools and applications*, *80*, 8091-8126.

Klein, J. P., Moeschberger, M. L., Klein, J. P., & Moeschberger, M. L. (2003). Nonparametric estimation of basic quantities for right-censored and left-truncated data. *Survival analysis: Techniques for censored and truncated data*, 91-138.

Lam, J. Y. J., & Banjevic, D. (2015). A myopic policy for optimal inspection scheduling for condition based maintenance. *Reliability Engineering & System Safety*, *144*, 1-11.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*. John Wiley & Sons.

Li, T., Rezaeipanah, A., & Tag El Din, E. M. (2022). An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2022.04.010

Liu, H., & Makis, V. (1996). Cutting-tool reliability assessment in variable machining conditions. *IEEE Transactions on Reliability*, *45*(4), 573-581.

Liu, W., Li, A., Fang, W., Love, P. E., Hartmann, T., & Luo, H. (2023). A hybrid data-driven model for geotechnical reliability analysis. *Reliability Engineering & System Safety*, *231*, 108985.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.

Maletič, D., Maletič, M., Al-Najjar, B., & Gomišček, B. (2020). An analysis of physical asset management core practices and their influence on operational performance. *Sustainability*, *12*(21), 9097.

Mascarenhas, W. F. (2004). The BFGS method with exact line searches fails for non-convex objective functions. *Mathematical Programming*, *99*(1), 49.

Mikhail, M., Ouali, M.-S., & Yacout, S. (2024). A data-driven methodology with a non-parametric reliability method for optimal condition-based maintenance strategies. *Reliability Engineering & System Safety*, *241*, 109668.

Mofolasayo, A., Young, S., Martínez, P., & Ahmad, R. (2022). How to adapt lean practices in SMEs to support Industry 4.0 in manufacturing. *Procedia Computer Science*, *200*, 934-943. https://doi.org/10.1016/j.procs.2022.01.291

Nehring, M., Knights, P., Kizil, M., & Hay, E. (2018). A comparison of strategic mine planning approaches for in-pit crushing and conveying, and truck/shovel systems. *International Journal of Mining Science and Technology*, *28*(2), 205-214.

Newby, M. (1988). Accelerated failure time models for reliability data analysis. *Reliability Engineering & System Safety*, *20*(3), 187-197.

Nguyen, V.-T., Do, P., Vosin, A., & Iung, B. (2022). Artificial-intelligence-based maintenance decision-making and optimization for multi-state component systems. *Reliability Engineering & System Safety*, *228*, 108757.

Paul, D., Su, R., Romain, M., Sébastien, V., Pierre, V., & Isabelle, G. (2017). Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Computerized Medical Imaging and Graphics*, *60*, 42-49.

Peiravi, A., Ardakan, M. A., & Zio, E. (2020). A new Markov-based model for reliability optimization problems with mixed redundancy strategy. *Reliability Engineering & System Safety*, *201*, 106987.

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1994). Rational Function Interpolation and Extrapolation, Numerical Recipes in C.

Rajendra, P., Girisha, A., & Gunavardhana Naidu, T. (2022). Advancement of machine learning in materials science. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2022.04.238

Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X., & Gu, L. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing*, *74*, 634-642.

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. *2002 IEEE international conference on acoustics, speech, and signal processing*, *4*, IV-4072.

Ridgeway, G. (1999). The state of boosting. *Computing science and statistics*, 172-181.

Safaei, N., & Jardine, A. K. (2018). Aircraft routing with generalized maintenance constraints. *Omega*, *80*, 111-122.

Sancho, A., Ribeiro, J., Reis, M., & Martins, F. (2022). Cluster analysis of crude oils with k-means based on their physicochemical properties. *Computers & Chemical Engineering*, *157*, 107633. https://doi.org/10.1016/j.compchemeng.2021.107633

Satten, G. A., & Datta, S. (2001). The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, *55*(3), 207-210.

Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, 1-6.

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of machine learning and artificial intelligence in oil and gas industry. *Petroleum Research*, *6*(4), 379-391. https://doi.org/10.1016/j.ptlrs.2021.05.009

Smith, R. L. (1991). Weibull regression models for reliability data. *Reliability Engineering & System Safety*, *34*(1), 55-76.

Steinley, D., & Brusco, M. J. (2011). Choosing the number of clusters in -means clustering. *Psychological Methods*, *16*(3), 285-297. https://doi.org/10.1037/a0023346

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, *45*(1), 89-103.

Syed, Z., & Lawryshyn, Y. (2020). Multi-criteria decision-making considering risk and uncertainty in physical asset management. *Journal of Loss Prevention in the Process Industries*, *65*, 104064.

Troccoli, E. B., Cerqueira, A. G., Lemos, J. B., & Holz, M. (2022). K-means clustering using principal component analysis to automate label organization in multi-attribute seismic facies analysis. *Journal of Applied Geophysics*, *198*, 104555. https://doi.org/10.1016/j.jappgeo.2022.104555

Vlok, P., Coetzee, J., Banjevic, D., Jardine, A. K., & Makis, V. (2002). Optimal component replacement decisions using vibration monitoring and the proportional-hazards model. *Journal of the operational research society*, *53*(2), 193-202.

Wächter, A., & Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, *106*, 25-57.

Wang, J., Zhao, Z., Liu, G., & Xu, H. (2022). A robust optimization approach of well placement for doublet in heterogeneous geothermal reservoirs using random forest technique and genetic algorithm. *Energy*, *254*, 124427.

Wilson, C. M., Li, K., Sun, Q., Kuan, P. F., & Wang, X. (2021). Fenchel duality of Cox partial likelihood with an application in survival kernel learning. *Artificial intelligence in medicine*, *116*, 102077.

Xinmin, G., Zong'an, X., Jun, Z., Falong, H., Jiangtao, L., ZHANG, H., Shuolong, W., Shenyuan, N., & Ji'er, Z. (2022). An unsupervised clustering method for nuclear magnetic resonance transverse relaxation spectrums based on the Gaussian mixture model and its application. *Petroleum Exploration and Development*, *49*(2), 339-348.

Xu, H., Ma, C., Lian, J., Xu, K., & Chaima, E. (2018). Urban flooding risk assessment based on an integrated k-means cluster algorithm and improved entropy weight method in the region of Haikou, China. *Journal of Hydrology*, *563*, 975-986. https://doi.org/10.1016/j.jhydrol.2018.06.060

Xu, M., Droguett, E. L., Lins, I. D., & das Chagas Moura, M. (2017). On the q-Weibull distribution for reliability applications: An adaptive hybrid artificial bee colony algorithm for parameter estimation. *Reliability Engineering & System Safety*, *158*, 93-105.

Yang, A., Qiu, Q., Zhu, M., Cui, L., Chen, W., & Chen, J. (2022). Condition based maintenance strategy for redundant systems with arbitrary structures using improved reinforcement learning. *Reliability Engineering & System Safety*, 108643.

Zheng, R., Chen, B., & Gu, L. (2020). Condition-based maintenance with dynamic thresholds for a system using the proportional hazards model. *Reliability Engineering & System Safety*, *204*, 107123.

Zheng, R., Najafi, S., & Zhang, Y. (2022). A recursive method for the health assessment of systems using the proportional hazards model. *Reliability Engineering & System Safety*, *221*, 108379.

Zheng, R., Wang, J., & Zhang, Y. (2023). A hybrid repair-replacement policy in the proportional hazards model. *European Journal of Operational Research*, *304*(3), 1011-1021.