UNIVERSIDAD EAFIT

Modelo de predicción funcional de demanda de postes de repuesto en redes de distribución de energía a partir de regresiones kernel

Diana Lisette Arango Cañas

Tesis

Asesor, docente
Santiago Ortiz

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería
Maestría en Ciencias de los Datos y Analítica
Medellín
2024

Resumen

La gestión de repuestos es crucial para la eficiencia operativa en el sector eléctrico, donde la falta de disponibilidad puede generar costos significativos y afectar la continuidad del servicio. Este estudio se centra en la estimación de la demanda de postes de repuesto en una empresa eléctrica colombiana mediante análisis de datos funcionales. Utiliza la regresión kernel para suavizar curvas y obtener una representación precisa de los datos, identificando factores externos e intrínsecos que influyen en la demanda. Se propone un modelo de regresión semi-funcional lineal parcial con variables exógenas para predecir la demanda anual de postes. Los resultados permiten una planificación de compras más eficiente, optimizando la gestión de inventario, reduciendo costos y asegurando la disponibilidad de repuestos.

Palabras Clave: Demanda de postes, Suavizadores no paramétricos, Regresión semifuncional parcial, Pronóstico, Intervalos de predicción..

1. Introducción

La gestión de repuestos es un pilar fundamental para la búsqueda de la eficiencia operativa de las empresas, numerosos estudios se centran en este tema debido al potencial de recuperación económica que representa. Al respecto, en un estudio realizado por Cakir and Canbolat (2008), se demostró que los costos de mantenimiento en una planta industrial son alrededor del 60 % de los costos totales, donde entre el 25 % y el 30 % de estos costos están ocasionados al manejo de repuestos. Las empresas del sector eléctrico en Colombia tienen un gran potencial de mejora en este rubro, debido a que garantizar la prestación del servicio de energía en diferentes geografías del territorio nacional implica tener una amplia diversidad de equipos y, por lo tanto, determinar los repuestos necesarios para dichos equipos adquiere gran relevancia.

Actualmente, una empresa encargada de distribuir la energía en varias regiones de Colombia enfrenta un desafío en su operación: una discrepancia significativa en la estimación anual de postes de repuesto requeridos. Esta situación no solo afecta la relación con los proveedores, al no poder anticipar las necesidades de compra, sino que también impacta la asignación presupuestaria de la empresa, al incurrir en costos adicionales por compras de emergencia o almacenamiento excesivo. La falta de repuestos disponibles también provoca demoras en la reparación, afectando la continuidad del servicio y la satisfacción del cliente. Abordar esta problemática permitirá optimizar la gestión de inventario, fortalecer la colaboración con proveedores, mejorar la planificación y asegurar la disponibilidad de repuestos, fortaleciendo así la cadena de suministro y aumentando la eficiencia operativa.

La necesidad de comprender el comportamiento subyacente de la demanda de postes, eliminando fluctuaciones atípicas y manteniendo una representación precisa, impulsa la aplicación del análisis de datos funcionales. Como lo menciona Wang et al. (2016), en el análisis de datos funcionales, se consideran los datos como la fuente de comprensión de un proceso estocástico subyacente. Por lo tanto, esta metodología permite obtener una estimación funcional continua que describe la demanda de postes a lo largo de un período determinado, a partir de datos discretos. Numerosos estudios han abordado la predicción de la demanda desde el análisis funcional, como se evidencia en Aneiros et al. (2013), Vilar et al. (2018) y Wagner-Muns et al. (2018). Siguiendo esta línea, este trabajo aplica modelos de regresión semi-funcional lineal parcial, incorporando variables exógenas para mejorar la predicción, tal como se exploró en Medina (2022). Además, se investigarán diversos métodos de suavizado de curvas para obtener la representación más precisa de los datos, enriqueciendo así el análisis y la capacidad predictiva del modelo.

Este documento está organizado de la siguiente manera. En la Sección 2 se presentan los métodos de suavizado de datos para obtención de curvas funcionales, incluyendo la selección del ancho de banda. El método de predicción semi-funcional lineal parcial, así como los intervalos de predicción. En la Sección 3 se detallan los datos reales de demanda de postes metálicos de una empresa de distribución de energía del sector eléctrico colombiano, desde el año 2021 al 2023, covariable funcional exógenas como la cantidad de eventos de falla que

se presentan en las redes de distribución de energía y cuya causa de falla haya sido el poste. En la Sección 4 se presentan los resultados de los métodos aplicados y la predicción de la demanda de postes. Finalmente, en la Sección 5 se muestran las conclusiones derivadas de este trabajo.

2. Modelos funcionales

Los modelos funcionales son una herramienta estadística pertinente para el análisis de datos que varían continuamente a lo largo de un dominio, como el tiempo o el espacio (Ramsay and Silverman, 2005). Conceptualmente, los datos funcionales están definidos por funciones continuas, esto permite capturar la naturaleza dinámica y la evolución de los fenómenos estudiados. Por supuesto, en la práctica se tienen comúnmente observaciones de valores discretos, pero esto no modifica la forma de analizarlos, ya que existen técnicas para la creación de una función continua que represente la relación subyacente entre los puntos de datos discretos.

Los modelos funcionales ofrecen diversas estrategias para abordar la resolución de problemas. Un enfoque común es el método lineal, ampliamente utilizado por su familiaridad y facilidad de implementación. Sin embargo, para adaptarse mejor a la complejidad de los fenómenos reales, se ha explorado el enfoque de la estadística no paramétrica, que brinda mayor flexibilidad y capacidad de ajuste a los datos observados. Dada la naturaleza de los datos, detallada en la Sección 3, este estudio adoptará un enfoque no paramétrico para el análisis. El primer paso en esta dirección será el suavizado de los datos, que se abordará en la siguiente subsección.

2.1. Suavizado de datos

La comprensión del ajuste de valores discretos de datos a modelos de datos funcionales se fundamenta en la estimación mediante técnicas como pueden ser mínimos cuadrados o máxima verosimilitud entre otros más, técnicas que establecen un vínculo crucial entre el análisis de datos funcionales y las herramientas del análisis de regresión múltiple (Ramsay and Silverman, 2005). Los métodos de regresión permiten estimar una variable dependiente a partir de una o varias variables independientes. Matemáticamente, si tenemos pares de observaciones $(x_1, y_1), \ldots, (x_n, y_n)$, la relación entre la variable dependiente Y y la variable independiente X se puede expresar mediante la ecuación:

$$y_i = r(x_i) + \epsilon_i \quad \text{donde} \quad E(\epsilon_i) = 0, \quad i = 1, \ldots, n. \tag{1}$$

La regresión lineal, al asumir una relación lineal entre la variable independiente y la dependiente, puede generar problemas e imprecisiones al intentar modelar relaciones más complejas presentes en los datos. Además, los métodos usualmente utilizados para estimar los coeficientes de esta relación, como la estimación de máxima verosimilitud o el método de mínimos cuadrados, se basan en supuestos que no siempre se cumplen en la realidad, como la linealidad, la independencia y la distribución normal de los residuos, y la ausencia

de multicolinealidad. La violación de estos supuestos puede llevar a resultados poco fiables (Peña, 2002).

A diferencia de la regresión paramétrica, la regresión no paramétrica no impone una forma predefinida al predictor. En su lugar, el estimador se construye directamente a partir de la relación observada entre la variable dependiente y la predictora, lo que le confiere una mayor flexibilidad para adaptarse a patrones complejos en los datos.

2.1.1. Regresión kernel

La regresión kernel es una técnica no paramétrica para estimar la esperanza condicional de una variable aleatoria. El objetivo es encontrar la relación entre X y Y, donde la esperanza condicional de una variable Y respecto a una variable X se describe como:

$$\mathrm{E}(Y \mid X) = m(X). \tag{2}$$

Donde $m(\cdot)$ es una función de valor real predefinida. Esta función o estimadores no paramétricos permiten por sus características adaptarse a los datos, con la ventaja adicional de reducir los supuestos distribucionales necesarios. En este contexto se tienen los estimadores de regresión no paramétricos: Nadaraya-Watson (Nadaraya, 1964; Watson, 1964), Priestley-Chao (Priestley and Chao, 1972) y Gasser-Müller (Gasser and Müller, 1979). Estos tres estimadores se exploran en este trabajo y los resultados se presentan en la Sección 4.

Los métodos kernel, ampliamente reconocidos por su capacidad para ponderar localmente los datos, emplean una función kernel $K(X, h)$ simétrica respecto a la vecindad, y no negativa, que asigna pesos a los puntos de datos. En este contexto, la función kernel se centra en cada valor de la variable independiente y asigna pesos a los valores de la variable dependiente cercanos. En la literatura existen diversas propuestas de funciones kernel, entre las que destacan la triangular, gaussiano, coseno, Epanechnikov y tricubo. Debido a su eficiencia y practicidad, el kernel Gaussiano es uno de los más utilizados. Por esta razón, se ha seleccionado para este trabajo el kernel Gaussiano, definido como:

$$k(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2}\right). \tag{3}$$

Otro aspecto fundamental de la regresión kernel es la elección del ancho de banda, comunmente nombrada como h, ya que controla la suavidad de la curva de regresión estimada y, por ende, la capacidad del modelo para capturar patrones relevantes sin sobreajustarse.

Estimador Nadaraya-Watson

El estimador de Nadaraya-Watson es un método de estimación localmente ponderado que asigna mayor relevancia a las observaciones más cercanas al punto de interés. Esto se logra mediante la asignación de pesos a cada observación, donde los pesos disminuyen a medida que aumenta la distancia entre dicha observación y el punto donde se desea estimar la función de regresión. El estimador de Nadaraya-Watson está dado por la Expresioón (4),

donde h es el ancho de banda y se cumple que $h > 0$ y $K(\cdot)$ es una función de kernel. Este estimador se define como:

$$\hat{m}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i \quad \text{donde} \quad l_i(x) = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} K(\frac{x-x_i}{h})}. \tag{4}$$

Si bien el estimador de Nadaraya-Watson ha sido ampliamente utilizado en trabajos relacionados con series de tiempo (Cai, 2001), presenta la desventaja de ser sensible a la presencia de valores atípicos. Esta sensibilidad puede dificultar la obtención de una curva adecuada para la demanda de postes. Por esta razón, en este trabajo exploramos un estimador robusto de Nadaraya-Watson, propuesto por Osorio et al. (2021), con el objetivo de mitigar la variabilidad que tienen los datos expuestos en la Sección 3.

- **Estimador robusto de Nadaraya-Watson:** La distancia de Mahalanobis es una medida de distancia estadística que tiene en cuenta la correlación entre variables, ponderando las diferencias entre puntos en función de la varianza y covarianza de las variables. Por este motivo en Osorio et al. (2021), se propone modificar $l_i(x)$ para incluir una medida que indique que tan extremos son los datos usando la distancia de Mahalanobis al cuadrado de las observaciones w_i=(x_i, Y_i), mediante:

$$MD_R^2(w_i) = (w_i - \hat{\mu}_R)^T \, \hat{S}_R^{-1}(w_i - \hat{\mu}_R), \tag{5}$$

donde $\hat{\mu}_R$ y $\hat{S}_R^{-1}$ son estimadores robustos multivariantes de localización y dispersión. El objetivo de la aplicación de la distancia expresada en la ecuación 5, es establecer una medida de profundidad de Mahalanobis de la forma $M_hD(w_i) = (1 + MD_R^2)^{-1}$. En este sentido, en Osorio et al. (2021) se propone usar $M_hD(w_i)$ como indicador de la atipicidad de los puntos, de ese modo, introducir una función de peso $d(w_i)$ que depende del rango de la medida de profundidad para cada w_i. Formalmente, esta propuesta se define como:

$$d(w_i) = M_hD(w_i) \left(\sum_{j=1}^{n} M_hD(w_j) \right)^{-1}, \tag{6}$$

donde $d(w_i) : R^2 \to [0, 1]$, siguiendo el procedimiento introducido en Liu et al. (1999); de modo que las observaciones con poca profundidad, o gran atipicidad, obtienen pesos pequeños (Stahel, 1981; Donoho and Gasko, 1992). $d(w_i)$ permite definir un nuevo peso para $l_i^*(x)$ basado en una estimación robusta de la profundidad de Mahalanobis. Por lo tanto, aplicando (6) en (4), la propuesta para el estimador robusto de Nadaraya-Watson se define como:

$$\hat{m}^*(x) = \sum_{i=1}^{n} l_i^*(x) Y_i \quad \text{donde} \quad l_i^*(x) = \frac{d(w_i) K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} d(w_j) K(\frac{x-x_i}{h})}. \tag{7}$$

Dado que $d(w_i)$ depende de estimadores robustos de localización y dispersión, se exploran en este trabajo algunos estimadores robustos de dispersión multivariante, estos son: COMEDIAN (Falk, 1997) y Fast-MCD (Rousseeuw and Van Driessen, 1999).

Estimador Priestley-Chao

El estimador de Priestley-Chao es un método no paramétrico utilizado en estadística para la estimación de la función de densidad de probabilidad de una variable aleatoria. Derivado del estimador de regresión de kernel, este estimador se concentra en un diseño aleatorio representado por una variable incondicional distribuida uniformemente. A diferencia de otros métodos que solo modelan la media condicional, el estimador de Priestley-Chao ofrece una visión más completa al estimar la totalidad de la distribución de probabilidad (Konečná, 2018).

En su formulación, el estimador incorpora un término diseñado para capturar la diferencia sucesiva entre observaciones temporales. Esta característica resulta fundamental para estabilizar las estimaciones en situaciones donde las observaciones no se realizan con una frecuencia temporal constante. Al considerar estas diferencias, el modelo se adapta a la variabilidad en los intervalos de tiempo entre mediciones, asegurando una mayor precisión y robustez para estos casos. Se define como:

$$\hat{m}_{PC}(x) = h^{-1} \sum_{i=2}^{n} (x_i - x_{i-1}) K\left(\frac{x - x_i}{h}\right) y_i. \tag{8}$$

Estimador Gasser-Müller

El estimador de Gasser-Müller, al igual que los estimadores mencionados anteriormente, es una herramienta potente en estadística no paramétrica, que proporciona una manera efectiva de estimar funciones de densidad y curvas de regresión sin necesidad de asumir una forma específica para la distribución de los datos. En este estimador, los pesos ponderados se obtienen al calcular la integral de la función kernel $K(\cdot)$, escalada por el ancho de banda h, en el intervalo definido por los puntos medios de los valores adyacentes a cada punto x. El área bajo la curva del núcleo $K(\cdot)$ en este intervalo determina la influencia de cada observación en la estimación final. El estimador se define mediante la siguiente formula:

$$\hat{m}_{GM}(x) = h^{-1} \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} K\left(\frac{x - u}{h}\right) du \right] y_i \quad \text{donde} \quad s_i = \frac{x_{i-1} - x_i}{2}. \tag{9}$$

2.1.2. Ancho de banda

El ancho de banda en la regresión kernel actúa como una ventana móvil que determina qué puntos de datos se consideran para la estimación local en cada punto de la curva de regresión. La elección de este parámetro, siempre mayor a cero, es crucial, ya que un valor elevado implica la inclusión de numerosos puntos en el cálculo de la media ponderada, lo que resulta en una curva de regresión más suave. Si bien esto reduce la varianza de la estimación y disminuye la sensibilidad a valores atípicos, un ancho de banda excesivamente grande puede introducir sesgo y ocultar detalles importantes de la dinámica real del fenómeno subyacente. En el caso opuesto, un valor pequeño para h implica que solo se consideran unos pocos puntos de datos, lo que produce una curva de regresión más ajustada a los datos.

Esto reduce el sesgo de la estimación, lo que significa que la curva se adapta mejor a la forma real de la relación, pero aumenta la varianza y por lo tanto se convierte más sensible a las fluctuaciones de los datos. Se consideraron diferentes métodos para la selección del ancho de banda:

- Validación cruzada: Este método encuentra el ancho de banda óptimo, basado en la subdivisión de los datos y en la elección de aquel valor que minimiza el error de predicción en los datos restantes. Se explora en este trabajo lo realizado por Li and Racine (2004), Racine and Li (2004) y Quintela del Río and Vilar Fernández (1992).
- Función genérica de densidad: Calcula las estimaciones de densidad del kernel. El método por defecto lo hace con el kernel gaussiano y para datos univariados (Sheather and Jones, 1991).

Para comparar y elegir el mejor ancho de banda se utilizan varios índices de medición que se describen en la siguiente sección.

2.1.3. Índices de medición

El enfoque de este trabajo se centra en elegir el modelo que disminuya de mejor manera el error en la predicción, es por esto que para evaluar la precisión de los modelos de predicción y validar los métodos de selección del ancho de banda, se utilizarán métricas de desempeño ampliamente reconocidas, como el error porcentual medio absoluto (MAPE), el error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE). Estas métricas nos permitirán comparar objetivamente el rendimiento de cada modelo y determinar su capacidad para reflejar los datos reales. Dado que el MAE es una métrica intuitiva y fácil de interpretar, se ha elegido como el indicador principal para comunicar los resultados de manera clara y comprensible al usuario final.

- Error absoluto medio (MAE): Es el promedio del valor absoluto de la diferencia entre el valor predicho y el valor observado. Se expresa mediante la fórmula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

- Error porcentual medio absoluto (MAPE): Es el promedio de los errores porcentuales absolutos de las predicciones. Se expresa mediante la fórmula:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{11}$$

- La raíz del error cuadrático medio (RMSE): Es la raíz cuadrada del promedio de la diferencia entre el valor predicho y el valor real elevado al cuadrado. Este indicador tiene la característica de penalizar en mayor medida los errores, ya que los

eleva al cuadrado antes de promediarlos, otorgando así mayor peso a las desviaciones significativas en las predicciones. Se expresa mediante la fórmula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

2.2. Modelo semi-funcional lineal parcial SFLP

Los modelos de regresión, diseñados para descubrir la relación entre una variable dependiente y una o más variables independientes, pueden extenderse al ámbito de los datos funcionales, dando lugar a la regresión funcional. Según las características de las variables y la respuesta, se distinguen tres tipos principales de modelos funcionales:

- Modelo totalmente funcional: tanto la respuesta como los regresores son observaciones funcionales.
- Modelo de respuesta escalar: respuesta escalar y regresores funcionales.
- Modelo de respuesta funcional: respuesta funcional y regresores escalares.

En este trabajo, se empleará un modelo de regresión funcional en el cual tanto la variable predictora como la variable a predecir serán observaciones funcionales. Específicamente, la variable predictora será la curva que representa los eventos con daños de postes a lo largo de las 52 semanas del año, mientras que la variable a predecir será la curva que describe la demanda de postes en el mismo período. Este enfoque fue propuesto en un entorno de curvas independiente en Ferraty and Vieu (2006). Este mismo modelo fue aplicado para predecir las curvas de demanda residual de energía eléctrica en España Aneiros et al. (2013).

Partiendo de la propuesta de Vilar et al. (2018), en la que se implementa un modelo de regresión semifuncional lineal parcial con variables exógenas escalares, este trabajo presenta una extensión del mismo para abordar casos con respuesta funcional y variable exógena funcional. En este tipo de escenarios, resulta intuitivo extender el modelo de regresión mediante la inclusión de un componente lineal con p variables exógenas escalares en la función, el modelo SFLP se construye como:

$$\zeta_{i+1}(t) = X_{t+1}^T \beta_t + m_t(\zeta_i) + \xi_{t,i+1}, \quad i = 1, \ldots, n \tag{13}$$

donde $X_{t+1}^T = (x_{i+1,1}, \ldots, x_{i+1,p})$ es un vector de p variables exógenas funcionales y $\beta_t = (\beta_{t,1}, \ldots, \beta_{t,p})^T$ es un vector de parámetros desconocidos a ser estimado y $\xi_{t,i+1}$ es el error funcional aleatorio con media cero. Se propone los estimadores para β_t y $m_t(\cdot)$ basados en los mínimos cuadrados ordinarios y en suavización por kernel, sus expresiones son:

$$\widehat{\beta}_{t,h} = (\widetilde{\mathrm{X}}_h^T \widetilde{\mathrm{X}}_h)^{-1} \widetilde{\mathrm{X}}_h^T \widetilde{\zeta}_{t,h}$$

y

$$\widehat{m}_{t,h}^{SFLP}(\zeta) = \sum_{i=1}^{n} \omega_h(\zeta, \zeta_i)(\zeta_{i+1}(t) - X_{t+1}^T \widehat{\beta}_{t,h}),$$

Tenga en cuenta que se denota $\widetilde{\mathrm{X}}_h = (I - W_h)X$ y $\widetilde{\zeta}_{t,h} = (I - W_h)\zeta_t$, donde $W_h = (\omega_h(\zeta_i, \zeta_j))_{i+1,j+1}, X = (x_{i+1,j})_{i+1,1\leq j\leq p}$ y $\zeta_t = (\zeta_{i+1}(t))_{i+1}$. Ahora, la predicción de $\zeta_{N+1}(t)$ se obtiene como:

$$\widehat{\zeta}_{N+1}(t) = X_{N+1}^T \widehat{\beta}_{t,h} + \widehat{m}_{t,h}^{SFLP}(\zeta_N)$$

2.2.1. Intervalos de predicción

La estimación de valores futuros de una variable conlleva un grado inherente de imprecisión o incertidumbre en los resultados. Los intervalos de predicción ofrecen un rango estimado probable para una observación futura con un cierto nivel de confianza, lo que reduce la incertidumbre en la predicción, pues en ocasiones las estimaciones puntuales no son suficientes cuando se tienen datos con gran variabilidad. Los intervalos de predicción se centran en estimar una respuesta en lugar del estimador de regresión, lo que significa que contienen tanto la variabilidad derivada de la estimación como el error del modelo.El punto de partida para construir un intervalo de predicción es buscar un intervalo (a, b) tal que, $P((\zeta_{N+1}(t)|\zeta_N) \in (a, b)) = 1 - \alpha$ (Vilar et al., 2018).

Para hallar los intervalos de predicción se implementa en este trabajo el procedimiento bootstrap utilizado en Vilar et al. (2018). Este método propone cálculos iterativos para aproximar las distribuciones de $m_t(\zeta_N) - \widehat{m}_{t,h}^{SFLP}(\zeta_N)$ y $\xi_t|\zeta_N$ que, en la práctica, son desconocidas. Para el modelo SFLP la muestra será S, donde $S = \{(X_i, Y_i), \ldots, (X_n, Y_n)\}$. Basado en la muestra para el modelo SFLP el cual es asumido de la ecuación (13). La variable predictora para $\zeta_{N+1}(t)/\{X_{N+1}, \zeta_N\}$ es $\widehat{\zeta}_{N+1}(t)/\{X_{N+1}, \zeta_N\} = X_{N+1}^T \widehat{\beta}_{t,h} + \widehat{m}_{t,h}^{SFLP}(\zeta_N)$. Ahora la distribución está enfocada en $\zeta_{N+1}(t)$ condicionada por las covariables funcionales no paramétricas X_{N+1} y ζ_N. Se tiene la siguiente descomposición:

$$\begin{aligned}\zeta_{N+1}(t)/\{X_{N+1}, \zeta_N\} = X_{N+1}^T \widehat{\beta}_{t,h} + X_{N+1}^T(\beta - \widehat{\beta}_{t,h}) + \widehat{m}_{t,h}^{SFLP}(\zeta_N) \\ +(m_t(\zeta_N) - \widehat{m}_{t,h}^{SFLP}(\zeta_N)) + (\xi_{t,N+1}|\{X_{N+1}, \zeta_N\})\end{aligned}$$

Como los valores verdaderos de la función de regresión $m(t)$ y vector de parámetros β_t son desconocidos en la práctica, es necesario aproximar $(\beta_t - \widehat{\beta}_{t,h})$ y $(m_t(\zeta_N) - \widehat{m}_{t,h}^{SFLP}(\zeta_N))$ y el término del error $\xi_{t,N+1}|\{X_{N+1}, \zeta_N\}$. El intervalo de predicción bootstrap $(1 - \alpha)$ para $\zeta_{N+1}(t)|\{X_{N+1}, \zeta_N\}$ se construye como:

$I^*_{\{X_{N+1},\zeta_N\},1-\alpha} = (L_t, U_t)$ donde,

$L_t = X_{N+1}^T \widehat{\beta}_{t,h} + \widehat{m}_{t,h}^{SFLP}(\zeta_N) + q^*_{t,\alpha/2}(X_{N+1}, \zeta_N)$

y

$U_t = X_{N+1}^T \widehat{\beta}_{t,h} + \widehat{m}_{t,h}^{SFLP}(\zeta_N) + q^*_{t,1-\alpha/2}(X_{N+1}, \zeta_N)$

los cuantiles bootstrap $q^*_{t,p}(X_{N+1}, \zeta_N)$ son calculados mediante el siguiente algoritmo:

Paso 1: Calcular $\widehat{\beta}_{t,b}$ y $\widehat{m}_{t,b}^{SFLP}(\zeta_i), i = 1, \ldots, n$ sobre el conjunto de datos S

Paso 2: Calcular los residuales $\widehat{\xi}_{t,i+1} = \zeta_{i+1}(t) - X_i^T \widehat{\beta}_{t,b} - \widehat{m}_{t,b}^{SFLP}(\zeta_i)$ donde $i = 1, \ldots, n$.

Paso 3: Aplicar el procedimiento de bootstrap para obtener los errores: Crear n variables aleatorias i.i.d $\xi_1^*, \ldots, \xi_n^*$ a partir de la función de distribución empírica de $(\widehat{\xi}_{1,b} - \widehat{\xi}_b, \ldots, \widehat{\xi}_{n,b} - \widehat{\xi}_b)$, donde $\widehat{\xi}_b = n^{-1} \sum_{i=1}^n \widehat{\xi}_{i,b}$.

Paso 4: Calcular $\zeta_{i+1}^*(t) = X_{i+1}^T \widehat{\beta}_{t,b} + \widehat{m}_{t,b}^{FNP}(\zeta_i) + \xi_i^*, i = 1, \ldots, n$ y los estimadores boostrap $\widehat{\beta}_{t,b}^* = (\tilde{X}_b^T \tilde{X}_b)^{-1} \tilde{X}_b^T \tilde{\zeta}_{t,b}^*$
y
$$\widehat{m}_{t,hb}^{SFLP*}(\zeta_N) = \sum_{i=1}^{n} \omega_h(\zeta_i, \zeta_N)(\zeta_{i+1}^* - X_{i+1}^T \widehat{\beta}_{t,b}^*)$$

Paso 5: Repetir B veces los pasos 3-4, obteniendo B estimaciones $\left\{\widehat{\beta}_{t,b}^{*,r}\right\}_{r=1}^{B}$ y $\left\{\widehat{m}_{t,hb}^{*,r}(\zeta_N)\right\}_{r=1}^{B}$.

Paso 6: Crear B variables aleatorias i.i.d $\tilde{\xi}^1, \ldots, \tilde{\xi}^B$ a partir de la función de distribución empírica de los residuos centrados en el Paso 2. $\tilde{\xi}$ se aproxima al error del modelo.

Paso 7: Calcular el conjunto de errores bootstrap:
$$E_{boot} = \left\{ X^T(\widehat{\beta}_{t,b} - \widehat{\beta}_{t,b}^{*,r}) + (\widehat{m}_{t,b}(\zeta_N) - \widehat{m}_{t,hb}^{*,r}(\zeta_N)) + \tilde{\xi}_{t,r} \right\}_{r=1}^{B}$$

Paso 8: Calcular el cuantil bootstrap $q_{t,p}^*(X_{N+1}, \zeta_N)$ del cuantil de orden p de E_{boot}.

3. Datos

El presente estudio aborda la predicción de la demanda de postes de repuesto en una empresa de distribución de energía del sector eléctrico colombiano, donde se evidencia una problemática de desajuste entre la cantidad de postes requeridos y los planificados. Esta demanda presenta fluctuaciones influenciadas por diversos factores, por un lado, la modernización de infraestructuras obsoletas y eventos climáticos extremos generan picos de demanda a corto plazo, al requerir reemplazos urgentes, mientras que avances tecnológicos y cambios normativos moldean la demanda a largo plazo.

Los postes de energía son estructuras fundamentales en la infraestructura eléctrica, ya que soportan cables y otros equipos necesarios para la distribución de electricidad hasta los consumidores finales. Por ello, este estudio se basa en los datos de necesidades de postes metálicos para todo el sistema de distribución de energía de una empresa del sector eléctrico colombiano, agrupado por semana para los años 2021, 2022 y 2023. Adicionalmente, se tiene información de la variable exógena, daños en postes, la cuál contiene la cantidad de eventos de falla en el sistema de distribución de energía que hayan requerido el cambio de al menos un poste. Se tiene estos datos para los años 2022 y 2023, agrupados por semana.

En este trabajo se tiene como variable dependiente la demanda de postes, para esto se usan los datos del año 2022 como datos de entrenamiento y el año 2023 como datos de testeo, con el fin de corroborar la precisión de la predicción. Para obtener dicha predicción, se tiene como variable predictora los daños en postes para los mismos periodos de tiempo. En el desarrollo de este trabajo, se aplica suavizado de datos para convertir los datos de la variable a predecir y la variable predictora de 52 semanas, a objetos funcionales utilizando diversas técnicas como los estimadores de regresión no paramétricos de Nadaraya-Watson, Priestley-Chao and Gasser-Müller.

4. Resultados

En esta sección, se exponen los resultados derivados de la aplicación de las técnicas descritas en las Sección 2. Considerando que el propósito de este trabajo es pronosticar la demanda de postes para el próximo año, con el fin de facilitar las compras necesarias de estos elementos. Se utilizó un modelo de regresión no paramétrica que incorpora datos de la demanda de postes durante las 52 semanas del año anterior, así como una variable exógena que refleja la cantidad de eventos relacionados con cambios de postes en el sistema de distribución de energía durante ese mismo período. En la implementación de la regresión se exploran tres estimadores no paramétricos descritos en 2.1.1, Nadaraya-Watson, Priestley-Chao y Gasser-Müller, con el objetivo de capturar de la mejor forma el comportamiento de la demanda de postes en una empresa del sector eléctrico colombiano. La curva resultante de la aplicación del suavizado de datos será el insumo para predecir la demanda del año siguiente mediante un modelo de regresión no paramétrica semifuncional lineal parcial, con la incorporación de la variable exógena.

En la Figura 1 se presentan los datos históricos de demanda de postes para los años mencionados. Se observa una gran variabilidad en los datos y la presencia de valores asociados a picos de demanda en todos los años. Además, no se identifica una tendencia ni una estacionalidad marcada en los datos.

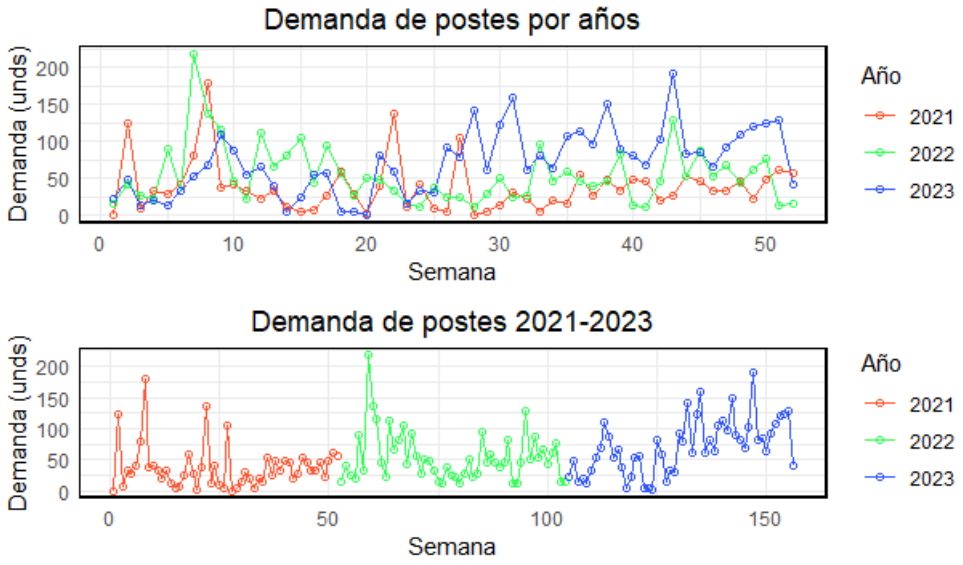

Figura 1: Demanda de postes por años

En la Figura 2 se presentan los datos históricos de eventos de falla con cambio de postes para los años 2022 y 2023. Igualmente estos datos presentan variabilidad y datos atípicos. Adicionalmente, no se identifica una tendencia ni una estacionalidad evidente en los datos.

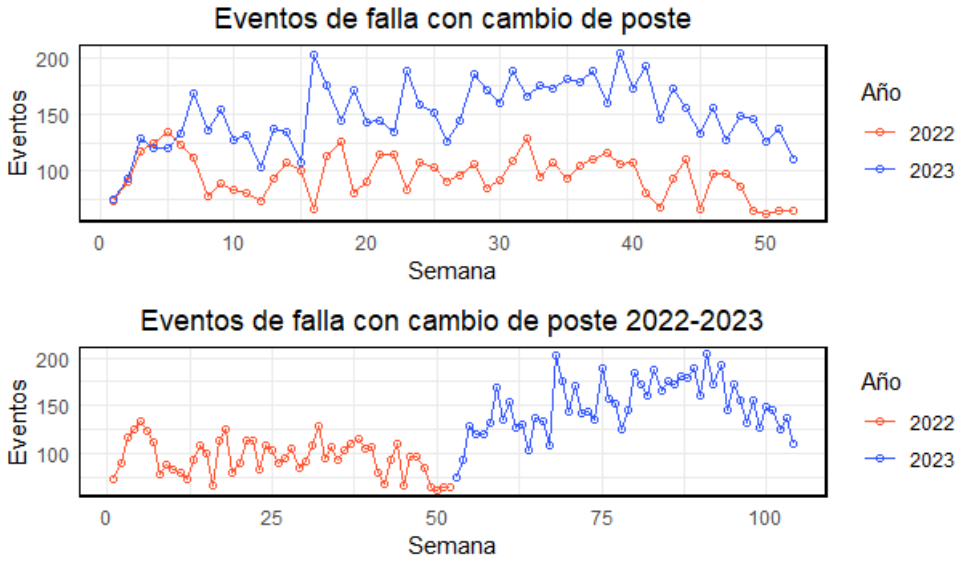

Figura 2: Daños postes por años

La serie temporal de demanda de postes considera observaciones de tiempo discreto de un proceso estocástico de tiempo continuo, $\{\zeta(t)\}_{t \in \mathbb{R}}$, las unidades para t son semanas del año, es decir, $\tau = 52$ para la predicción de la demanda del siguiente año. Para obtener las curvas funcionales de los datos se debe determinar el valor adecuado para el parámetro de

suavizado h. A continuación, se presentan los resultados obtenidos al aplicar los métodos descritos en la Sección 2.1.2 y las herramientas del lenguaje de programación R expuestas en Hayfield and Racine (2008) a los datos de demanda de postes correspondientes al año 2022:

- hnp **y** hnp_{epa}**:** Usando el paquete np de R, específicamente la función npregbw, basada en los trabajos de Racine and Li (2004), Li and Racine (2004) y Hurvich et al. (1998).
- hst**:** Usando la función genérica densidad para calcular las estimaciones de la densidad del kernel.
- hst_{sj}**:** Usando la función $bw.SJ$ en R, basada en el trabajo de Sheather and Jones (1991).
- hnp_{gcv}**:**Usando la función genérica de validación cruzada $np.gcv$ en modelos de regresión no paramétrica.

En la Figura 3, se muestran los resultados de selección del ancho de banda h para la demanda de postes del año 2022. Tal como se había mencionado en la Sección 2.1.2, se observa en la gráfica como al aumentar el valor de h la curva tiene un mayor suavizado, lo que la hace menos sensible a la variabilidad de los datos. De acuerdo a las métricas definidas en la Sección 2.1.3 y a los resultados de la Cuadro 1 se selecciona el ancho de banda $h = 1{,}74$ debido a su menoCuadro valor en los indicadores MAPE, MAE y RMSE. Este ancho de banda seguirá siendo usado para los cálculos de comparación de los estimadores no paramétricos.

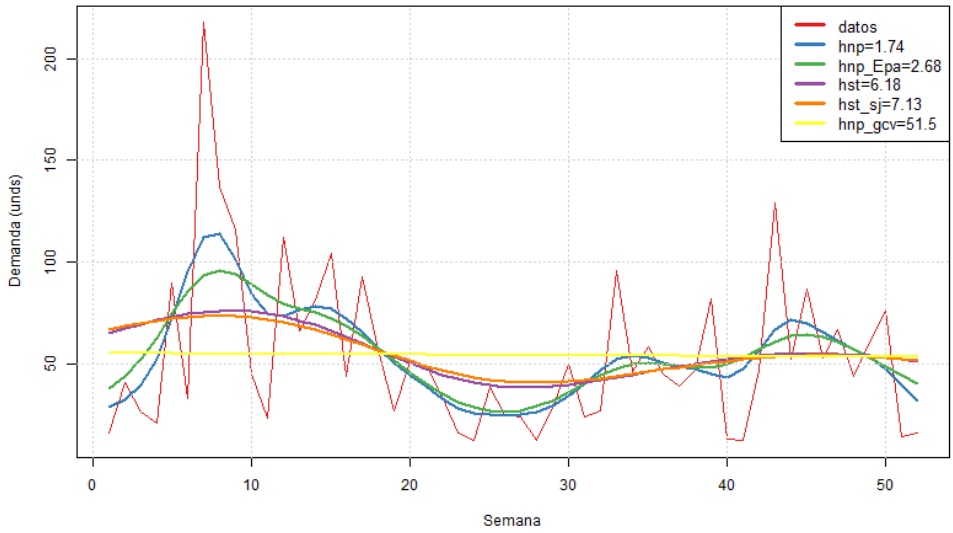

Figura 3: Selección del ancho de banda demanda de postes

Ancho de banda	MAPE	MAE	RMSE
hnp	56.07	20.32	27.86
hnp_epa	64.25	22.36	31.01
hst	78.40	25.74	35.30
hst_sj	80.01	26.19	35.90
hnp_gcv	86.52	28.86	39.06

Cuadro 1: Resultados ancho de banda demanda de postes

Siguiendo el mismo procedimiento, se realizó la selección del ancho de banda para la variable exógena, cantidad de eventos relacionados con cambios de postes en el sistema de distribución de energía durante las 52 semanas del año 2022. Los resultados se observan en la Figura 4. Para estos datos se tiene un resultado particular, el ancho de banda $hnp = 0{,}18$ es tan pequeño que la curva resultante se ajusta casi por completo a los datos originales, lo que provoca que ambas líneas se superpongan en la Figura 2. Esto también se evidencia en la Cuadro 2, donde los resultados de las métricas para este ancho de banda son valores muy pequeños.

Debido a esto, se ha determinado que el ancho de banda óptimo para estos datos es $h = 2{,}97$. Este valor no solo produce métricas aceptables, sino que también se ajusta adecuadamente a la gran variabilidad presente en los datos de eventos de postes sin generar sobreajuste.

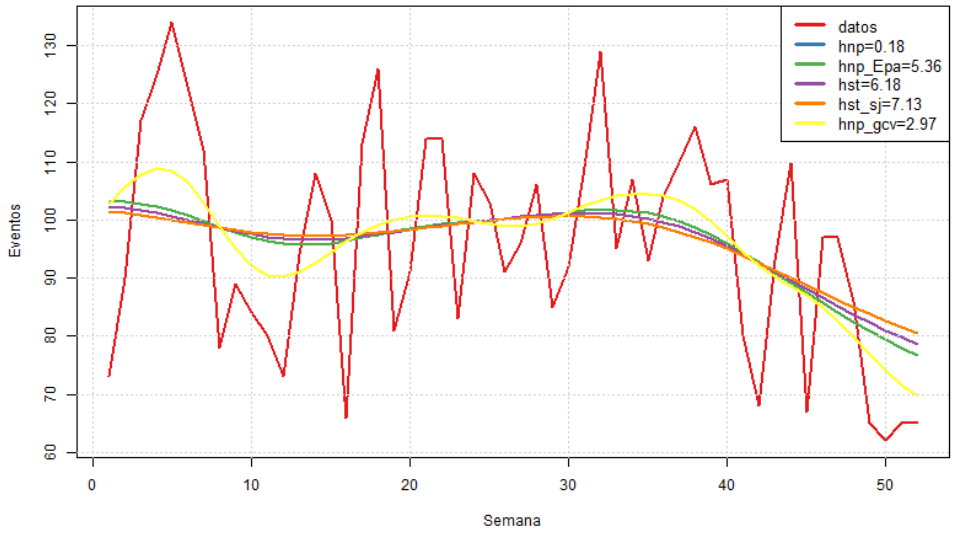

Figura 4: Selección del ancho de banda eventos de postes

Ancho de banda	MAPE	MAE	RMSE
hnp	7.52e-06	6.81e-06	8.91e-06
hnp_epa	15.36	13.89	15.90
hst	15.62	14.10	16.15
hst_sj	15.85	14.28	16.40
hnp_gcv	13.78	12.49	14.46

Cuadro 2: Resultados ancho de banda eventos postes

El rendimiento de los estimadores propuestos en la 2.1.1 se debe identificar en función de la capacidad que tiene de ajustarse a la trayectoria de los datos y reflejar el comportamiento de los mismos. Dado el contexto de negocio, es relevante tener presente al momento de seleccionar el estimador la perdida de valor empresarial en el caso hipotético de presentar un requerimiento de un poste de energía y no se tenga la disponibilidad del mismo. Esta realidad genera influencia en la elección del estimador, donde se debe buscar minimizar la influencia de valores atípicos y al mismo tiempo capturar las variaciones que se presentan en los datos. El rendimiento del estimador de Nadaraya-Watson y los estimadores robustos derivados de su intervención se observan en la Figura 5

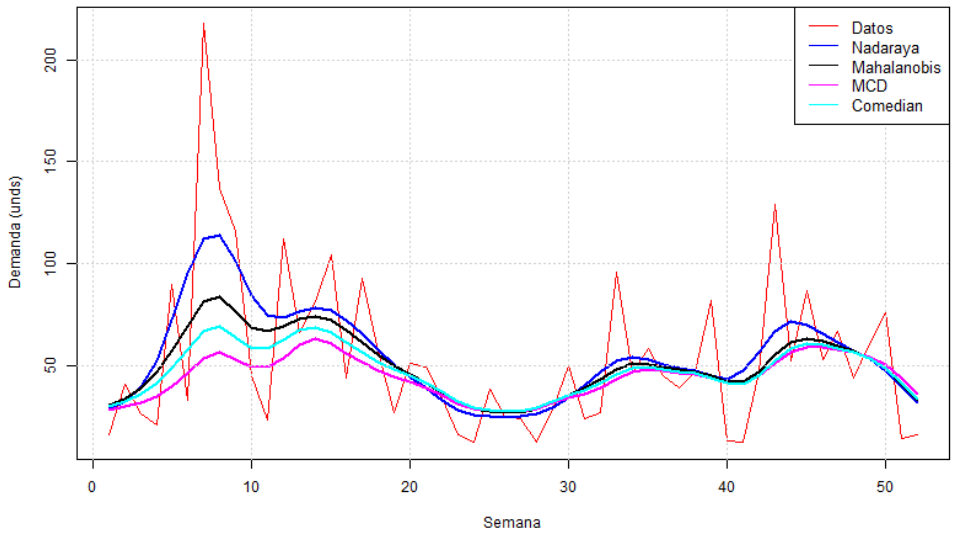

Figura 5: Nadaraya-Watson métodos robustos

Considerando lo expuesto en el párrafo anterior y las métricas de la Cuadro 3, se elige el estimador de Nadaraya-Watson en su formulación original debido a su capacidad para capturar de manera más precisa la variabilidad inherente de la demanda de postes. Los suavizados robustos, en cambio, tienden a aplanar la curva, alejando los datos reales de la curva funcional y perdiendo así información relevante sobre las fluctuaciones de la demanda presentándose así el escenario indeseado mencionado en el parráfo anterior.

Suavizado	MAPE	MAE	RMSE
NW	56.07	20.32	27.86
NW_mahalanobis	54.10	21.30	30.88
NW_MCD	50.46	22.51	35.65
NW_Comedian	52.26	21.87	33.11

Cuadro 3: Resultados Nadaraya-Watson métodos robustos

Finalmente, para definir la forma funcional de la demanda de postes para el año 2022, se compara el estimador de Nadaraya-Watson seleccionado en la etapa anterior con los estimadores de Priestley-Chao y Gasser-Müller. En la Figura 6 se observa como el estimador de Gasser-Müller es el más sensible a la presencia de valores atípicos, lo que genera que la curva intente seguir los datos y por lo tanto un sesgo en la definición de la forma funcional. Los estimadores de Nadaraya-Watson y Priestley-Chao tiene un comportamiento muy semejante, sin embargo al tener este último mejores indicadores de desempeño, se selecciona el mismo para realizar los cálculos restantes en búsqueda de la predicción de la demanda de postes para el año 2023.

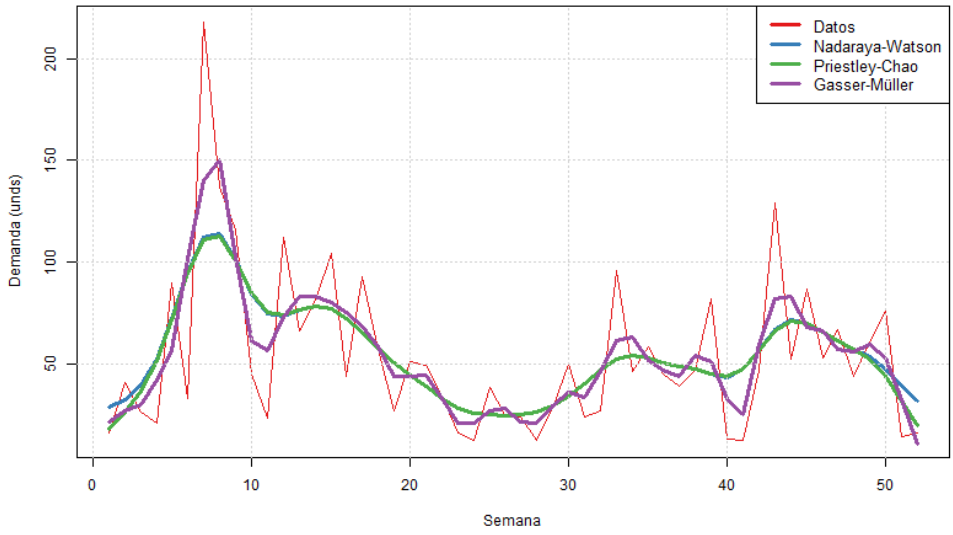

Figura 6: Comparación de estimadores

Estimadores	MAPE	MAE	RMSE
Nadaraya-Watson	56.07	20.32	27.86
Priestley-Chao	52.61	19.95	27.84
Gasser-Müller	41.17	16.95	23.10

Cuadro 4: Resultados estimadores

Se emplea el modelo SFLP presentado en la Sección 2.2 para predecir el comportamiento de la demanda de postes a lo largo de las 52 semanas del año 2023. Este modelo se basa en la curva funcional del comportamiento de la demanda en el año 2022 y en la incorporación de la curva funcional de los eventos de daños de postes como variable exógena durante el mismo período. En Figura 7 se observa como los resultados de la predicción son bastantes satisfactorios, ya que la curva de demanda resultante captura las fluctuaciones de los datos discretos de la demanda de postes para el año 2023, lo que genera una alta certeza en la predicción y por ende una métricas de desempeño con valores pequeños. En el Cuadro 5 se evidencian los resultados con un valor del error medio absoluto (MAE) de 3.68, indicando que en promedio se tiene un error de 4 postes en la predicción de la demanda de postes para las 52 semanas del año siguiente. Esta predicción aporta un gran valor a la empresa al facilitar la planificación de la compra de postes de repuesto y optimizar la gestión de inventario. De esta manera, se evitan gastos innecesarios por almacenamiento ocioso y se garantiza la disponibilidad de los postes necesarios para asegurar la continuidad del servicio de energía.

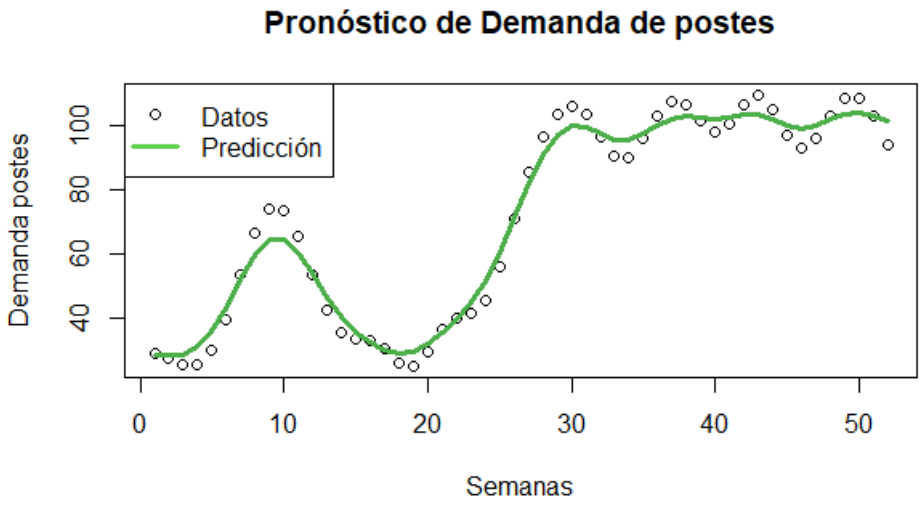

Figura 7: Predicción demanda de postes

Métrica	Valor
MAPE	6.15
MAE	3.68
RMSE	4.36

Cuadro 5: Métricas predicción

Para aumentar la fiabilidad del pronóstico, se incluyó un intervalo de predicción según lo definido en la Sección 2.2.1. Este intervalo, con un margen del 90 %, abarca todos los datos reales de la demanda de postes para el año 2023, como se observa en la Figura 8.

Esta cobertura total de los datos reales aumenta la confianza en los resultados del modelo. Sin embargo, es importante destacar que, al ser un intervalo teórico, puede presentar discrepancias con la realidad del negocio.

En este caso particular, se observan valores negativos de demanda de postes para algunas semanas, lo cual es evidentemente incoherente en un contexto real, luego es suficiente con asumir un valor de cero en estos casos. Adicionalmente, las diferencias entre el valor máximo del intervalo y los datos reales en algunos puntos de la curva pueden llegar a ser de alrededor de 50 postes, una discrepancia inaceptable para la operación de la empresa debido al aumento del inventario de repuestos y los consiguientes gastos de almacenamiento. Estos resultados sugieren la necesidad de proponer un intervalo de predicción más ajustado. A pesar de sus limitaciones, el intervalo de predicción sigue siendo una herramienta valiosa para la toma de decisiones, siempre y cuando se interprete con cautela y se consideren sus posibles limitaciones.

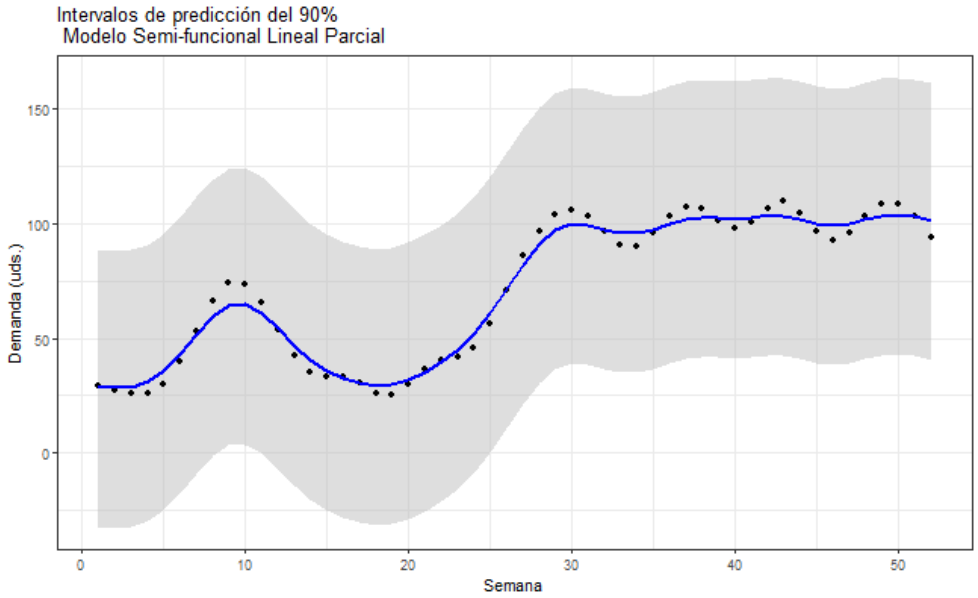

Figura 8: Intervalo de predicción

La capacidad de proyectar con precisión las necesidades de repuestos a lo largo del año, teniendo en cuenta las fluctuaciones y tendencias, permite una planificación más eficaz tanto desde el punto de vista logístico como económico. La demanda de postes entre los años 2021 y 2023, como se observa en la Figura 1, presenta gran variabilidad en los datos, con semanas donde se requirieron más de 200 postes metálicos de repuesto y otras donde esta necesidad no superaba las 10 unidades. Implementar modelos de datos funcionales que capturan con exactitud el comportamiento de la demanda, ignorando anomalías que podrían causar picos o caídas representativas influenciadas por factores externos como cambios climáticos o falta de repuestos, es crucial. Estos factores, al analizar los datos históricos, se evidencian como datos atípicos que no caracterizan la realidad a lo largo del año y han dificultado generar

estimaciones precisas en los presupuestos anuales de la empresa.

Al abstraer la curva de demanda de eventos atípicos, se obtiene una visión más clara y consistente del consumo de postes, lo que permite generar predicciones más precisas. En el desarrollo de este trabajo, se logró una desviación promedio de tan solo cuatro unidades en la predicción semanal para las 52 semanas del año 2023. Esta precisión representa una mejora sustancial en comparación con la diferencia actual entre la cantidad de postes planificados y los realmente requeridos por la empresa. Considerando que estos elementos se compran por miles cada año, una desviación de cuatro unidades en la predicción semanal es un resultado sumamente valioso, ya que permite optimizar los recursos y evitar costos innecesarios por exceso o falta de stock.

5. Conclusión

En este trabajo de grado, se desarrolló un modelo de predicción funcional de la demanda de postes de repuesto en redes de distribución de una empresa del sector eléctrico en Colombia. El modelo implementado se basa en estimadores kernel, los cuales permiten obtener una representación fiel de la curva de demanda a partir de datos discretos. La metodología empleada consideró variables exógenas y aplicó técnicas avanzadas de suavizado de curvas para mejorar la precisión y confiabilidad de las predicciones. Se exploraron diversos estimadores no paramétricos, comparándolos con los métodos robustos de Nadaraya-Watson, que suelen ser los más utilizados.

Los hallazgos del estudio revelan que el modelo de regresión no paramétrica semifuncional lineal parcial es efectivo para predecir la demanda de postes metálicos, incluso considerando la variabilidad y los eventos atípicos en los datos históricos. La selección adecuada del ancho de banda fue crucial para el suavizado de datos, lo que permitió una representación más fiel de la demanda y los eventos de falla. Estadísticamente, los resultados muestran un ajuste a los intervalos de predicción y por lo tanto se concluye que el modelo puede prever las necesidades de postes de repuesto, permitiendo a la empresa eléctrica planificar con mayor eficiencia sus operaciones y recursos, mejorando así la gestión de la cadena de suministro.

En cuanto a consideraciones futuras, este trabajo deja abiertas varias áreas para exploración adicional. En primer lugar, es recomendable incluir en estudios futuros variables exógenas adicionales como la expansión de las redes de energía que implica un crecimiento en la cantidad de postes instalados y factores climáticos, que pueden afectar la demanda de postes. Además, la inclusión de modelos de confiabilidad o resultados de cálculos de condición del activo permitiría comprender las tasas de deterioro y, por ende, el fin de la vida útil de los postes. Esta información, como variable exógena, contribuiría significativamente a la predicción de la cantidad de postes requeridos. Finalmente, este modelo podría extenderse a otras referencias de postes utilizados en la distribución de energía eléctrica, como postes de madera o fibra de vidrio, para validar aún más los resultados obtenidos y

ampliar su aplicabilidad.

Referencias

Aneiros, G., Vilar, J. M., Cao, R., and Muñoz San Roque, A. (2013), "Functional Prediction for the Residual Demand in Electricity Spot Markets," *IEEE Transactions on Power Systems*, 28, 4201–4208.

Cai, Z. (2001), "Weighted Nadaraya–Watson regression estimation," *Statistics Probability Letters*, 51, 307–318.

Cakir, O. and Canbolat, M. S. (2008), "A web-based decision support system for multi-criteria inventory classification using fuzzy AHP methodology," *Expert Systems with Applications*, 35, 1367–1378.

Donoho, D. L. and Gasko, M. (1992), "Breakdown properties of location estimates based on halfspace depth and projected outlyingness," *Ann. Stat.*, 20, 1803–1827.

Falk, M. (1997), "On mad and comedians," *Annals of the Institute of Statistical Mathematics*, 49, 615–644.

Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis*, Springer Series in Statistics, New York, NY: Springer, 2006 edition.

Gasser, T. and Müller, H. G. (1979), *Smoothing Techniques for Curve Estimation*, Heidelberg: Springer.

Hayfield, T. and Racine, J. S. (2008), "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Hurvich, C. M., Simonoff, J. S., and Tsai, C. L. (1998), "Smoothing Parameter Selection in Nonparametric Regression using an improved Akaike information criterion," *Journal of the Royal Statistical Society Series B*, 60, 271–293.

Konečná, K. (2018), "The Priestley-Chao estimator of conditional density with uniformly distributed random design," *Statistika*, 98, 283 – 294.

Li, Q. and Racine, J. (2004), "Cross-validated local linear nonparametric regression," *Statistica Sinica*, 14, 485 – 512. Cited by: 294.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999), "Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh)," *Ann. Stat.*, 27, 783–858.

Medina, M. (2022), *Predicción de tipo funcional para la demanda de productos de maquinaria agrícola a partir de suavizados robustos Nadaraya-Watson*, Master's thesis, Universidad EAFIT.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory Probab. Appl.*, 9, 141–142.

Osorio, P., Ortiz, S., and Laniado, H. (2021), "Weighted Nadaraya-Watson Kernel Regression Based on a Robust Mahalanobis-Depth Estimation," *Este artículo pertenece a las Memorias del Simposio Internacional de Estadística XXX Versión Evento-Virtual Septiembre 21 a 24 de 2021*, 272–277.

Peña, D. (2002), *Análisis de datos multivariante*, McGraw-Hill Interamericana de España S.L.

Priestley, M. B. and Chao, M. T. (1972), "Non-parametric function fitting," *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 385–392.

Quintela del Río, A. and Vilar Fernández, J. (1992), "A local cross-validation algorithm for dependent data," *Test*, 1, 123 – 153. Cited by: 9.

Racine, J. S. and Li, Q. (2004), "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99–130.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer Series in Statistics, New York, NY: Springer, 2 edition.

Rousseeuw, P. J. and Van Driessen, K. (1999), "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, 41, 212–223.

Sheather, S. J. and Jones, M. C. (1991), "A reliable data-based bandwidth selection method for kernel density estimation," *J. R. Stat. Soc.*, 53, 683–690.

Stahel, W. A. (1981), *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*, Ph.D. thesis, Universität Zürich.

Vilar, J., Aneiros, G., and Raña, P. (2018), "Prediction intervals for electricity demand and price using functional data," *International Journal of Electrical Power Energy Systems*, 96, 457–472.

Wagner-Muns, I. M., Guardiola, I. G., Samaranayke, V., and Kayani, W. I. (2018), "A Functional Data Analysis Approach to Traffic Volume Forecasting," *IEEE Transactions on Intelligent Transportation Systems*, 19, 878 – 888. Cited by: 51.

Wang, J. L., Chiou, J. M., and Müller, H. G. (2016), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhya, Ser. A*, 26, 359–372.