vol. 20, n.° 2, julio-diciembre de 2022
ISSN impreso: 1657-6527 • ISSN en línea: 2145-8502

FUENTES
EL REVENTÓN ENERGÉTICO

Metodología para la caracterización energética de procesos industriales basada en modelos de regresión bayesianos. Caso de implementación

Methodology for energetic characterization of industrial processes based on bayesian regression models. Implementation case study

Carlos Jeyson Camargo Fiorillo[1]; Carlos Humberto García Rincón[2]; Gustavo Andrés Valle Tamayo[1]

[1]Ecopetrol S.A., Colombia. Correo electrónico: gustavo.valleta@ecopetrol.com.co; carlosje.camargofi@ecopetrol.com.co

[2]Universidad del Atlántico, Colombia. Correo electrónico: chgarcia@mail.uniatlantico.edu.co

Recibido: 23 de junio, 2022. Aceptado: 23 de junio, 2022. Versión final: 18 agosto, 2022.

RESUMEN

Este trabajo presenta el diseño y desarrollo de una novedosa metodología basada en técnicas estadísticas, que permite realizar una caracterización energética de procesos industriales, cumpliendo con los lineamientos de la norma internacional NTC ISO 50001:2019, cuya implementación es recomendada por la Unidad de Planeación Minero-Energética en la versión 2022-2030 del Plan Indicativo del Programa de Uso Racional y Eficiente de Energía en Colombia. La norma ISO 50001 exige tener una referencia cuantitativa (línea de base energética) del desempeño energético en los procesos que se propone calcular a través de modelos de regresión bayesianos. Además, esta metodología también permite identificar las variables o eventos que tienen mayor relevancia en su eficiencia energética, para en una etapa posterior implementar control sobre ellos y, de este modo, controlar el desempeño energético del proceso mediante la manipulación de estas variables.

Palabras claves: eficiencia energética, sistemas de bombeo, regresión bayesiana, selección de modelos.

Como citar: Camargo Fiorillo, C., García Rincón, C. & Valle Tamayo, G. (2022). Metodología para la caracterización energética de procesos industriales basada en modelos de regresión bayesianos. Caso de implementación. Revista Fuentes: El Reventón Energético, 20(2), 7-22. https://doi.org/10.18273/revfue.v20n2-2022002

ABSTRACT

This work presents the design and development of a novel methodology based on statistical techniques, which allows to perform an energy characterization of industrial processes in compliance with the guidelines of the international standard NTC ISO 50001:2019, its implementation is recommended by the Mining-Energy Planning Unit. in the 2022-2030 version of the Indicative Plan of the Program for the Rational and Efficient Use of Energy in Colombia. The ISO 50001 standard requires having a quantitative reference (energy baseline) for the energy performance of the process. This paper proposes to calculate the energy baseline through Bayesian regression models. This methodology also allows to identify the variables or events that have greater relevance in the energy efficiency of the process, to implement control over them at a later stage and thus improve the energy performance of the process by manipulating these variables.

Key words: Energy efficiency, Pumping systems, Bayesian Regression, Model Selection.

1. Introducción

La eficiencia energética definida por la norma (NTC-ISO 50001, 2019) es la proporción u otra relación cuantitativa entre un resultado de desempeño, servicio, producto, materia prima o energía y una entrada de energía. Esta es de vital importancia en procesos industriales debido a los altos consumos energéticos, por esto se busca minimizarlos manteniendo las mismas condiciones en los productos finales. Optimizar los consumos energéticos tiene un impacto económico y ambiental sobre los procesos industriales que se alinean a las metas nacionales e internacionales de reducción de emisiones de CO_2 (Minambiente, 2020).

Para establecer sistemas de gestión de la energía, cuya finalidad es obtener mejoras en la eficiencia energética en procesos industriales, es común seguir los requisitos definidos por la norma internacional (NTC-ISO 50001, 2019), donde dos de los principales requisitos corresponden a establecer una línea de base energética e identificar las variables relevantes a la eficiencia energética. Una línea de base energética es una referencia cuantitativa que proporciona la base para la comparación de la eficiencia energética, esta permite establecer ahorros o sobreconsumos de energía en un periodo de seguimiento respecto a un periodo de referencia. Una variable relevante es un factor cuantificable que impacta en forma significativa en la eficiencia energética y cambia de forma rutinaria. Los cambios en estas variables explican los ahorros y sobreconsumos de energía cuantificados con la ayuda de la línea base (NTC-ISO 50001, 2019).

Existen diferentes formas de construir una línea de base energética, autores internacionales como Chen y Therkelsen (2019) establecen el uso de modelos de regresión lineal simple para establecer un modelo que predice la energía que se debería consumir en función de una variable productiva; para validar este modelo, los autores verifican la significancia estadística del modelo (p-valor) y el R^2 obtenido. Otros autores nacionales como Caicedo et al. (2019) utilizan modelos de regresión lineal simple o multivariable y definen la significancia estadística del modelo y la selección de variables relevantes basados en el p-valor, sin validar los supuestos del modelo.

Este trabajo propone una metodología para la construcción de líneas de base energética y selección de variables relevantes a la eficiencia energética mediante el uso de regresiones lineales con un enfoque bayesiano, en donde se conoce todo el espacio de modelos posibles**.** Esta metodología propuesta se encuentra enmarcada en la norma NTC ISO 50001:2019, que desde el año 2022 es recomendada por el Plan de Acción Indicativo de Eficiencia Energética (PAI PROURE) 2022 -2030, como un método para la optimización energética del consumo en la industria de los hidrocarburos en Colombia (PAI PROURE, 2022). El criterio de información BIC (*bayesian information criteria*) es utilizado para calcular las probabilidades *a posteriori* de cada modelo (Beier *et al*., 2001), asumiendo una distribución uniforme de probabilidad para cada variable y una distribución normal gaussiana multivariable *a priori* para el modelo.

Se presenta un caso de estudio de un sistema de bombeo para inyección de agua, donde el mejor modelo para la línea de base energética es seleccionado utilizando el factor de Bayes, que es una relación entre las verosimilitudes marginales calculadas a través de BIC (Hastie *et al*., 2008). El modelo seleccionado incluye la variable energía hidráulica, y posteriormente se presenta la selección de variables relevantes a la eficiencia energética obteniendo variables como recirculación y presión de succión, las cuales son típicas de estos sistemas.

2. Metodología para la caracterización energética

Esta sección presenta cada uno de los pasos por seguir en la metodología para la caracterización energética explicando las técnicas, conceptos y formulaciones estadísticas utilizadas para desarrollarlos.

La presente metodología comprende tres grandes grupos que cumplen con los requisitos de la norma NTC ISO 50001(NTC-ISO 50001, 2019): i) numeral 6.3, que exige las variables relevantes y la medición del desempeño actual; y ii) numeral 6.5, establece que la organización debe determinar una o varias líneas base para los usos significativos de energía.

2.1. Análisis previo y preparación de datos

El grupo de actividades tiene como objetivo principal preparar los datos para la realización de la línea de base energética y la identificación de variables relevantes a la eficiencia energética. Comprende los siguientes pasos:

A. Selección del periodo base

Es el periodo que se utilizará de referencia para la comparación de la eficiencia energética del proceso objeto de estudio. Este periodo debe incluir ciclos normales de operaciones típicas, no haber presentado grandes cambios tecnológicos, energéticos u operativos y no debe contener eventos intermitentes e irregulares (Caicedo *et al.*, 2019). Durante este periodo se debe garantizar una adecuada calidad de medición y almacenamiento de datos.

B. Definición de variables de ingeniería

La selección de las variables de ingeniería se realiza en función de la naturaleza del sistema. Estas se definen como todas aquellas que tengan alguna influencia física sobre el comportamiento del proceso, por lo tanto, podrían potencialmente afectar el desempeño energético. Este paso se basa en criterios técnicos de ingeniería para identificar variables que teóricamente tienen relación con el consumo de energía y el desempeño energético, y que son posibles candidatas para participar en la construcción de la línea base o ser variables relevantes en el proceso de estudio. Lo anterior indica que la selección inicial de variables de ingeniería es una entrada que permite determinar el universo de estas, que será utilizado para establecer las variables relevantes como las define el numeral 6.6 de la norma NTC ISO 50001(NTC-ISO 50001, 2019).

C. Recopilación de información histórica

Corresponde a la descarga de los registros históricos del periodo base de cada una de las variables de ingeniería definidas, con una frecuencia establecida según disponibilidad de datos o interés del estudio energético (segundos, minutos, horas, días, etc.).

D. Selección del tamaño mínimo de la muestra

Es la estimación inicial del tamaño mínimo de muestra necesario para representar de manera confiable el comportamiento del proceso. Para definir este tamaño se propone utilizar los métodos de tamaño de efecto y potencia estadística, en los cuales se utiliza una muestra piloto, tomada de manera aleatoria de la recopilación de datos históricos, para calcular el coeficiente de correlación R^2 ajustado. Este coeficiente se utiliza en estos métodos como tamaño del efecto (Ellis, 2010).

El cálculo del número mínimo de muestra para un modelo de regresión lineal simple o múltiple que sea capaz de identificar un tamaño de efecto dado por el R^2 de un grupo de variables B en una variable Y, se realiza mediante la siguiente ecuación (Cohen, 1988):

$$N = \frac{\lambda(1 - R^2_{Y \cdot B})}{R^2_{Y \cdot B}} \qquad (1)$$

es un factor que depende del tamaño del efecto, el número de variables y los grados de libertad del número de observaciones.

$$\lambda = \frac{(1 - R^2_{Y \cdot B})(u + v + 1)}{R^2_{Y \cdot B}} \qquad (2)$$

Debido a que λ depende de N, se deben usar las tablas de λ para una prueba de Fisher dadas por Cohen (1988), para determinar un número mínimo de muestra que cubra el tamaño de efecto y potencia deseada.

E. Filtrado de datos

Se deben filtrar los datos con comportamientos atípicos y sacarlos de la muestra, para esto se realizan dos tipos de filtrados:

- **Físico:** empleado para identificar datos faltantes, valores nulos, errores en la obtención de información (*e. g.* fallas en la comunicación del sistema; remisión de valores "override"; lecturas no adecuadas del sensor, tipo negativos/ letras). También se deben tener en cuenta criterios termodinámicos y balances de energía (*e. g.* eficiencias mayores al 100 %, presiones de vacío en descarga de bombas, balance de energías negativos); para identificar datos que no representan el comportamiento real del proceso.

- **Estadístico:** se propone emplear un método multivariado para la identificación de valores atípicos. Autores como De Maesschalck et al. (2000) utilizan la distancia de Mahalanobis para esta aplicación en el área de minería de datos y construcción de algoritmos como clusterización y clasificación. La distancia de Mahalanobis es una distancia entre las observaciones que puede ser calculada en el espacio real o en los componentes principales y que tiene en cuenta la correlación existente en la matriz de datos (Li *et al*., 2019).

La distancia de Mahalanobis se diferencia de la distancia euclidiana clásica debido a que tiene en cuenta la correlación entre las variables, esto al considerar una relación de la varianza y la covarianza de un conjunto de datos p-dimensional $x = (x_1, x_2, x_3, \dots, x_p)^T$, con centroide en $\mu = (\mu_1, \mu_2, \mu_3, \dots, \mu_p)^T$.

La ecuación siguiente (Cohen, 1988) ilustra la forma en que se calcula la distancia de Mahalanobis, $D_M(x)$, donde Σ es la matriz de covarianza.

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \qquad (3)$$

Una vez se tiene calculada la distancia de Mahalanobis, es posible definir una elipse de confianza con una distribución chi-cuadrado con p grados de libertad, así se define el criterio para clasificar un valor como atípico cuando la distancia sea mayor que el valor crítico de la distribución chi-cuadrado, como se muestra a continuación:

$$d > {\chi_p}^2(1-\alpha) \qquad (4)$$

La figura 1 muestra un ejemplo de la identificación de datos atípicos mediante el uso de la distancia de Mahalanobis, donde se observan dos gamas de colores, amarillo para datos atípicos y azul para datos típicos de operación.

F. Estadística descriptiva

Se debe realizar un reconocimiento de la información a utilizar mediante la elaboración de estadísticos descriptivos de las variables disponibles en el proceso, como son los valores máximos, mínimo, media, mediana, rangos intercuartílicos, desviaciones estándar, entre otros. Al describir cada variable con estos estadísticos se tiene una idea global del comportamiento que tuvo cada variable durante el periodo base. Al cruzar estos valores con el conocimiento ingenieril de cada variable, se obtienen unas primeras conclusiones sobre el comportamiento energético del proceso de estudio.

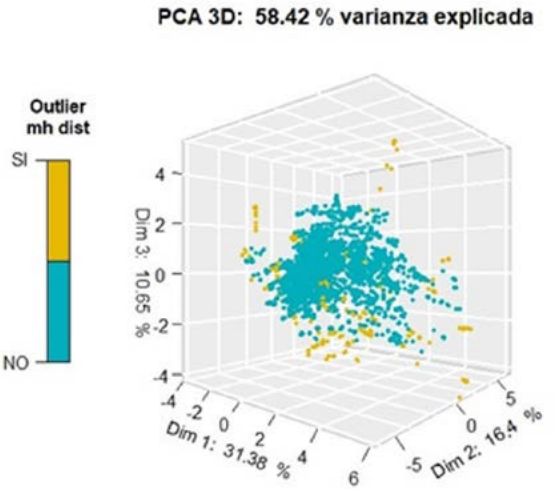


Figura 1. Representación gráfica de *outliers* identificados con la distancia de Mahalanobis.

2.2. Determinación de la línea de base energética

Para determinar la línea de base energética se proponen los siguientes pasos:

A. Definición de variables de ingeniería para el modelo de línea de base energética

Las variables que van a participar en la elaboración del modelo corresponden a variables relevantes que no son controlables por la operación y mantenimiento, y se deben normalizar en el modelo para poder realizar la comparativa del desempeño energético (Campos, 2017).

B. Cálculo y selección del mejor modelo

Una vez se tiene la información preparada y las variables de ingeniería seleccionadas, se realiza una regresión bayesiana con el objetivo de conocer todos los posibles modelos a construir basados en las diferentes combinaciones de variables disponibles. Para elegir el mejor modelo, se utiliza el factor de Bayes, que compara las verosimilitudes marginales de cada

modelo y sus probabilidades "*a posteriori*", estas son calculadas a partir del criterio de información BIC, asumiendo una distribución de Gauss multivariada "*a priori*" (Hastie *et al*., 2008).

El BIC es un criterio de información similar al AIC (*akaike information criteria*), los cuales son muy útiles para realizar selección de modelos bajo un supuesto de máxima verosimilitud. El AIC busca elegir el conjunto de datos con la mayor probabilidad de explicar el modelo, penalizando el número de variables utilizadas para evitar sesgo por un alto número de predictores. Este puede ser calculado de la siguiente forma:

$$AIC = -2 \cdot loglik + 2d \tag{5}$$

Donde *lik* es un estimador de máxima verosimilitud y es el número de variables.

Paralelamente, se puede realizar una aproximación bayesiana de este criterio llamada BIC, el cual es proporcional al AIC, pero con una mayor penalización a modelos complejos (Hastie *et al*., 2008); este puede ser calculado de la siguiente manera:

$$BIC = -2 \cdot loglik + \log(N) \cdot d \tag{6}$$

Donde N es el número de datos.

Para el caso de modelos lineales con parámetros obtenidos por mínimos cuadrados, el BIC puede ser aproximado de la siguiente manera:

$$BIC = \sum_i \frac{\left(y_i - \hat{f}(x_i)\right)^2}{\sigma_\varepsilon^2} + \log(N) \cdot d \tag{7}$$

El principal diferenciador entre AIC y BIC es que a partir del término de BIC se pueden estimar las probabilidades *a posteriori* de cada modelo (dentro de un marco de análisis bayesiano, y se calculan de la siguiente manera:

$$\Pr(\mathcal{M}_m|Z) = \frac{e^{-\frac{1}{2}BIC_m}}{\sum_{l=1}^{M} e^{-\frac{1}{2}BIC_l}} \tag{8}$$

Para calificar la bondad del modelo es de común uso el factor de Bayes, que consiste en comparar la probabilidad marginal de un modelo de prueba y un modelo de contraste. El factor de Bayes se calcula de la siguiente manera:

$$BF(Z) = \frac{\Pr(Z|\mathcal{M}_m)}{\Pr(Z|\mathcal{M}_l)} \tag{9}$$

Teniendo en cuenta que puede ser aproximado a BIC (Hastie *et al*., 2008), y que los modelos se asumen con igual probabilidad, es decir, una distribución de Bernoulli; el factor de Bayes sería:

$$BF = \frac{e^{-\frac{1}{2}\sum_i \frac{\left(y_i - \hat{f}(x_i)\right)^2}{\sigma_\varepsilon^2}}}{e^{-\frac{1}{2}\sum_j \frac{\left(y_j - \hat{f}(x_j)\right)^2}{\sigma_\varepsilon^2}}} \tag{10}$$

A través del factor de Bayes se puede determinar cuál será el mejor modelo para la línea de base energética, usando criterios para la selección como el de Jeffreys (Nagin, 1999), donde se busca que el factor de Bayes, en esta aplicación, sea igual a 1, ya que se compara cada modelo con el modelo de mayor verosimilitud.

Con el mejor modelo seleccionado, se realiza una regresión de mínimos cuadrados (OLS) y se realiza el análisis de varianza para el modelo de regresión (Walpole *et al*., 2012). Resultado de este modelo lineal, se obtendrá la línea de base energética que tendrá la siguiente forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{11}$$

En caso de obtener un modelo multivariable, se deben utilizar los coeficientes de la regresión para obtener una variable equivalente con la que se construirá un modelo de regresión lineal simple equivalente.

C. Filtrado de residuos y cálculo de línea de base energética

Los residuos de este modelo de línea de base energética corresponden a las desviaciones del consumo de energía real y el valor esperado de consumo de energía, lo que significa que estos nos sirven para evaluar el desempeño energético. Al momento de correlacionar las variables operacionales con estas desviaciones en los próximos pasos, se busca que estas estén lo más centradas posibles para aumentar la probabilidad de que las variables realmente expliquen su comportamiento. Por esta razón, se debe realizar un filtrado intercuartil de los residuos de la regresión ajustando el factor n con valores entre 1 y 2 para analizar las desviaciones de energía típicas.

D. Validación de potencia de modelo y supuestos

La potencia estadística se valida con los datos finales del modelo seleccionado, donde se calcula el R^2, el número de datos y variables utilizadas. Esta indica la probabilidad de que este modelo entrege resultados estadísticamente significativos (Cohen, 1988).

Se deben usar las tablas de potencia para una prueba de Fisher dadas por Cohen (1988), para determinar la potencia real del modelo, calculando como se muestra en el paso "D", utilizando el número de datos, el número de variables y el coeficiente R^2 del modelo de regresión.

Para garantizar la validez del ANOVA (*analysis of variance*) de la regresión realizada se deben validar los siguientes supuestos:

- **Supuesto de independencia:** validado de manera visual mediante un gráfico de residuos contra el número de observaciones, buscando encontrar un comportamiento aleatorio en los residuos y evitando obtener patrones en los datos. Analíticamente, se comprueba mediante la prueba de Durbin-Watson, la cual entrega un valor que se debe encontrar dentro de un rango de valores críticos (1,5 y 2,5) (Rodríguez, 2011)

La ecuación para la prueba de Durbin-Watson es la siguiente (CFI, 2021):

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \quad (12)$$

Donde t es el índice de la observación, T es el total de observaciones y e_t es el error de la observación t.

- **Supuesto de normalidad:** validado de manera visual mediante un gráfico de probabilidad normal (Q-Q Plot) y de manera analítica mediante la prueba de Shapiro Wilk, la cual cuantifica un valor de probabilidad que los residuos provienen de una distribución normal (Flores Muñoz *et al*., 2019)

Para una muestra aleatoria X_1, X_2,…, X_n, el estadístico de prueba se calcula de la siguiente manera:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (13)$$

Donde b se define en la siguiente ecuación:

$$b = \sum_{i=1}^{n} a_i[X_{(n-i-1)} - x_i] \quad (14)$$

Donde a_i es un valor tabulado en los manuales para cálculo de Shapiro Wilks y $[X_{(n-i-1)} - x_i]$ son las diferencias obtenidas restando el primer valor del último valor, el segundo al antepenúltimo y así sucesivamente.

El valor crítico tabulado de esta prueba es definido como, que no sigue ninguna distribución conocida, y sus valores críticos son calculados a través de simulaciones de Monte Carlo (S. Shapiro & M. Wilk, 2015). Este valor se compara con el calculado, y se rechaza la hipótesis nula (los datos no se distribuyen de manera normal) si.

- **Supuesto de homocedasticidad:** validado de modo visual mediante un gráfico de residuos estandarizados contra el valor predicho, donde se busca que los puntos se encuentren de forma aleatoria (Santibáñez, 2018). De manera analítica se valida mediante la prueba de Breush-Pagan, la cual cuantifica mediante un valor de probabilidad que los datos se comporten con igualdad de varianza en una regresión lineal (Santibáñez, 2018).

La prueba de Breusch-Pagan se puede entender como una prueba de varianza de los residuos y las variables independientes del modelo (Breusch & Pagan, 1979), para esto se propone que la suma de varianza aplicada en una regresión de $g_i = \hat{\varepsilon}_i^2/\hat{\sigma}^2$ contra las variables independientes Z, se distribuyen asintóticamente con p-1 grados de libertad en una distribución χ^2, cuando la hipótesis nula es verdadera; en el caso que existe homocedasticidad. $\hat{\sigma}^2$ se calcula como:

$$\hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2/N \quad (15)$$

Donde $\hat{\varepsilon}_i$ es el residuo de la i-ésima observación, N es el número de observaciones. El estadístico de Breusch-Pagan se calcula de la siguiente manera:

$$LM = \frac{1}{2}[g'Z(Z'Z)^{-1}Z'g - N^{-1}(jg)^2 \quad (16)$$

Es claro que la expresión representa la suma de varianza explicada en la regresión; como se puede tomar de Wood (2010), *j* es un vector de N x 1 unitario. Por lo que la ecuación puede ser reescrita como:

$$LM = \frac{1}{2}[TSS - SSR] \quad (17)$$

Donde TSS es la suma de cuadrados totales y SSR es la suma de cuadrados de los residuos.

2.3. Identificación de variables relevantes a la eficiencia energética

Para identificar las variables relevantes a la eficiencia energética se proponen los siguientes pasos:

A. Definición de intervalos de confianza del modelo y categorización de residuos

Las bandas de confianza representan el rango en el que la verdadera línea de regresión podría ubicarse a un cierto nivel de confianza. Las bandas ilustran todos los intervalos para cada posible valor de la variable independiente y son más cerrados donde se encuentra la mayor densidad de datos (RPubs, 2021). Definir estas bandas es de mucha importancia porque con ellas se tiene una menor incertidumbre al momento de clasificar un ahorro o sobreconsumo de energía para la misma condición operativa.

El intervalo de confianza de la línea de regresión es calculado de la siguiente manera:

$$\hat{y}_h \pm t_{\alpha/2,n-2}\,(s.e.)_y \quad (18)$$

Donde $(s.e.)$ es definido como el error estándar de la línea de regresión, multiplicado por el error estándar de la estimación en . Su calculo se presenta a continuación (RPubs, 2021):

$$(s.e.)_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad (19)$$

La figura 2 muestra una representación gráfica de las bandas de confianza a lo largo de un modelo de regresión lineal, donde se observan los datos utilizados para calcularlas en color gris y en color azul los intervalos de confianza.

Los residuos de la regresión se deben clasificar en tres posibles categorías, las cuales están dadas por su posición respecto a las bandas de confianza. Los puntos dentro de las bandas de confianza corresponden a un desempeño energético igual al periodo base; los puntos por encima de la banda superior serán sobreconsumos de energía, y los puntos por debajo de la banda inferior corresponden a ahorros de energía. En la figura 3 se presenta un gráfico que muestra los residuos clasificados según su posición en el plano, con respecto a las bandas de confianza.

B. Identificar variables relevantes

Se deben tomar los residuos de la regresión y realizar una nueva regresión bayesiana con las posibles variables relevantes a la eficiencia energética del proceso; de este modo, se observan todas las posibles combinaciones de variables que representan de mejor manera el comportamiento de los residuos. El mejor modelo, seleccionado por el factor de Bayes, contiene las variables relevantes a la eficiencia energética. Para esta selección se realiza un procedimiento similar al realizado en el modelo de línea de base energética (Moreno & Girón, 2008).

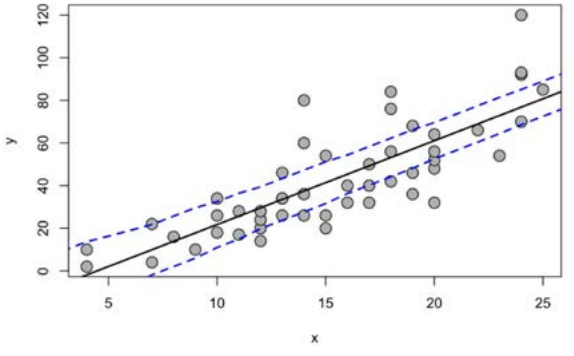


Figura 2. Representación gráfica de bandas de confianza de un modelo lineal (RPubs, 2021).

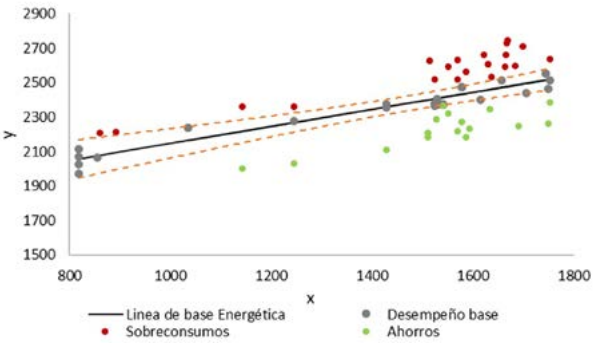


Figura 3. Representación gráfica de la clasificación de residuos.

C. Definir límites de variables relevantes

Los límites de las variables relevantes se calculan mediante regresiones lineales de cada una de estas contra los residuos de la regresión de la línea base, utilizando el valor del intercepto del modelo que explica el cambio del comportamiento de ahorro a sobreconsumo o viceversa.

La figura 4 representa el cambio de la variable relevante y sus variaciones de energía, la cual es de utilidad para calcular el intercepto de la ecuación lineal, que entrega el límite de eficiencia energética para esta variable.

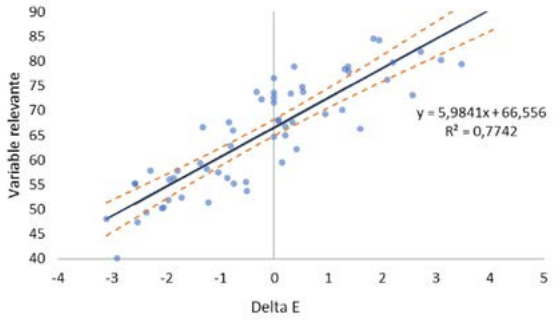


Figura 4. Representación gráfica de modelo lineal de delta de energía versus variable relevante.

3. Caso De Estudio: Sistema De Bombeo En Planta De Inyección De Agua

A continuación, se presenta un caso de estudio real, donde se aplica la metodología propuesta en el apartado anterior. Este caso corresponde a un sistema de bombeo en una planta de inyección ubicada en un campo petrolero en Santander, Colombia, donde se muestra el resultado de cada uno de los pasos para establecer una línea de base energética e identificar las variables operacionales controlables que impactan en el desempeño energético. Para desarrollar toda la metodología se utilizó el *software* Microsoft Excel y el lenguaje de programación R en el entorno de desarrollo "R Studio".

3.1. Selección del periodo base

El estudio energético se realizó durante el primer semestre del año 2021, por lo tanto, el periodo base utilizado corresponde al año inmediatamente anterior (2020), donde se tuvo una operación estable del proceso, sin cambios significativos en los equipos instalados.

3.2. Definición de variables de ingeniería

Debido a la naturaleza del proceso de bombeo, las variables de ingeniería a considerar corresponden a las de un sistema hidráulico, donde las más relevantes son:

- Consumo de energía
- Flujos (descarga y recirculación)
- Propiedades del fluido bombeado
- Presiones (descarga y succión)
- Apertura de válvulas (descarga y recirculación)

3.3. Selección del tamaño mínimo de la muestra

Para este paso se utilizó la librería "pwr" en "R Studio", la cual permite calcular el número mínimo de muestra a partir de un análisis de potencia estadística, utilizando la función "pwr.f2.test", cuyos parámetros de entrada son los grados de libertad del número de variables, el R^2 de una muestra piloto (correspondiente a una muestra aleatoria del 20 % de la información recopilada) y la potencia deseada para la regresión, establecida en 80 %.

3.4. Recopilación de información histórica

De los sistemas de información de la compañía se recibieron datos históricos del año 2020, que contienen variables operacionales desde el 1 de enero de 2020 hasta el 9 de noviembre de 2020 para 12 bombas, con un total de 50 variables, cada una con 314 registros y una frecuencia diaria.

Al ejecutar esta función se obtuvo un número mínimo de muestra de 20 datos. Debido a que el total de registros es mucho mayor que el número mínimo de muestra (314>>20), se puede asegurar que se obtendrá una potencia igual o mayor a la requerida.

3.5. Filtrado de datos

Se realizaron los dos tipos de filtrado:

- **Físico:** eliminación de registros sin datos de energía correspondiente al 12 % de los iniciales. También se excluyeron 25 variables, debido a su baja influencia física sobre el consumo energético (*e.g.* temperaturas de devanados y vibraciones); conforme con criterios técnicos y a fenómenos de termo-fluidos que ocurren en los sistemas de bombeo. Otras variables fueron desestimadas considerando su correspondencia con variables de bombas no operativas durante el periodo.

- **Estadístico:** se realizó un filtrado con criterio multivariable utilizando la distancia de Mahalanobis y estableciendo nubes de confianza a través de una distribución Chi cuadrado. Para esto se utilizó la función "mahalanobis" en "R Studio", que calcula las distancias de mahalanobis y se define un valor critico de chi-cuadrado a través de la función "qchisq" con un alfa de 0,05. Los puntos atípicos son identificados cuando la distancia de mahalanobis supera el valor crítico de chi-cuadrado.

Al ejecutar esta función se eliminaron 36 registros atípicos, correspondiente al 13,1 % de los datos. La figura 5 muestra una nube de datos representada en 3 componentes principales que explican el 44 % de la variabilidad total, donde se clasifican los puntos atípicos identificados. Se observa que los puntos atípicos se encuentran alejados del centro de la nube de datos, lo cual es de esperarse debido a que son puntos que no caracterizan el comportamiento normal del proceso.

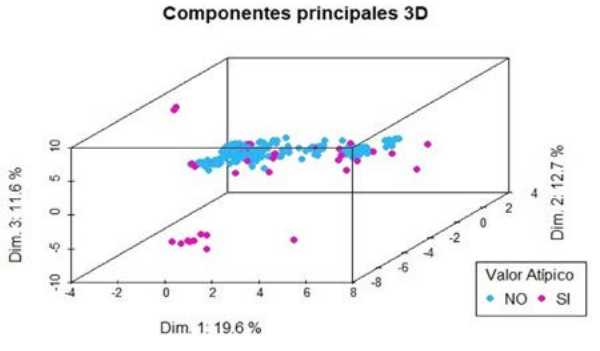


Figura 5. Representación gráfica de puntos atípicos en el *dataset*.

3.6. Estadística descriptiva

Con el grupo de datos obtenido luego de realizar el filtrado de datos, se presentan estadísticos descriptivos en la tabla 1 correspondientes a las variables objeto de estudio. Con la información ya filtrada se identifican algunas variables que no presentaron variación significativa (resaltadas en la tabla 1), cuyo valor máximo y mínimo se encontraban muy cercano a 0, por esta razón no serán tenidas en cuenta en el análisis y serán excluidas del *dataset*.

La variable "consumo de energía" tiene una variabilidad similar a la de "flujo" con valores de 11 % y 12 % respectivamente, como es de esperarse, debido a que la naturaleza física de las máquinas hidráulicas indica que la potencia eléctrica demandada por el sistema es directamente proporcional a la carga hidráulica que se maneja.

Se destaca también el alto coeficiente de varianza (CV) de la variable "recirculación" con un valor de 68 %, siendo esta la de mayor variabilidad. Esto es un indicio de que la variable durante el periodo base no se encontraba en control obteniendo valores máximos de 11.670 BWPD y mínimo de 542 BWPD.

Por otra parte, la variable "presión de descarga" tiene el menor CV, con un valor de 1 %, lo que indica que es una variable controlada y de vital importancia para el proceso, que debe ser tenida en cuenta al momento de construir la línea de base energética.

3.7. Definición de variables de ingeniería para el modelo de línea de base energética

Una vez filtrados los datos y seleccionadas las variables de estudio, estas se clasifican con un criterio técnico para conocer qué variables son requerimientos del proceso y cuáles pueden ser gestionadas por la operación y mantenimiento para mejorar el desempeño energético. Esta clasificación se presenta en la tabla 2, donde es importante observar que las variables "flujo" y "presión de descarga" fueron utilizadas para calcular una nueva variable llamada "energía hidráulica", la cual indica la cantidad de energía que posee el fluido bombeado. Esta variable es creada, ya que está muy relacionada con el consumo.

3.8. Cálculo y selección del mejor modelo

Utilizando la librería "BAS" en "R Studio", la función "bas.lm" con un modelo de probabilidad *a priori* BIC y asignando igual probabilidad en cada modelo

(distribución uniforme), se calcula las probabilidades *a posteriori* de cada modelo y su factor de Bayes, obteniendo como resultado un *ranking* de modelos presentado en la figura 6.

El modelo con la mayor probabilidad *a posteriori* corresponde al modelo con factor de Bayes igual a 1, que en este caso corresponde al que incluye únicamente la variable "energía hidráulica". Al compararlo con el segundo mejor modelo, cuyo factor de Bayes es igual a 0,08, no hay evidencia que respalde este segundo modelo según la escala de Jeffrey (Nagin, 1999).

Tabla 1. Descriptivos de grupo de variables obtenidos posterior al filtrado de datos

Variable	Máx.	Mín.	Media	Mediana	Desv. estand.	CV	Sesgo	Curtosis
Consumo de energía [kWh/día]	246.818	160.565	197.766	196.052	22.016,9	11 %	0,2	-0,9
Conductividad [us]	2.495	2.003	2.255	2.256	142,3	6 %	-0,1	-1,2
Nivel de PH	8	6	7	7	0,7	9 %	0,1	-0,6
Temp. agua [°C]	30	27	29	29	1,0	3 %	-0,1	-0,9
Presión de descarga [psig]	2.074	1.859	2.045	2.054	24,7	1 %	-2,8	13,6
Flujo [BWPD]	266.681	164.844	206.689	206.729	25.001,3	12 %	0,4	-0,6
Recirculación [BWPD]	11.670	542	3.979	3.432	2.716,7	68 %	0,8	0,0
Rec. 1 [%]	7	-4	-3	-4	1,4	-48 %	2,5	8,3
Rec. 2 [%]	101	0	79	87	33,9	43 %	-1,8	1,6
Presión succión 1 [psig]	57	0	42	48	17,8	43 %	-1,9	1,6
Rec. 3 [%]	0	-4	-3	-2	1,0	-37 %	0,8	0,3
Rec. 4 [%]	0	0	0	0	0,1	-44 %	1,8	4,2
Rec. 5 [%]	0	-4	-3	-4	2,1	-79 %	0,4	-1,8
Rec. 6 [%]	2	-4	-1	0	1,1	-170 %	-1,5	0,9
Rec. 7 [%]	-1	-1	-1	-1	0,1	-6 %	0,8	3,1
Descarga 1 [%]	100	90	99	99	0,9	1 %	-7,3	61,3
Presión succión 2 [psig]	51	40	44	44	1,3	3 %	2,2	10,1
Rec. 8 [%]	0	-1	-1	-1	0,3	-41 %	1,9	2,0
Descarga 2 [%]	101	0	86	101	35,3	41 %	-2,1	2,3
Rec. 9 [%]	-1	-2	-2	-2	0,1	-7 %	0,3	0,3
Descarga 3 [%]	96	88	95	96	0,8	1 %	-6,6	51,4
Presión succión 3 [psig]	45	30	35	34	3,1	9 %	0,6	-0,2
Rec. 10 [%]	0	-6	-5	-6	2,4	-53 %	1,4	-0,1
Descarga 4 [%]	102	0	78	101	41,5	53 %	-1,4	-0,1
Presión succión 4 [psig]	55	0	37	48	20,0	53 %	-1,3	-0,2

Rec.:Recirculación

Tabla 2. Clasificación de variables de ingeniería

Variable	Clasificación
Consumo de energía [kWh/día]	Variable respuesta
Conductividad [us]	Variables para la línea de base energética
Nivel de PH	
Temp. agua [°C]	
Energía hidráulica [kWh_H/dia]	
Recirculación [BWPD]	Variables controlables por operación y mantenimiento
Rec. 1 [%]	
Rec. 2 [%]	
Presión succión 1 [psig]	
Descarga 1 [%]	
Presión succión 2 [psig]	
Descarga 2 [%]	
Descarga 3 [%]	
Presión succión 3 [psig]	
Descarga 4 [%]	
Presión succión 4 [psig]	

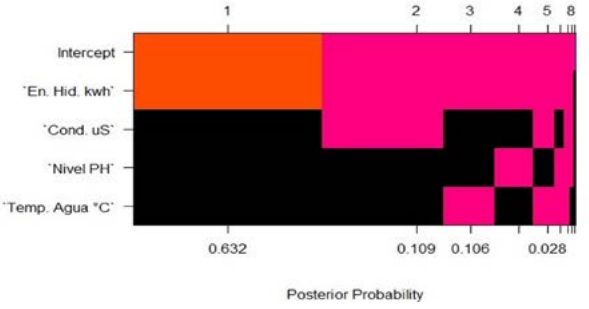


Figura 6. *Ranking* de modelos para la línea de base energética en función de probabilidad posterior de Bayes.

De la regresión OLS realizada con el mejor modelo, se tiene un coeficiente de correlación R^2 de 0,93, lo cual indica una alta relación entre las variables. También se valida la significancia de la variable de la regresión, con un p-valor mucho menor a 0,05 (1,2E-136).

La representación gráfica del modelo se presenta en la figura 7, donde se muestra la dispersión de puntos analizados.

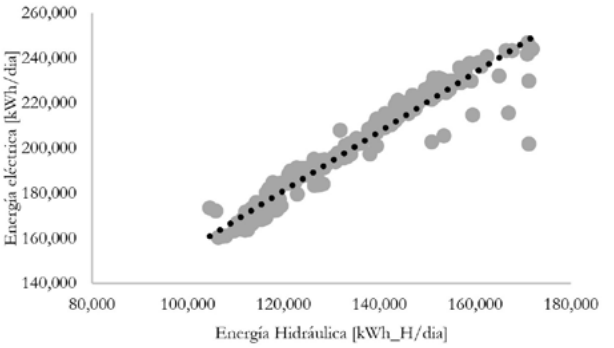


Figura 7. Modelo con la variable "energía hidráulica" para el sistema de bombeo del caso de estudio.

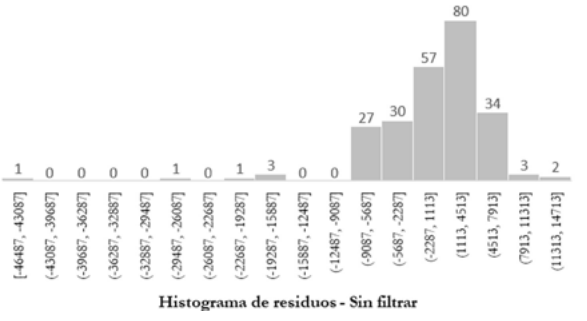
Se observan algunos puntos alejados del modelo, estos corresponden a condiciones operacionales reales alcanzables por el proceso, debido a que no fueron eliminados en el filtrado físico y estadístico realizado en la sección anterior, sin embargo, serán eliminados en la siguiente sección debido a que se busca tener las desviaciones centradas para su posterior análisis.

3.9. Filtrado de residuos y cálculo de línea de base energética

Al realizar el filtrado de los residuos mediante el uso de rango intercuartil, se eliminaron 29 registros obteniendo una distribución como se presenta en la figura 8.

Con el nuevo grupo de datos se realiza nuevamente una regresión OLS para calcular el modelo de line de base energética con el que se caracteriza el proceso, cuyo resumen se encuentra en la figura 9.

De la regresión OLS realizada con este modelo de residuos filtrados, se observa que se tiene un coeficiente de correlación R^2 de 0,98, lo cual indica que se tiene una alta relación entre las variables. También se valida la significancia de la variable de la regresión, con un p-valor mucho menor a 0,05 (5,7E-175).



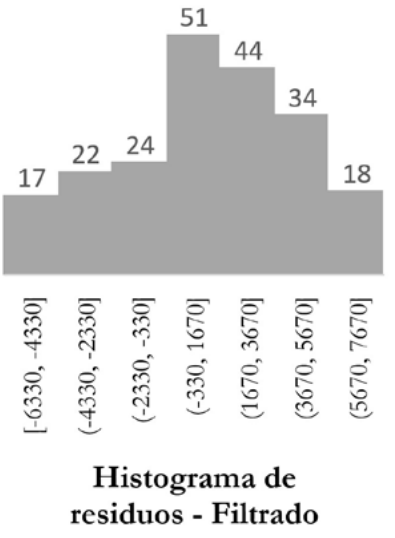


Figura 8. Distribución de residuos del modelo de regresión inicial vs. filtrados.

Resumen

Estadísticas de la regresión	
C. correlación múltiple	0,99
C. determinación R^2	0,98
R^2 ajustado	0,98
Error típico	3219,63
Observaciones	210

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	97174175728	97174175728	9374	5,7E-175
Residuos	208	2156129556	10366007		
Total	209	99330305285			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	14278,9	1919,9	7,4	2,6E-12	10494,0	18063,7	10494,0	18063,7
Energía Hidráulica [kWh_H/día]	1,4	0,0	96,8	5,7E-175	1,4	1,4	1,4	1,4

Figura 9. Resumen de regresión OLS para mejor modelo de línea de base energética.

La figura 10 muestra la línea de base energética con el filtrado intercuartil de los residuos. La formulación que explica el comportamiento entre estas dos variables y representa el modelo de línea de base energética se presenta en la siguiente ecuación:

$$EEb = 1{,}4 * EH + 14278{,}9 \qquad (20)$$

Donde EEb es la energía eléctrica base y EH es la Energía hidráulica del sistema.

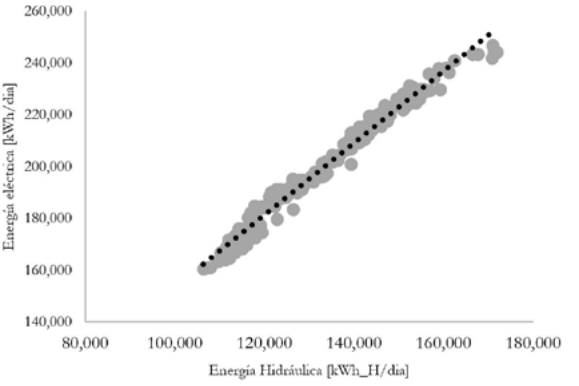


Figura 10. Modelo con la variable "energía hidráulica" para el sistema de bombeo del caso de estudio.

3.10. Validación de potencia de modelo y supuestos

- **Potencia de modelo:** se validó la potencia del modelo de manera análoga a la presentada en el paso D, obteniendo un valor de 100 %.

- **Normalidad:** se realiza una prueba de Shapiro-Wilk a los residuos del modelo al 95 % de confiabilidad, utilizando la función de R llamada "shapiro.test", obteniendo un p-valor de 0,054. Ya que este valor es mayor que 0,05, se concluye que los residuos tienen una distribución normal.

- **Homocedasticidad:** se efectúa una prueba de Breusch-Pagan a los residuos del modelo al 95 % de confiabilidad, utilizando la librería "lmtest" y función "bptest", obteniendo un p-valor de 0,4462. Ya que este valor es mayor que 0,05, se concluye que la varianza de los residuos se distribuye de igual manera.

- **Independencia:** se implementó una prueba de Drubin-Watson a los residuos del modelo al 95 % de confiabilidad, utilizando la librería "lmtest" y función "dwtest", obteniendo un p-valor menor que 2,2E-16. Ya que este valor es mucho menor que 0,05, se concluye que no hay independencia de los residuos. Autores como Mukherjee y Laha (2019) indican que en observaciones que son tomadas a través del tiempo, y que son afectadas por este, se detecta autocorrelación en los datos. Este es el caso del sistema de bombeo, el cual ha cambiado su forma de operación por periodos en los que las variables no son controladas. Esto no significa que las variables dependan del tiempo o tengan comportamientos cíclicos en los periodos, significa que el proceso tuvo cambios en variables

operacionales que no fueron controladas de manera inmediata. La autocorrelación afecta la eficiencia del estimador que se traduce en el nivel de error del mismo, sin embargo, no afecta la consistencia del estimador y además este continua siendo insesgado y normalmente distribuido (Mukherjee & Laha, 2019).

3.11. Definición de intervalos de confianza del modelo y categorización de residuos

Con el modelo de línea base establecido, se requiere clasificar los residuos en tres categorías, las cuales son: base, ahorros y sobreconsumos. Para esto es necesario construir primeramente las bandas de confianza del modelo lineal utilizando la formulación presentada por RPubs (2021). En la figura 11 se muestra la clasificación de los residuos.

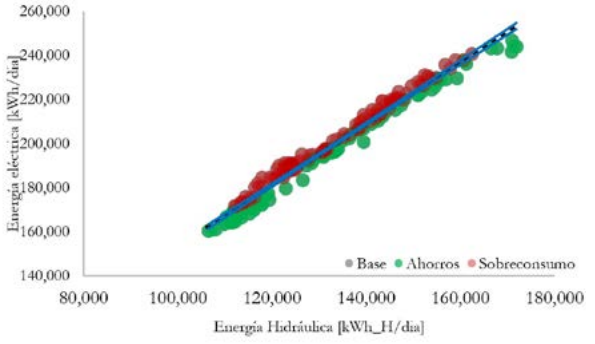


Figura 11. Bandas de confianza y categorización de residuos del modelo de línea de base energética.

Los puntos que se encuentran dentro de las bandas de confianza corresponden a la categoría “base”, la cual indica que el desempeño energético corresponde al valor esperado por el modelo. Los puntos por debajo de la banda inferior corresponden a operaciones con alto desempeño energético, y están en la categoría “ahorros”, y aquellos puntos por encima de la banda superior corresponden a operaciones con bajo desempeño energético, en la categoría “sobreconsumo”.

3.12. Identificar variables relevantes

Para identificar estas variables se utilizó la misma configuración explicada en el “paso H”, utilizando en lugar de la energía eléctrica los residuos de la regresión de línea base y las variables correspondientes identificadas en el “paso G”. La figura 12 muestra un *ranking* de modelos, donde se observan las variables que participan en cada uno de estos; siendo estas las variables relevantes que mejor explican el comportamiento de los residuos.

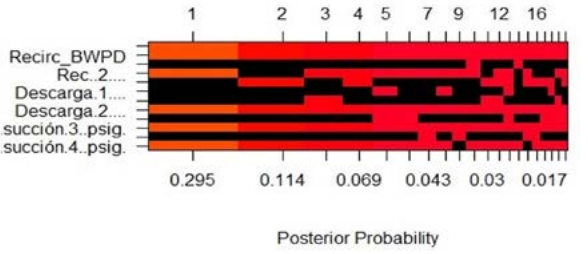


Figura 12. *Ranking* de modelos para las variables relevantes en función de probabilidad posterior de Bayes.

El factor de Bayes para los dos primeros modelos son 1 y 0,34, respectivamente. Debido a la escala de Jeffrey (Nagin, 1999) se observa que para valores en el rango de 1 > BF > 0,33 en los modelos a contrastar, no se tiene una diferencia significativa entre el comparado y el mejor. Por esta razón. cualquiera de los dos mejores modelos identificados puede ser utilizado.

El modelo que se selecciona es el primero por tener mayor probabilidad posterior, con un valor de 0,295, el cual contiene las variables de recirculación, presión succión 3, presión succión 4, recirculación 2 y descarga 2.

M. Definir límites de variables relevantes

Mediante una regresión lineal para cada una de las variables identificadas contra la desviación de Energía Eléctrica (residuos); se obtiene una ecuación cuyo intercepto corresponde al límite de la variable para pasar de Ahorros a Sobreconsumo o viceversa, según la naturaleza de la variable.

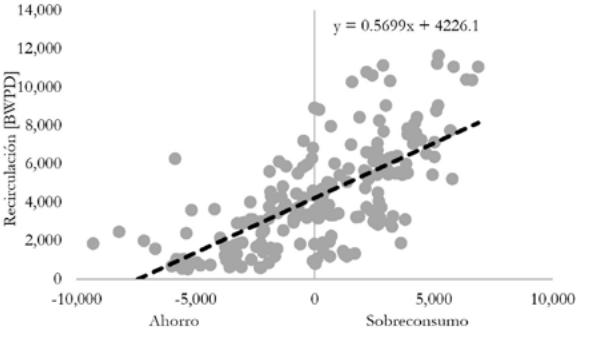


Figura 13. Modelo lineal de variable relevante (Recirculación) contra residuos del modelo de línea de base energética.

El comportamiento de la variable "recirculación" se observa en la figura 13, donde se evidencia un intercepto de 4.226 BWPD. Este valor representa el punto donde la variable ocasiona que el proceso cambie de ahorro a sobreconsumo, por lo que se prefiere la operación del proceso con esta variable con valores por debajo de este límite.

El comportamiento de la variable "presión succión 3" se observa en la figura 14, donde se evidencia un intercepto de 34,7 psig. Este valor representa el punto donde la variable ocasiona que el proceso cambie de ahorro a sobreconsumo, por lo que se prefiere la operación del proceso con esta variable con valores por encima de este límite. Esto valida la teoría de las máquinas hidráulicas, donde la potencia demandada es menor cuando se reduce la cabeza hidráulica debido a una mayor presión de succión.

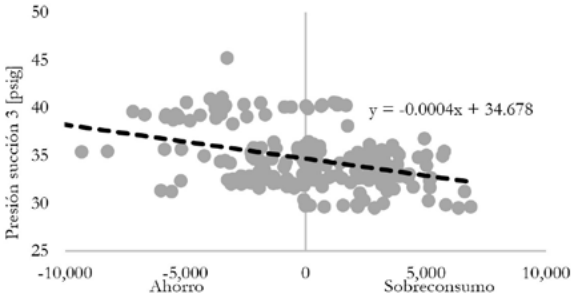


Figura 14. Comportamiento de la variable "descarga 3" contra residuos del modelo de línea de base energética.

Las variables presión succión 4, Rec. 2 y descarga 2 presentan condiciones operativas con intermitencias en la operación (toman valores de 0), lo cual no permite establecer un límite para la variable, aunque da un indicio de que al tomar ciertos valores (operativo o no operativo), el proceso tiende a tener ahorros de energía. Sin embargo, se debe realizar un análisis ingenieril más profundo, donde se incluyan otro tipo de variables de ingeniería que resulten en operaciones más eficientes para comprobar estos hallazgos, lo cual no hace parte del alcance del presente estudio.

4. Conclusiones

Del presente trabajo, cuyo objetivo es el desarrollo de una metodología para la caracterización energética de procesos industriales basada en modelos de regresión bayesianos alineada con la norma NTC ISO 50001, se puede concluir lo siguiente:

- Se desarrolló una metodología basada en diferentes técnicas estadísticas para lograr cumplir con requisitos establecidos en la normal NTC ISO 50001, como el establecimiento de una línea de base energética y la selección de variables que afectan la eficiencia energética de un proceso.

- La metodología mostró ser eficaz en el establecimiento de una línea base energética, siendo consistente con el modelo seleccionado en el caso de estudio a través de modelos de regresión bayesianos con probabilidad *a priori* basada en BIC, con lo que se espera obtener por la teoría de comportamiento de máquinas hidráulicas. También fue posible identificar variables relevantes en el desempeño energético a través del mismo tipo de modelos y estas son consecuentes y coherentes con la naturaleza física del proceso estudiado.

- El uso de regresión bayesiana como herramienta principal para la selección de modelos en la caracterización energética es una alternativa robusta frente a los métodos tradicionales de selección de modelos por eliminación hacia atrás, adelante o mixto que se encuentran en la literatura y presentan la ventaja de tener una menor incertidumbre a la hora de seleccionar el modelo final de línea de base energética o las variables relevantes al proceso, ya que se conoce el espacio completo de modelos y evitan el sesgo que genera la interacción de variables en el cálculo final del p-valor, que tiende a generar mayores errores tipo I (seleccionar una variable significativa cuando no lo es).

- La presente metodología garantiza el número mínimo de muestra, lo cual asegura la potencia deseada para el modelo disminuyendo la probabilidad de cometer errores tipo II (no seleccionar una variable significativa cuando lo era).

- Se destaca que técnicas como el filtrado multivariado de mahalanobis y el filtrado intercuartil para las desviaciones de energía son útiles para la obtención de una nube de datos que represente los comportamientos atípicos del proceso cuya implementación permite acelerar los tiempos de filtrado en comparación con el análisis univariado o bivariado en cada una de las variables.

- A pesar de ser una metodología robusta, esta es fácilmente implementable en lenguajes de código abierto como R, por lo cual es posible como un desarrollo futuro la creación de una herramienta basada en Shiny (lenguaje de interfaz R) que facilite el trabajo de gestión energética, acelerando los tiempos de cálculo, disminución de incertidumbres y niveles de error en el establecimiento de líneas de base energética e identificación de variables significativas.

Referencias

Beier, P., Burnham, K. P., & Anderson, D. R. (2001). Model selection and inference: A practical information-theoretic approach. In *The Journal of Wildlife Management* (Vol. 65, Issue 3). Springer. https://doi.org/10.2307/3803117

Breusch, T. S., & Pagan, A. R. (1979). A simple test for Heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287–1294.

Caicedo, O. F. P., Avella, J. C. C., Rodríguez, D. B. R., & Salas, A. P. (2019). *Implementación de un sistema de Gestión de la Energía Guía con base en la norma ISO*.

Campos, J. (2017). Línea base, indicadores de desempeño energético. *Aportes y Experiencias Practicas Del Grupo de Investigación Gestión Energética de La UA y de La ESCO E2 Energía Eficiente S.A.E.S.P.* http://guiaiso50001.cl/guia/wp-content/uploads/2017/05/linea-base-indicadores-de-desempeno-P-3-AP-1.pdf

CFI. (2021). *Durbin Watson statistic*. CFI Education Inc. https://corporatefinanceinstitute.com/resources/knowledge/other/durbin-watson-statistic/

Chen, Y., & Therkelsen, P. (2019). *The effect of linear regression modeling approaches on determining facility wide energy savings*. *April*, 37.

Cohen, J. (1988). *Power analysis for the behavioral sciences* (2nd ed.).

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7

Ellis, P. (2010). *The essential guide to effect sizes*. Cambridge.

Flores Muñoz, P., Muñoz Escobar, L., & Sánchez Acalo, T. (2019). Estudio de potencia de pruebas de normalidad usando distribuciones desconocidas con distintos niveles de no normalidad. *Perfiles*, *1*(21), 4–11. https://doi.org/10.47187/perf.v1i21.42

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning* (2nd ed.). Springer.

NTC-ISO 50001, (2019).

Li, X., Deng, S., Li, L., & Jiang, Y. (2019). Outlier detection based on robust mahalanobis distance and its application. *Open Journal of Statistics*, *09*(01), 15–26. https://doi.org/10.4236/ojs.2019.91002

Minambiente. (2020). *Colombia reducirá en un 51% sus emisiones de gases efecto invernadero para el año 2030 | Ministerio de Ambiente y Desarrollo Sostenible*. https://www.minambiente.gov.co/index.php/noticias/4877-colombia-reducira-en-un-51-sus-emisiones-de-gases-efecto-invernadero-para-el-ano-2030

Moreno, E., & Girón, F. J. (2008). Comparison of Bayesian objective procedures for variable selection in linear regression. *Test*, *17*(3), 491–492. https://doi.org/10.1007/s11749-008-0095-9

Mukherjee, A., & Laha, M. (2019). Problem of autocorrelation in linear regression detection and remedies. *International Journal of Multidisciplinary Research and Modern Education*, *5*(1), 105–110.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, *4*(2), 139–157. https://doi.org/10.1037/1082-989X.4.2.139

Plan de acción Indicativo del PROURE 2022-2030, 2022.

Rodríguez, M. D. (2011). Regresión lineal simple. *Estadística Inferencial Aplicada*. https://doi.org/10.2307/j.ctvswx88n.8

RPubs. (2021). *Linear regression confidence intervals*. https://rpubs.com/aaronsc32/simple-linear-regression

S. Shapiro, & M. Wilk. (2015). The Shapiro-Wilk and related test for normality. *Statistics*, *1*, 1–12.

Santibáñez, J. (2018). *Verificación del supuesto de homocedasticidad*. UNAM. http://sigma.iimas.unam.mx/jsantibanez/Cursos/Ciencias/2018_1/08_homocedasticidad.html

Walpole, R., Myers, R., & Myers, S. (2012). *Probabilidad y estadística para ingeniería y ciencia*. Pearson.

Wood, F. (2010). *Matrix approach to linear regresssion*. Columbia University. http://www.stat.columbia.edu/~fwood/Teaching/w4315/Fall2009/lecture_11