ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELOS ESTADÍSTICOS PARA LA DETECCIÓN DE PATRONES EN MEDIO AMBIENTE Y FINANZAS

UN NUEVO MARCO BIO-INSPIRADO PARA EL PRONÓSTICO DE EMISIONES DE CO2 EN ECUADOR

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

FRANCIS CÉSAR PROAÑO CARBO

francis.proano@epn.edu.ec

DIRECTOR: PH. D. MIGUEL ALFONSO FLORES SÁNCHEZ

miguel.flores@epn.edu.ec

DMQ, FEBRERO 2022

CERTIFICACIONES

Yo, FRANCIS CÉSAR PROAÑO CARBO, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

Francis César Proaño Carbo

Certifico que el presente trabajo de integración curricular fue desarrollado por Francis César Proaño Carbo, bajo mi supervisión.

Ph. D. Miguel Alfonso Flores Sánchez
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

Francis César Proaño Carbo

Ph. D. Miguel Alfonso Flores Sánchez

RESUMEN

La contaminación ambiental por emisiones de gases de efecto invernadero (GEI) son consideradas como un desafío para muchos investigadores. Entre los gases de efecto invernadero, el dióxido de carbono (CO2) ha generado efectos masivos en la calidad del aire de grandes ciudades y se considera como un problema alarmante en todo el mundo ya que las emisiones de CO2 también afectan a la salud humana, la vida natural y la agricultura. El objetivo principal de este trabajo de integración curricular es estudiar, y estimar las emisiones de CO2 que son provenientes de distintas fuentes de consumo de energía en el Ecuador. Se utilizó un conjunto de datos recopilado durante los años 1965 y 2020. La estimación y pronóstico de las emisiones de CO2 se realiza mediante el desarrollo de un Modelo de Regresión Lineal Múltiple (RLM) utilizando el método de mínimos cuadrados, un Modelo de Regresión Potencial Generalizado (RPG) utilizando el algoritmo de optimización por enjambre de partículas (OEP) y de un Modelo Aditivo Generalizado (MAG) utilizando splines con penalizaciones (p-splines). El rendimiento de dichos modelos se midió mediante el uso de parámetros de calidad estadísticos y los resultados obtenidos indican que el MAG brinda estimaciones más precisas.

Palabras clave: Emisiones de CO2, Contaminación del aire, Estimaciones, Regresión Lineal Múltiple, Optimización por Enjambre de Partículas, Modelo Aditivo Generalizado.

ABSTRACT

Environmental pollution from greenhouse gas (GHG) emissions is considered a challenge for many researchers. Among the greenhouse gases, carbon dioxide (CO2) has generated massive effects on air quality in large cities and is considered as an alarming problem worldwide since CO2 emissions also affect human health, natural life and agriculture. The main objective of this research work is to study and estimate CO2 emissions from different sources of energy consumption in Ecuador. A data set collected during the years 1965 and 2020 was used. The estimation and forecasting of CO2 emissions is performed by developing a Multiple Linear Regression Model (MLR) using the least squares method, a Generalized Potential Regression Model (GPR) using the Particle Swarm Optimization (PSO) algorithm and a Generalized Additive Model (GAM) using penalty splines (p-splines). The performance of these models was measured using statistical quality parameters and the results obtained indicate that the MAG provides more accurate estimates.

Keywords: CO2 emissions, Air pollution, Estimates, Multiple Linear Regression, Particle Swarm Optimization, Generalized Additive Model.

Índice general

Índice de figuras

Capítulo 1

Descripción del componente desarrollado

La creciente dependencia energética de la sociedad ha hecho que la energía sea considerada como un factor influyente en el crecimiento y desarrollo de un país; por tanto, la energía juega un papel destacado en varios sectores de la economía. Con el crecimiento económico de un país y la expansión simultánea de la producción y el consumo de energía, surgen amenazas como la contaminación y en general, la destrucción de los recursos naturales, creando así una paradoja entre el aumento de la producción y la calidad ambiental; esto provoca un mayor interés tanto por parte de los académicos como de los responsables políticos en el desarrollo de un modelo de estimación y de predicción en el ámbito de las emisiones de gases de efecto invernadero [1].

La contaminación climática por la emisión de carbono se convirtió en un problema importante y grave que afecta a los países desde los diferentes aspectos, salud, clima, agricultura, economía y turismo [3]. De acuerdo con la Organización Meteorológica Mundial (OMM), hace aproximadamente 3 y 5 millones de años fue la última vez que ocurrió en la Tierra una concentración de dióxido de carbono (CO2) como la actual, una concentración donde la temperatura era de 2 a 3 grados más cálida y el nivel del mar era entre 10 y 20 metros superior al actual [23]. Ajustar las políticas energéticas es un proceso necesario para eliminar el problema de la contaminación y mantener la atmósfera limpia. Muchos países tienen hoy compromisos entre ellos para reducir la emisión de gases de

efecto invernadero, como el protocolo de Kioto y el acuerdo de Naciones Unidas (ONU) de esta manera, se logrará seguir controlando el porcentaje de emisión de CO2 en la atmósfera para reducirlo a los niveles deseados [22]. Toda la lectura futura, indica el aumento de las emisiones de CO2 y gases de efecto invernadero (GEI); por todo ello, es necesario desarrollar un marco que permita estimar y pronosticar la emisión de CO2.

En este estudio se realizó una comparación entre los modelos de Regresión Lineal Múltiple mediante Mínimos Cuadrados Ordinarios (RLM), Regresión Potencial Generalizada mediante Optimización por Enjambre de Partículas (OEP) y un Modelo Aditivo Generalizado mediante splines con penalización (MAG) para determinar qué modelo se ajusta mejor a nuestros datos. Para este caso, se exploró el efecto de cuatro variables de entrada: Consumo de derivados del Petróleo, de Electricidad, de Gas natural y de Energía Primaria en la estimación y predicción de emisiones de CO2.

1.1. Objetivo general

Estimar y pronosticar las emisiones de CO2 en Ecuador mediante la aplicación de técnicas estadísticas y de optimización tales como Regresión Lineal Múltiple mediante Mínimos Cuadrados Ordinarios (RLM), Regresión Potencial Generalizada mediante Optimización por Enjambre de Partículas (OEP) y un Modelo Aditivo Generalizado mediante splines con penalización (MAG); a los datos históricos correspondientes de diversas fuentes de consumo de energía.

1.2. Objetivos específicos

1. Recolectar los datos correspondientes a las emisiones de CO2 en Ecuador y al consumo de energía a partir de diversas fuentes de consumo; para poder organizar y procesar la información necesaria que nos permita realizar un diseño utilizado como base en la estructuración del proyecto.

2. Formular un Modelo de RLM mediante el uso del método de mínimos

cuadrados ordinarios que nos permita conocer la relación de dependencia entre una variable dependiente, variables independientes y un término de error; y así, se obtendrá la estimación y predicción de emisiones de CO2 en Ecuador a partir de diversas fuentes de consumo de energía.

3. Formular un Modelo de RPG mediante el uso del algoritmo de OEP; para conocer la relación no lineal entre los predictores y la variable dependiente; y así realizar una estimación y predicción para las emisiones de CO2 en Ecuador a partir de diversas fuentes de consumo de energía.

4. Formular un MAG mediante el uso de splines con penalización, el cual permita ajustar funciones polinómicas en sus predictores; y así obtener la estimación y predicción de emisiones de CO2 en Ecuador a partir de diversas fuentes de consumo de energía.

5. Comparar los modelos anteriores mediante el análisis de estadísticos de calidad y la visualización de los resultados obtenidos por cada uno de los modelos.

1.3. Alcance

Por todo lo anteriormente mencionado, se vuelve necesario implementar nuevas tecnologías que permitan llevar un registro y control sobre las emisiones de CO2 en el país. Es así que mediante la aplicación de técnicas estadísticas como la RLM mediante Mínimos Cuadrados Ordinarios [17]; la Regresión Potencial Generalizada mediante OEP, técnica perteneciente al área de la inteligencia artificial que está inspirada en el comportamiento social de los seres vivos y es muy eficiente en la búsqueda global de soluciones para el problema planteado [17], [28]; y los MAG mediante p-splines [33], se busca estimar y pronosticar las emisiones de CO2 en Ecuador a partir de las diversas fuentes de consumo de energía como lo son el Petróleo, Gas natural, Electricidad y el resto de Energía Primaria (Leña, productos de caña, solar, etc) [7].

En el caso ecuatoriano, una limitante a tener en cuenta en esta investigación es que Ecuador cuenta con poca información oficial sobre el

consumo de energía y las emisiones de CO2 a nivel sectorial; las fuentes oficiales más completas de información sobre las emisiones de GEI y la política ambiental son la Tercera Comunicación Nacional de Ecuador a la Convención Marco de las Naciones Unidas sobre el Cambio Climático Cambio Climático [26] y el Balance Energético Nacional desarrollado por el Ministerio de Energía y Recursos Naturales no Renovables, el cual consolida la información energética del país en su totalidad [20].

1.4. Marco teórico

1.4.1. Antecedentes

El clima mundial está cambiando y eso supone riesgos cada vez más graves para los ecosistemas, la salud humana y la economía. Estos cambios se producen porque se liberan a la atmósfera grandes cantidades de gases de efecto invernadero (GEI) como consecuencia de muchas actividades humanas en todo el mundo, entre las que destaca la quema de combustibles fósiles para la generación de electricidad, la calefacción y el transporte [3]. La bibliografía existente sobre economía de la energía se centra generalmente en la relación que existe entre el crecimiento económico, el consumo de energía y los indicadores de calidad medioambiental, como las emisiones de CO2 [19].

Mientras que varios de los estudios desarrollados alrededor de esta temática examinan la validez de la relación de la curva en U invertida entre el crecimiento económico y las emisiones de CO2, como sugiere la hipótesis de la curva de Kuznets ambiental [18] , otros estudios amplían el análisis con la inclusión de aspectos extras como el desarrollo financiero, la urbanización y la asignación de recursos [13]. En general, los resultados de estos estudios muestran que la interacción entre el consumo de energía, el crecimiento de las actividades económicas y las emisiones de CO2 varían entre los países debido a las diferencias en las condiciones económicas, tecnológicas, institucionales, políticas y geográficas. En el caso particular de las economías dependientes del petróleo, como es el caso de Ecuador, la incorporación de la influencia de la riqueza petrolera en este tipo de análisis ofrecería una mejor explicación a la comprensión

de la relación existente entre el crecimiento de actividades económicas y sociales, el consumo de energía y la calidad ambiental en estas economías [19].

1.4.2. Consumo de energía por fuente y emisiones de CO2

Según Bernard Looney, director ejecutivo de British Petroleum: "El consumo de energía primaria disminuyó un 4,5 % el año pasado (2019), el primer descenso del consumo energético desde 2009. El descenso fue impulsado en gran medida por el petróleo (-9,7 %), que representó casi tres cuartas partes de la disminución. El gas natural y el carbón también experimentaron descensos significativos. El consumo de todos los combustibles disminuyó, salvo el de las renovables (+9,7 %) y el de la energía hidráulica (+1,0 %). El consumo se redujo en todas las regiones, con los mayores descensos en Norteamérica (-8,0 %) y Europa (-7,8 %). El menor descenso se produjo en Asia-Pacífico (-1,6 %) debido al crecimiento de China (+2,1 %), el único país importante donde el consumo de energía aumentó en 2020. En las demás regiones, el descenso del consumo osciló entre el -7,8 % de América del Sur y Central y el -3,1 % de Oriente Medio".

El carbón es el segundo combustible más importante en 2020, con un 27,2 % del consumo total de energía primaria, lo que supone un ligero aumento respecto al 27,1 % del año anterior. La cuota del gas natural y de las energías renovables ha aumentado hasta alcanzar máximos históricos del 24,7 % y el 5,7 % respectivamente. Las energías renovables han superado a la nuclear, que sólo representa el 4,3 % de la combinación energética. La cuota de energía hidráulica aumentó 0,4 puntos porcentuales el año pasado, hasta el 6,9 %, el primer aumento desde 2014 [6].

En Ecuador, el aporte respecto al consumo de energía es de apenas el 2 % en América Latina y el Caribe; sin embargo, dicho consumo ha crecido a una tasa media del 6 %. Este crecimiento en el consumo de energía es razonable debido a que los países en vías de desarrollo tienden a presentarlo [25]. Es por ello por lo que, basándonos en la evidencia empírica presentada (Ver Figura 1.1) y en el hecho de que aún existen lagunas en el conocimiento para la matriz energética ecuatoriana, la presente inves-

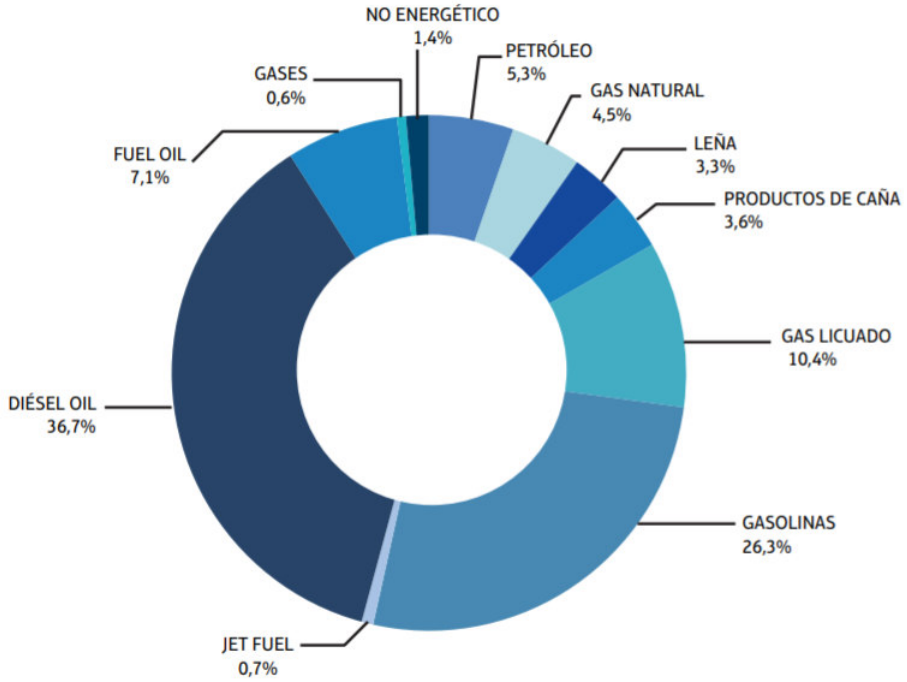

Figura 1.1: Emisiones de GEI por fuente (%) [20].

tigación busca analizar la relación que existe entre el consumo de energía y las emisiones de dióxido de carbono en Ecuador.

1.4.3. Fundamentos teóricos

En esta sección del capítulo se presentan definiciones y técnicas de la minería de datos; estas permitirán proceder con los cálculos en el ámbito relacionado a la estimación y predicción de datos.

De acuerdo con Alasadi y Bhaya (2017) en su publicación “Revisión de las Técnicas de Preprocesamiento de Datos en la Minería de Datos”; la minería de datos es el proceso de extracción de patrones y modelos útiles sobre un enorme conjunto de datos, los cuales desempeñan un papel eficaz en la toma de decisiones. Además, mencionan que la minería de datos depende básicamente de la calidad de los datos; los datos en bruto suelen ser susceptibles de contener datos atípicos, datos incoherentes y/o perdidos. Por lo tanto, es necesario que sean preprocesados antes de ser minados (ver Figura 1.2) [2].

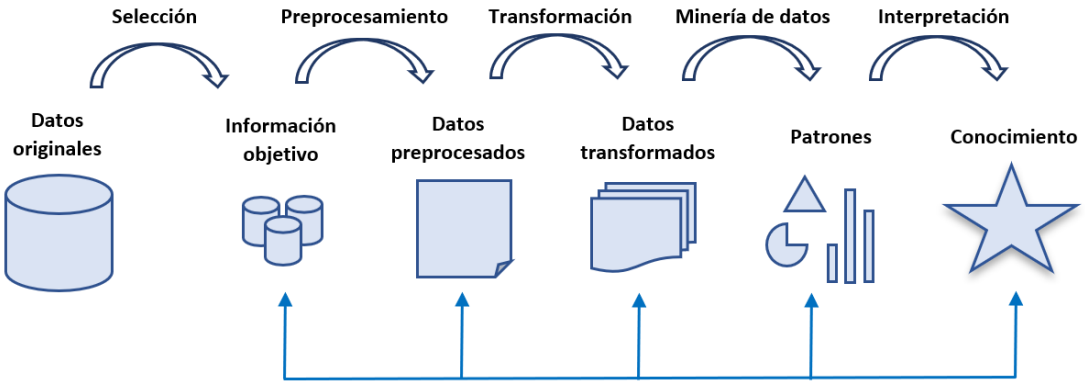

Figura 1.2: Elaboración propia

Regresión Lineal Múltiple mediante MCO

La Regresión Lineal Múltiple (RLM) es una técnica estadística clásica que posee varias ventajas, entre ellas están la interpretabilidad, la simplicidad y la posibilidad de ajuste sobre las transformaciones de las variables. En la mayoría de los problemas de investigación en los que se aplica el análisis de regresión se necesita más de una variable independiente para el modelo de regresión. Es por ello por lo que, el modelo de RLM y su estimación mediante Mínimos Cuadrados Ordinarios es sin duda una herramienta muy popular y útil en diferentes campos de la ciencia, dicho modelo permite estimar la relación entre una variable predictoria (dependiente) y un conjunto de variables explicativas (independientes).

Sea la variable cuantitativa Y y sea el conjunto de $k \in \mathbb{N}$ variables predictoras x_1, x_2, $\cdots$, x_k; el Modelo de RLM supone que la media de Y determina los valores de las variables predictoras en una combinación lineal:

$$E(\Upsilon) = \mu_{\Upsilon|x_1,x_2,\ldots,x_k+} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

donde,

β_0: es la ordenada en el origen. El valor de la variable dependiente y cuando todos los predictores son cero.

β_i: para todo $i \in \{1, \cdots, k\}$. Es el efecto promedio del incremento en

una unidad de x_i sobre la variable dependiente y. Se conocen como coeficientes de regresión.

La respuesta estimada de $E(\Upsilon)$ se obtiene a partir de la ecuación de regresión muestral

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k,$$

donde cada β_i se estima por medio de b_i, a partir de los datos muestrales, usando el método de mínimos cuadrados ordinarios. Por ejemplo, si $E(\Upsilon)$ es una función de las variables x_1 y x_2, se puede escoger una aproximación plana a la verdadera respuesta media haciendo uso del modelo lineal $E(\Upsilon) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ de esta manera, $E(\Upsilon)$ sería una función lineal de β_0, β_1, β_2 y representa un plano en el espacio y, x_1, x_2 (ver Figura 1.3) [32].

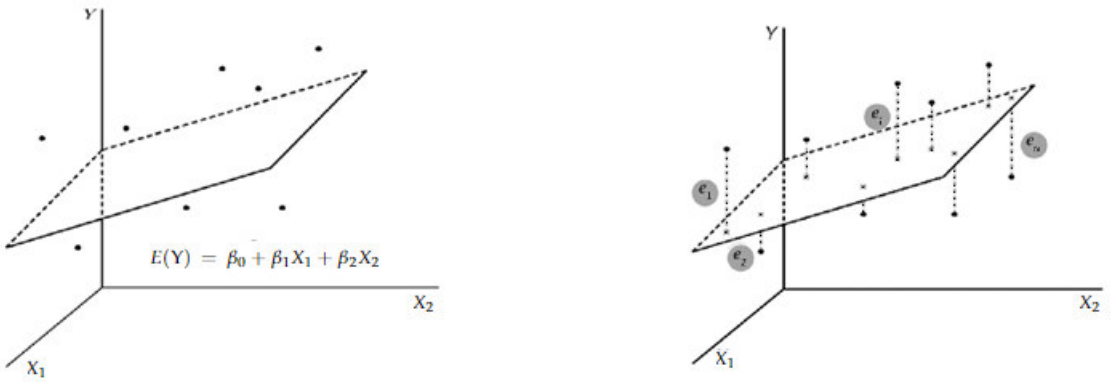

Figura 1.3: Modelo de RLM con dos predictores [9]

Estimación de coeficientes (Método de Mínimos Cuadrados Ordinarios)

El método de Mínimos Cuadrados Ordinarios es uno de los procedimientos para estimar los parámetros de cualquier modelo lineal dado que ajusta un hiperplano que pasa por un conjunto de n puntos. Dicho método es intuitivo debido a que buscamos que las diferencias entre los puntos correspondientes en el hiperplano ajustado y los valores observados sean lo menor posible; una manera eficaz de lograrlo y de que se obtengan buenas propiedades, es minimizar la suma de cuadrados de las desviaciones verticales mediante el hiperplano ajustada (ver Figuras 1.3) [32].

Sean $k \in \mathbb{N}$ y el modelo estadístico lineal

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_i \qquad \forall i \in \{1, \cdots, k\},$$

Haciendo $n \in \mathbb{N}$ observaciones independientes, $y_1,\ y_2,\ \cdots,\ y_n$, en Y. Se puede escribir la observación y_i de la siguiente manera

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i,$$

o bien,

$$y_i = \hat{y} + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,$$

donde,

x_{ij}: es el ajuste de la j-ésima variable independiente para la i-ésima observación.

ϵ_i: es el error aleatorio asociado con la respuesta y_i.

e_i: es el residual asociado con el valor ajustado $\hat{y}$.

De manera similar al caso de la RLM, se supone que los ϵ son independientes y siguen una distribución normal σ^2. Ahora, usando las definiciones del método de Mínimos Cuadrados Ordinarios para calcular $b_0,\ b_1,\ \cdots,\ b_k$, procedemos a diferenciar la siguiente expresión respecto a $b_0,\ b_1, \cdots,\ b_k$

$$SCE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{2i} - b_1 x_{2i} \cdots - b_k x_{ki}\right)^2$$

e igualamos a cero; es decir,

$$\frac{\partial SCE}{\partial b_i} = 0 \quad \forall i \in \{1, \cdots, k\}.$$

De esta manera se generan $k + 1$ ecuaciones normales para la regresión

lineal múltiple

$$\begin{array}{ccccccccccc}
nb_0 & + & b_1\sum_{i=1}^{n} x_{1i} & + & b_2\sum_{i=1}^{n} x_{2i} & + & \cdots & + & b_k\sum_{i=1}^{n} x_{1i} & = & \sum_{i=1}^{n} y_i \\
b_0\sum_{i=1}^{n} x_{1i} & + & b_1\sum_{i=1}^{n} x_{1i}^2 & + & b_2\sum_{i=1}^{n} x_{1i}x_{2i} & + & \cdots & + & b_k\sum_{i=1}^{n} x_{1i}x_{ki} & = & \sum_{i=1}^{n} x_{1i}y_i \\
\vdots & \vdots & \vdots & + & \vdots & + & \ddots & + & \vdots & = & \vdots \\
b_0\sum_{i=1}^{n} x_{ki} & + & b_1\sum_{i=1}^{n} x_{ki}x_{1i} & + & b_2\sum_{i=1}^{n} x_{ki}x_{2i} & + & \cdots & + & b_k\sum_{i=1}^{n} {x_{ki}}^2 & = & \sum_{i=1}^{n} x_{ki}y_i
\end{array}$$

Finalmente, los valores de b_0, b_1,$\cdots$, b_k se obtienen al resolver el sistema de ecuaciones lineales. Una manera práctica de manejar los resultados y deducciones de la Regresión Lineal Simple a la RLM es mediante álgebra de matrices; es por ello por lo que a continuación se usarán matrices para expresar los resultados en torno a la estimación de parámetros mediante el método de Mínimos Cuadrados Ordinarios. Sea $x_0 = 1$

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ {}_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} * \begin{bmatrix} 1 & x_{11} & \cdots & {}_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} {x_{1i}}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} {x_{ki}}^2 \end{bmatrix},$$

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}.$$

$$\Rightarrow (X'X)b = X'Y$$

$$\Rightarrow b = (X^{'}X)^{-1}X^{'}Y$$

Regresión potencial mediante OEP

Siguiendo la misma idea que en la sección anterior, consideremos el Modelo de RLM:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i,$$

este modelo de relación lineal entre y_i y x_{ki} es una versión restringida del modelo

$$y_i = \beta_0 + \beta_1 x_{1i}^{\alpha_1} + \beta_2 x_{2i}^{\alpha_2} + \cdots + \beta_k x_{ki}^{\alpha_k} + \epsilon_i,$$

donde los $\alpha_i = 1$. A este nuevo modelo lo denominaremos Modelo de Regresión Potencial (RPG). Como extensión directa del caso lineal, el modelo potencial es adecuado para estudiar posibles no linealidades entre dos variables, cuando ya se ha analizado el modelo lineal [15]. También se puede utilizar para estudiar la no linealidad del efecto de algunas variables explicativas o predictores x_i sobre y_i en una regresión múltiple. Además, el uso de este tipo de modelos reduce la influencia de los valores atípicos y, en consecuencia, aumenta la robustez del análisis (Box y Tiao, 1962; West, Liu y Wang, 2007) [5], [16].

Estimación de coeficientes (Método de Optimización por Enjambre de Partículas)

En términos matemáticos, el problema de optimizar (minimizar/maximizar) una función sobre una región se puede enunciar de la siguiente manera:

$$f : \Omega \subseteq \mathbb{R}^n \longrightarrow \mathbb{R},$$

$$\Omega = \{x \in \mathbb{R}^n : h_i(x) = 0,\ i = 1, \cdots M \wedge g_j(x) = 0,\ j = 1, \cdots N\}$$

Ω representa el conjunto de todos los puntos de $\mathbb{R}^n$ que satisface ciertas condiciones de igualdad y desigualdad.

Entonces se denota el problema de minimización de $f(x)$ como

$$\min_{x \in \Omega} \quad f(x) \tag{1.1}$$

La matemática ofrece una serie de métodos que mediante condiciones necesarias y suficientes, caracterizan lo puntos que optimizan una función. Dichas condiciones, que casi siempre requieren calcular derivadas, requieren del tipo de problema. El conjunto Ω juega un papel crucial porque el problema en cuestión podría ser catalogado como irrestricto si $\Omega = \mathbb{R}^n$; así mismo, puede ser catalogado como problema sin restricciones en caso contrario ($\Omega \neq \mathbb{R}^n$). [35]

Los algoritmos convencionales como la programación lineal, la programación no lineal y los métodos de puntos interiors se han utilizado ampliamente durante décadas para resolver problemas de optimización. Sin embargo, estos métodos convencionales necesitan varios supuestos matemáticos, entre ellos están las propiedades diferenciales de las funciones objetivo y mínimos únicos existentes en los dominios del problema y a menudo atrapan en soluciones óptimas locales [8].

En los últimos años, algunos métodos de inteligencia artificial artificial, como el algoritmo genético, el recocido simulado y la programación evolutiva evolutiva [4], se han aplicado a problemas de optimización para llegar al óptimo global. Un nuevo método de computación evolutiva, denominado Optimización por Enjambre de Partículas (OEP), ha sido propuesto por Russell Eberhart y James Kennedy en 1995 y ha demostrado ser un competidor en el campo de la optimización. OEP, inspirado en el comportamiento de los pájaros o de los peces, es una herramienta de optimización basada en la población. La idea que se sigue es: iniciar con un conjunto de puntos elegidos al azar en la región que se busca el mínimo de una función, a ese conjunto de puntos se lo denomina *enjambre* y a cada uno de los puntos *partículas*. A lo largo de cada iteración del algoritmo, dichas partículas se moverán con una cierta componente de aleatoriedad; esta componente no es totalmente aleatoria porque las partículas también serán atraídas por la que tenga el mínimo valor global en la iteración. A la larga, ocurrirá que la función no puede minimizarse más o que se han llevado a cabo un número máximo permitido de itera-

ciones. Finalmente, se arroja como solución la partícula que en la última iteración asume el mínimo valor global [14].

Dentro del enjambre, cada partícula recorre el espacio de búsqueda multidimensional con una determinada velocidad; esta, se actualiza constantemente por la mejor posición de la partícula ($pbest$) y por la mejor posición global ($gbest$) encontrada por todo el grupo hasta el momento [34]. Los grupos alcanzarán su mejor posición a través de la comunicación de los miembros del enjambre que ya obtuvieron una mejor condición; lo que conlleva converger al estado que sea más conveniente para todos los miembros del enjambre. Este proceso se repetirá hasta que se descubra la mejor solución (ver Figura 1.4) [27]. La velocidad de cada partícula está determinada por el conocimiento intelectual de la partícula y el conocimiento social. La posición actual de cada partícula $i \in \{1, \cdots, n\}$ $\forall n \in \mathbb{N}$, está representada por

$$X_i = (x_{i1} + x_{i2} + \cdots + x_{iD}),$$

donde:

D: es la dimension del espacio de exploración.

La mejor posición personal de cada partícula está representada por

$$pbest_i = (pbest_{i1} + pbest_{i2} + \cdots + pbest_{iD}),$$

y la velocidad actual de cada partícula viene dada por

$$V_i = (v_{i1} + v_{i2} + \cdots + v_{iD}).$$

La mejor posición global encontrada por el enjambre hasta cierto momento es

$$best = (gbest_1 + gbest_2 + \cdots + gbest_D).$$

En el algoritmo OEP, los vectores de velocidad gobiernan la forma en que las partículas se mueven en el espacio de búsqueda D, estos vectores

se componen por la contribución de tres términos. El primero es conocido como la inercia o el impulso, el cual evita que la partícula cambie drásticamente de dirección, al mantener la dirección del flujo anterior. El segundo está basado en la experiencia personal que ha desarrollado durante su desplazamiento en el espacio de búsqueda. Este comportamiento es conocido como *componente cognitivo*, es denotado por $(p_{id} - x_{id})$; y se define como la posición devuelta por el mejor valor de la función objetivo de entre todas las posiciones visitadas por la partícula en cuestión. Y finalmente el tercero está relacionado con la experiencia de todas las partículas. Este comportamiento es conocido como *componente social*, es denotado por $(p_{gd} - x_{id})$; y se define como la posición devuelta por el mejor valor de la función objetivo de entre todas las posiciones visitadas por todas las partículas [28].

Sea $d \in D$, las ecuaciones de actualización del método OEP son,

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1}$$

$$v_{id}^{t+1} = v_{id}^{t} + \phi_1 \cdot c_1 \cdot (p_{id}^{t} - x_{id}^{t}) + \phi_2 \cdot c_2 \cdot (p_{gd}^{t} - x_{id}^{t}),$$

donde,

ϕ_1, ϕ_2: son constantes de aceleración que indican la confianza individual y la confianza social, respectivamente. Eso significa que tanto la componente cognitiva como la social tienen una influencia estocástica en la regla de actualización de la velocidad.

c_1, c_2: representan a matrices diagonales con números que siguen una distribución uniforme en $[0, 1]$.

p_{id}^{t}: es la mejor posición de la partícula alcanzada hasta el momento t.

p_{gd}^{t}: es la posición global de la mejor partícula alcanzada hasta el momento t.

En resumen, la nueva posición de la partícula i en la posición $t + 1$ es una combinación lineal entre la dirección hacia la mejor partícula en

la iteración t y la mejor dirección de la partícula i en esa iteración. Todo esto, con una influencia de aleatoriedad, el efecto de v_{id}^t es mantener a la partícula en su posición actual.

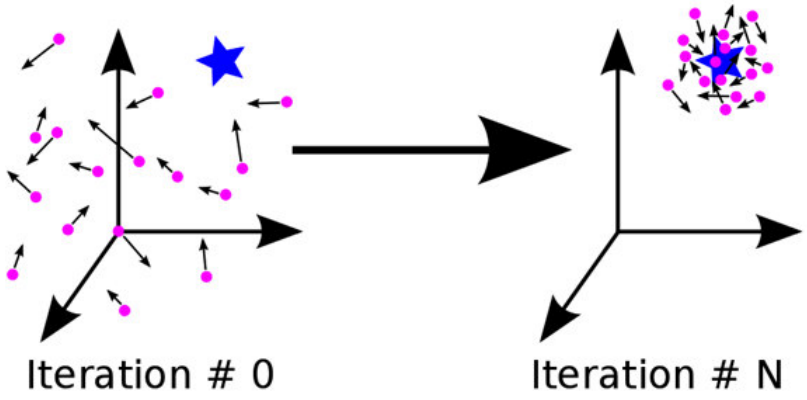

Figura 1.4: Comportamiento del algoritmo OEP [37].

Aplicamos este método para encontrar las estimaciones de $\beta_0,\ \beta_1,\ \cdots,\ \beta_n$ como solución del problema de optimización

$$\min_{x\in\mathbb{R}^{\times}} \quad \sum_{i=1}^{n}(y_i - b_0 - b_1x_{1i} - \cdots - b_kx_{ki}) \tag{1.2}$$

Modelo Aditivo Generalizado mediante p-splines

De manera similar a la sección anterior, presentamos al Modelo Aditivo Generalizado (MAG) como una extensión del Modelo de Regresión Lineal Múltiple. Los MAG fueron desarrollados por Hastie y Tibshirani (1986, 1990); y en estos se presenta la posibilidad en la que el efecto de las variables predictoras sobre la variable dependiente sea de forma desconocida y suave, ya no únicamente lineal; llevándonos a obtener más información entre la variable objetivo y las variables independientes.

Desde este punto de vista, la estructura del modelo podría ser planteada de la siguiente manera:

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_2(x_{3i}) + f_4(x_{4i}) + \cdots, \quad i = 1, 2, \cdots, n$$

donde,

$\mu_i = E(Y_i)$.

Y_i: es una variable de respuesta la cual sigue una distribución de la Familia Exponencial.

X_i^*: i-ésima fila de la matriz correspondiente a las covariables que definen las componentes paramétricas del modelo.

Para $j = 1,\ 2,\ \cdots,\ k, \quad f_j$: son funciones suaves no pre-especificadas (provienen de splines con penalización).

θ: vector de coeficientes de regresión [33].

En esta subsección se esclarece cómo los MAGs pueden representarse utilizando splines de regresión penalizados, estimados por métodos de regresión penalizados, y cómo el grado apropiado de suavidad para el $f_j, \quad j = 1, 2, \cdots, k$ puede estimarse a partir de los datos utilizando la validación cruzada.

Una pregunta que surge es **¿qué es un spline y por qué utilizarlos?**

De acuerdo con Subana Shanmuganathan y Sandhya Samarasinghe (2016), "El término 'spline' se utiliza para referirse a una amplia clase de funciones que se utilizan en aplicaciones que requieren interpolación y/o suavización de datos. Los datos pueden ser unidimensionales o multidimensionales " [30].

De manera general, al especificar el MAG sólo en términos de "funciones suaves", en lugar de relaciones paramétricas detalladas, es posible evitar algunos tipos de modelos poco manejables o tediosos. Esta flexibilidad tiene como contrapartida dos nuevos problemas. Por un lado es necesario representar las funciones suaves de alguna manera, y por otro elegir cuán suaves deben ser.

El *primer problema* es resuelto mediante las técnicas de estimación de f_j, por ejemplo, se puede usar: función de la media, Mínimos Cuadrados Ordinarios, estimación tipo kernel o splines [10]. Lo que sigue, es definir a las funciones suaves.

Lo que sigue es definir una base para cada función f_j; por ello definimos las bases de funciones.

Bases de Funciones

Comencemos nuestro análisis desde la perspectiva del Modelo de Regresión Lineal Simple; es decir, tendremos en cuenta una sola variable predictora X. Estaríamos busando una función f que satisfaga todos los supuestos ya conocidos, y la mejor manera es considerando un modelo que contenga una de estas funciones como covariables; es decir

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \cdots, n.$$

Si deseamos aplicar los métodos estadísticos mencionados hasta ahora, sería necesario definir una base de funciones b_j conocidas de dimensión n, de la cual f forme parte. En (Wood, 2006) se propone realizar un ajuste polinomial; es decir, tomar funciones básicas y combinarlas

$$b_0(x) = 1, \quad b_1(x) = x, \quad b_2(x) = x^2, \quad \cdots, \quad b_m(x) = x^m,$$

con $n \in \mathbb{N}$ de tal forma que definiendo $i \in [1, m]$ y si $b_i(x)$ es la i-ésima función de la base, tenemos la siguiente clase de bases de funciones, donde

$$f : [a, b] \longrightarrow \mathbb{R}$$

es un spline polinómico de grado m, si satisface:

1. $f(x)$ es $m - 1$ veces continuamente diferenciable
2. $f(x)$ es un polinomio de grado m con $x \in [n_j, n_{j+1}), \quad j = 1, 2, \cdots, i - 1$

Además, con la ayuda de un vector de parámetros β se puede representar cada spline polinómico por una base $n = i + m - 1$ funciones como se sigue

$$f(x) = \sum_{j=1}^{n} b_j(x)\beta_j$$

lo que junto a la función y_i, da lugar a la representación de cada función suave f de la siguiente manera

$$y_i = \sum_{j=1}^{n} b_j(x_i)\beta_j + \epsilon_i, \quad i = 1, 2, \cdots, n$$

Se podría considerar una base simple donde los polinomios estén formados por el orden que se considera f; sin embargo, se presentan algunos problemas cuando f está definida en el dominio $[0, 1]$ En esta investigación se hará uso de bases formadas por P**-Splines** para la construcción de las funciones f. Comenzamos definiendo una base formada por Basic-splines (B**-splines**), estas bases de grado m son resultado de fusionar $m + 1$ polinomios de grado m en los $m - 1$ nodos del spline.

Por otro lado, los P-splines son una herramienta flexible para el suavizado de funciones, están basados en la regresión con funciones de base local: B-splines. En 1986, O'Sullivan se dio cuenta de que si modelamos una función como una suma de B-splines, la conocida medida de rugosidad, la segunda derivada cuadrada integrada, puede expresarse como una función cuadrática de los coeficientes. Las P-splines van un paso más allá: utilizan B-splines igualmente espaciadas y descartan la derivada por completo. La rugosidad se expresa como la suma de los cuadrados de las diferencias de los coeficientes. Las diferencias son muy fáciles de calcular y la generalización a órdenes superiores es sencilla.

El *segundo problema* trata sobre cuán suave es la función, para ello, la dimensión de la base es fijada en un tamaño un poco mayor al que se crea necesario. Para controlar la suavidad del modelo, se añaden penalizaciones por posibles "ondulaciones" a la función de ajuste por Mínimos Cuadrados Ordinarios

$$||y - X\beta||^2 + \lambda \int_0^1 [f''(x)]^2 dx$$

donde,

$\int_0^1 [f''(x)]^2 dx$: es un término de penalización; el modelo es penalizado por ondulaciones.

λ: parámetro suavizante. Ilustra la compensación entre el modelo ajustado y el modelo suavizado.

Sea S una matriz de coeficientes conocidos y sea la $||\cdot||$ la norma usual, como f es lineal en los parámetros β_i, el término de penalización puede ser escrito como

$$\int_0^1 [f''(x)]^2 dx = \beta' S \beta,$$

donde $S = \int b_j''(x)[b_j''(x)]' dx$.

Entonces, para ajustar un spline de regresión penalizado, se debe minimizar

$$||y - X\beta||^2 + \lambda \beta' S \beta.$$

Así, la tarea de estimar el grado de suavidad del modelo es resumido en estimar el parámetro suavizante λ. Además, es relevante considerar la estimación de β dado λ

$$\hat{\beta} = (X'X + \lambda S)^{-1} X' y.$$

Estimación del parámetro de suavización λ (validación cruzada)
El valor de λ es importante porque si se hace una buena elección, la función del spline ($\hat{f}$) será lo más cercana posible a f. Un valor alto de λ hará que los datos se suavicen en exceso, provocando sobreajuste; por el lado contrario, si λ es pequeño, les faltará ajuste a los datos.

Un correcto criterio para elegir λ es la mininización del estimador M

$$M = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i(x_i) - f_i(x_i))^2,$$

M puede utilizarse directamente porque f no es conocida, pero sí es posible obtener una estimación del error cuadrático esperado de la nueva variable $E(M)+\sigma^2$- Sea $\hat{f}^{-i}$ el modelo que ajusta a todos los datos excepto a y_i y sea el score ordinario de validación cruzada:

$$\vartheta_0 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2.$$

Dicho score es el resultado de omitir cada dato sucesivamente, mediante el ajuste del modelo a los datos restantes y de calcular el promedio de las diferencias al cuadrado entre el dato que falta y su valor predicho. Esta técnica se basa en dos tipos de Splines, los Splines de Regresión y los Splines de suavizado. Reemplazando $y_i = f_i + \epsilon_i$, tenemos

$$\begin{aligned}\vartheta_0 &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2 - (\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2.\end{aligned}$$

Dado que ϵ y $\hat{f}_i^{[-i]}$ son independientes, que $E(\epsilon_i) = 0$; si tomamos la esperanza de la igualdad anterior

$$E(\vartheta_0) = \frac{1}{n}E\left(\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2\right) + \sigma^2$$

Dado que para n suficientemente grande, $\hat{f}_i^{[-i]} \approx f_i$; entonces, $E(\vartheta_0) = E(M) + \sigma^2$. Por tanto, elegir λ para para minimizar ϑ_0 es un enfoque razonable si el ideal sería minimizar M. Elegir λ para minimizar ϑ_0 se conoce como *validación cruzada ordinaria* [33].

Capítulo 2

Metodología

Teniendo en consideración que el tema a investigar tiene suficiente sustento teórico, el presente trabajo de investigación curricular será de tipo **no experimental** ya que se observará el fenómeno de las emisiones de CO2 en su contexto natural a partir de variables, categorías y conceptos; es decir, observaremos situaciones ya existentes y posteriormente a ello, se procederá al análisis de las mismas. Esto implica que no se podrá manipular, controlar o en general, alterar las observaciones; ya que este tipo de investigación buscará llegar a interpretaciones o conclusiones a partir del estudio de cómo exactamente ocurrieron los fenómenos.

Dentro de los distintos tipos de diseños de investigación no experimental, el tipo de **diseño longitudinal** es el más adecuado para nuestra investigación puesto que los datos con los que trabajaremos son históricos; es decir, se han obtenido en distintos períodos de tiempo y esto nos permitirá la realización de análisis respecto al desarrollo de las relaciones que existen entre las variables y a los cambios a lo largo del tiempo. Asimismo, se busacará conocer sus determinantes y consecuencias principales.

Por otro lado, el presente trabajo será diseñado bajo el planteamiento metodológico del **enfoque cuantitativo**. Para probar hipótesis establecidas y/o contestar preguntas de investigación, se hace uso del enfoque

cuantitativo porque en este, se utiliza la recolección y el análisis de datos. Examinaremos datos de manera numérica, se establecerán patrones de comportamiento y en general, estudiaremos las relaciones existentes entre los menncionados datos. Por estas razones, se deduce que el enfoque cuantitativo se adapta a las características y necesidades de la investigación.

Según lo mencionado por Hernández, Fernández y Baptista en su libro Metodología de la investigación (sexta edición), la investigación no experimental cuantitativa es aquella "que se realiza sin manipular deliberadamente variables. Es decir, se trata de estudios en los que no hacemos variar en forma intencional las variables independientes para ver su efecto sobre otras variables."[11].

2.1. Recolección de información

Del enfoque cuantitativo presentado anteriormente, se tomará la técnica de análisis de bases de datos para la estimación de emisiones de CO2 en Ecuador, debido a que la información se obtendrá del Balance Energético Nacional (BEN) 2020 documento donde se consolida toda la información referente a energía en Ecuador. El BEN se crea por la necesidad de disponer de una herramienta actualizada la cual sirva como referencia para la toma de decisiones en torno a los acontecimientos del sector energético ecuatoriano. Según Juan Carlos Bermeo Calderón, Ministro de Energía y Recursos Naturales No Renovables, para la elaboración de este reporte se considera la metodología propuesta por la Organización Latinoamericana de la Energía (OLADE) publicada en 2017 [20].

El proceso para la recopilación de información se realizó con las instituciones pertinentes y encargadas de registrar los dato socioeconómicos y energéticos del Estado ecuatoriano; por otro lado, la Agencia de Regulación y Control de Energía y Recursos Naturales No Renovables fue la encargada de la integración y procesamiento de datos en las planillas y la validación de la información energética. El documento correspondiente al

BEN contiene las principales tendencias de producción, transformación, emisiones y consumo del sector energético; asimismo, se pueden evidenciar las posibles consecuencias de factores externos, como la pandemia originada por el Covid-19.

Dentro del BEN se especifica, como notas metodológicas, que el documento no incluye pérdidas del sector hidrocarburífero, esto debido a que hasta el momento no se encuentra de información disponible. Asimismo, se realiza una constante revisión y actualización de la metodología para la estimación de las series históricas correspondientes al consumo de diésel oil y fuel oil en el sector comercial. Desde los despachos industriales se extrae la cantidad estimada de diésel oil y fuel oil en el sector comercial; por tanto, las series históricas de consumo referentes al sector industrial también son actualizadas.

2.2. Formulación del problema

En el presente trabajo de investigación se utilizará información correspondiente a las emisiones anuales de CO2 en Ecuador a partir de diferentes fuentes de consumo, dicha información es considerada desde el año 1965 hasta el año 2020 (ver Anexo 1). Según Munim, Hakim y Abdullah (2010), los principales indicadores del consumo de energía por fuente son: Consumo de Carbón, consumo de productos derivados del Petróleo, consumo de Gas Natural y consumo de Electricidad [21] y [29].

Según lo publicado en el BEN 2020, entre los años 2010 y 2020, los combustibles fósiles han representado un 81 % correspondiente a los requerimientos energéticos en el Ecuador. Por otro lado, se ha evidenciado un crecimiento del 55,1 % que corresponde a la demanda de energía eléctrica. Finalmente, el gas licuado de petróleo (GLP) presenta un crecimiento durante el mismo período de 28,8 % [20].

Para el caso ecuatoriano se propone la siguiente clasificación (ver Figura 2.1):

1. Consumo de derivados del Petróleo ($Consumo_{DP}$):

- Gasolina
- Jet fuel
- Diésel oil
- Fuel oil
- Gas licuado
- Otros derivados de hidrocarburos

2. Consumo de Energía Primaria ($Consumo_{EP}$):
   - Leña
   - Bagazo de caña
   - Melaza y jugo de caña
   - Otras primarias
3. Consumo de Electricidad ($Consumo_{EL}$):
   - Hidráulica
   - Térmica
   - No convencionales
4. Consumo de Gas Natural ($Consumo_{GN}$)

Cada categoría presentada anteriormente corresponde a una variable predictora: $Consumo_{DP}$, $Consumo_{EP}$, $Consumo_{EL}$ y $Consumo_{GN}$. Cada una de estas variables de estudio contiene datos del consumo anual en Miles de Barriles Equivalentes de Petróleo ($kBOE$), unidad de energía que equivale a la energía liberada en el proceso de quema de aproximadamente un barril de petróleo crudo de 42 galones estadounidenses o 158,987 litros. Esta unidad de $kBOE$ representa aproximadamente:

- $0,000\ 146$ Megatonelada equivalente de petróleo ($Mtoe$).
- $1'\ 699\ 406,\ 44$ Kilovatios hora (kWh).
- $6'\ 117\ 863,\ 20$ Megajulios (MJ).
- $5'\ 798\ 615\ 480,\ 70$ Unidades Térmicas Británicas (BTU).

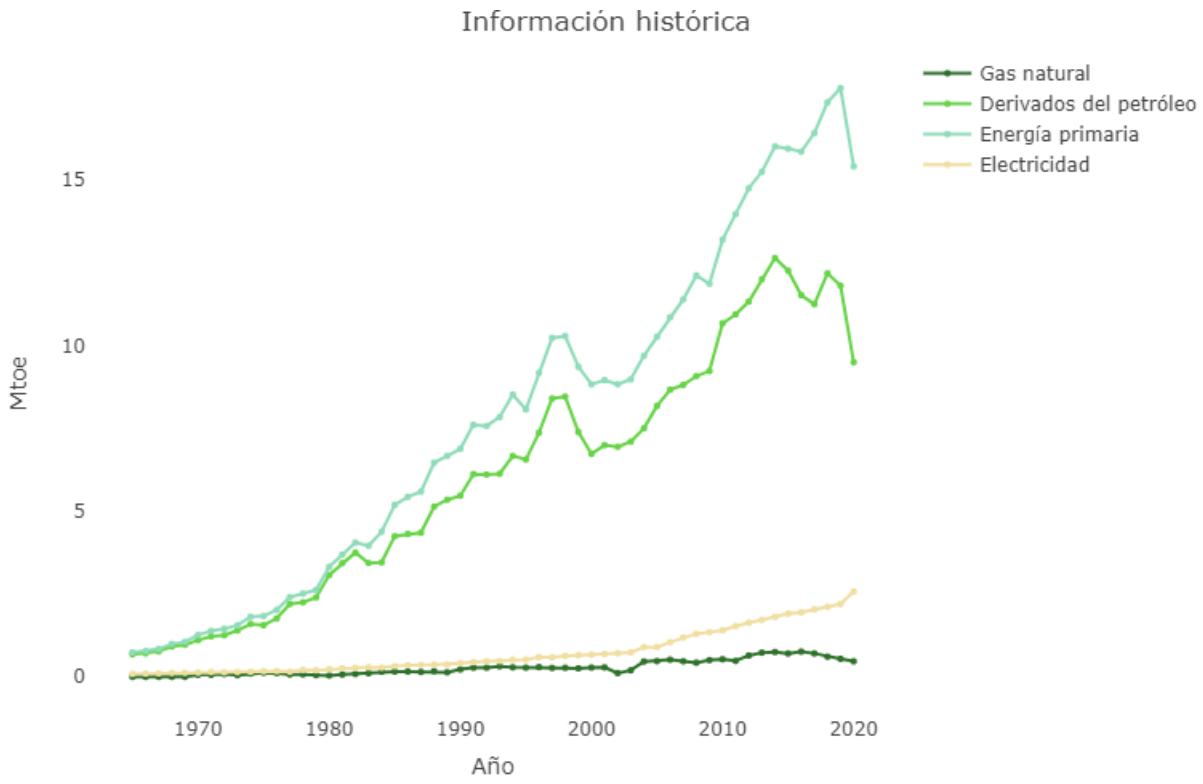

Figura 2.1: Información histórica del consumo de energía por fuente (Elaboración propia)

Para una mejor visualización y mejor manejo de los datos, se realizará la conversión de unidades $kBOE$ a $Mtoe$. Esta última se emplea en la economía y en la industria para poder expresar grandes cantidades de energía, es equivalente a la energía que rinde un millón de toneladas de petróleo; y es utilizada como parámetro de comparación de los niveles de emisión de CO2 a la atmósfera generados por la quema de diversos combustibles [31].

Dado que los objetivos a cumplir; en general, buscan la estimación de las emisiones de CO2 a partir de distintas fuentes de consumo de energía, se plantearán tres tipos de modelos estadísticos: Regresión Lineal Múltiple, Regresión Potencial Generalizada y Modelo Aditivo Generalizado.

2.3. Estimación de las emisiones de CO2

La estimación de emisiones de CO2 se lleva acabo mediante una ecuación de forma lineal aplicando RLM con Mínimos Cuadrados Ordinarios y dos ecuaciones de forma no lineal aplicando RPG con optimización por enjambre de partículas y GAM con splines de penalización. Los modelos se basarán en los cuatro indicadores socioeconómicos, a saber, el consumo de derivados del petróleo, consumo de gas natural, consumo de electricidad y consumo de energía primaria; debido a que estos indicadores se consideran como los máximos emisiones de CO2 en el país [6].

A partir de esta sección se hará uso del software libre R, el cual permite realizar análisis estadístico. R es utilizado bajo un entorno de de desarrollo integrado conocido como RStudio ya que posee características y funcionalidades que permitirán implementar las bases teóricas de una manera más sencilla en comparación al entorno original de R.

2.3.1. Tamaño de la muestra

Para la estimación de las emisiones de CO2, los datos anuales fueron tomados desde 1965 hasta 2020. Por otro lado, nos basamos en lo expuesto por Abu-Mostafa, Magdon-Ismail y Lin (2012) en su libro *Learning from data* para hallar el tamaño de la muestra de entrenamiento. Dichos autores desarrollan una teoría matemática que sirve como base para hallar los tamaños de entrenamiento y de prueba. Se introduce la **complejidad de la muestra**, la cual denota el número de observaciones que se deben considerar en el conjunto de entrenamiento para lograr un determinado rendimiento de generalización en el modelo. El rendimiento se especifíca mediante dos parámetros: ε y δ [36]; la ecuación para el tamaño de la muestra de entranmiento viene dado por la ecuación **(9)**.

$$N \geq \frac{8}{\varepsilon^2} \ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right), \qquad \textbf{(9)}$$

donde,

ε: es la tolerancia de error. Determina el error de generalización per-

mitido.

δ: es el parámetro de confianza. Determina el nivel de confianza.

$\mathcal{H}$: es el conjunto de hipótesis. Determinado por todas las funciones que el algoritmo podría escoger para dar la mejor estimación de los valores observados.

$m_{\mathcal{H}}(N)$: es el máximo número de dicotomías que se pueden generar en N puntos.

Para el propósito de este análisis, nos centraremos en funciones objetivo binarias, por lo que cada $h \in \mathcal{H}$ ubica al espacio de observaciones ($\mathcal{X}$) en {-1, +1}; esto debido a que cada función h puede o no ser la que mejor se ajuste a nuestros datos. Si $h \in \mathcal{H}$ se aplica a una muestra finita $x_1, \cdots, x_N \in \mathcal{X}$ obtenemos una N-tupla $h(x_1), \cdots, h(x_N)$ de $\pm$ 1's. Dicha N-tupla se denomina *dicotómica* ya que divide $x_1, \cdots, x_N$ en dos grupos: aquellospuntos para los que h es -1 y aquellos para los que h es +1. Las dicotomías generadas por $\mathcal{H}$ en los puntos $x_1, \cdots, x_N \in \mathcal{X}$ están definidas por

$$\mathcal{H}(x_1, \cdots, x_N) = \{(h(x_1), \cdots, h(x_N)) \mid h \in \mathcal{H}\}$$

Un $\mathcal{H}(x_1, \cdots, x_N)$ significa que H es más 'diverso' generando dicotomías sobre $x_1, \cdots, x_N$.

Listing 2.1: Código R - Tamaño para los conjuntos de Entrenamiento y Validación

```
error<- c(0.47,0.47,0.47,0.48,0.48,0.49,0.49)
nivel<- c(0.9,0.95,0.98,0.9,0.95,0.9,0.95)
N<- ceiling( (8/error^2)*log(4/nivel)  )

train<- N/(56)
validacion<- 1-train

tabla1<- data.frame( Error = error, Nivel=nivel, Entrenamiento =
    train,  Validacion = validacion)
tabla1
```

Error	Confianza	% Entrenamiento	% Validación
0.47	0.90	0.9821429	0.01785714
0.47	0.95	0.9464286	0.05357143
0.47	0.98	0.9107143	0.08928571
0.48	0.90	0.9285714	0.07142857
0.48	0.95	0.8928571	0.10714286
0.49	0.90	0.8928571	0.10714286
0.49	0.95	0.8571429	0.14285714

Cuadro 2.1: Tamaño en % del conjunto de entrenamiento y validación

Debido a que abordaremos un problema de verificación; es decir, queremos conocer si la hipótesis empleada se aproxima de forma correcta a la función objetivo, el valor de $m_{\mathcal{H}}(2N)$ es igual a 1. De esta manera tendríamos una dicotomía. Por otro lado, considerando un error de $\varepsilon = 0.48$ y un nivel de confianza $\delta = 0.95$, emplearemos muestreo 90-10; es decir, el 90 % de los datos serán empleados para entrenar el modelo y el 10 % para su validación. Esto implica que los datos comprendidos entre los años 1965 y 2014 serán utilizados para entrenamiento mientras que los correspondientes a 2015 - 2020 serán utilizados como datos de validación.

2.3.2. Estimación RLM mediante MCO

Modelo de Regresión Lineal Múltiple mediante Mínimos Cuadrados Ordinarios

Técnica estadística conocida por su capacidad para el análisis y modelado de datos. La función a obtener corresponde a la combinación lineal de los parámetros del modelo. La ecuación de RLM se da en **(1)**

$$Y_{RLM} = a \cdot X_1 + b \cdot X_2 + c \cdot X_3 + d \cdot X_4 + e \qquad \textbf{(1)}$$

donde,

Y_{RLM}: es la variable respuesta e indica las estimaciones de la emisión de CO2.

X_1: es una variable predictora que indica el consumo de derivados del Petróleo.

X_2: es una variable predictora que indica el consumo de Gas Natural.

X_3: es una variable predictora que indica el consumo de Electricidad.

X_4: es una variable predictora que indica el consumo de Energía Primaria.

$a,\ b,\ c,\ d$: son los coeficientes de regresión.

e: es el término de error y se trata como una variable aleatoria, denota la variación no explicada de la variable respuesta.

Se aplica la función $lm()$ del paquete $stats$ para crear un modelo de regresión con una fórmula dada, en nuestro caso tenemos 4 variables predictoras que corresponden al consumo de derivados del Petróleo, de Gas Natural, de Electricidad y de Energía Primaria. El modelo estimado se muestra en el Cuadro 2.2 de la página 30.

Listing 2.2: Código R - Ajuste del modelo RLM

```
datos <- zData %>%
  dplyr::select(c(`Derivados del petr leo`, `Energ a primaria
     `, `Gas natural`, Electricidad, CO2))

set.seed(1234)
entrenamiento <- sample_frac(datos, .9)
val <- setdiff(datos, entrenamiento)
set.seed(1234)
modelo1 <- lm(CO2 ~ . , data = entrenamiento)
summary(modelo1)
```

Validación del modelo

- Relación lineal entre la variable dependiente y los predictores

Si la relación entre la variable dependiente y los predictores es lineal, los residuos deben distribuirse aleatoriamente en torno a 0 con una variabilidad constante a lo largo del eje x

Call:
lm(formula = CO2 ~., data = entrenamiento)

Residuals:

Min	1Q	Median	3Q	Max
-8.3358	-1.8911	-0.3673	1.6896	8.3562

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.8682	0.9152	3.134	0.00303	**
'Derivados del petroleo'	3.8031	1.4125	2.693	0.00993	**
'Energia primaria'	-1.1607	1.2071	-0.962	0.01141	*
'Gas natural'	7.5139	6.3696	1.180	0.04434	*
Electricidad	2.6531	3.0326	0.875	1.18629	

—
Residual standard error: 2.961 on 45 degrees of freedom
Multiple R-squared: 0.9449, Adjusted R-squared: 0.94
F-statistic: 192.8 on 4 and 45 DF, p-value: <2.2e-16

Cuadro 2.2: Estadísticas del modelo RLM estimado

Si en realidad los datos no están bien representados por un modelo lineal, los parámetros y sus interpretaciones carecerían de sentido; esto conlleva a que incluso las predicciones no sean acertadas. Para este caso, vemos que los residuos se distribuyen aleatoriamente en torno a 0; por lo tanto, se cumple el supuesto de linealidad para los predictores.

- Residuos con distribución normal

El **gráfico Q-Q**, o **gráfico de cuantiles a cuantiles**, es un gráfico que prueba la conformidad entre la distribución empírica y la distribución teórica dada. Es uno de los métodos utilizados para verificar la normalidad de los errores de un modelo de regresión es construir una gráfica Q-Q de los residuos.

Listing 2.3: Código R - Gráfico QQ, test SW (*Modelo RLM*)

```
## qqplot
qqnorm(modelo1$residuals)
qqline(modelo1$residuals)
## Test de Saphiro-Wilk (SW)
```

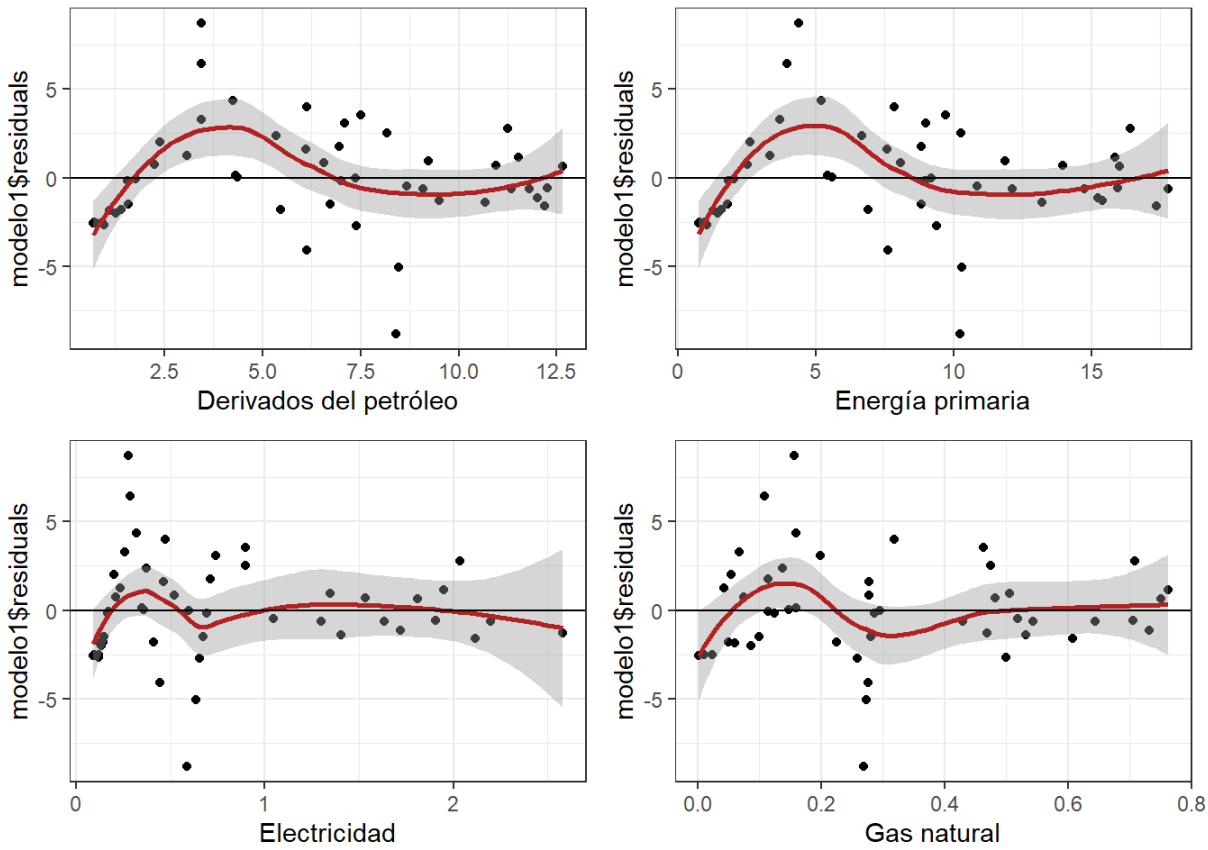

Figura 2.2: Relación lineal entre la variable respuesta y los predictores numéricos (*Modelo RLM*)

```
shapiro.test(modelo1$residuals)
```

A simple vista podemos ver que los puntos están alineados en la línea $x = y$, lo que nos da indicios de que los residuos tienen distribución normal.

Además, aplicando el **Test de normalidad de Shapiro-Wilk**

```
Shapiro-Wilk normality test

data: modelo1$residuals
W = 0.96304    p-value = 0.1192
```

El método de Shapiro-Wilk es ampliamente recomendado para la prueba de normalidad y proporciona una mejor potencia que el K-S. Se basa

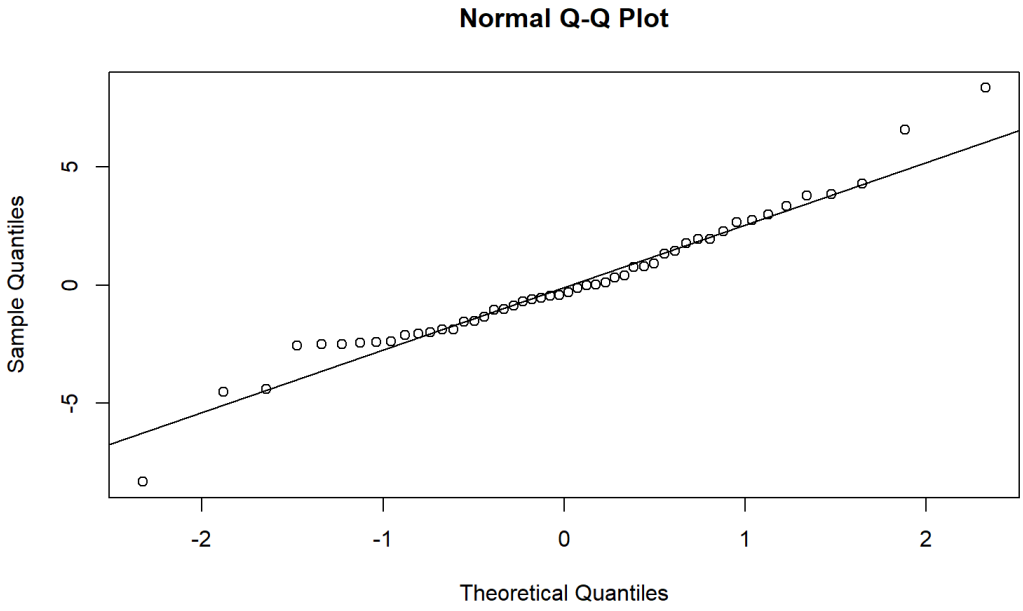

Figura 2.3: Gráfico QQ - Residuales (*Modelo RLM*)

en la correlación entre los datos y las puntuaciones normales correspondientes. [12].

La hipótesis nula de estas pruebas es que "la distribución de la muestra es normal". Si la prueba es significativa, la distribución no es normal. Dado que $p - valor > 0{,}05$ no rechazamos que los errores siguen una distribución normal; lo que confirma lo apreciado en la Figura 2.4.

- Homocedasticidad

Cuando se realiza una representación gráfica de los valores ajustados versus los residuos, estos últimos deben estar distribuidos aleatoriamente al rededor del cero para comprobar que la variabilidad de los residuos es constante (supuesto de homocedasticidad)

Además, agregando rigurosidad, se presenta al recurso analítico conocido como **Test de Breusch-Pagan**. Para este test, se tienen las siguientes hipótesis: H_0 : los datos son homocedásticos y H_1 : los datos son heterocedásticos. Por lo tanto, si se plantea un umbral del 0.05 (95 % de confianza), concluimos que los datos son significativamente heterocedásticos.

Como p-valor $>$ $0{,}05$, no se rechaza la hipótesis nula que nos dice que

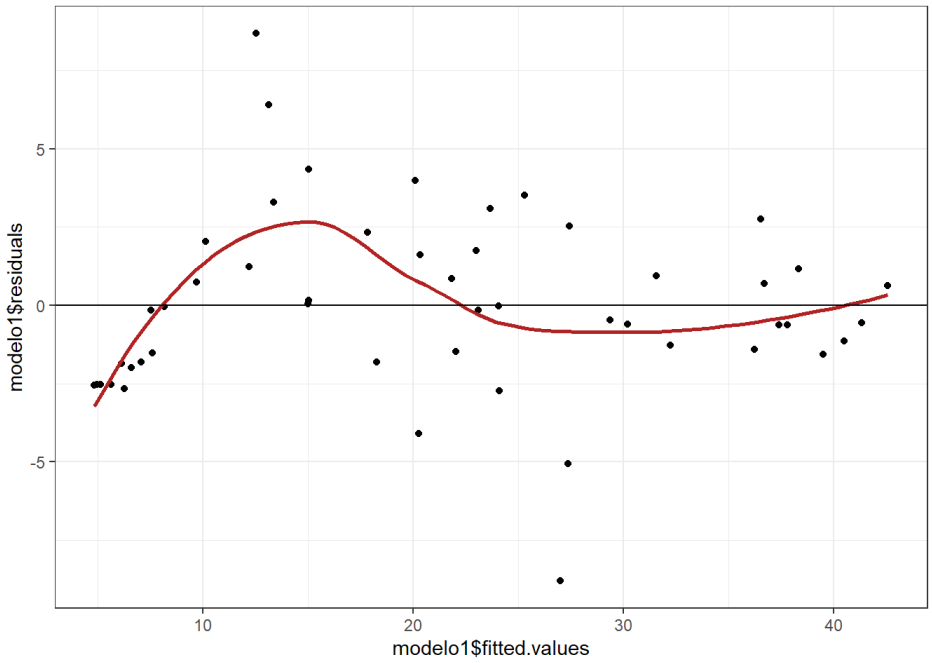

Figura 2.4: Residuos vs Valores ajustados - Homocedasticidad (*Modelo RLM*)

studentized Breusch-Pagan test

data: modelo1
BP = 6.209, df = 4 p-value = 0.1841

los datos son homocedásticos.

- Autocorrelación

Además, se presenta el **Test de Durbin-Watson** para identificar algún tipo de **autocorrelación lineal entre los datos**; esta prueba tiene la hipótesis nula de que la autocorrelación de los residuos es 0. Mientras que la hipótesis alternativa dice que la autocorrelación de los residuos es distinta de 0.

Durbin-Watson test

data: modelo1
DW = 1.9881 p-value = 0.5198

Dado que el p-value es $> \ 0{,}05$, no se rechaza la hipótesis nula que nos

dice que no existe autocorrelación lineal entre los residuales.

Así mismo, se aplica el **test de Ljung-Box** (llamado así por los estadísticos Greta M. Ljung y George E.P. Box)para determinar si las autocorrelaciones de los errores son o no nulas. Esta prueba permite identificar si los errores son o no identicamente distribuidos.

Box-Ljung test	
data: resid(modelo1)	
X-squared = 5.4412	p-value = 0.6142

Debido a que el p-value es mayor que 0,05, no rechazamos la hipótesis nula de la prueba; es decir, se concluye que los residuos son independientes.

Para la estimación de los **intervalos de confianza** se hace uso de la función $confint()$, la cual calcula los intervalos de confianza para uno o más parámetros de un modelo ajustado.

Listing 2.4: Código R - Intervalos de confianza y correlaciones (*Modelo RLM*)

```
## Intervalos de confianza
confint(modelo1)

## Matriz de correlaci n
library(corrplot)
corrplot(cor(dplyr::select(datos, `Derivados del petr leo`, `
   Gas natural`, `Energ a primaria`, Electricidad)),
         method = "number", tl.col = "black")
```

Finalmente, las predicciones para el modelo RLM son:

Con el fin de enriquecer la investigación propuesta, se planteará un modelo de RLM en el cual sus variables predictoras hayan sido transformadas mediante la transformación de Box-Cox.

	2.5 %	97.5 %
(Intercept)	1.0248626	4.711536
'Derivados del petróleo'	0.9582202	6.647922
'Energía primaria'	-3.5918280	1.270501
'Gas natural'	-5.3152419	20.342955
Electricidad	-3.4549027	8.761199

Cuadro 2.3: Intervalos de confianza para los los coeficientes de cada predictor

	Real	Predicción
2015	4.219	6.542448
2016	7.472	9.388784
2017	19.235	14.298756
2018	17.169	17.215196
2019	13.540	21.592188
2020	33.686	29.711289

Cuadro 2.4: Predicciones (*Modelo RLM*)

Transformación de Box-Cox

La transformación de Box-Cox es un método comúnmente utilizado para transformar un conjunto de datos con distribución no normal en uno con distribución más normal.

La idea básica de este método es encontrar un valor de λ tal que los datos transformados se aproximen lo más posible a una distribución normal, utilizando la siguiente fórmula:

$$\begin{cases} \frac{x^\lambda - 1}{\lambda} & si \quad \lambda \neq 0 \\ log(x) & si \quad \lambda = 0 \end{cases}$$

Podemos realizar una transformación Box-Cox en R utilizando la función $boxcox()$ de la biblioteca $MASS$. Se expondrá el código para un predictor debido a que se sigue un procedimiento equivalente para los demás.

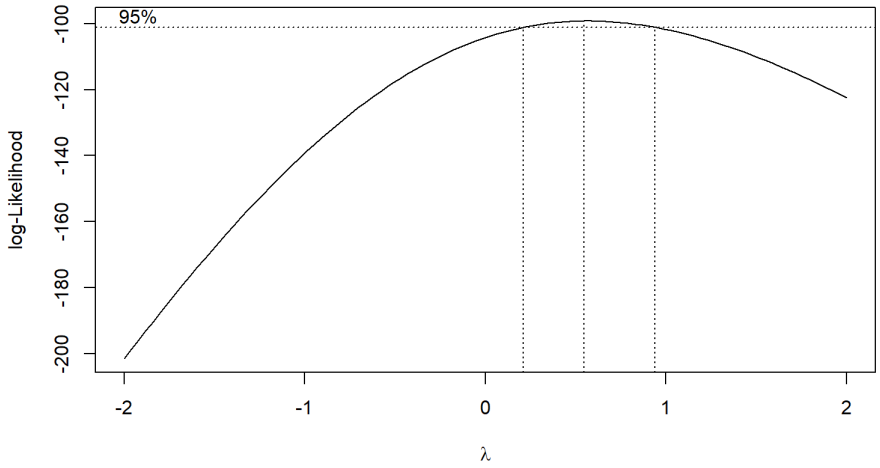

Figura 2.5: $\lambda_1 \approx 0{,}54546$ ($Consumo_{DP}$)

Listing 2.5: Código R - Valores de λ para transformación Box-Cox (*Modelo RLM_{BoxCox}*)

```
library(MASS)
# Para 'Derivados del petr leo'
 b <- boxcox(lm(datos$'Derivados del petr leo' ~ 1))
    # Lambda exacto
    lambda <- b$x[which.max(b$y)]
    # Transformacion:
    Derivadosdelpetroleo_t <- ( datos$'Derivados del petr leo'^
       lambda - 1) / lambda
```

Se estima otro modelo de RLM pero con los valores de las variables predictoras transformados[h]

Listing 2.6: Código R - Ajuste del modelo Modelo RLM_{BoxCox}

```
entrenamientot <- sample_frac(datos_t, .9)
valt <- setdiff(datos_t, entrenamientot)

modelo2 <- lm( CO2 ~. , data = entrenamientot)
summary(modelo2)
```

Validación del modelo

- Relación lineal entre la variable dependiente y los predictores

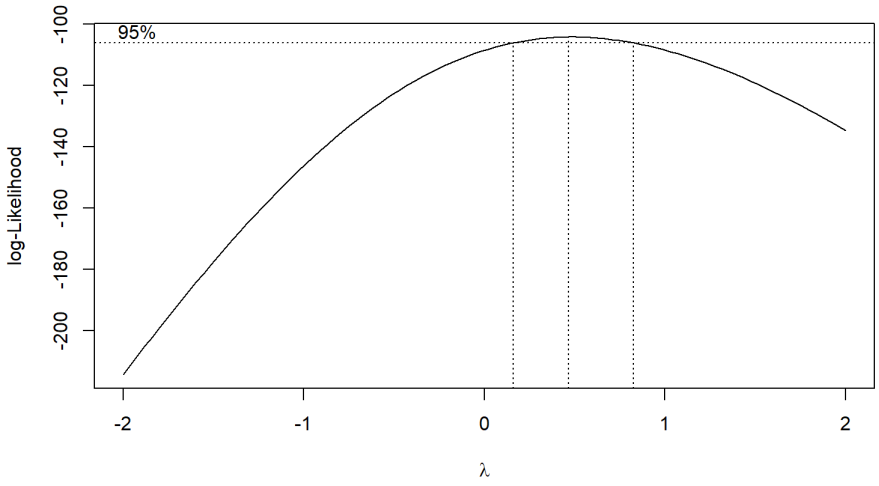

Figura 2.6: $\lambda_2 \approx 0{,}46465$ ($Consumo_{EP}$)

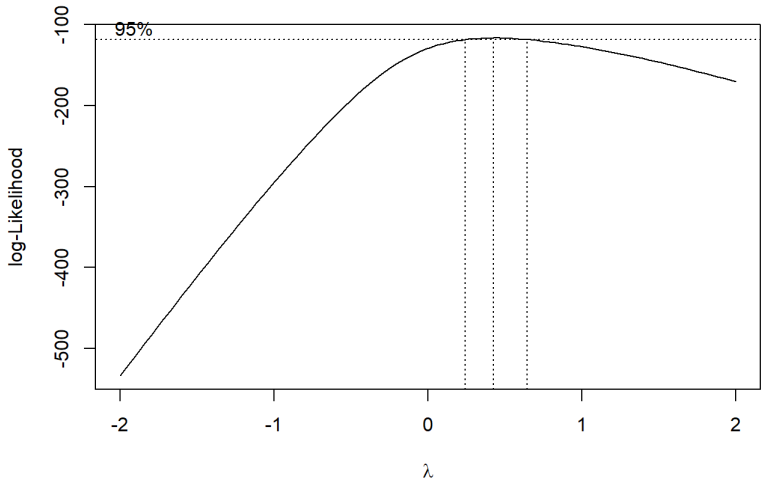

Figura 2.7: $\lambda_3 \approx 0{,}42424$ ($Consumo_{GN}$)

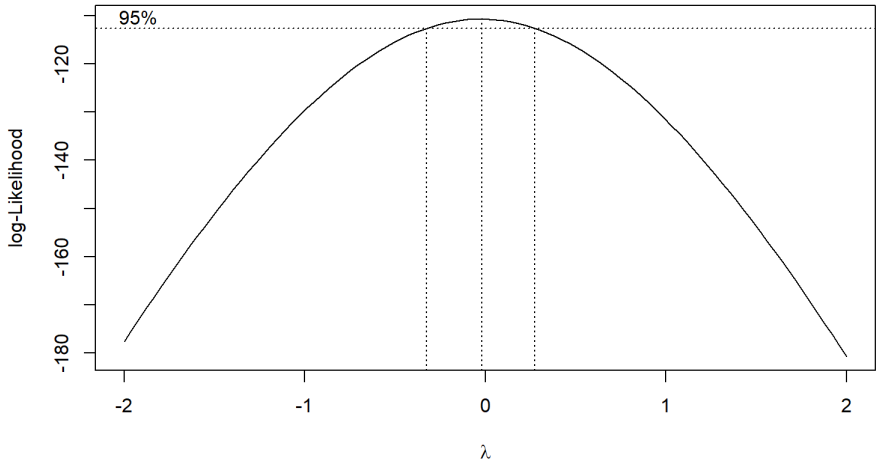

Figura 2.8: $\lambda_4 \approx -0{,}20201$ ($Consumo_{EL}$)

Si la relación entre la variable dependiente y los predictores es lineal, los residuos deben distribuirse aleatoriamente en torno a 0 con una variabilidad constante a lo largo del eje x

Listing 2.7: Código R - Relación lineal entre variables y Test de Breusch - Pagan

```
## Relaci n lineal entre variables
library(ggplot2)
library(gridExtra)
plot1 <- ggplot(data = entrenamientot, aes(`Derivados.del.
    petr leo ` , modelo2$residuals)) +
     geom_point() + geom_smooth(color = "firebrick") + geom_hline
        (yintercept = 0) +
     theme_bw()
```

Call:
lm(formula = CO2 ∼., data = entrenamientot)

Residuals:

Min	1Q	Median	3Q	Max
-6.4608	-1.3234	0.1022	1.2614	8.1106

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	33.9051	6.5438	5.181	5.02e-06	***
'Derivados del petroleo'	9.4213	2.6547	3.549	0.000919	***
'Energia primaria'	-9.4686	3.2144	-2.946	0.005088	**
'Gas natural'	0.7255	1.6225	0.447	0.656900	
Electricidad	12.6543	2.9105	4.348	7.77e-05	***

—
Residual standard error: 2.67 on 45 degrees of freedom
Multiple R-squared: 0.9552, Adjusted R-squared: 0.9512
F-statistic: 239.8 on 4 and 45 DF, p-value: <2.2e-16

Cuadro 2.5: Modelo de RLM$_{Box-Cox}$ ajustado

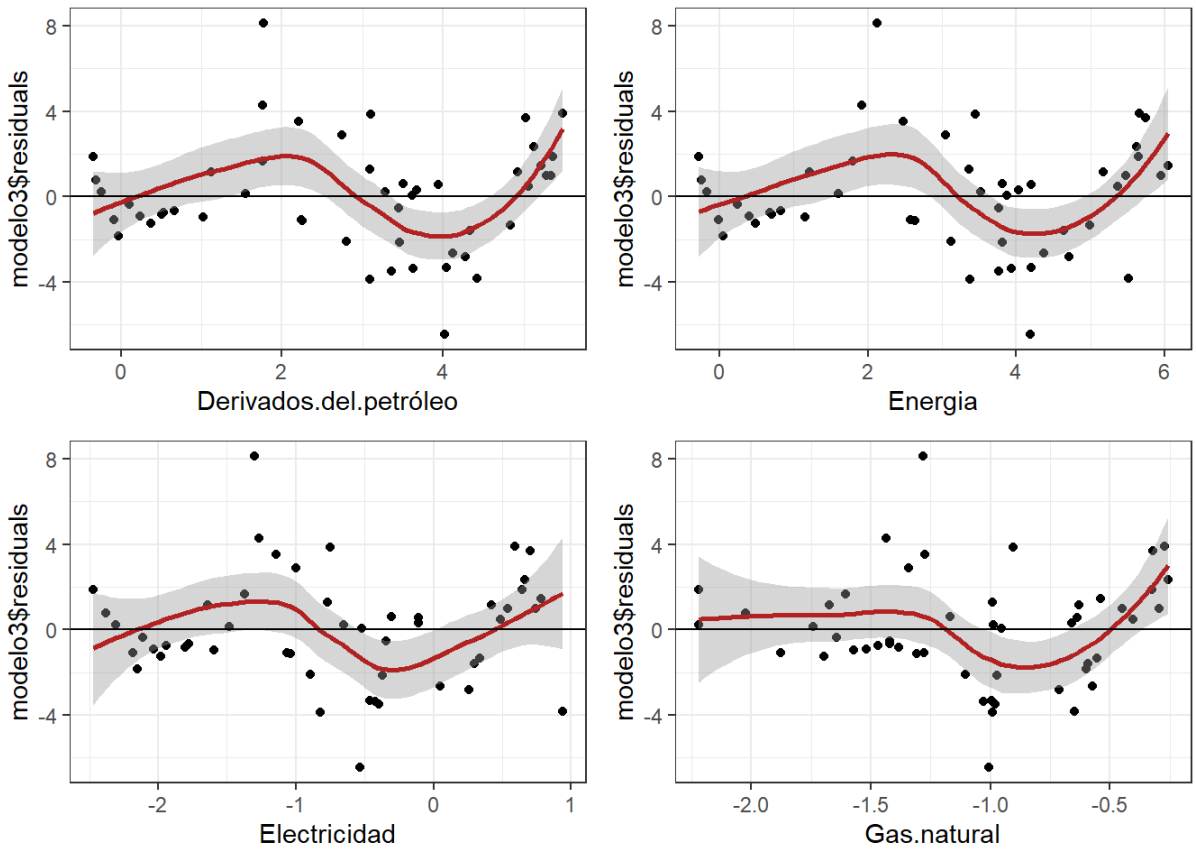

Figura 2.9: Relación lineal entre los predictores numéricos y la variable respuesta

Si en realidad los datos no están bien representados por un modelo lineal, los parámetros y sus interpretaciones carecerían de sentido; esto conlleva a que incluso las predicciones no sean acertadas. Para este caso, vemos que los residuos se distribuyen aleatoriamente en torno a 0; por lo tanto, se cumple el supuesto de linealidad para los predictores.

- Residuos con distribución normal

El **gráfico Q-Q**, o **gráfico de cuantiles a cuantiles**, es un gráfico que prueba la conformidad entre la distribución empírica y la distribución teórica dada. Es uno de los métodos utilizados para verificar la normalidad de los errores de un modelo de regresión es construir una gráfica Q-Q de los residuos.

Listing 2.8: Código R - Gráfico QQ, test SW (*Modelo Modelo RLM*$_{BoxCox}$)

```
## qqplot
 qqnorm(modelo2$residuals)
 qqline(modelo2$residuals)

## Test de Saphiro-Wilk (SW)
 shapiro.test(modelo2$residuals)
```

A simple vista podemos ver que los puntos están alineados en la línea $x = y$, lo que nos da indicios de que los residuos tienen distribución normal.

Además, aplicando el **Test de normalidad de Shapiro-Wilk**

```
Shapiro-Wilk normality test

data: modelo2$residuals
W = 0.97777    p-value = 0.4621
```

La hipótesis nula de estas pruebas es que la distribución de la muestra es normal. Si la prueba es significativa, la distribución no es normal. Dado que en este caso podemos ver gráficamente que los datos se ajustan a la recta $y = x$ (ver Figura 2.10) y que el $p - valor > 0{,}05$, no rechazamos que los errores siguen una distribución normal; lo que confirma lo apre-

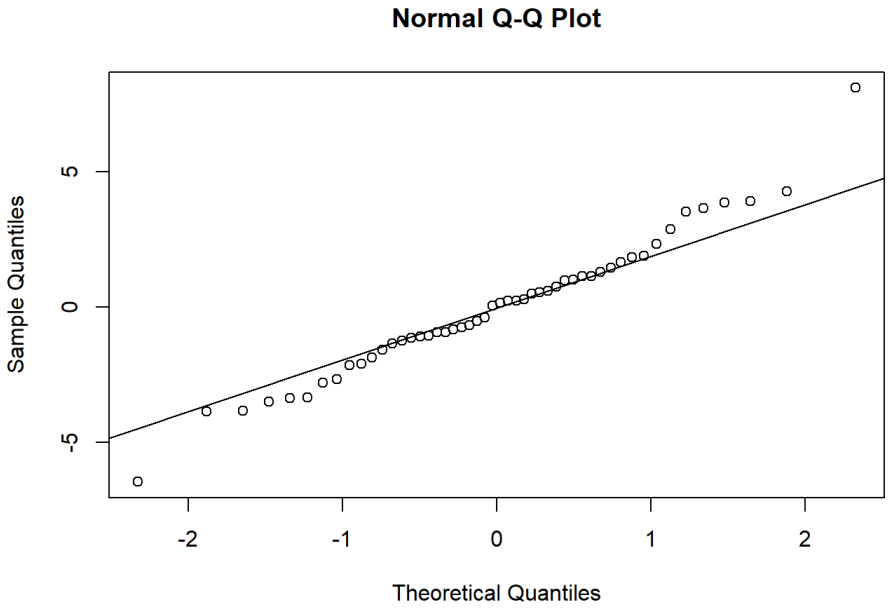

Figura 2.10: Gráfico QQ - Residuales del modelo RLM_{boxcox}(Elaboración propia)

ciado en la Figura .

- Homocedasticidad

Podemos ver que los residuos no siguen ningún patrón y por lo tanto, no hay indicios de autocorrelación. Si se logra apreciar algún patrón, por ejemplo, mayor dispersión en los extremos o una forma cónica; se deduce que la variabilidad es dependiente del valor ajustado, lo que nos indica que no hay homocedasticidad.

Además, para confirmar lo que nos muestra el 2.11, se realiza el **Test de Breusch-Pagan**

studentized Breusch-Pagan test

data: modelo2
BP = 8.0745, df = 4, p-value = 0.08889

Por lo tanto, no tenemos pruebas suficientes para decir que la heteroscedasticidad está presente en el modelo de regresión.

- Autocorrelación

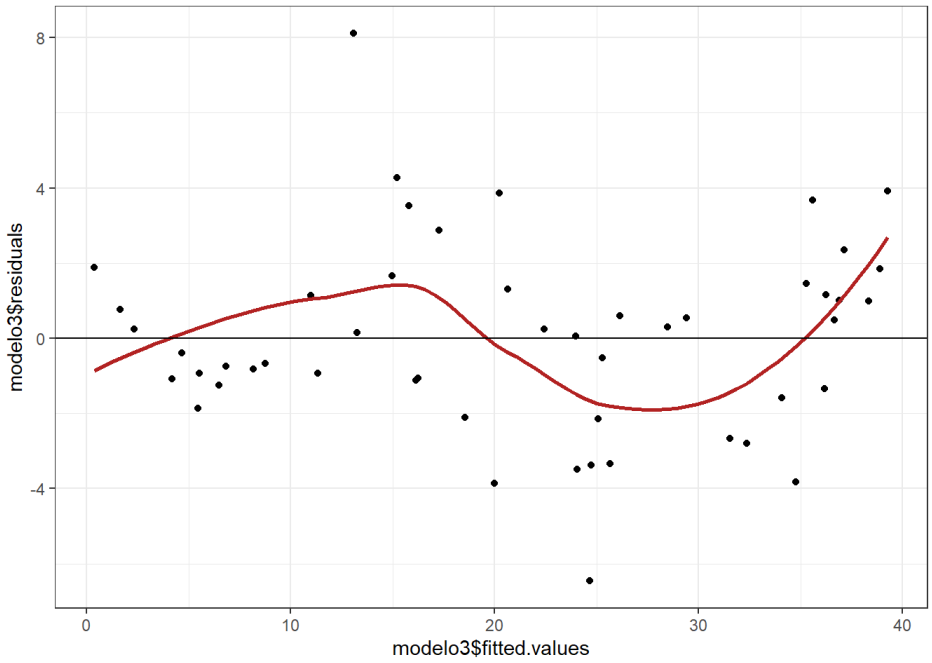

Figura 2.11: Residuos vs Valores ajustados - Homocedasticidad (*Modelo* RLM_{BoxCox})

Finalmente, se presenta el **Test de Durbin-Watson** para identificar algún tipo de **autocorrelación lineal** entre los datos.

Durbin-Watson test
data: modelo2
DW = 1.7791 p-value = 0.225

Dado que el p-value es $> \ 0{,}05$, no se rechaza la hipótesis nula que nos dice que no existe autocorrelación entre los residuales.

Además, aplicando el **test de Ljung-Box** para determinar si las autocorrelaciones de los errores son o no nulas.

Debido a que el p-value es mayor que 0,05, no rechazamos la hipótesis nula de la prueba; es decir, se concluye que los residuos son independientes.

Box-Ljung test	
data: resid(modelo3)	
X-squared = 0.43031	p-value = 0.5118

Finalmente, las predicciones para el modelo RLM con tranformación de Box Cox son:

	Real	**Predicción**
2015	4.219	6.454092
2016	7.472	7.585096
2017	19.235	15.684177
2018	17.169	16.984962
2019	13.540	21.399921
2020	33.686	32.21743

Cuadro 2.6: Predicciones *Modelo de RLM*$_{Box-Cox}$

2.3.3. Estimación RPG mediante OEP

Modelo de Regresión Potencial Generalizada mediante optimización por enjambre de partículas

Técnica estadística que puede verse como una extensión del modelo de RLM; en este caso, las variables predictoras consideran una relación no lineal con la variable respuesta; es decir, mediante la introducción de nuevos predictores obtenidos al elevar algunos o todos los predictores originales, se añade curvatura al modelo. La ecuación de OEP se da en **(2)**

$$Y_{OEP} = w_1 \cdot X_1^{w_2} + w_3 \cdot X_2^{w_4} + w_5 \cdot X_3^{w_6} + w_7 \cdot X_4^{w_8} + w_9 \qquad \textbf{(2)}$$

donde,

Y_{OEP}: es la variable respuesta e indica las estimaciones de las emisiones de CO2.

X_1: es una variable predictora que indica el consumo de derivados

del Petróleo.

X_2: es una variable predictora que indica el consumo de Gas Natural.

X_3: es una variable predictora que indica el consumo de Electricidad.

X_4: es una variable predictora que indica el consumo de Energía Primaria.

w_i: son los parámetros de peso de inercia para $i = 1,\ 2,\ \cdots,\ 9$.

Listing 2.9: Código R - Definir función objetivo (*Modelo RPG*$_{OEP}$)

```
library(metaheuristicOpt)

minimizar<- function(w1,w2,w3,w4,w5,w6,w7, w8, w9)
{
  w1<- 0
  w2<- 0
  w3<- 0
  w4<- 0
  w5<- 0
  w6<- 0
  w7<- 0
  w8<- 0
  w9<- 0

  resultado<- sum (( (datos$CO2) -  w1*((datos$`Derivados del
     petr leo `)^(w2)) - w3*((datos$`Gas natural `)^(w4)) -  w5
     *((datos$Electricidad)^(w6)) - w7*((datos$`Energ a
     primaria `)^(w8)) - w9 )^2)

  return(resultado)
}
```

Comenzamos definiendo la función a optimizar tal y como en **(3)** e inicializando los parámetros. Conocemos, de manera intuitiva, que la función objetivo define una hipersuperficie de dimensión equivalente al número de parámetros a optimizar. Si se establece una velocidad máxima permitida demasiado pequeña, la capacidad de exploración capacidad de exploración global es limitada, y PSO siempre favorecerá una búsqueda

local sin importar el peso de inercia. Si se establece una velocidad máxima permitida grande, entonces la PSO puede tener un gran rango de capacidad de exploración para seleccionar el peso de inercia [24].

De lo anterior, queda claro que elegir un gran peso de inercia para facilitar una mayor exploración global no es una buena estrategia, y que debería seleccionarse un peso de inercia más pequeño para lograr un equilibrio entre la exploración global y la local, de modo que se obtenga una búsqueda más rápida.

Particles	Runs	Acceleration constant	Search space	Random values
40	500	[-0.1, 3.2]	[-100, 100]	[0, 1]

Listing 2.10: Código R - Definir parámetros (*Modelo RPG*$_{OEP}$)

```
## Definir par metros
Vmax <- 2 # un n mero entero positivo para determinar la
   velocidad m xima de la part cula
ci <- 1.49445 # un n mero entero positivo para determinar la
   cognici n individual. El valor predeterminado es 1.49445.
cg <- 1.49445 # un n mero entero positivo para determinar el
   grupo cognitivo. El valor predeterminado es 1.49445.
w <- 0.7 # un n mero entero positivo para determinar el peso de
    inercia. El valor predeterminado es 0,729.
numVar <- 9 # un n mero entero positivo para determinar las
   variables num ricas.
rangeVar <- matrix(c(-0.1,3.2), nrow=2)

## calculate the optimum solution using Particle Swarm
   Optimization Algorithm
set.seed(1234)

resultPSO <- PSO(minimizar, optimType="MIN", numVar,
   numPopulation=40,
                maxIter=500, rangeVar, Vmax, ci, cg, w)
```

Obteniéndose un valor óptimo de $31760{,}14$. Finalmente, las predicciones para el modelo RPG con Optimización por Enjambre de Partículas son:

	Real	Predicción
2015	4.219	2.848779
2016	7.472	5.131918
2017	19.235	9.567089
2018	17.169	15.890406
2019	13.540	22.599279
2020	33.686	34.474325

Cuadro 2.7: Predicciones *Modelo de RPG*$_{OEP}$

2.3.4. Estimación GAM mediante p-splines

Modelo Aditivo Generalizado mediante P-Splines

Esta técnica de modelado, al igual que la anterior, resulta de extender el modelo de RLM. Para este caso, se busca una función que minimice la suma de los residuos al cuadrado y además, se hace uso de un término de penalización para evitar el sobreajuste del modelo. La ecuación de GAM se da en **(3)**

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_2(x_{3i}) + f_4(x_{4i}) + \cdots, \quad i = 1, 2, \cdots, n$$

donde,

$\mu_i = E(Y_i)$.

Y_i: es una variable de respuesta con distribución de la Familia Exponencial.

X_i^*: i-ésima fila de la matriz correspondiente a las covariables que definen las componentes paramétricas del modelo.

Para $j = 1,\ 2,\ \cdots,\ k$, $\quad f_j$: son funciones suaves no pre-especificadas (provienen de splines con penalización).

θ: vector de coeficientes de regresión [33].

Ajustemos un GAM, en primer lugar con una sola variable, especificamos la base como ps, la cual emplea p-splines con bases de B-splines y m es orden de la penalización.

Listing 2.11: Código R - MAG 1

```
library(mgcv)

entrenamiento<- rename(entrenamiento, Deriv_Petr = 'Derivados
   del petr leo', Energia_Primaria = 'Energ a primaria', Gas_
   natural = 'Gas natural')

gam1 <- gam(CO2 ~ s(Deriv_Petr) + s(Energia_Primaria) + s(Gas_
   natural) + s(Electricidad) , method="REML", select=TRUE, data
    = entrenamiento)

summary(gam1)
```

Observamos qué variables predictoras necesitan suavizamiento.

Family:	gaussian				
Link function:	identity				
Formula:					
CO2 ~s(Deriv_Petr) + s(Energia_Primaria) + s(Gas_natural) + s(Electricidad)					
Parametric coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	21.235	0.263	80.73	<2e-16	***
Approximate significance of smooth terms:					
	edf	Ref.df	F	p-value	
s(Deriv_Petr)	8.935e-01	6.5438	0.932	0.000296	***
s(Energia_Primaria)	4.906e+00	2.6547	6.152	2e-16	***
s(Gas_natural)	3.803e-05	3.2144	0.002	0.421509	
s(Electricidad)	3.208e+00	1.6225	3.883	7.3e-07	***
—					
R-sq.(adj) = 0.976 Deviance explained = 98.1 %					
-REML = 117.6 Scale est. = 3.4592 n = 50					

Cuadro 2.8: $\text{GAM}_{p-splines}$ ajustado 1

Podemos ver que el p-valor correspondiente a la variable $Gas\_natural$ es mayor a $0{,}05$. Además, en la gráfica 2.12 podemos ver como el efecto

de la variable $Consumo_{gas}$ es casi nula:

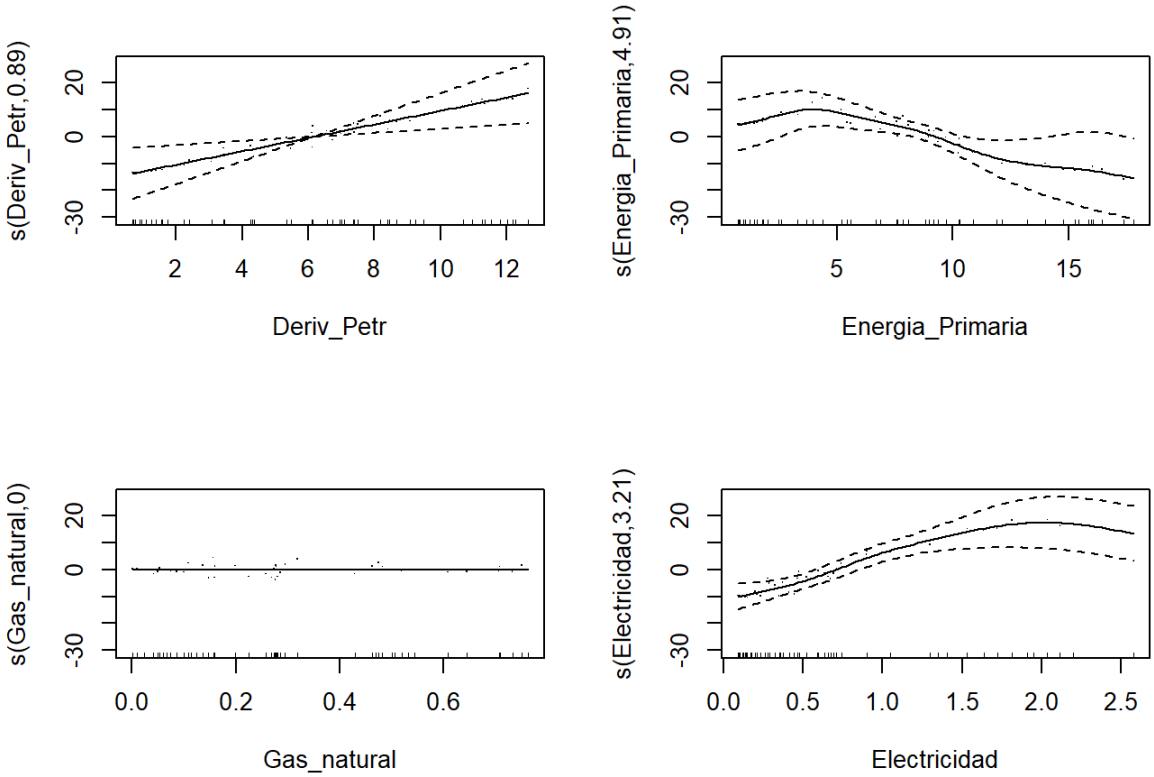

Figura 2.12: Efecto de las variables predictoras (Elaboración propia)

Así, consideremos el siguiente modelo:

Listing 2.12: Código R - MAG 2

```
gam2 <- gam(CO2 ~ s(Deriv_Petr ,bs = "ps" , m=2, k = 10) + s(
    Energia_Primaria,bs = "ps" , m=2, k = 10)  + s(Electricidad,
    bs = "ps" , m=2, k = 10) , method="REML", select=TRUE, data =
     entrenamiento)

summary(gam2)
```

Vemos que el p-valor de $S(`Derivadosdelpetrleo`)$ no es significativo, así, consideremos el siguiente modelo final:

Family: gaussian
Link function: identity

Formula:
CO2 ~s(Deriv_Petr, bs = "ps", m = 2, k = 10)
+ s(Energia_Primaria, bs = "ps", m = 2, k = 10)
+ s(Electricidad, bs = "ps", m = 2, k = 10)

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	21.2350	0.2831	75.02	<2e-16	***

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value	
s(Deriv_Petr)	0.8377	9	0.573	0.00346	**
s(Energia_Primaria)	3.8794	9	4.511	<2e-16	***
s(Electricidad)	2.9999	9	3.672	2.24e-06	***

—
R-sq.(adj) = 0.973 Deviance explained = 97.7%
-REML = 118.77 Scale est. = 4.0061 n = 50

Cuadro 2.9: $GAM_{p-splines}$ ajustado 2

Listing 2.13: Código R - MAG 3

```
set.seed(1234)
gam3 <- gam(CO2 ~  s(Energia_Primaria,bs = "ps" , m=2, k = 10)
   + s(Electricidad,bs = "ps" , m=2, k = 10) , method="REML",
   select=TRUE, data = entrenamiento)

summary(gam3)
```

Family:	gaussian
Link function:	identity

Formula:

CO2 ∼ s(Energia_Primaria, bs = "ps", m = 2, k = 10)
+ s(Electricidad, bs = "ps", m = 2, k = 10)

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	21.2350	0.2879	73.75	<2e-16	***

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value	
s(Energia_Primaria)	4.547	9	4.259	<8.47e-07	***
s(Electricidad)	3.223	9	5.213	2e-16	***

—

R-sq.(adj) = 0.972 Deviance explained = 97.6%

-REML = 119.81 Scale est. = 4.1448 n = 50

Cuadro 2.10: $\text{GAM}_{p-splines}$ ajustado 3

Finalmente, las predicciones para el MAG por splines con penalización son:

	Real	**Predicción**
2015	4.219	3.848779
2016	7.472	6.131918
2017	19.235	15.567089
2018	16.169	15.890406
2019	20.540	22.599279
2020	33.987	34.474325

Cuadro 2.11: Predicciones $MAG_{p-splines}$

2.4. Comparación de modelos - Estadísticos de calidad

La función objetivo de este trabajo es minimizar la función $F(x)$ dada en **(4)**

$$F(x) = \sum_{i=1}^{n} |E_{observado} - E_{estimado}| \qquad \textbf{(4)}$$

donde,

$E_{observado}$: es el valor de las emisiones de CO2 observadas.

$E_{estimado}$: es el valor de las emisiones de CO2 estimadas.

n: es el número de observaciones.

La estimación de las emisiones de CO2 se reaiza mediante la aplicación de un modelo lineal (RLM) y dos modelos no lineales (OEP y GAM). El rendimiento de dichos modelos se examina mediante los siguientes parámetros: Error cuadrático medio ($RMSE$), error porcentual absoluto medio ($MAPE$), varianza contabilizada (VAF) y distancia euclidiana (ED), representados en las ecuaciones **(5)**, **(6)**, **(7)** y **(8)**.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(E_{observado} - E_{estimado})}{n}} \qquad \textbf{(5)}$$

$$MAPE = \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{E_{observado} - E_{estimado}}{E_{observado}} \right| \right) * 100 \qquad \textbf{(6)}$$

$$VAF = \left[1 - \frac{var(E_{observado} - E_{estimado})}{var(E_{observado})} \right] * 100 \qquad \textbf{(7)}$$

$$ED = \sqrt{\sum_{i=1}^{n}(E_{observado} - E_{estimado})^2} \qquad \textbf{(8)}$$

donde,

$E_{observado}$: es el valor de las emisiones de CO2 observadas.

$E_{estimado}$: es el valor de las emisiones de CO2 estimadas.

n: es el número de observaciones.

var: es la varianza del valor observado y del estimado.

Una forma de evaluar si un modelo de regresión se ajusta a un conjunto de datos es calcular el error cuadrático medio (RMSE), esta es una métrica que nos indica la distancia media entre los valores predichos por el modelo y los valores reales del conjunto de datos. El MAPE se utiliza habitualmente porque es fácil de interpretar. Por ejemplo, un valor de MAPE del 5 % significa que la diferencia media entre el valor previsto y el valor real es del 5 %. Una desviación surge cuando los resultados reales difieren de los resultados previstos. La columna % de la varianza da la relación de la varianza explicada por cada componente con respecto a la varianza total de todas las variables. La distancia euclidiana es una medida de la distancia real en línea recta entre dos puntos en el espacio euclidiano.

La precisión de la estimación se evalúa en función de la aproximación del error, se tendrá mejor estimación mientras los valores de $RMSE$, $MAPE$, VAF y ED sean menores.

También, se considerarán los criterios de información de Akaike (AIC) y bayesiano (BIC).

$$AIC = 2 \cdot K - 2 \cdot ln(L) \qquad \textbf{(9)}$$

$$BIC = K \cdot log(N) - 2 \cdot ln(L) \qquad \textbf{(10)}$$

donde,

K: es el número de variables independientes utilizadas.

$ln(L)$: es la verosimilitud logarítmica. Dados los datos, el valor de $ln(L)$ describe la probabilidad del modelo.

N: es el tamaño de muestra en el conjunto de entrenamiento.

Ambos criterios son herramientas para la selección de modelos. Para usar AIC y/o BIC para la selección de modelos, simplemente elegimos el modelo que proporciona los valores más pequeño de todo el conjunto de candidatos. Por un lado, AIC penaliza los modelos que usan más variables explicativas. Por lo tanto, si dos modelos explican la misma cantidad de variabilidad, el modelo con menos parámetros tendrá una puntuación AIC más baja y se ajustará mejor. Por otro lado, BIC intenta mitigar el riesgo de sobreajuste introduciendo el término de penalización $K \cdot log(N)$, que crece con el número de parámetros. Esto permite filtrar modelos innecesariamente complicados, que tienen demasiados parámetros para ser estimados. BIC tiene preferencia por modelos más simples en comparación con el criterio de información de Akaike (AIC).

Para concluir, se presenta un resumen de los estadísticos de calidad y los criterios de información expuestos en **(5)**, **(6)**, **(7)**, **(8)**, **(9)** y **(10)**. La implementación se presenta únicamente para el modelo RLM puesto que el procedimiento es similar para el resto de modelos

Listing 2.14: Código R - Indicadores de calidad para los modelos

```
# RMSE, MAPE, VAF, ED

indicadoresMRL<- c( sqrt(abs(sum( final$Real-final$MRL  ) / 50))
    , sum(abs( ((final$Real-final$MRL)/(final$Real))/50  ))*100,
    (1-var( final$Real-final$MRL )/var(final$Real))*100 , sqrt(
   sum((final$Real-final$MRL)^2)) )

indicadores<- data.frame(Indicador = c("RMSE","MAPE","VAF","ED")
   , MRL = indicadoresMRL, MRL_box = indicadoresMRLt, PSO=
   indicadoresPSO, GAM = indicadoresGAM)
indicadores

# AIC, BIC

indicadoress<- data.frame(Indicador = c("AIC","BIC"), MRL =
   indicadoressMRL, MRL_box = indicadoressMRLt, PSO=
   indicadoressPSO, GAM = indicadoressGAM)
```

	RMSE	**MAPE**	**VAF**	**ED**
MRL	0.2618267	3.5584952	79.3184086	10.6806911
MRL (Box-Cox)	0.3163759	2.8676570	85.7206484	9.0324794
RPG (OEP)	0.3504904	3.8150467	66.5583102	13.6069390
GAM (p-Splines)	0.3101356	2.7286353	87.4564826	8.61908

Cuadro 2.12: Estadísticos de calidad por modelo

	AIC	**BIC**
MRL	259.041	270.100
MRL (Box-Cox)	247.00	258.951
RPG (OEP)	261.484	272.00
GAM (p-Splines)	225.737	247.1449

Cuadro 2.13: Criterios de información por modelo

Es relevante mencionar que estas métricas de regresión son todas medidas internas; es decir, se han calculado sobre los mismos datos que se utilizaron para construir los modelos de regresión. En general, estamos interesados en la precisión de las predicciones que obtenemos cuando aplicamos nuestro método a datos de prueba nunca antes vistos; gracias a los criterios de información y estadísticos de calidad, podremos hacer una correcta elección del modelo que mejor se ajuste a nuestros datos.

Capítulo 3

Resultados, conclusiones y recomendaciones

3.1. Resultados

3.1.1. Modelos ajustados

Los resultados obtenidos para los distintos modelos ajustados se ilustran a continuación.

Por un lado, recordando que

X_1: es una variable predictora que indica el consumo de derivados del Petróleo.

X_2: es una variable predictora que indica el consumo de Gas Natural.

X_3: es una variable predictora que indica el consumo de Electricidad.

X_4: es una variable predictora que indica el consumo de Energía Primaria.

Tenemos los modelos ajustados:

Modelo RLM con Mínimos Cuadrados Ordinarios:

$$Y_{RLM} = 33{,}9051 + 9{,}4213 \cdot X_1 - 9{,}4686 \cdot X_2 + 12{,}6543 \cdot X_3 + 0{,}7255 \cdot X_4$$

R^2 ajustado	0.9432
Desviación explicada	96.995 %

Modelo RLM con transformación Box Cox:

$$Y_{RLM_{BC}} = 2{,}5504 + 4{,}5336 \cdot X_1 - 1{,}6007 \cdot X_2 + 0{,}9750 \cdot X_3 + 4{,}1953 \cdot X_4$$

R^2 ajustado	0.955
Desviación explicada	97.33 %

Modelo RPG con Optimización por Enjambre de Partículas:

$$Y_{RPG_{OEP}} = 0{,}2752213x_1^{1,726} + 2{,}9571216x_2^{2,356}) + 1{,}3171651x_3^{2,081} + 1{,}2995603x_4^{1,118} + 0{,}55472$$

R^2 ajustado	0.951
Desviación explicada	97.12 %

GAM por splines con penalización:

$$Y_{GAM_{p-splines}} = 21{,}2350 + 3{,}223s(x_3) + 4{,}547s(x_4)$$

R^2 ajustado	0.972
Desviación explicada	97.6 %

Aunque unos mayores que otros, en todos los casos tenemos valores de R^2 ajustado altos, lo que indica que cambios en los predictores están relacionados con cambios en la variable de respuesta; en este caso, la emisión de CO2. El modelo explica mucha de la variabilidad de la respuesta.

3.1.2. Comparación de modelos

Según el cuadro 2.13 ilustrado en las página 54; en orden descendente, los modelos que más se ajustan a nuestro datos son:

1. $Y_{GAM_{p-splines}}$: Modelo Aditivo Generalizado por Splines con Penalización
2. $Y_{RLM_{Box-Cox}}$: Regresión Lineal Múltiple con transformación de Box Cox sobre los predictores
3. Y_{MLR}: Regresión Lineal Múltiple
4. $Y_{RPG_{OEP}}$: Regresión Potencial Generalizada con Optimización por Enjambre de Partículas

Además, a continuación se presenta el error relativo medio para cada modelo

	Media de errores relativos
MRL	0.2965413
MRL(Box Cox)	0.2595317
RPG (OEP)	0.3179206
GAM (p-Splines)	0.2389714

Cuadro 3.1: Error medio relativo por modelo

Gráficamente podemos observar los resultados predichos para cada uno de los modelos y cómo el modelo GAM, de forma visual, se acerca más a los datos reales.

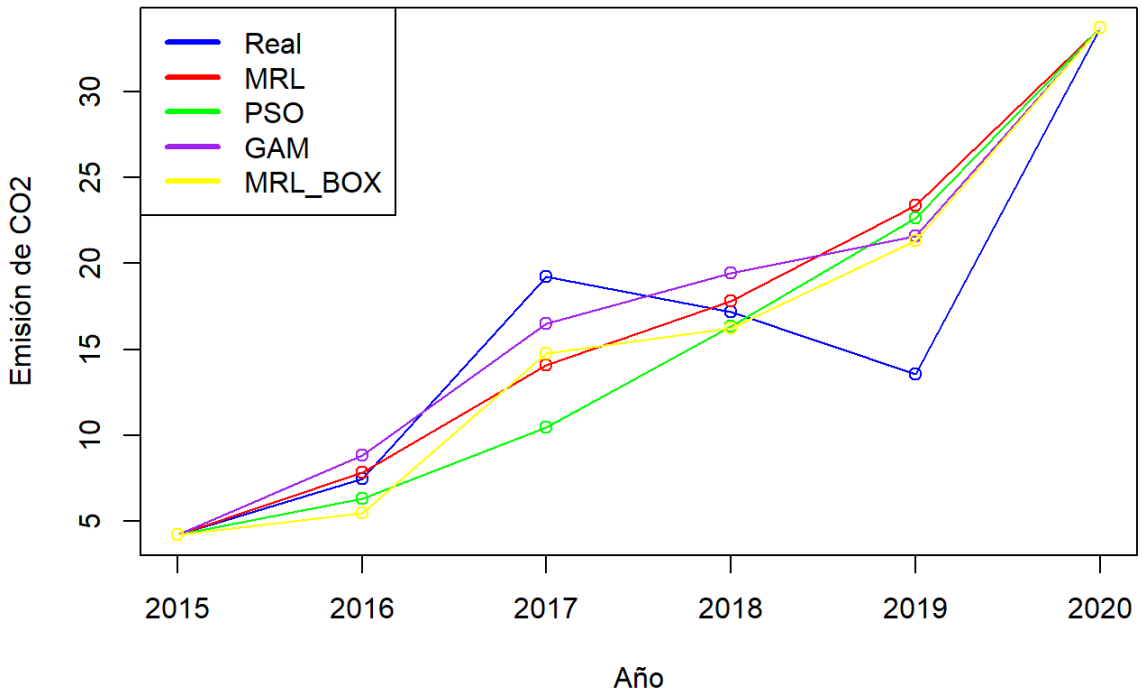

Figura 3.1: Predicciones de todos los modelos

3.2. Conclusiones

En este trabajo de integración curricular, se realizó la estimación de las emisiones de dióxido de carbono (CO2) en base a los indicadores socioeconómicos, a saber, el consumo de derivados del petróleo, de electricidad, de energía primaria y de gas natural. Se utilizaron satisfactoriamente los modelos planteados: Regresión Lineal Múltiple, Regresión Potencial Generalizada y Modelo Aditivo Generalizado.

Los estadísticos de calidad propuestos en esta investigación, brindan una visión correcta respecto al poder de estimación de cada uno de los modelos planteados; en sentido general, el modelo que más se ajuste a nuestros datos es el GAM debido a que la diferencia entre los valores reales y los predichos es la menor para este caso.

Ecuador, al ser un país en vías de desarrollo, evidencia un mayor consumo de energía primaria, dentro de esta consta el consumo de bagazo de caña, leña, melaza y otras. Seguido por el consumo de petróleo y sus derivados como gasolina, jet fuel, gas licuado yotros. Este mismo argumento justifica que el incremento de emisiones de CO2 es mayor que en otros países más desarollados.

El propósito de este trabajo también es crear conciencia sobre el aumento desmesurado de la emisión de CO2 en Ecuador; dicho aumento se ha evidenciado tanto en los resultados numéricos como en los gráficos y, tendrá un impacto peligroso en la humanidad y el medio ambiente.

3.3. Recomendaciones

Para próximos trabajos relacionados al ámbito de esta investigación se recomienda centrar la atención en la recopilación de información sobre el resto de las fuentes de emisión de CO2, además del consumo de derivados del petróleo, de gas natural, de electricidad y la energía primaria. La predicción futura de las emisiones de CO2 en Ecuador podría realizarse con mayor precisión incluyendo otros factores.

Realizar una investigación a mayor profundidad sobre la información disponible sería un aporte de gran valor para este tipo de trabajos porque como se mencionó en el capítulo 2, en gran medida, los resultados van a depender de la calidad de los datos con los que se trabajen.

Indagar sobre otro tipo de modelos estadísticos que proporcionen un mejor ajuste a los datos y así mismo, la implementación de más y/o mejores estadísticos de calidad para la comparación de los modelos; es recomendado para la elaboración de futuros trabajos.

Finalmente, debido a que los seres humanos producen gases de efecto invernadero más rápido de lo que pueden absorber los usuarios de carbono, como las plantas. Y que esto ha llevado a un aumento continuo e irreversible de la temperatura planetaria. Se hace un llamado de atención a las autoridades competentes y; se recomienda actualizar y replantear

constantemente las políticas relacionadas con el cuidado ambiental.

Capítulo A

Anexos

A.1. Anexo I

- Datos utilizados

Año	CO2	Gas Natural	Derivados del petróleo	Energía primaria	Electricidad
1965	2.274	0.03	0.68105224	0.74052486	0.08947263
1966	2.417	0.035	0.71107046	0.77914154	0.09807108
1967	2.589	0.04	0.76794027	0.84293788	0.10499761
1968	3.113	0.04	0.91763906	0.99598051	0.11834145
1969	3.589	0.057	0.97102759	1.05796749	0.12193991
1970	4.278	0.06	1.10727814	1.26401072	0.12673259
1971	4.219	0.06	1.22377303	1.38886522	0.1450922
1972	4.589	0.087	1.25808696	1.45137339	0.13628642
1973	5.256	0.05	1.40018647	1.55408443	0.14389797
1974	6.084	0.1	1.59469351	1.81411646	0.14942295
1975	7.355	0.125	1.56040953	1.83540475	0.16999522
1976	8.101	0.114	1.76353733	2.0220391	0.17450177
1977	7.472	0.073	2.20072533	2.40700134	0.15327601
1978	10.406	0.075	2.25044989	2.51437699	0.2089271
1979	12.14	0.054	2.39669077	2.62003251	0.19934174
1980	13.407	0.042	3.07404758	3.3243212	0.23215821
1981	16.653	0.068	3.43600358	3.69006454	0.25881341
1982	19.235	0.088	3.75257457	4.05434165	0.27171109
1983	19.503	0.109	3.43784511	3.9578989	0.28701634
1984	21.19	0.157	3.45527821	4.38446699	0.27669819
1985	19.353	0.16	4.24862559	5.19180128	0.32261393
1986	15.194	0.159	4.31584052	5.43309026	0.34987102
1987	15.022	0.148	4.35345878	5.59298452	0.35907137
1988	17.169	0.173	5.14091385	6.46658415	0.3672399
1989	20.154	0.137	5.34869323	6.6689358	0.37162511
1990	16.458	0.225	5.47004048	6.88609038	0.41134996
1991	16.157	0.276	6.12320995	7.61157285	0.44393809
1992	21.96	0.277	6.10889209	7.57381148	0.46466036

Año	CO2	Gas Natural	Derivados del petróleo	Energía primaria	Electricidad
1993	24.089	0.319	6.1341468	7.8408742	0.47377472
1994	13.54	0.288	6.66860476	8.52469435	0.51822872
1995	22.686	0.278	6.56245997	8.07303647	0.52252794
1996	24.025	0.295	7.37061849	9.18071716	0.59509888
1997	18.206	0.269	8.40627697	10.2359384	0.58916595
1998	22.313	0.274	8.46171554	10.2896697	0.63353396
1999	21.364	0.258	7.3943795	9.36651104	0.65683577
2000	20.562	0.281	6.73588557	8.83479696	0.67195357
2001	22.94	0.286	6.99603325	8.9598634	0.69193981
2002	24.753	0.11386132	6.94760175	8.83541187	0.71059501
2003	26.747	0.1986978	7.10154476	8.98227465	0.74030241
2004	28.786	0.46250206	7.50539723	9.69380548	0.8984043
2005	29.97	0.47486664	8.18068519	10.2712124	0.9004196
2006	28.895	0.51886671	8.67044722	10.8613625	1.04718917
2007	33.686	0.47018991	8.81366871	11.3982127	1.18955288
2008	29.573	0.42888348	9.07997319	12.1200922	1.29914359
2009	32.503	0.50547629	9.23414984	11.8702293	1.34867068
2010	34.826	0.53216014	10.6729775	13.1975141	1.40653482
2011	37.398	0.48216367	10.9454427	13.973041	1.53200688
2012	37.154	0.64399709	11.3304914	14.7537455	1.63545486
2013	39.362	0.73188456	12.0056617	15.2487101	1.72073087
2014	43.207	0.75025972	12.6456028	16.0162518	1.81218917
2015	40.759	0.70596149	12.266559	15.9472554	1.9080877
2016	39.504	0.76208864	11.5254489	15.8551322	1.94933964
2017	39.276	0.70785642	11.2531065	16.4157818	2.0363233
2018	37.919	0.60811169	12.1840345	17.3455494	2.11562752
2019	36.758	0.54346596	11.8100024	17.7764171	2.19740929
2020	30.932	0.46931733	9.4979278	15.4095107	2.57846432

A.2. Anexo II

- Código de R completo

Listing A.1: Librerías

```
# LIBRERIAS

library(ggplot2)
library(gridExtra)
library(dplyr)
library(tidyr)
library(readxl)
library(plotly)
library(rmdformats)
library(ggplot2)
library(kableExtra)
library(pander)
library(MASS)
library(lmtest)
```

Listing A.2: Datos

```
# DATOS

zData <- read_excel('00_Data.xlsx', sheet = 2)
names(zData) <- c('Fecha', 'CO2', 'Gas natural', 'Derivados del
    petroleo', 'Energia primaria', 'Electricidad', 'Poblacion'
    , 'PIB')

## Grafico
plot_ly(zData, type = 'scatter', mode = 'lines') %>%
  add_trace(x = ~Fecha, y = ~zData[[3]], name = names(zData)[3],
      line = list(color = '#2e712a', width = 2),
            marker = list(color = '#2e712a', size = 4)) %>%
  add_trace(x = ~Fecha, y = ~zData[[4]], name = names(zData)[4],
      line = list(color = '#67d448', width = 2),
            marker = list(color = '#67d448', size = 4)) %>%
  add_trace(x = ~Fecha, y = ~zData[[5]], name = names(zData)[5],
      line = list(color = '#8edcb9', width = 2),
                  marker = list(color = '#8edcb9', size = 4))
                     %>%
  add_trace(x = ~Fecha, y = ~zData[[6]], name = names(zData)[6],
      line = list(color = '#f0dfa1', width = 2),
```

```
                    marker = list(color = '#f0dfa1', size = 4))
                        %>%
          layout(title = 'Informaci n hist rica',
                 xaxis = list(zerolinecolor = '#ffff',
                              zerolinewidth = 2,
                              gridcolor = 'ffff',
                              title = 'A o'),
                 yaxis = list(zerolinecolor = '#ffff',
                              zerolinewidth = 2,
                              gridcolor = 'ffff',
                              title = 'Mtoe'))


## An lisis por variable

**- Derivados del petr leo:**
data.frame('Media' = mean(Sub_oil),
           'Desv Est' = sd(Sub_oil))

Cuartiles
summary(Sub_oil)


**- Energ a primaria:**
data.frame('Media' = mean(Sub_ep),
           'Desv Est' = sd(Sub_ep))

Cuartiles
```{r message=FALSE, warning=FALSE, echo = FALSE}
summary(Sub_ep)


**- Electricidad:**
data.frame('Media' = mean(Sub_ele),
           'Desv Est' = sd(Sub_ele))

Cuartiles
```{r message=FALSE, warning=FALSE, echo = FALSE}
```

```
summary(Sub_ele)


**- Gas natural:**
data.frame('Media' = mean(Sub_gn),
           'Desv_Est' = sd(Sub_gn))

Cuartiles
summary(Sub_gn)
```

Listing A.3: RLM

```
# RLM

datos <- zData %>%
  dplyr::select(c(`Derivados del petr leo`, `Energ a primaria
     `, `Gas natural`, Electricidad, CO2))

set.seed(1234)
entrenamiento <- sample_frac(datos, .9)
val <- setdiff(datos, entrenamiento)
set.seed(1234)
modelo1 <- lm(CO2 ~ . , data = entrenamiento)
summary(modelo1)

## Validaci n del modelo

### Relaci n lineal entre predictores y variable independiente
plot11 <- ggplot(data = entrenamiento, aes(`Derivados del
   petr leo` , modelo1$residuals)) +
    geom_point() + geom_smooth(color = "firebrick") + geom_hline
       (yintercept = 0) +
    theme_bw()
plot22 <- ggplot(data = entrenamiento, aes(`Energ a primaria`,
   modelo1$residuals)) +
    geom_point() + geom_smooth(color = "firebrick") + geom_hline
       (yintercept = 0) +
    theme_bw()
plot33 <- ggplot(data = entrenamiento, aes(Electricidad, modelo1
```

```
   $residuals)) +
    geom_point() + geom_smooth(color = "firebrick") + geom_hline
       (yintercept = 0) +
    theme_bw()
plot44 <- ggplot(data = entrenamiento, aes(`Gas natural`, modelo
   1$residuals)) +
    geom_point() + geom_smooth(color = "firebrick") + geom_hline
       (yintercept = 0) +
    theme_bw()
grid.arrange(plot11, plot22, plot33, plot44)

### Dist normal de los residuos
qqnorm(modelo1$residuals)
qqline(modelo1$residuals)

Test de Shapiro
Aplicando el test de Shapiro:

shapiro.test(modelo1$residuals)

Dado que $p-valor > 0.05$ NO rechazamos que los errores siguen
   una distribuci n normal.

### Homocedasticidad
ggplot(entrenamiento, aes(modelo1$fitted.values,modelo1$
   residuals))+ geom_point() +
  geom_smooth(color = 'firebrick', se=FALSE)+ geom_hline(
     yintercept = 0)+ theme_bw()

library(lmtest)
bptest(modelo1)


### Autocorrelacion
Test Durbin-Watson
library(lmtest)
dwtest(modelo1)
```

```
### Intervalos de confianza:
confint(modelo1)
```

Listing A.4: RLM - Box Cox

```
# Transformaci n de Box y Cox

# Para 'Derivados del petr leo'
 b <- boxcox(lm(datos$'Derivados del petr leo' ~ 1))
 # Lambda exacto -->
 lambda <- b$x[which.max(b$y)]
 # Transformacion:  -->
 Derivadosdelpetroleo_t <- ( datos$'Derivados del petr leo'^
    lambda - 1) / lambda

#Para ep -->
 b1 <- boxcox(lm(datos$'Energ a primaria' ~ 1))
 # Lambda exacto -->
 lambda1 <- b1$x[which.max(b1$y)]
 # Transformacion:  -->
EnergiaPrimaria_t <- ( datos$'Energ a primaria'^lambda1 - 1) /
   lambda1


 b3 <- boxcox(lm(datos$Electricidad ~ 1))
 # Lambda exacto -->
lambda3 <- b3$x[which.max(b3$y)]
 # Transformacion:  -->
 Electricidad_t <- ( datos$Electricidad^lambda3 - 1) / lambda3

  b2 <- boxcox(lm(datos$'Gas natural' ~ 1))
 # Lambda exacto -->
lambda2 <- b2$x[which.max(b2$y)]
 # Transformacion:  -->
 GasNatural_t <- ( datos$'Gas natural'^lambda2 - 1) / lambda2

 datos_t<- data.frame('Derivados del petr leo' =
    Derivadosdelpetroleo_t, Energia = EnergiaPrimaria_t,
```

```
    Electricidad = Electricidad_t, `Gas natural` = GasNatural_t
    , CO2 = datos$CO2)

set.seed(1234)
entrenamientot <- sample_frac(datos_t, .9)
valt <- setdiff(datos_t, entrenamientot)


modelo3<- lm( CO2 ~. , data = entrenamientot)
summary(modelo3)


## Validaci n del modelo
### Relaci n lineal entre los predictores num ricos y la
   variable respuesta

plot1 <- ggplot(data = entrenamientot, aes(`Derivados.del.
   petr leo` , modelo3$residuals)) +
    geom_point() + geom_smooth(color = "firebrick") + geom_hline
       (yintercept = 0) +
    theme_bw()
plot2 <- ggplot(data = entrenamientot, aes(Energia, modelo3$
   residuals)) +
    geom_point() + geom_smooth(color = "firebrick") + geom_hline
       (yintercept = 0) +
    theme_bw()
plot3 <- ggplot(data = entrenamientot, aes(Electricidad, modelo3
   $residuals)) +
    geom_point() + geom_smooth(color = "firebrick") + geom_hline
       (yintercept = 0) +
    theme_bw()
plot4 <- ggplot(data = entrenamientot, aes(`Gas.natural`, modelo
   3$residuals)) +
    geom_point() + geom_smooth(color = "firebrick") + geom_hline
       (yintercept = 0) +
    theme_bw()
grid.arrange(plot1, plot2, plot3, plot4)
```

```
### Dist normal de los residuos
 qqnorm(modelo3$residuals)
 qqline(modelo3$residuals)

Test de Shaphiro
####Aplicando el test de Shapiro:
shapiro.test(modelo3$residuals)


### Homocedasticidad
ggplot(entrenamiento, aes(modelo3$fitted.values,modelo3$
   residuals))+ geom_point() +
  geom_smooth(color = 'firebrick', se=FALSE)+ geom_hline(
     yintercept = 0)+ theme_bw()

####Test de Breusch-Pagan
bptest(modelo3)

Dado que $p-valor >0.05$ no rechazamos la hip tesis de
   homocedasticidad.

### autocorrelacion
Test Durbin-Watson
library(lmtest)
dwtest(modelo3)
```

Listing A.5: RPG

```
# RPG

library(metaheuristicOpt)
minimizar<- function(w1,w2,w3,w4,w5,w6,w7, w8, w9)
{
  w1<- 0
  w2<- 0
  w3<- 0
  w4<- 0
  w5<- 0
```

```
  w6<- 0
  w7<- 0
  w8<- 0
  w9<- 0

  resultado<- sum (( (datos$CO2) -  w1*((datos$`Derivados del
     petr leo `)^(w2)) - w3*((datos$`Gas natural `)^(w4)) -  w5
     *((datos$Electricidad)^(w6)) - w7*((datos$`Energ a
     primaria `)^(w8)) - w9 )^2)

  return(resultado)
 }


## Definir par metros
Vmax <- 2 # un n mero entero positivo para determinar la
   velocidad m xima de la part cula
ci <- 1.5 # un n mero entero positivo para determinar la
   cognici n individual. El valor predeterminado es 1.49445.
cg <- 1.5 # un n mero entero positivo para determinar el grupo
   cognitivo. El valor predeterminado es 1.49445.
w <- 0.7 # un n mero entero positivo para determinar el peso de
    inercia. El valor predeterminado es 0,729.
numVar <- 9 # un n mero entero positivo para determinar las
   variables num ricas.
rangeVar <- matrix(c(-0.1,3.2), nrow=2)

## calculate the optimum solution using Particle Swarm
   Optimization Algorithm
set.seed(1234)

resultPSO <- PSO(minimizar, optimType="MIN", numVar,
   numPopulation=40,
                maxIter=500, rangeVar, Vmax, ci, cg, w)

minimizar(resultPSO)
resultPSO
```

Listing A.6: GAM

```
# GAM

library(mgcv)

entrenamiento<- rename(entrenamiento, Deriv_Petr = `Derivados
   del petr leo`, Energia_Primaria = `Energ a primaria`, Gas_
   natural = `Gas natural`)

gam1 <- gam(CO2 ~ s(Deriv_Petr) + s(Energia_Primaria) + s(Gas_
   natural) + s(Electricidad) , method="REML", select=TRUE, data
    = entrenamiento)
summary(gam1)

par(mfrow=c(2,2))
plot(gam1,residuals=TRUE)

gam3 <- gam(CO2 ~ s(Deriv_Petr ,bs = "ps" , m=2, k = 10) + s(
   Energia_Primaria,bs = "ps" , m=2, k = 10)  + s(Electricidad,
   bs = "ps" , m=2, k = 10) , method="REML", select=TRUE, data =
    entrenamiento)
summary(gam3)

set.seed(1234)
gam4 <- gam(CO2 ~  s(Energia_Primaria,bs = "ps" , m=2, k = 10)
   + s(Electricidad,bs = "ps" , m=2, k = 10) , method="REML",
   select=TRUE, data = entrenamiento)

summary(gam4)
```

Listing A.7: Comparación

```
# COMPARACI N

val<- rename(val, Energia_Primaria = `Energ a primaria`, Gas_
   natural = `Gas natural`)

pred_mod1<- predict(modelo1,val)
```

```
pred_mod2<- predict(modelo2,val)

pred_PSO<- resultPSO[1]*(val$`Derivados del petr leo`)^(
   resultPSO[2]) + resultPSO[3]*(val$`Gas natural`)^(resultPSO[4
   ]) + resultPSO[5]*(val$Electricidad)^(resultPSO[6]) +
   resultPSO[7]*(val$`Energ a primaria`)^(resultPSO[8]) +
   resultPSO[9]

predGAM <- predict.gam(gam4,val)

final<- data.frame(Real = val$CO2, MRL = pred_mod1 , MRLt =
   predmod3, PSO = pred_PSO )
final<- final %>% mutate(GAM=predGAM,MRL_BOX = predmod3)

final<- final %>% mutate( Error_relativo_MRL = abs( Real-MRL ) /
    Real , Error_relativo_PSO = abs( Real-PSO ) / Real, Error_
   relativo_GAM = abs( Real-GAM ) / Real, Error_relativo_MRL_BOX
    = abs( Real-MRL_BOX ) / Real )

## Indicadores
error<- data.frame( MediaMRL =  mean(final$Error_relativo_MRL) ,
    MediaPSO = mean(final$Error_relativo_PSO), MediaGAM = mean(
   final$Error_relativo_GAM))
error

# Calculo del RMSE y MAPE

# RMSE, MAPE, VAF, ED

indicadoresMRL<- c( sqrt(abs(sum( final$Real-final$MRL  ) / 50))
    , sum(abs( ((final$Real-final$MRL)/(final$Real))/50  ))*100,
    (1-var( final$Real-final$MRL )/var(final$Real))*100 , sqrt(
   sum((final$Real-final$MRL)^2)) )

indicadoresMRLt <- c( sqrt(abs(sum( final$Real-final$MRL_BOX) /
   50)) , sum(abs( ((final$Real-final$MRL_BOX)/(final$Real))/50
    ))*100, (1-var( final$Real-final$MRL_BOX)/var(final$Real))*1
   00 , sqrt(sum((final$Real-final$MRL_BOX)^2)) )
```

```
indicadoresPSO<- c( sqrt(abs(sum( final$Real-final$PSO  ) / 50))
    , sum(abs( ((final$Real-final$PSO)/(final$Real))/50  ))*100,
    (1-var( final$Real-final$PSO )/var(final$Real))*100 , sqrt(
   sum((final$Real-final$PSO)^2)) )

indicadoresGAM<- c( sqrt(abs(sum( final$Real-final$GAM  ) / 50))
    , sum(abs( ((final$Real-final$GAM)/(final$Real))/50  ))*100,
    (1-var( final$Real-final$GAM )/var(final$Real))*100 , sqrt(
   sum((final$Real-final$GAM)^2)) )

indicadores<- data.frame(Indicador = c("RMSE","MAPE","VAF","ED")
   , MRL = indicadoresMRL, MRL_box = indicadoresMRLt, PSO=
   indicadoresPSO, GAM = indicadoresGAM)
indicadores


## Resultado, predicciones
final<- final  %>% mutate(Fecha= c(2015,2016,2017,2018,2019,2020)
    )


plot(x=final$Fecha, y = final$Real,xlab = "A o",ylab="Emisi n␣
   de␣CO2" , type="o",col="blue")
par(new=TRUE)
plot(x=final$Fecha, y = final$MRL, xlab="", ylab="" ,type="o",
   col="red",axes=FALSE)
par(new=TRUE)
plot(x=final$Fecha, y = final$PSO, xlab="", ylab="" ,type="o",
   col="green",axes=FALSE)
par(new=TRUE)
plot(x=final$Fecha, y = final$GAM, xlab="", ylab="" ,type="o",
   col="purple",axes=FALSE)
par(new=TRUE)
plot(x=final$Fecha, y = final$MRL_BOX, xlab="", ylab="" ,type="o
   ",col="yellow",axes=FALSE)

legend("topleft",legend = c("Real","MRL","PSO","GAM","MRL_BOX"),
```

```
lwd=3,col=c("blue","red","green","purple","yellow"))
```

Referencias bibliográficas

[1] Udara Abeydeera, Wadu Jayantha, and Tharushi Samarasinghe. Global research on carbon emissions: A scientometric review. *Sustainability*, 11:3972, 07 2019.

[2] Suad Alasadi and Wesam Bhaya. Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences 12*, 16:2–7, 2017.

[3] Atef Saad Alshehry and Mounir Belloumi. Energy consumption, carbon dioxide emissions and economic growth: The case of saudi arabia. *Renewable and Sustainable Energy Reviews*, 41:237–247, 2015.

[4] Alves da Silva A. P. and Ahrao P. J. Application of evolutionary computation in electric power systems. *Proceedings of International Congress on Evolutionary Computation*, pages 1057– 1062, 2002.

[5] Box, G. E. P. and Tiao, G. C. A robust sequential bayesian method for identification of differentially expressed genes. *Biometrika*, 49:419–432, 1962.

[6] British Petroleum. Statistical Review of World Energy. *BP Statistical Review of World Energy*, 70:4–106, 2021.

[7] Edwin Buenaño, Emilio Padilla, and Vicent Alcántara. Relevant sectors in CO2 emissions in Ecuador and implications for mitigation policies. *Energy Policy*, 158:4–7, 2021.

[8] C. Zhou, H. B. Gao, L. Gao, and W. G. Zhang. Particle swarm optimization (pso) algorithm. *Application Research of Computers*, 12:7–11, 2003.

[9] Arys Carrasquilla, Alfonso Chacon-Rodriguez, Montero Núñez, Olman Gómez-Espinoza, Johnny Valverde, and Maritza Guerrero. Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. *Revista Tecnología en Marcha*, 29:33, 12 2016.

[10] Trevor Hastie and Robert Tibshirani. *Statistical Science Generalized Additive Models*, volume 1. Institute of Mathematical Statistics, Estados Unidos, 1986.

[11] Roberto Hernández, Carlos Fernández, and Pilar Baptista. *Metodología de la investigación: Sexta Edición*. Mc Graw Hill Education, México D.F., 2014.

[12] Alboukadel Kassambara. Comparing groups: Numerical variables. *Practical Statistics in R*, 1:172–186, 11 2019.

[13] Salih Katircioglu and Nigar Taspinar. Testing the moderating role of financial development in an environmental kuznets curve: Empirical evidence from turkey. *Renewable and Sustainable Energy Reviews*, 68:572–586, 02 2017.

[14] Kennedy J. and Eberhart R. C. Particle swarm optimization. *Proceedings of International Conference on Neural Networks*, pages 1942–1948, 1995.

[15] Ravindra Khaiwal, Preety Rattan, Suman Mor, and Ashutosh Nath. Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment International*, 132:13–15, 08 2019.

[16] Liang, F., Liu, C., and Wang, N. A robust sequential bayesian method for identification of differentially expressed genes. *Statist. Sinica*, 17:571–597, 2007.

[17] Yin Libao, Yao Tingting, Zhou Jielian, Liu Guicai, Liao Yanfen, and Ma Xiaoqian. Prediction of co2 emissions based on multiple linear regression analysis. *Energy Procedia*, 105:4222–4228, 2017.

[18] Cosimo Magazzino. The relationship between real gdp, co2 emissions, and energy use in the gcc countries: A time series approach. *Cogent Economics Finance*, 4, 12 2016.

[19] Marques, A. C., Fuinhas, J. A., and Nunes, A. R. Electricity generation mix and economic growth: What role is being played by nuclear sources and carbon dioxide emissions in France. *BP Statistical Review of World Energy*, 92:7–19, 2016.

[20] Ministerio de Energía y Recursos Naturales No Renovables. Consumo de energía por sector y fuente. *Balance Energético Nacional 2020*, 9:1–13, 2020.

[21] Munim, J.M.A, Hakim, M.M., and Abdullah-Al-Mamun, M. Analysis of energy consumption and indicators of energy use in bangladesh. *Econ Change Restruct*, 43:275–302, 2010.

[22] I.M. Nazarov, Yu.A. Izrael, M.L. Gitarskii, Alexander Nakhutin, and A.F. Yakovlev. The problem of anthropogenic climate change and the kyoto protocol. pages 96–104, 01 2004.

[23] Organización Meteorológica Mundial. Se alcanzan niveles récord de concentración de gases de efecto invernadero en la atmósfera. *Noticias ONU*.

[24] Magnus Erik Hvass Pedersen. Good parameters for particle swarm optimization. *Hvass Laboratories Technical Report no. HL1001*, pages 69 – 73, 2010.

[25] Kathia Pinzón. Dynamics between energy consumption and economic growth in ecuador: A granger causality analysis. *Economic Analysis and Policy*, 57, 09 2017.

[26] Programa de las Naciones Unidas para el Desarrollo. Ecuador lanza su tercera comunicación nacional sobre cambio climático. *Noticias Programa de las Naciones Unidas para el Desarrollo*.

[27] Dian Palupi Rini, Siti Mariyam Shamsuddin, and Siti Sophiyati Yuhaniz. Particle swarm optimization: Technique, system and challenges. *Int. J. Comput. Appl.*, 14, 2011.

[28] Sangeetha, A. and Amudha, T. A novel bio-inspired framework for co 2 emission forecast in india. *Procedia Computer Science*, 125:367–375, 01 2018.

[29] Lee Schipper, Fridtjof Unander, Scott Murtishaw, and Mike Ting. Indicators of energy use and carbon emissions: Explaining the energy economy link. *Annual Review of Energy and the Environment*, 26:49–81, 2001.

[30] Subana Shanmuganathan and Sandhya Samarasinghe. *Artificial Neural Network Modelling*. Studies in Computational Intelligence, Estados Unidos, 2016.

[31] Sociedad Nuclear Española. Diccionario nuclear. https://www.sne.es/diccionario-nuclear/tonelada-equivalente-de-petroleo/. Obtenido: 2021-12-26.

[32] Dennis Wackerly, William Mendenhall, and Richard Scheaffer. *Estadística matemática con aplicaciones: Séptima Edición*. Cengage Learning Editores, S.A, Argentina, 2010.

[33] Wood, Simon N. Generalized additive models: An introduction with r. *Chapman and Hall/CRC*, page 121, 2006.

[34] Jin Xin, Liang Yongquan, Tian Dongping, and Zhuang Fuzhen. Particle swarm optimization using dimension selection methods. *Applied mathematics and computation*, 219:5185–5197, 2013.

[35] Xin-She Yang. Introduction to mathematical optimization. from linear programming to metaheuristics. pages 29 – 33, 01 2008.

[36] Yaser, S, Malik Magdon Ismail, and Hsuan Tien Lin. *Learning from data - A short course*. AML book, Estados Unidos, 2012.

[37] Rui Zou, Vijay Kalivarapu, Eliot Winer, James Oliver, and Sourabh Bhattacharya. Particle swarm optimization based source seeking. *Iowa State University*, 12 2015.