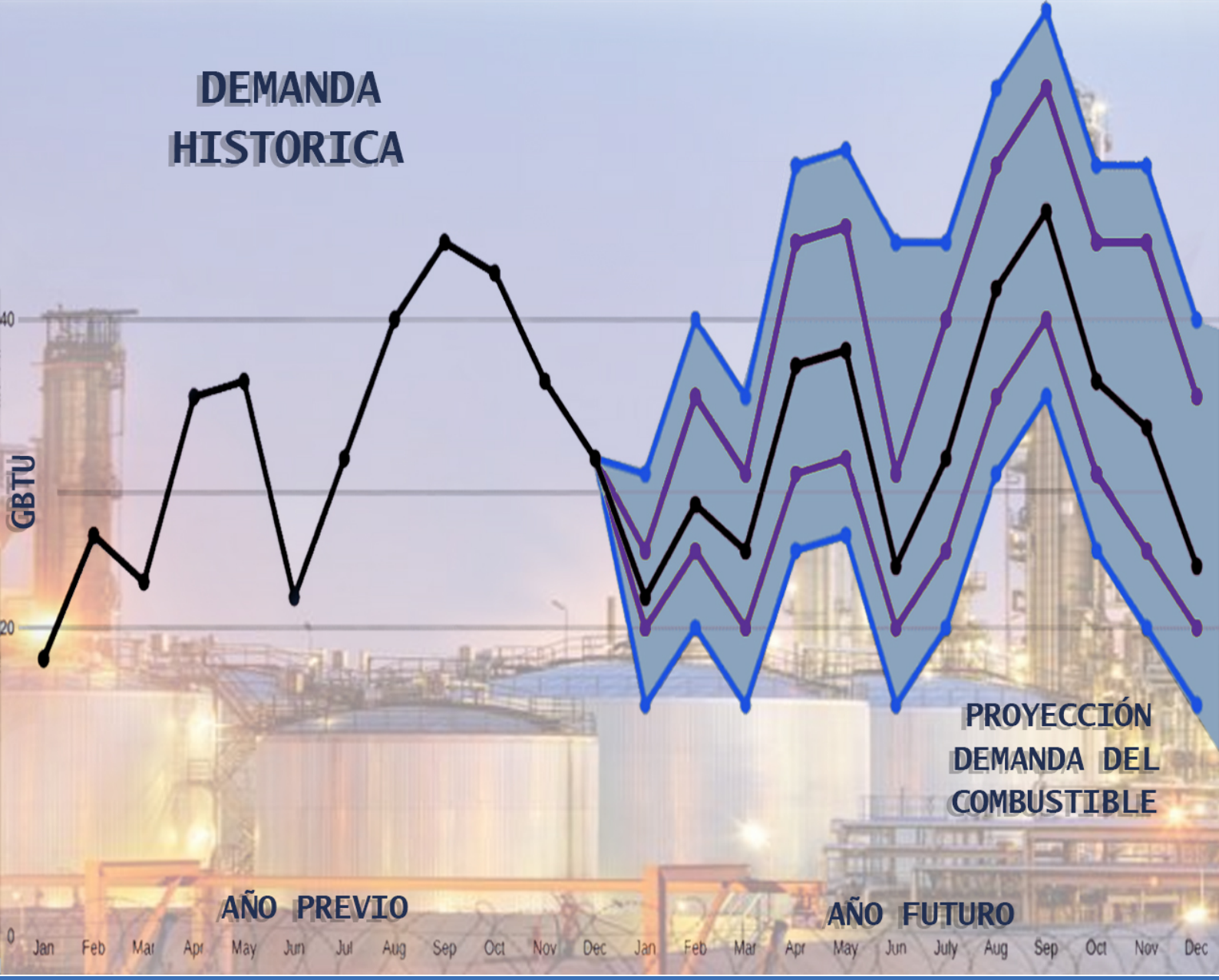

Proyecto UPME: Proyecciones de la demandas de combustibles líquidos y GLP

Entregable # 3

Autores: Mauricio Lopera, Jorge Iván Pérez, Noé Mesa, John Esteban Londoño, Carlos Osuna, Adiel Restrepo y Diego Mejía Giraldo

Institución: Universidad de Antioquia

Fecha de entrega: Diciembre 19, 2021

Versión: 1.0

Estado: : En Revisión

UNIVERSIDAD DE ANTIOQUIA

upme

Índice general

1. Objetivo del proyecto **3**

2. Revisión de literatura de proyecciones de demanda de energía eléctrica y gas natural **4**

2.1. Energía Eléctrica 4

2.2. Gas Natural 13

3. Metodologías para la proyección de la demanda de energía eléctrica y gas natural usadas por la UPME **21**

3.1. Metodología de proyección de demanda de energía eléctrica 21

3.2. Metodología de proyección de demanda de gas natural 24

3.2.1. Sector Termoeléctrico 26

4. Recomendaciones y conclusiones sobre metodología UPME para proyectar energía eléctrica y gas natural **27**

4.1. Introducción 27

4.2. Variables explicativas 28

4.2.1. Energía eléctrica 29

4.2.2. Gas Natural 31

4.3. Modelos alternativos 33

4.3.1. Aplicaciones 34

4.3.2. Resultados casos de estudio 61

4.4. Observaciones adicionales 70

4.4.1. Pronósticos demanda sector termoeléctrico y refinerías 70

4.4.2. Realización de pronósticos bajo los diferentes escenarios del PIB 70

4.4.3. Uso de los pronósticos del PIB en la demanda de la energía eléctrica . . 71

4.4.4. Demanda de energía eléctrica de grandes consumidores 71

4.4.5. Proyecciones de vehículos eléctricos 71

4.4.6. Escenarios de generación distribuida 72

4.5. Conclusiones 72

5. Revisión de literatura de combustibles líquidos y GLP **73**

5.1. Ecopetrol 73

5.2. Diésel . . . 76
5.3. Fuel Oil . . . 79
5.4. GLP . . . 80
5.5. Gasolina . . . 85
5.6. Jet Fuel . . . 94
5.7. Queroseno . . . 97
5.8. Conclusiones . . . 100

6. Metodología propuesta de proyección de combustibles líquidos 101
6.1. Introducción . . . 101
6.2. Modelo de Regresión Lineal Múltiple (MLR) . . . 102
6.3. Modelo Aditivo Generalizado (GAM) . . . 103
6.4. Regresión lineal con Shrinkage (LASSO) . . . 104
6.5. Regresión Spline Adaptativa Multivariante (MARS) . . . 106
6.6. Modelo de Redes Neuronales Recurrentes de Corta y Larga Memoria (LSTM) . . . 107
6.7. Metodología de combinación de pronósticos . . . 109
6.8. *Bootstrap* en Modelos de Regresión . . . 110
6.9. El *Wild Bootstrap* . . . 111
6.10. Variables explicativas . . . 112
6.11. Metodologías Clásicas de Predicción y sus Restricciones en la Modelación de Combustibles Líquidos . . . 122
6.11.1. Modelos de Vectores Autorregresivos y Modelos de Corrección de Error . . . 122
6.11.2. Modelos Estacionales Autorregresivos de Media Móvil con Variables Externas el Modelo SARIMAX . . . 124
6.11.3. Modelo SARIMAX para la demanda de gasolina . . . 125
6.12. Medidas de bondad de ajuste y comparación de modelos . . . 126
6.12.1. Error Medio (ME) . . . 127
6.12.2. Error Porcentual Medio (MPE) . . . 128
6.12.3. Error Absoluto Medio (MAE) . . . 128
6.12.4. Error Absoluto Porcentual Medio (MAPE) . . . 128
6.12.5. Suma de Cuadrados del Error (SSE) . . . 129
6.12.6. Error Cuadrático Medio (MSE) . . . 129
6.12.7. Raíz del Error Cuadrático Medio (RMSE) . . . 130
6.12.8. Criterio de Información de Akaike (AIC) . . . 130
6.12.9. Criterio de Información de Akaike Corregido (AICc) . . . 131

6.12.10.Criterio de Información Bayesiano (BIC) . . . 131
6.12.11.Criterio de Información Hannan-Quinn (HQC) . . . 132
6.12.12.Error de Predicción Final (FPE) . . . 132

7. Resultados combustibles líquidos **133**
7.1. ACPM/Diésel . . . 134
7.1.1. ACPM/Diésel: Escenario 1 . . . 134
7.1.2. ACPM/Diésel: Escenario 2 . . . 140
7.1.3. ACPM/Diésel: Escenario 3 . . . 146
7.2. Fuel oil . . . 151
7.2.1. Fuel Oil: Escenario 1 . . . 152
7.2.2. Fuel Oil: Escenario 2 . . . 157
7.3. Gas Licuado de Petróleo (GLP) . . . 162
7.3.1. GLP: Escenario 1 . . . 162
7.3.2. GLP: Escenario 2 . . . 168
7.4. Gasolina Motor (GM) . . . 174
7.4.1. GM: Escenario 1 . . . 174
7.4.2. GM: Escenario 2 . . . 179
7.5. Jet Fuel . . . 183
7.5.1. Jet Fuel: Escenario 1 . . . 184
7.5.2. Jet Fuel: Escenario 2 . . . 189
7.6. Conclusiones y recomendaciones . . . 195

A. Manual Usuario **207**
A.1. Conexión Google Colab—Drive . . . 207
A.2. Funciones . . . 207
A.3. Cargar Módulos y Base de Datos . . . 207
A.4. Nombre y Descripción de Variables . . . 208
A.5. Información de entrada para modelar . . . 208
A.5.1. Ajustes generales . . . 208
A.5.2. Modelos a estimar . . . 209
A.5.3. Ajuste base de datos . . . 209
A.6. Cálculo de Información de Entrada . . . 210
A.7. Entrenamiento y Validación de los modelos . . . 210
A.7.1. Ajuste modelo e Hiperparámetros . . . 210

A.7.2. Gráficos Entrenamiento, Validación y Proyección 210

B. Modelos de proyección de combustibles **211**

B.1. Modelo de Regresión Lineal (LRM) 211

B.2. Modelo Aditivo Generalizado 212

B.3. Un Ejemplo con Datos Simulados 216

B.4. El problema del sobre-ajuste 219

B.5. Estimación penalizada 221

B.6. Redes Neuronales Feed-Forward 221

B.7. Redes Neuronales Recurrentes (RNN) 224

B.8. Redes Neuronales Recurrentes de Corta y Larga Memoria. LSTM 226

B.9. Regresión Spline Adaptativa Multivariante (MARS) 228

Índice de figuras

3.1. Metodología actual para la proyección de la energía eléctrica de la UPME 22
3.2. Metodología actual para la proyección del gas natural de la UPME 25

4.1. Serie original y modificada de la demanda de energía eléctrica para el periodo 2009-1 a 2019-12 37
4.2. Serie original y modificada del PIB para el periodo 2009-1 a 2019-12 37
4.3. Función de autocorrelación para la demanda de energía eléctrica y el PIB, para un total de 60 rezagos. 39
4.4. Serie original la demanda de energía eléctrica y PIB para el periodo 2009-1 a 2019-8 43
4.5. Función de autocorrelación para la demanda de energía eléctrica y PIB, para un total de 60 rezagos. 44
4.6. Serie de la demanda de gas natural del sector industrial y del IPG para el periodo 2009-1 y 2019-8 47
4.7. Función de autocorrelación para la demanda de gas natural del sector industrial y el IPG, para un total de 60 rezagos. 48
4.8. Serie y ACF de la demanda de gas natural agregada para el periodo 2009-1 y 2017-12 52
4.9. Serie precios promedio de gas natural para usuarios regulados y usuarios no regulados para el periodo 2009-1 y 2017-12 56
4.10. ACF precios promedio de gas natural para usuarios regulados y usuarios no regulados para el periodo 2009-1 y 2017-12 57
4.11. Ajuste pronóstico modelos Caso 1: Demanda energía eléctrica bajo escenario desimulación del COVID-19. 65
4.12. Ajuste pronóstico modelos Caso 2: Demanda energía eléctrica al incluir campaña "Apagar-Paga" y efecto delfenómeno del niño-niña. 65
4.13. Ajuste pronóstico modelos Caso 3: Demanda gas natural para el sector industrial. 66
4.14. Ajuste pronóstico modelos Caso 4: Demanda gas natural agregada escenario base. 66
4.15. Ajuste pronóstico modelos Caso 5: Demanda gas natural agregada al incluir el efecto del fenómeno del niño. 67
4.16. Ajuste pronóstico modelos Caso 6: Demanda gas natural agregada al incluir el precio promedio del gas natural para usuarios regulado. 67
4.17. Ajuste pronóstico modelos Caso 7: Demanda gas natural agregada al incluir el precio promedio del gas natural para usuarios no regulados. 68

4.18. Ajuste pronóstico modelos Caso 8: Demanda gas natural agregada al incluir el precios promedio del carbón. 68
4.19. Ajuste pronóstico modelos Caso 9: Demanda gas natural agregada al incluir el precios promedio del GLP. 69

5.1. Resumen de la metodología propuesta por Correira et al. (2020) 83

6.1. Salidas del Paquete R para un Modelo ARIMAX 126
6.2. Salidas del Paquete Python para un Modelo ARIMAX 126

7.1. Ajuste del modelo de combinación de pronósticos a la demanda de ACPM/Diésel para el periodo 2010/01 - 2035/12. 135
7.2. Ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda ACPM/Diésel para el periodo 2010/01 - 2020/09. 136
7.3. Ajuste del modelo de combinación de pronósticos para datos de validación de la demanda ACPM/Diésel para el periodo 2020/10 - 2021/09. 138
7.4. Proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésel para el periodo 2021/10 - 2035/12. 139
7.5. Ajuste del modelo de combinación de pronósticos a la demanda de ACPM/Diésel para el periodo 2010/01 - 2035/12. 141
7.6. Ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda ACPM/Diésel para el periodo 2010/01 - 2020/09 142
7.7. Ajuste del modelo de combinación de pronósticos para datos de validación de la demanda ACPM/Diésel para el periodo 2020/10 - 2021/09. 143
7.8. Proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésel para el periodo 2021/10 - 2035/12 145
7.9. Ajuste del modelo LSTM a la demanda de ACPM/Diésel para el periodo 2010/01 - 2035/12 147
7.10. Ajuste del modelo LSTM para datos de entrenamiento de la demanda ACPM/-Diésel para el periodo 2010/01 - 2020/09 148
7.11. Ajuste del modelo LSTM para datos de validación de la demanda ACPM/Diésel para el periodo 2020/10 - 2021/09 149
7.12. Proyecciones del modelo LSTM para la demanda de ACPM/Diésel para el periodo 2021/10 - 2035/12 150
7.13. Ajuste del modelo LSTM a la demanda de Fuel Oil para el periodo 2010/01 - 2035/12 153
7.14. Ajuste del modelo LSTM para datos de entrenamiento de la demanda Fuel Oil para el periodo 2010/01 - 2020/09 154

7.15. Ajuste del modelo LSTM para datos de validación de la demanda Fuel Oil para el periodo 2020/10 - 2021/09 155
7.16. Proyecciones del modelo LSTM para la demanda de Fuel Oil para el periodo 2021/10 - 2035/12 156
7.17. Ajuste del modelo de combinación de pronósticos de pronósticos a la demanda de Fuel Oil para el periodo 2010/01 - 2035/12 158
7.18. Ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda Fuel Oil para el periodo 2010/01 - 2020/09 159
7.19. Ajuste del modelo de combinación de pronósticos para datos de validación de la demanda Fuel Oil para el periodo 2020/10 - 2021/09 160
7.20. Proyecciones del modelo de combinación de pronósticos para la demanda de Fuel Oil para el periodo 2021/10 - 2035/12 161
7.21. Ajuste del modelo GAM a la demanda de GLP para el periodo 2010/01 - 2035/12 163
7.22. Ajuste del modelo GAM para datos de entrenamiento de la demanda GLP para el periodo 2010/01 - 2020/09 164
7.23. Ajuste del modelo GLP para datos de validación de la demanda GLP para el periodo 2020/10 - 2021/09 165
7.24. Proyecciones del modelo GAM para la demanda de GLP para el periodo 2021/10 - 2035/12 167
7.25. Ajuste del modelo GAM a la demanda de GLP para el periodo 2010/01 - 2035/12 169
7.26. Ajuste del modelo GAM para datos de entrenamiento de la demanda GLP para el periodo 2010/01 - 2020/09 170
7.27. Ajuste del modelo GLP para datos de validación de la demanda GLP para el periodo 2020/10 - 2021/09 171
7.28. Proyecciones del modelo GAM para la demanda de GLP para el periodo 2021/10 - 2035/12 172
7.29. Ajuste del modelo MARS a la demanda de GM para el periodo 2010/01 - 2035/12 175
7.30. Ajuste del modelo MARS para datos de entrenamiento de la demanda GM para el periodo 2010/01 - 2020/09 176
7.31. Ajuste del modelo MARS para datos de validación de la demanda GM para el periodo 2020/10 - 2021/09 177
7.32. Proyecciones del modelo MARS para la demanda de GM para el periodo 2021/10 - 2035/12 178
7.33. Ajuste del modelo de combinación de pronósticos a la demanda de GM para el periodo 2010/01 - 2035/12 179
7.34. Ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda GM para el periodo 2010/01 - 2020/09 180

7.35. Ajuste del modelo de combinación de pronósticos para datos de validación de la demanda GM para el periodo 2020/10 - 2021/09 181
7.36. Proyecciones del modelo de combinación de pronósticos para la demanda de GM para el periodo 2021/10 - 2035/12 182
7.37. Ajuste del modelo LSTM a la demanda de Jet Fuel para el periodo 2010/01 - 2035/1 184
7.38. Ajuste del modelo LSTM para datos de entrenamiento de la demanda Jet Fuel para el periodo 2010/01 - 2020/09 185
7.39. Ajuste del modelo LSTM para datos de validación de la demanda Jet Fuel para el periodo 2020/10 - 2021/09 187
7.40. Proyecciones del modelo LSTM para la demanda de Jet Fuel para el periodo 2021/10 - 2035/12 188
7.41. Ajuste del modelo LSTM a la demanda de Jet Fuel para el periodo 2010/01 - 2035/1 190
7.42. Ajuste del modelo LSTM para datos de entrenamiento de la demanda Jet Fuel para el periodo 2010/01 - 2020/09 191
7.43. Ajuste del modelo LSTM para datos de validación de la demanda Jet Fuel para el periodo 2020/10 - 2021/09 192
7.44. Proyecciones del modelo LSTM para la demanda de Jet Fuel para el periodo 2021/10 - 2035/12 194

B.1. Simulación Propia. Los puntos corresponden a un conjunto generado por el verdadero proceso simulado. El verdadero proceso se simuló como la concatenación de tres funciones distintas en los rangos de Age, $(0, 50)$, $(50, 100)$ y $(100, 150)$ La linea azul es la regresión lineal simple de y_t con Age_t 213
B.2. Tomado de Gareth et al. (2021). La figura de la parte superior izquierda muestra el ajuste de dos polinomios cúbicos en los rangos de Age entre $(10, 50)$ y $(50, 80)$. Note como en el nudo $Age = 50$, se genera una discontinuidad. La figura superior derecha muestra los mismos polinomios de la parte superior izquierda, pero con la restricción de continuidad. Sin embargo, las pendientes en $Age = 50$ son distintas. La figura de la parte baja izquierda es un spline cúbico que corresponde a la figura superior derecha agregando las restricciones de que en $Age = 50$, las primeras y segundas derivadas son continuas. La figura inferior derecha es un spline lineal. 215
B.3. Modelo Lineal 217
B.4. Modelo GAM 217
B.5. Modelo Polinomial 218

B.6. Modelos Ajustados y Datos Reales. Note como el modelo lineal sub ajusta los datos, el modelo polinomial de grado 10 tiene más variabilidad que el GAM, lo que indica un sobre ajuste de los datos. El Modelo GAM es una función suave. Cálculos propios 219

B.7. Tomado de Gareth et al. (2021). El modelo verdadero es la linea negra. La linea naranja es la regresión lineal de y_t sobre x_t. La linea azul es la regresión spline de y_t sobre x_t. La linea verde que sobre ajusta el modelo es una función polinomial de grado 10 que sobre ajusta los datos. El mejor modelo es la función spline ya que tiene el menor error de prueba en la linea roja en la parte derecha. Note como el error de entrenamiento en la linea gris de la parte derecha, siempre disminuye con el número de términos en el modelo, mientras que el error de prueba disminuye al principio, alcanza un mínimo en la regresión spline y aumenta con la regresión polinomial que tiene más términos de los necesarios para dar buenas predicciones. 220

B.8. Tomado de Gareth et al. (2021). La capa de entrada en el vector x_t. La capa oculta tiene cinco unidades de activación A_i. Las unidades de activación de la capa oculta alimenta la capa de salida $f(x)$. 222

B.9. Tomado de Gareth et al. (2021). La función de activación sigmoidal es la curva verde, La función de activación ReLU es la linea negra 223

B.10. Tomado de Gareth et al. (2021). La capa de entrada es x_t, la primera capa oculta recibe la información de la capa de entrada y la procesa con sus L_1 unidades de activación. La información de las L_1 unidades de activación de la primera capa alimentan las unidades de activación de la segunda capa oculta. Las L_2 unidades de activación de la segunda capa oculta alimentan cada una de las funciones de la capa de salida $f_i(x_t)$. 224

B.11. Tomado de Gareth et al. (2021). La capa de entrada ahora es secuencial X_1 alimenta a A_1; X_2 y A_1 alimentan a A_2; X_3, A_1 y A_2 alimentan a A_3. AsíA_3 secuencialmente se alimentan las unidades de activación en la RNN. 225

B.12. Tomado de Gareth et al. (2021). Las funciones de activación σ son sigmoidales y deciden que información de x_t y h_{t-1} se mantiene para procesar la celda de estado c_t. Si las funciones sigmoidales están cerca de cero, la información en x_t y h_{t-1} se desecha; si están cerca de uno, se mantiene. Las funciones tangente hiperbólica al estar en el rango $(-1, 1)$ deciden si se toma la información de x_t y h_{t-1} de forma positiva o negativa en la celda de estado. La celda de estado alimenta la capa oculta h_t y esta última alimenta a y_t. 226

Índice de cuadros

2.1. Variables explicativas y modelos empleados en la literatura para el pronóstico de la demanda de Energía Eléctrica. 12

2.2. Variables explicativas y modelos empleados en la literatura para el pronostico de la demanda de la Gas Natural. 20

4.1. Variables explicativas sugeridas para el planteamiento de modelos de pronóstico de la demanda de la Energía Eléctrica en Colombia. 31

4.2. Variables explicativas sugeridas para el planteamiento de modelos de pronóstico de proyección de Gas Natural en Colombia. 33

4.3. Prueba HEGY demanda energía eléctrica y PIB 40

4.4. Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p) 40

4.5. Prueba de Johansen con traza para las variables de la demanda de energía eléctrica y el PIB 41

4.6. Prueba HEGY demanda energía eléctrica y PIB 45

4.7. Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p) 45

4.8. Prueba de Johansen con traza para las variables de la demanda de energía eléctrica y el PIB 46

4.9. Prueba HEGY demanda gas natural agregada e IPG 49

4.10. Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p) 50

4.11. Prueba de Johansen con traza para las variables de la demanda de gas natural para el sector industrial y el IPG 50

4.12. Prueba HEGY demanda gas natural agregada 53

4.13. Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p) 53

4.14. Prueba de Johansen con traza para las variables de la demanda de gas natural agregada y el IPG 54

4.15. Prueba HEGY precio promedio gas natural usuarios regulados y no regulados . 58

4.16. Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p) 58

4.17. Prueba de Johansen con traza para las variables de la demanda de gas natural agregada, el IPG y el precio promedio gas natural, para usuarios regulados (Caso 6) y usuarios no regulados (Caso 7) 59
4.18. MAPE modelos de demanda de energéticos para los diferentes casos de estudio . 62

5.1. Resumen revisión de literatura sobre diésel 78
5.2. Resumen revisión de literatura sobre fuel oil 80
5.3. Resumen revisión de literatura sobre GLP 84
5.4. Resumen revisión de literatura sobre gasolina 90
5.5. Resumen revisión de literatura sobre jet fuel 97
5.6. Resumen revisión de literatura sobre queroseno 99

6.1. Variables explicativas empleadas en los modelos de demanda de combustibles líquidos. 119

7.1. Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda de ACPM/Diésel 136
7.2. Escala de precisión de los pronósticos MAPE (Lewis, 1982), (Melikoglu, 2017). . 137
7.3. Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de validación de la demanda de ACPM/Diésel 139
7.4. Encabezado de validación y proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésel 140
7.5. Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda de ACPM/Diésel 142
7.6. Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de validación de la demanda de ACPM/Diésel 144
7.7. Encabezado de validación y proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésell 146
7.8. Medidas de bondad de ajuste del modelo LSTM para datos de entrenamiento de la demanda de ACPM/Diésel 148
7.9. Medidas de bondad de ajuste del modelo LSTM para datos de validación de la demanda de ACPM/Diésel 149
7.10. Encabezado proyecciones modelo LSTM para la demanda de ACPM/Diésel . . . 151
7.11. Medidas de bondad de ajuste del modelo LSTM para datos de entrenamiento de la demanda de Fuel Oil 154
7.12. Medidas de bondad de ajuste del modelo LSTM para datos de validación de la demanda de Fuel Oil 155
7.13. Encabezado proyecciones modelo LSTM para la demanda de Fuel Oil 157

7.14. Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda de Fuel Oil 159
7.15. Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de validación de la demanda de Fuel Oil 160
7.16. Encabezado proyecciones modelo de combinación de pronósticos para la demanda de Fuel Oil 161
7.17. Medidas de bondad de ajuste del modelo GAM para datos de entrenamiento de la demanda de GLP 165
7.18. Medidas de bondad de ajuste del modelo GAM para datos de validación de la demanda de GLP 166
7.19. Encabezado proyecciones modelo GAM para la demanda de GLP 168
7.20. Medidas de bondad de ajuste del modelo GAM para datos de entrenamiento de la demanda de GLP 170
7.21. Medidas de bondad de ajuste del modelo GAM para datos de validación de la demanda de GLP 172
7.22. Encabezado proyecciones modelo GAM para la demanda de GLP 173
7.23. Medidas de bondad de ajuste del modelo MARS para datos de entrenamiento de la demanda de GM 176
7.24. Medidas de bondad de ajuste del modelo MARS para datos de validación de la demanda de GM 177
7.25. Encabezado proyecciones modelo GAM para la demanda de GM 178
7.26. Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda de GM 180
7.27. Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de validación de la demanda de GM 181
7.28. Encabezado de validación y proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésell 182
7.29. Medidas de bondad de ajuste del modelo LSTM para datos de entrenamiento de la demanda de Jet Fuel 186
7.30. Medidas de bondad de ajuste del modelo LSTM para datos de validación de la demanda de Jet Fuel 188
7.31. Encabezado proyecciones modelo LSTM para la demanda de Jet Fuel 189
7.32. Medidas de bondad de ajuste del modelo LSTM para datos de entrenamiento de la demanda de Jet Fuel 192
7.33. Medidas de bondad de ajuste del modelo LSTM para datos de validación de la demanda de Jet Fuel 193
7.34. Encabezado proyecciones modelo LSTM para la demanda de Jet Fuel 195

B.1. El MSE de entrenamiento disminuye con la Flexibilidad. El MAPE de prueba disminuye del Modelo lineal (Flexibilidad 2) al GAM (Flexibilidad=6) y aumenta con el Modelo Polinomial de grado 10 (Flexibilidad=10). 219

Resumen ejecutivo

El informe aquí planteado está construido de la siguiente manera. En el Capítulo 1 se plantean los objetivos generales y específicos que se establecieron para la realización del proyecto.

En el Capítulo 2 se presenta una revisión bibliográfica en la que se examinaron tanto artículos científicos, como reportes, tesis de grado y otros trabajos técnicos, con el fin de ilustrar algunas metodologías alternativas a los modelos VAR y VEC que se emplean en la literatura para el pronóstico de la energía eléctrica y el gas natural. Es de anotar que entre las metodologías empleadas en la literatura también se destacan los modelos de regresión, ARIMA, redes neuronales recurrentes, modelos GAM, entre otros. Adicionalmente se pudo evidenciar entre las metodologías más recientes (último año) que algunos trabajos incluían variables relacionadas con el efecto de la pandemia causada por el COVID-19; con el fin de encontrar una explicación al comportamiento atípico que tuvo la demanda de energía eléctrica y gas. En el mismo sentido se logra identificar también algunas variables recurrentes que se empleaban en los diferentes trabajos de la literatura, en donde se destacan las variables macroeconómicas: Ingreso per-cápita, Población, Precio de la energía eléctrica, Precio del gas natural; y variables meteorológicas: Temperatura (mínima, media y máxima), "Heating degree days" y "Cooling degree days". También, se ha encontrado que variables asociadas a la pandemia del COVID-19 han sido empleadas recientemente en la literatura internacional para modelos de pronóstico de demanda de energía eléctrica.

En el Capítulo 3 se realiza una descripción sobre las metodologías que emplea actualmente la UPME para la proyección de la demandas de energía eléctrica y el gas natural, en donde el objetivo es ilustrar los procedimientos empleados por la UPME para la estimación de los diferentes modelos de pronóstico, resaltando las variables y metodologías empleadas para tal fin.

Finalmente, en el Capítulo 4 se plantean una serie de recomendaciones a las metodologías empleadas por la UPME para la proyección de demanda de energía electricidad y de gas natural, las cuales se elaboran una vez se conoce de forma detallada los procedimientos empleados por la UPME para realizar sus proyecciones. Es de anotar que las recomendaciones realizadas pueden clasificarse en dos categorías, corto plazo y mediano plazo, en donde las primeras se elaboran con el fin de aportar ciertas mejoras a las metodologías actuales de la UPME, mientras que, las segundas se elaboran con vista a cambios más significativos de las metodologías, puesto que se plantean modelos alternativos a los usados en la actualidad por la UPME.

Posteriormente, este informe se dedica al análisis de metodologías para la proyección de combustibles líquidos y GLP para el horizonte de 2035. En el Capítulo 5 se presenta una revisión de trabajos científicos publicados en revistas nacionales e internacionales, con las metodologías y estrategias empleadas en la proyección de demanda de combustibles líquidos, tales como el ACPM/Diesel, Fuel Oil, GLP, Gasolinas (GMC y GME), Jet Fuel y Queroseno. Entre los

resultados más destacados del capítulo se identifica en la revisión de la literatura, que las variables macroeconómicas, tales como el PIB, la población, el precio de los combustibles y la demanda de combustibles alternativos son las variables más recurrentes encontrados en la literatura. A su vez, se ha observado que la utilización de modelos de econométricos de series de tiempo tales como ARCH, ARIMA, GARCH, SARIMA, VAR, VEC y Datos de Panel, junto con metodologías de aprendizaje de máquina como modelos basados en regresión y redes neuronales artificiales, han sido los modelos predominantes en la implementación de estrategias de pronóstico para los combustibles líquidos y GLP.

En el Capítulo 6 se describen las modelos que se han implementado para la proyección de la demanda de los combustibles líquidos: modelo de regresión lineal múltiple (MLR), modelo aditivo generalizado (GAM), regresión lineal con contracción (LASSO), regresión spline adaptativa multivariada (MARS), redes neuronales recurrentes de memoria de corto y largo plazo (LSTM), y una metodología ampliamente usada en la literatura para combinación de pronósticos basada en minimización de la varianza del error de los pronósticos. Adicionalmente, se expone la metodología Bootstrap usada para calcular los intervalos de confianza de las proyecciones. También, en este capítulo se realiza un pequeño resumen sobre las variables con las cuales se tiene disponibilidad para la implementación de los modelos estadísticos, en donde se presenta la fuente de consecución de las variables y la descripción de las mismas. Es de anotar que en su mayoría las variables empleadas para la obtención de los resultados fueron suministrados principalmente por la UPME; sin embargo, variables adicionales como población, temperatura, ONI, proporción de días laborales, entre otras fueron construcciones propias o tomadas de fuentes alternativas.

Finalmente, en el Capítulo 7 se presentan los ejercicios de proyección a 15 años para ACPM/Diésel, Fuel Oil, GLP, Gasolina Motor (GM) y Jet Fuel. Allí se han realizado múltiples procesos de ajuste y validación de modelos con el fin de encontrar la mezcla de variables explicativas y rezagos que proporcionen los mejores ajustes por fuera de muestra, cuantificando dicho ajuste a través de los MAPE de validación. Para cada combustible se han planteado diferentes casos de estudio que buscan combinar diferentes variables explicativas que puedan, desde un punto de vista económico, explicar la demanda del combustible. Además, para cada caso de estudio, se deriva una serie de escenarios en los que se combinaban rezagos para la variable dependiente y algunas variables explicativas, con el fin de localizar el modelo que ofrezca mejores ajustes en términos de validación junto con los hiperparámetros que permiten dicho ajuste. De esta manera se logra extraer las máximas cualidades de cada uno de los modelos en el ajuste de los combustibles. La implementación y optimización de hiperparámetros de los modelos, y las proyecciones mensuales de demandas con horizonte a 15 años fueron realizadas usando el lenguaje de programación Python.

Capítulo 1 Objetivo del proyecto

Objetivo general

Desarrollar una metodología para las proyecciones de demanda de combustibles líquidos y GLP.

Objetivos Específicos

(I) Evaluar integralmente las metodologías de proyección actualmente utilizadas en la UPME para energía eléctrica y gas natural.

(II) Desarrollar una propuesta para la proyección de demanda de combustibles líquidos y GLP, así como un primer ejercicio de estimación a 15 años.

(III) Socializar al personal de la UPME y al público externo la metodología de formulación y estimación de los modelos.

Capítulo 2 Revisión de literatura de proyecciones de demanda de energía eléctrica y gas natural

Tener un conocimiento cercano al comportamiento futuro de la demanda de energía eléctrica y gas natural, permite realizar ajustes continuos a los planes de expansión de los diferentes sectores en el país. Para lograrlo, es esencial tener en cuenta la posible evolución de variables demográficas, financieras, económicas, sociales, meteorológicas, entre otras. Esto significa que la precisión en la aplicación de cualquier metodología de pronóstico debe estar fundamentada en un conjunto de variables que sean de fácil obtención y/o predicción. Estas variables deben ser consideradas por diferentes estrategias matemáticas, i.e., modelos que aplican un procesamiento particular de las variables con el fin de tener la mejor proyección futura de las demandas de gas natural y electricidad.

Pronosticar la demanda de energía eléctrica ha sido tema de investigación y desarrollo por muchos años. Debido al surgimiento continuo de modelos matemáticos y de significativas mejoras tecnológicas encaminadas a un alto desempeño computacional, la temática de pronósticos es cada vez más vigente. Por tanto, este equipo de trabajo consideró que tener una revisión de literatura siempre será útil para comprender el contexto actual de la problemática. Más aún cuando hay eventualidades como la pandemia del COVID-19, que ha generado grandes cambios en los patrones de las variables, de todo índole, que describen a las sociedades.

Por tanto, en este capítulo se presenta una revisión detallada de literatura científica dedicada a las metodologías y modelos de proyección de demandas de energía eléctrica y gas natural. Para esto, se han revisado más de 25 publicaciones en revistas indexadas y reportes técnicos en los que se evidencian nuevos modelos y metodologías, y diferentes estrategias de la manipulación de las variables importantes del problema.

En particular, el análisis ha sido principalmente sobre aplicaciones de proyecciones de largo plazo, metodologías y modelos recientes; aplicaciones sobre algunos mercados internacionales, e impactos de la pandemia causada por el COVID-19. Con el fin de ilustrar de una manera organizada este informe, esta revisión se ha dividido en dos áreas de demandas: energía eléctrica y gas natural.

2.1 Energía Eléctrica

La literatura relacionada con la predicción de la demanda de energía, en términos generales, se basa en metodologías probadas y que han dado resultados útiles, tales como VAR, VEC, regresión lineal múltiple, descomposición de Fourier, redes neuronales, machine learning, AR/-

MA, entre otros. Para el caso particular, se busca realizar un modelo que considere no sólo las variables que actualmente se han utilizado, sino que además integre otras variables reportadas en la literatura especializada, de tal forma que se aproxime la solución del modelo deseado y se compare con la efectividad del modelo actual.

Esta revisión de literatura inicia explorando investigaciones realizadas a nivel nacional. El pronóstico oficial de la demanda de energía eléctrica a largo plazo en Colombia es realizado por la UPME. En el corto plazo, con horizontes de pocas semanas, la demanda de energía eléctrica con frecuencia horaria es pronosticada también por XM, entidad que opera el mercado de energía eléctrica en Colombia, XM (XM, 2021). Sin embargo, los fines con los cuales XM pronostica la demanda de energía eléctrica son de carácter operativo; y que últimas busca garantizar que la generación del sistema, en cada instante, satisfaga la demanda.

La metodología empleada por la UPME es un referente para otros estudios a nivel nacional como (Pérez, 2020); (Cervantes, 2018); (Medina y Marulanda, 2017); (Vera, 2016); (Franco et al., 2008); (Medina et al., 2011); (Rey, 2018); (Grimaldo et al., 2016). En las propuestas realizadas, dado el componente de autocorrelación significativo en la series de tiempo de demanda de energía eléctrica, se utilizan los históricos de la demanda de energía en (Vera, 2016); (Franco et al., 2008); (Medina et al., 2011), siendo esta última la que mejor respuesta tiene, comparativamente con los datos medidos en el SIN en el corto plazo, teniendo en cuenta que sólo utiliza tres semanas para realizar el pronóstico.

Por otra parte, en la propuesta de (Grimaldo et al., 2016) se realiza un pronóstico por sector económico en el país, complementando la serie de la demanda de energía con el PIB sectorial y el total, para luego computar los resultados como la suma de los pronósticos individuales; mientras que (Medina y Marulanda, 2017) desarrollaron una propuesta en la que no sólo se incluye el PIB, sino también el crecimiento poblacional que arroja la UPME. Esta propuesta se compara con los datos de la UPME, para los siguientes 15 años, dando resultados cercanos (5 %) a los presentados en el plan de expansión.

Asimismo, (Cervantes, 2018); (Rey, 2018) proponen incluir como variables de entrada en el modelo la temperatura de las regiones (Caribe y Bogotá-región, respectivamente), con el fin de obtener una aproximación al uso de electrodomésticos para climatización. En ambos casos los errores son inferiores al 3 %, aunque se hace énfasis en que fenómenos como El Niño que se dio en el país en el año 2015, pueden hacer que los pronósticos no tengan un buen desempeño.

Los trabajos presentados por Pérez (2020); Rey (2018), proponen la reducción de las variables y la inclusión de el gas, como complemento al uso de la energía eléctrica. En el primero de los casos, los pronósticos permiten realizar aproximaciones a la realidad en el corto plazo, mientras que en el segundo caso se plantea el consumo del gas como una variable exógena que justificaría la disminución del consumo eléctrico.

El trabajo de Jiménez et al. (2019) presenta una metodología basada en redes neuronales artificiales para el pronóstico de la demanda de la región Caribe en Colombia. Los autores observaron que existe baja correlación de la demanda mensual de energía con la temperatura,

humedad y velocidad del viento entre 2005 y 2016. Sin embargo, el fenómeno del Niño incrementó la demanda en el segundo semestre de 2015, y el fenómeno de la Niña la disminuyó durante el segundo semestre de 2016. La metodología que plantean los autores realiza una preselección de variables macroeconómicas a través de análisis factorial que emplea 29 variables donde se incluían variables del mercado como generación hidráulica, térmica, precio de bolsa, entre otros precios de commodities. Al final, la red neuronal es entrenada con 9 variables de entrada como histórico de demanda de energía eléctrica (rezagos de 4 meses), proporción de demanda de días laborales en el mes, proporción de demanda de domingos y festivos, indicador del mes a pronosticar, número de días del mes, número de días festivos y domingos del mes, crecimiento anual del precio del oro del año a pronosticar, crecimiento anual del PIB del año a pronosticar, crecimiento anual de la población del año a pronosticar, crecimiento anual del precio del crudo del año a pronosticar. Adicionalmente, para incorporar efectos y/o correcciones al pronóstico por escenarios meteorológicos, se empleó otra RN. Según los autores, el pronóstico de la demanda de energía para 2017, disminuye del 2.2 % al 1.6 % de MAPE cuando es ajustado por la segunda red neuronal (fenómeno del Niño y la Niña). Este tipo de modelos presenta un desempeño bastante satisfactorio. Sin embargo, el MAPE es calculado para pronosticar un año donde las proyecciones de las variables exógenas, probablemente, tienen menor incertidumbre.

Recientemente, (Jimenez et al., 2021) presentan una metodología para definir de manera satisfactoria los hiperparámetros de una red neuronal para pronosticar la demanda de energía horaria. La metodología de pronóstico es analizada sobre la serie de tiempo de la demanda del Atlántico. La metodología planteada usa variables exógenas como el tipo de día y variables meteorológicas. Sin embargo, los autores no especifican qué variables meteorológicas, en particular, emplean. Los resultados presentados por los autores son comparados contra los pronosticados por el operador de red local; y como conclusión, muestran que la metodología propuesta llega a mejores resultados.

En resumen, las variables observadas en el contexto colombiano son el PIB Medina y Marulanda (2017); Grimaldo et al. (2016); Rey (2018); Cervantes (2018); Jiménez et al. (2019), la población (Grimaldo et al., 2016); (Medina y Marulanda, 2017); (Rey, 2018); (Jiménez et al., 2019), la temperatura (Rey, 2018); (Cervantes, 2018), el fenómeno del Niño (Cervantes, 2018), el precio del gas (Rey, 2018), el precio de la energía eléctrica (Rey, 2018), efecto calendario y laboral (Jiménez et al., 2019), crecimiento del precio del oro (Jiménez et al., 2019) y crecimiento del precio del crudo (Jiménez et al., 2019). Los trabajos de (Franco et al., 2008); (Medina et al., 2011); (Vera, 2016) sólo empleen la historia de la demanda de energía eléctrica.

Es importante resaltar que fenómenos meteorológicos como el del Niño no han sido comúnmente empleados en las metodologías de pronóstico de demanda de energía eléctrica en el país. El efecto meteorológico en la demanda se logra percibir cuando el pronóstico se realiza a nivel desagregado donde se pueda analizar de manera individual una región (o ciudad) con condiciones meteorológicas similares (como por ejemplo el Valle del Cauca, o Barracabermeja, o Barranquilla, entre otros); y no un país como convencionalmente se realiza. Al analizar el país, en donde hay múltiples regiones/ciudades con condiciones meteorológicas diversas, los modelos

no necesariamente capturan el impacto meteorológico sobre la demanda.

Sin embargo, un escenario en el que sí se podría presentar un impacto significativo en la demanda a nivel país, es cuando la generación solar distribuida se instale masivamente. En este caso, la radiación solar y la temperatura ambiente serían determinantes en la producción solar distribuida, y por ende en la demanda neta a satisfacer por el sistema.

Luego de la aparición del COVID-19, la demanda de energía eléctrica ha sufrido variaciones según lo pronosticado; tal es el caso del pronóstico a corto plazo en el estado de Ontario en Canadá realizado por Abu-Rayash y Dincer (2020), en donde se muestra que la demanda de energía pasó de 12–17 GWh en 2019, a 10–14 GWh en 2020, lo cual representa una disminución en la demanda del 14 %; traducida en una reducción de 40000 ton de CO_2e. Cabe señalar que la variación en la demanda también tuvo repercusiones en los días de consumo máximo y mínimo, pues durante la pandemia los días con mayor demanda fueron lunes y martes, mientras que los de menor demanda fueron jueves y viernes, un comportamiento completamente contrario a lo ocurrido en las semanas pre-pandemia.

Un análisis similar fue presentado por Li et al. (2021), en donde los autores incluyeron dentro de los pronósticos de demanda de energía a corto plazo, el número de infecciones diarias, el número de muertes diarias, y un indicador que mide el nivel de restricciones impuestas fruto de la pandemia, con el fin de evitar desviaciones. Por su parte, Norouzi et al. (2020) analizó el impacto de la pandemia en la demanda de electricidad y petróleo en China, usando períodos mensuales a través de métodos autorregresivos (AR) y redes neuronales artificiales (RNN); las variables utilizadas fueron PIB, demanda de petróleo, índice de severidad de la pandemia de la OMS[1], acumulado de personas infectadas por coronavirus, purchasing managers index (PMI) de manufactura, ingresos por exportaciones, inversión extranjera, productividad industrial, índice de la bolsa. Es de anotar que para su estudio, Norouzi et al. (2020) construyen los escenarios de pronóstico de demanda asumiendo ciertas tendencias en algunas de las variables de entrada, a saber, infectados, inversión extranjera, índice de la bolsa y exportaciones.

En esta misma línea de predicción a corto plazo, Ivanin y Direktor (2018) plantean un modelo con redes neuronales para realizar el pronóstico de la demanda diaria, en redes con consumidores conectados en zonas apartadas, pero dependientes del sistema conectado; en otras palabras, zonas residenciales, comerciales y/o industriales que sólo tienen una línea para el suministro de energía eléctrica. Para ello, los datos de entrenamiento usados fueron la población, la temperatura, las características de las cargas eléctricas (industriales, residenciales, comerciales), la hora del día, la temporada del año y la intensidad lumínica percibida del sol.

Por su parte, en el trabajo desarrollado por la IEA (2020) se establece que la demanda de energía global ha caído aproximadamente un 6 %, siendo ésta ésta la cifra más alta que se ha registrado en los últimos 70 años. De hecho, esta estimación es aproximadamente 7 veces mayor que la presentada durante la crisis económica de 2008 (IEA, 2020). Esto sugiere que los aspectos socio-culturales se deben tener en cuenta a la hora de realizar proyecciones. Por

[1] medido por el número de muertes diarias y el peor escenario de muertes diarias

ejemplo, los pronósticos hechos por Jiang et al. (2021) consideran diferentes escenarios según el manejo que se le dio a la pandemia, en donde los autores ilustran que dependiendo del nivel de cuarentena implementado se dieron reducciones en la demanda, en donde bajo restricciones limitadas, la demanda de energía semanal cayó aproximadamente 9 %, en aislamiento parcial aproximadamente 17 %, y bajo aislamiento total aproximadamente 24 %. A pesar de esto, hay sectores en los que la demanda de energía ha crecido fruto del aumento del teletrabajo y telemedicina durante el confinamiento.

En Australia, la demanda residencial ha aumentado un 14 %, al igual que en algunos sectores de la economía debido a la producción masiva de tapabocas, antibacteriales, alcoholes, entre otros. La producción masiva de vacunas (aprox. 1000 millones), su transporte (15000 vuelos) y enfriamiento (15 millones de cajas de enfriamiento) también supone un incremento que no es despreciable en la demanda de energía (DHL, 2020).

Ahora bien, la evaluación de los impactos de la pandemia por COVID-19 en el sector económico, financiero, educativo, salud, industrial, financiero, energía eléctrica, hidrocarburos, desempleo, medio ambiente en la India, se analizan en Kumar et al. (2021). Para predecir los impactos de la pandemia, los autores emplearon el modelo de suavización exponencial de Holt-Winter, una regresión lineal y una regresión lineal con estacionalidad, empleando como variables explicativas e número de pacientes COVID-19, número de casos encontrados diariamente (tasa de pacientes COVID-19), número de muertes por COVID-19, PIB y cifras de desempleo.

Por otra parte, previo a la COVID-19, los modelos para el pronóstico de la demanda eléctrica se centran en aplicaciones de regresión lineal, como por ejemplo Liu et al. (2017), He et al. (2017), Trotter et al. (2016), Pessanha y Leon (2015). En estos casos, se usan variables econométricas tales como PIB, crecimiento poblacional, histórico de la demanda eléctrica, entre otros. Además, en He et al. (2017) se propone realizar un modelo individual que contemple el crecimiento de la demanda en cada subsector de la industria, por ejemplo, industria alimenticia, agrícola, química, transporte, manufactura, entre otros. Según la proyección de crecimiento industrial de cada uno de estos subsectores se propone una correlación y un modelo que permita determinar su crecimiento y, como producto final, se integre el resultado mediante una sumatoria de crecimientos, considerando algunos factores de simultaneidad que tengan en cuenta que no toda el crecimiento será simultáneo. Para estos factores se consideran históricos de crecimiento de cada subsector.

Desde el punto de vista del cambio climático, Perwez et al. (2015), Suhono (2015) y Mirjat et al. (2018) proponen el uso del programa LEAP (Long range Energy Alternatives Planning[2]), que es un software desarrollado por el Instituto Ambiental de Estocolmo, el cual permite realizar un pronostico energético, a mediano y largo plazo, de la demanda de energía mediante la proyección de variables econométricas tales como PIB, crecimiento poblacional, factores de eficiencia de energía, pero además se considera la cantidad de emisiones de CO_2 que se podrían emitir en cada uno de los escenarios, según los tipos de energéticos utilizados para la generación

[2] https://openei.org/wiki/Long-Range_Energy_Alternatives_Planning_System_(LEAP)

de energía que satisfaga la demanda.

Cabe señalar que las proyecciones de Trotter et al. (2016) se realizan en escenarios a muy largo plazo, en específico 84 años, con una periodicidad diaria, la cual se complementa con ajustes anuales que permiten generar las políticas energéticas nacionales que cubran el crecimiento propuesto. Por su parte, Perwez et al. (2015), Suhono (2015), Sakunthala et al. (2018) y Mirjat et al. (2018), han propuesto iniciativas de análisis que combinan variables econométricas de país, además de las meteorológicas y de la información que se tiene sobre el crecimiento de la demanda, para proponer predicciones a 25, 29, 30 y 35 años, en algunas de las cuales se han adicionado técnicas de procesamiento de la información como redes neuronales, recocido simulado, algoritmos genéticos, entre otros. En todas ellas se plantean ajustes anuales a la predicción, buscando tener en cuenta el cambio climático, las políticas de regulación actuales y las emergentes.

Vale la pena incluir dentro de este análisis de literatura lo referenciado por el IPCC (*Intergovernmental Panel on Climate Change*) con respecto al incremento de temperatura promedio en la tierra, del orden de 1.5ºC, producidos, en buena medida, por emisiones de CO_2 a la atmósfera (IPCC, 2021). Este análisis, que puede complementarse con el presentado en (IEA, 2021b), lanza un reto para la planeación de la generación, teniendo en cuenta la reducción de las emisiones que se busca. Sin embargo, según lo reportado en la literatura consultada, el incremento en la temperatura a causa del cambio climático sólo será evidente en aquellas regiones donde culturalmente se busca disminuir los efectos del calentamiento con el energético que tienen disponible. Es decir, para la región de la costa atlántica del país, (Cervantes, 2018) propone tener en cuenta la temperatura en esta región cuando se pronostique su demanda, ya que este incremento, incluso sin que se haya presentado el fenómeno de El Niño, hará que la demanda de energía eléctrica aumente por el consumo que tienen los aires acondicionados, neveras, enfriadores, entre otros. Un fenómeno similar ocurre en la región pacífica, donde los análisis reportados por algunos consultores coinciden en que se presente un aumento en el consumo de energía cuando hay incrementos de temperatura, debido, entre otros, al aumento en el uso de los sistemas de riego para los cultivos de caña. Por otra parte, para el caso del centro del país, (Rey, 2018) reporta que no se presenta un incremento en el consumo de energía eléctrica cuando se dan altas temperaturas en la región, pero sí un aumento en el consumo de gas para calefacción. De igual forma, (Velásquez et al., 2009); (Jiménez et al., 2017); (Jimenez et al., 2021); (Mariño et al., 2021), proponen modelos para el caso de Colombia en el que se tienen en cuenta las variables meteorológicas y días de la semana obteniendo errores que están alrededor del 3 %, para los análisis hechos en la zona atlántica del país. En otras palabras, al igual que (Cervantes, 2018), se evidencia la necesidad de variables exógenas para obtener mejores resultados en los pronósticos. Por otro lado, a pesar de que desde 2009 con los acuerdos de Copenhague donde se aprobó la creación de estrategias que permitieran la disminución de CO_2, responsables en buena medida del calentamiento global (IPCC, 2021), las emisiones no han disminuido lo esperado. En (Mirjat et al., 2018) se ha planteado usar las emisiones de CO_2 para pronosticar la demanda de energía eléctrica. Sin embargo, en Colombia no se ha observado

tal análisis. Además, no creemos que sea de alta utilidad puesto que, por el contrario, cambios en la demanda (y generación) conllevan a una cantidad de emisiones del sector eléctrico en particular.

Para 2015 las Naciones Unidas, en París [3], nuevamente firmaron acuerdos que buscan el mismo objetivo de reducción de emisiones. Esto ha hecho que se generen campañas de disminución en el uso de combustibles fósiles que, de alcanzarse, llevarían a disminución en el consumo de energía En caso de no alcanzar las metas propuestas, el calentamiento global continuaría, y la demanda de energía eléctrica en el país tendería a aumentar en aquellas regiones que serán mayormente golpeadas por altas temperaturas.

Ahora bien, con el fin de presentar de forma resumida la información presentada en la revisión de la literatura, se presenta la Cuadro 2.1 con las variables y modelos empleados en la literatura revisada para el pronóstico de la demanda de energía eléctrica.

Autores	**Frecuencia**	**Modelo**	**Variables**
Franco et al. (2008)	A, M	SEMU	Demanda nacional mensual
Medina et al. (2011)	S	NN	Demanda de energía, con tres semanas de rezago
Pessanha y Leon (2015)	A	MRL	Consumo promedio de energía, tasa de electrificación y el número de abonados.
Perwez et al. (2015)	A	LEAP	Crecimiento de las industrias, PIB, crecimiento poblacional.
Suhono (2015)	A	LEAP	Crecimiento de la población, consumo de las casas, tasa de electrificación e intensidad de la demanda.
Trotter et al. (2016)	D	MRL	Crecimiento poblacional. Variables climatológicas. Crecimiento industrial.
Grimaldo et al. (2016)	A	MRL	PIB sectorial, la población y el PIB total
Vera (2016)	M	MRL	Demanda nacional mensual
He et al. (2017)	D	FSQRM	Consumo de energía según el tipo de industria.
Liu et al. (2017)	A	MRL	Producto interno bruto (PIB), crecimiento de la población, las importaciones y exportaciones.

[3] https://www.un.org/es/content/summits2021/

Medina y Marulanda (2017)	A,M	MARS	Demanda de energía residencial, PIB, Crecimiento poblacional
Mirjat et al. (2018)	A	LEAP	Tecnologías disponibles para generación, disponibilidad y costo del carbón, políticas de eficiencia energética, recursos potenciales, parámetros tecnoeconómicos de las plantas de generación y emisiones de CO2.
Sakunthala et al. (2018)	A	MRL	Indicadores económicos del estado, crecimiento de la población, PIB, demanda anual máxima y el ingreso per cápita.
Rey (2018)	M	MRL, MRLDL	Datos históricos de temperatura superficial, Producto Interno Bruto, ingresos per cápita, suscriptores, consumo y precio de la de energía eléctrica y precio del servicio de gas natural
Cervantes (2018)	A,M	NN	PIB, temperatura media histórica de la región Caribe, El Niño
Ivanin y Direktor (2018)	D	NN	Demanda de energía, población, temperatura
Jiménez et al. (2019)	M	NN	Histórico de demanda de energía eléctrica (rezagos de 4 meses), proporción de demanda de días laborales en el mes, proporción de demanda de domingos y festivos, indicador del mes a pronosticar, número de días del mes, número de días festivos y domingos del mes, precio del oro, PIB, crecimiento poblacional, crecimiento del precio del crudo. Fenómeno del Niño y de la Niña.

Angelopoulos et al. (2019)	A	ORA	Producto interno bruto, meteorología, procesos de eficiencia energética, tasa de desempleo, población, Heating Degree Days (HDD), Cooling Degree Days (CDD), precio de la electricidad y del gas natural, light fuel oil price y un indicador de eficiencia energética.
Jiang et al. (2020)	D	AFD	Histórico de demanda.
Pérez (2020)	M	SARIMAX	IPC, históricos de demanda de energía
Norouzi et al. (2020)	M	AR, ANN	PIB,demanda de petroleo,indice de severidad de la pandemia,acumulado de personas infectadas
Abu-Rayash y Dincer (2020)	H	N/A	Demanda de energía
Mariño et al. (2021)	M	SARIMA, Holt Winters	Demanda de energía
Li et al. (2021)	D	VAR	Histórico de demanda, discriminando los días hábiles y no hábiles.
Özbay y Dalcali (2021)	D	LSTM, NARX-ANN	Día, semana, mes, temperatura, medida de precauciones ante una pandemia
Jiang et al. (2021)	S	N/A	Demanda de energía, escenarios COVID-19
Kumar et al. (2021)	D	ESM, MRL	Demanda de energía, número de pacientes COVID-19,casos encontrados diariamente, nÚmero de muertes por COVID-19, PIB y desempleo

H: Horaria **D**: Diario; **M**: Mensual; **T**: Trimestral; **A**: Anual

Cuadro 2.1: Variables explicativas y modelos empleados en la literatura para el pronóstico de la demanda de Energía Eléctrica.

2.2 Gas Natural

En el caso del gas natural, el planteamiento de modelos de pronóstico de la demanda de gas natural es fundamental para saber cuanta es la cantidad del energético que debe importarse para suplir la demanda interna mensual, poder anticiparse a aumentos o caídas a causa de efectos estacionales, para evitar de esta manera la escasez de suministros.

Por esta razón, la revisión de literatura en esa sección se enfoca en los principales trabajos en proyecciones de la demanda de gas natural, tanto para el caso de Colombia como para el caso internacional, mencionando tanto las principales metodologías de proyección como las principales variables explicativas que se utilizan para pronosticar la demanda de gas natural.

Según Li et al. (2021), la teoría de modelos predictivos para el consumo de gas natural se ha investigado durante casi 70 años, en donde al revisar la literatura se encuentra que gracias al desarrollo que ha tenido de la informática y la tecnología de la inteligencia artificial, hoy día los modelos de corto plazo (hora, día o semana) son los más populares en las investigaciones.

En Li et al. (2021) también se señala que en el caso de las proyecciones a largo plazo, la demanda de gas natural puede verse afectada por la producción, la población y las variables económicas; a mediano plazo por variables económicas y temperatura; y a corto plazo principalmente por las condiciones meteorológicas y el efecto calendario. En cuanto a modelos, Li et al. (2021) sugiere que los modelos de series de tiempo como SARIMAX y de regresión son los mejores para la proyección a largo plazo. Para el mediano plazo los modelos de aprendizaje estadístico (GAM, BOOSTING, SVM, entre otros) son los preferidos. Y para el corto plazo, los modelos basados en inteligencia artificial (como RNN y LSTM) han presentado el mejor desempeño.

Con el objetivo de brindar una presentación más ordenada para los trabajos que se han realizado, se presenta inicialmente aquellos trabajos en los cuales se plantean modelos que dependen solamente de la variable de demanda gas natural para realizar los pronósticos, luego se listan los trabajos que usan otras variables explicativas para pronosticar la demanda de gas natural, y finalmente se presentan aquellos trabajos que hayan tenido el efecto de la pandemia dentro de sus estimaciones.

Un trabajo que emplea solamente la demanda de gas natural como variable para realizar proyecciones de la demanda anual en España es el planteado por Gutiérrez et al. (2005), el cual presenta un proceso de difusión de innovación tipo Gompertz como un modelo de crecimiento estocástico (SGIDP), empleando para el proceso de estimación más de 24 años de información histórica reportada para la variable de gas natural. Entre las conclusiones obtenidas en su trabajo, Gutiérrez et al. (2005) muestran que el modelo SGIDP tiene mejor desempeño al hacer proyecciones de corto y mediano plazo, cuando se compara con un modelo de innovación logística estocástica y un modelo lognormal de crecimiento estocástico basado en proceso de innovación sin difusión.

Una aplicación realizada en Irán para pronosticar la demanda anual de gas natural residen-

cial y comercial a corto plazo es presentada en Forouzanfar et al. (2010), en donde los autores emplean un modelo de optimización de programación no lineal (NLP) y un algoritmo genético (GA) para optimizar los parámetros asociados a una curva logística característica. Según los autores, los dos modelos de optimización llegan a parámetros similares al estimar la demanda que se tendrá para diferentes estaciones del año cuando se usa la curva logística, además de que ambos ofrecen pronósticos cercanos de la demanda de gas natural, al compararlos con los valores reales observados.

En el caso de Turquía, la demanda de gas natural ha sido pronosticada mensualmente hasta 2025 por Huseyin (2021), debido a que Turquía es un país que cuenta con pocas fuentes de producción de gas natural y, en consecuencia, debe importar cerca del 99 % de la demanda de dicho energético que se consume en el país. En su trabajo, Huseyin (2021) plantean un modelo de GREY estacional SGM(1,1), un modelo SARIMA, y además desarrollan un nuevo enfoque para el modelo SGM(1,1), llamado modelo de GREY estacional ajustado ASGM(1,1)[4]. Para el caso del modelo SARIMA, los autores encuentran que el mejor modelo SARIMA que se ajusta a la demanda de gas natural en Turquía es un SARIMA(1,1,0)(1,1,0)^{1}2. Por su parte, para el modelo SGM(1,1) se encuentra que para una estacionalidad de $s = 12$ meses, el coeficiente de ajuste que mejor se adapta a los datos es de $\beta = 0.01$; mientras que, para el modelo ASGM(1,1), el coeficiente de ajuste y el número de rezagos óptimo que mejor se adapta a los datos son respectivamente, $\beta = 0.01$ y $k = 9$. A la hora del proceso de pronóstico y verificación de los modelos, los autores concluyen que el modelo que presenta el mejor desempeño predictivo medido a través del MAE, RMSE, y MAPE, es el ASGM(1,1) con $\beta = 0.01$ y $k = 9$, seguido por el SGM(1,1) con $\beta = 0.01$, y el modelo SARIMA(1,1,0)(1,1,0)^{1}2.

En el caso de los modelos que emplean variables explicativas para realizar las predicciones, se tiene el trabajo propuesto por Liu y Lin (1991), en donde los autores emplea un modelo SARIMAX a datos mensuales con intervenciones, para pronosticar la demanda de gas natural empleando como variable de explicativa la temperatura.

La temperatura, al igual que el número de clientes, consumo industrial de gas y la variable diferencia de la temperatura real con respecto a la esperada, también fueron empleadas en Suykens et al. (1996). Se destaca el uso del precio de la gasolina en el modelo, ya que la gasolina es un sustituto del gas no solo para transporte, sino también para generación eléctrica. Los autores, mediante simulación y no sobre hechos reales, señalan que las redes neuronales feedforward superan (en desempeño predictivo) a los modelos que solo usan la temperatura como variable predictora, como es el caso de Electrabel SA de Bélgica, que usa una función compuesta de la variable temperatura para hacer predicciones de la demanda de gas.

El trabajo de Huntington (2007) emplea un modelo de regresión autorregresivo para estimar la función demanda de gas en el sector industrial, utilizando variables como el consumo de gas natural en el sector industrial, precio de la gasolina, consumo de energía en el sector de

[4]Los autores señalan que la ventaja que posee el modelo ASGM(1,1) sobre el modelo SGM(1,1) se debe a que este último logra capturar los patrones estacionales de las series que no capturados con el modelo SGM(1,1).

industrial, capacidad utilizada de los sectores industrial y de la manufactura, las variables meteorológicas HDD y CDD, precio del gas, precio del petróleo y precio del carbón.

Por su parte, Kamrani (2010) usa en sus modelos de pronóstico de la demanda de gas natural variables como: temperatura, población, ingreso nacional, producto nacional bruto, índice de precios al consumidor y demanda de gas un periodo atrás, emplea modelos de regresión autorregresivos, con un solo rezago de tiempo para la variable dependiente, AR(1), y llega a resultados predictivos muy acertados para una frecuencia anual, debido a que éstos registran un MAPE bastante bajo.

Özmen et al. (2018) argumenta que los modelos MARS (Multivariate Adaptative Regression Splines) y CMARS (Conic Multivariate Adaptative Regression Splines) proporcionan resultados importantes para el pronóstico diario de consumo de gas natural residencial para periodos de un año. Los autores realizan una comparación de estos modelos con las redes neuronales y con la regresión lineal múltiple. En estos modelos se obtuvieron un MAPE de 5.7 % y 7.9 %, respectivamente. El MAPE de los modelos MARS y CMARS se situó en 4.8 %. De modo que, los modelos MARS y CMARS están en el centro del análisis predictivo debido a que su estructura semi-paramétrica permite que el modelo tenga cierto grado de interpretabilidad y sea útil para realizar inferencias sobre los parámetros del modelo y, además, permitan realizar previsiones sobre el comportamiento futuro de la demanda de gas para periodos de corto y mediano plazo, a diferencia de los modelos basados en redes neuronales que terminan siendo completamente una caja negra.

Continuando con su linea de trabajos, en Özmen (2021) los autores plantean la estimación de un modelo MARS para para pronosticar la demanda de gas natural a partir de funciones de base no lineal[5] y encuentran que éste modelo supera en desempeño predictivo la regresión LASSO y al modelo de regresión lineal múltiple tanto a corto como a largo plazo, en donde, a corto plazo, el modelo MARS presenta un error promedio de pronóstico del 4.8 %; pero a medida que aumenta el horizonte de tiempo del pronóstico, se encuentran con que su desempeño predictivo disminuye situándose en promedio en 13.4 %.

En el trabajo propuesto por Anđelković y Bajatović (2020), se consideran dos escenarios para el pronóstico de la demanda de gas natural horario, a saber, el primero llamado régimen de calefacción y el segundo llamado régimen sin calefacción, y para cada uno de ellos se plantea un modelo basado en un algoritmo de sistema de inferencia Neuro-Difusa Adaptativa (ANFIS), y emplea como variables del modelo el tipo de día (día de trabajo y de no trabajo), valores mínimos de calóricos del gas natural[6] y variables meteorológicas tales como: temperatura del aire, radiación solar global y humedad relativa. Una vez realizado el proceso de estimación,

[5]Es de anotar que el modelo MARS estima parámetros funcionales para cada una de las variables explicativas que se incluyan, de tal forma que para un gran volumen de covariables la estimación del modelo puede resultar computacionalmente compleja.

[6]El valor mínimo calórico o valor mínimo de calefacción, es una medida de energía térmica que se produce en la combustión de combustible, medida en unidades de energía, unidades de masa o volumen de una sustancia (Dincer, 2018)

pronóstico y validación de los modelos, los autores concluyen que al integrar los pronósticos de las variables meteorológicas para la estimación del modelo ANFIS, se logra un mejor desempeño predictivo que cuando dichas variables no son consideradas.

Entre los trabajos más recientes, se presentan el desarrollado por Gaweł y Paliński (2021) en donde los autores emplean para el pronóstico de la demanda de gas natural a largo plazo, un método de pronóstico análogo espacio-temporal en combinación con un enfoque de árbol de decisión difuso (FAM). Entre las variables exógenas se consideran la demanda anual de gas natural total y la relación entre el consumo anual de energía y el PIB, para 79 países con un horizonte de tiempo entre 1965 y 2019. Adicionalmente, tienen en cuenta datos macroeconómicos para cada país, tales como: PIB per cápita, población, PIB industrial, emisiones de CO_2 como una medida de adaptación del cambio climático, índice de días de calor para temperaturas inferiores a los 15^oC y participación del consumo de gas natural en el consumo anual total de energía primaria. Una vez ajustado el modelo FAM, Gaweł y Paliński (2021) deciden comparan el desempeño predictivo del modelo respecto a otras metodologías[7] y concluyen que su modelo ofrece resultados prometedores para realizar pronósticos a largo plazo, pues la precisión de sus pronósticos es mayor que la obtenida con las demás metodologías.

En Li et al. (2021) los autores presentan una revisión de la historia de modelos de demanda de gas hasta la fecha. Allí se comenta que junto con la capacidad de cómputo, los modelos han evolucionado desde los SARIMAX para largo plazo hasta los modelos de aprendizaje estadístico y aprendizaje profundo, para mediano y corto plazo. Adicionalmente, los autores clasifican los modelos según el horizonte. En modelos de corto plazo, los principales factores son los meteorológicos, tipo de día y frecuencia del dato. En los de mediano plazo las variables explicativas son heating degree days (HDD), cooling degree days (CDD), precio del gas, combustóleo residual, precio del gas residual, ingreso per-cápita, temperatura, diferencia entre temperatura real y esperada, precio de la gasolina, número de consumidores, consumo de la industria, consumo histórico, condiciones meteorológicas, día de la semana, población, PIB, índice estacional. Y en los de largo plazo, las principales variables son producción, población y variables económicas.

Yucesan et al. (2021) realizan un sólido marco comparativo sobre la aplicación de diferentes métodos en la previsión del consumo diario de gas natural, en el cual se evidencia la gran potencia que ofrecen los métodos híbridos para predecir el comportamiento futuro de la demanda de gas natural. Los modelos ARIMA son excelentes para capturar patrones lineales, mientras que los modelos de Redes Neuronales Artificiales (ANN) capturan de una forma más apropiada los patrones no lineales, de forma tal que la precisión de las estimaciones puede incrementar al utilizar conjuntamente estos métodos. Los mejores modelos fueron el ARIMAX-ANN (con 9 neuronas) y el SARIMAX-ANN(con 7 neuronas) los cuales consiguieron un MAPE del 0.5 % y del 0.35 %, respectivamente. Finalmente, los autores recomiendan que para mejorar los resultados futuras investigaciones agregar variables explicativas como la temperatura, la velocidad

[7] las metodologías que usan los autores para realizar la comparación son: ARIMA, enfoque Naïve, NARR, regresión lineal, tendencia lineal individual para cada país y modelos híbridos de árbol de decisión nítido, nítido con término de error, y difuso

del viento, el índice de producción industrial, la sustitución del gas natural, el precio unitario del gas natural y la disponibilidad de nuevas reservas.

En el contexto de la UPME, en el reporte UPME (2021) se presenta un análisis descriptivo sobre el comportamiento del gas en los últimos cinco años, los autores señalan que para el consumo de gas entre enero y marzo del 2020 la demanda de gas aumentó en el sector termoeléctrico (debido a restricciones eléctricas). Pero la demanda de gas disminuyó en los sectores terciario, petrolero, transporte e industria. Esto último, según estos los autores, llevó a que en general, entre enero y marzo del 2020, el consumo de gas disminuyera en la mayoría de los sectores y en la mayoría de las regiones, sobre todo en las más grandes. Sin embargo, a partir del mes de abril del 2020 se presentó un aumento general hasta el mes de marzo del 2021 y, a partir de allí, una leve disminución debido posiblemente al paro nacional. De acuerdo con los modelos utilizados por estos autores, en este trabajo de referencia para nuestros objetivos, la demanda proyectada para todos los sectores agregados prevé un crecimiento positivo para el periodo posterior a la pandemia de COVID-19, que es inferior a la proyectada en el periodo anterior al COVID-19. Señalan en este trabajo que para el periodo 2021–2035 el sector industrial, según sus proyecciones, tendrá un consumo estable en los sectores residencial, transporte, terciario, compresores y petroquímico creciente, mientras que el sector petrolero presentará una proyección que tiende a la baja en promedio y el termoeléctrico eléctrico también tiende a la baja en promedio. El trabajo de (UPME, 2021), se basa en la combinación de pronósticos de varios modelos: VAR con todas las variables endógenas, VAR con variables exógenas y VEC. Específicamente proyectan todos los modelos en el periodo 2021 a 2135, y aplican métodos de combinación de pronósticos para sacar sus proyecciones finales.

Además de la UPME, la empresa Concentra realiza proyecciones de demanda de gas natural en diferentes sectores productivos en Colombia. En Concentra (2021) se hacen proyecciones de la demanda de gas utilizando modelos de vectores autorregresivos (VAR) y se emplean variables como la demanda de gas natural por parte del sector residencial, precio del GLP en tanque de 40 lbs y tarifa de gas natural residencial, para estratos 3 y 4, mientras que, para el sector productivo emplea, el IPC y el IPP, entre otras dependiendo del sector demandante de gas natural.

En el reporte de Concentra (2021) se realizan pronósticos utilizando modelos de vectores autorregresivos (VAR). Utilizan variables como la demanda de gas natural, por parte del sector residencial; precio del GLP, en tanque de 40 lbs; tarifa de gas natural residencial, para estratos 3 y 4, en distintos sectores productivos; el IPC y el IPP, entre otras dependiendo del sector demandante de gas natural.

En cuanto a la literatura internacional de modelos de demanda de gas que tengan en cuenta el efecto COVID-19, solo se han encontrado a la fecha modelos de corto plazo, entre los que podemos mencionar el Norouzi et al. (2020) , Özbay y Dalcali (2021) que estima dos modelos de redes neuronales diferentes (LSTM y NARX-ANN) para pronosticar el consumo de gas 30 días adelante, en este modelo se usan variables de efecto calendario, la temperatura

y un índice al cual el autor hace referencia como precauciones pandémicas para capturar el efecto de la pandemia. Sin embargo, estos autores no son muy claros en especificar que es tal índice. Por otro lado, podemos mencionar el trabajo de Norouzi et al. (2020) donde se estima la elasticidad de las demandas de gasolina y electricidad con respecto a varias variables, entre las que podemos mencionar la epidemiológica, medida por el número de muertes diarias en el mundo de acuerdo con la Organización Mundial de la Salud (OMS).

De acuerdo con la IEA (2021a) después de una caída récord en la demanda global de alrededor de 75 mil millones de metros cúbicos en 2020, los mercados de gas natural experimentaron tensiones significativas entre la oferta y la demanda en los primeros meses de 2021. Temperaturas más frías de lo esperado y una oferta más ajustada llevaron a repuntes y picos de precios, primero en el noreste de Asia en enero y luego en América del Norte en febrero. Las tormentas invernales proporcionaron cierto apoyo a corto plazo al crecimiento de la demanda de gas natural, pero el comportamiento de las variables macroeconómicas del mercado para 2021 a nivel mundial siguen siendo frágiles. La IEA en su informe IEA (2021a), espera que la demanda mundial de gas se recupere a su nivel de 2019, pero con incertidumbres con respecto a la trayectoria de recuperación en los mercados de rápido crecimiento en comparación con las regiones más maduras.

La IEA (2021a) argumenta que las fuertes temporadas de frío, unida a una limitada capacidad de almacenamiento en Japón, China, Korea, varios países de Europa y USA, dispararon los precios del gas a principios del 2021. En USA, los fuertes fríos congelaron los yacimientos de gas, lo que limitó aún más su producción y conllevó a un aumento en el precio. El pronóstico de la IEA prevé un aumento interanual del 3.2 % en la demanda mundial de gas, suficiente para compensar las pérdidas de 2020, pero sujeto a la incertidumbre relacionada con la salud mundial y la recuperación económica.

Del informe de la IEA (2021a) se puede sospechar que, si Colombia importa gas de manera significativa, los precios del gas en el país estarían expuestos y sujetos a las alzas en los meses de invierno de los países productores. Sin embargo, como lo señala también la IEA (2021a), una política de almacenamiento podría ayudar a la estabilización de precios del gas natural.

Similar al caso de energía eléctrica, en la Cuadro 2.2 se presenta un resumen de las variables y modelos encontrados en esta revisión de literatura sobre la demanda de gas natural para diferentes horizontes de tiempo.

Autores	**Frecuencia**	**Modelo**	**Variables**
Liu y Lin (1991)	M, T	FT, SARIMA con Ajuste de Outliers	precio del gas, temperatura, variables de intervención que refleja las pocas variaciones del precio del gas en el período de análisis

Suykens et al. (1996)	M	NN	temperatura, precio del petróleo, número de clientes domésticos y consumo por industria.
Gutiérrez et al. (2005)	A	MCE	demanda de gas natural
Huntington (2007)	M	MRL	demanda de gas natural, precio de la gasolina, consumo de energía en el sector de manufactura, capacidad utilizada de los sectores industrial y de la manufactura, HDD, CDD, precio del gas, precio del petróleo, precio del carbón
Forouzanfar et al. (2010)	A	NLP	demanda de gas natural
Kamrani (2010)	A	MRL	población, ingreso nacional, Producto Nacional Bruto, IPC, temperatura
Özmen et al. (2018)	D, M, A	MARS, CMARS, LR, NN.	costo unitario del gas para usuarios residenciales, tipo de cambio del USD y las liras turcas, temperatura media, mínima y máxima, HDD, consumo de energía, precipitación, número de usuarios residenciales, días festivos nacionales y religiosos, dummy de días de la semana y fines de semana, precio del gas natural, humedad relativa, temperatura del suelo, velocidad promedio del viento.
Anđelković y Bajatović (2020)	D	ANFIS	Demanda de gas, tipo de día, valores mínimos del gas natural y variables meteorológicas
Anagnostis et al. (2020)	D	NN	temperatura, el tipo de mes y el tipo de día de la semana, días festivos
Huseyin (2021)	M	SGM, SARIMA, ASGM	Demanda de gas natural

Liu et al. (2021)	D, M, T, A.	TS, LR, ANN, SVM, GM, RM, DL	Producción Gas Natural, producción industrial/doméstica, clima, temperatura efectiva, precio del gas, población, ingresos por ventas, ingresos, calefacción, consumo histórico, temperatura promedio, datos de descubrimiento de gas, reservas de gas, número de consumidores, tasa de cambio, ingreso nacional, oferta de gas y GLP, PIB, producto industrial bruto.
Concentra (2021)	M	VAR, VECM	demanda de gas natural residencial, precio del GLP, tarifa de gas natural residencial
Özmen (2021)	D, A	MARS, LASSO, LR	temperatura media, temperatura máxima, temperatura mínima, días-grado de calefacción, humedad relativa, velocidad del viento, días de la semana, días festivos nacionales y religiosos, dummy de días laborales y fin de semana, número de usuarios residenciales, demanda de gas natural
Yucesan et al. (2021)	D, M, A	ARIMAX, SARIMAX, ANN, NARX, LSTM, ARIMAX-ANN, SARIMAX-ANN, GA-ANN y PSO-ANN	Consumo de gas natural diario, mes del año, día de la semana, días festivos, tasa de cambio.
Marziali et al. (2021)	D	ELM	tipo de día, días festivos, después de día festivo, temperatura

D: Diario; **M**: Mensual; **T**: Trimestral; **A**: Anual

Cuadro 2.2: Variables explicativas y modelos empleados en la literatura para el pronostico de la demanda de la Gas Natural.

Capítulo 3 Metodologías para la proyección de la demanda de energía eléctrica y gas natural usadas por la UPME

Con el fin de tener un contexto más amplio y acertado para el desarrollo del primer entregable del proyecto, se ha decidido presentar de manera gráfica la estructura metodológica empleada por la UPME para la elaboración de la proyección de la demanda de energía eléctrica y gas natural, a partir de las conversaciones que se han tenido con los funcionarios de la Entidad y el análisis de los informes técnicos que fueron suministrados para la contextualización de la metodología (UPME, 2014b); (UPME, 2014a).

En este sentido se decide plantear en la Sección 3.1 y la Sección 3.2 una descripción cualitativa de las etapas, supuestos, estrategias, variables y criterios empleados por la UPME para la elaboración de la proyección de la demanda de energía eléctrica y gas natural, respectivamente.

3.1 Metodología de proyección de demanda de energía eléctrica

La Figura 3.1 ilustra el procedimiento empleado por la UPME para pronosticar la demanda de energía eléctrica y la potencia máxima del sistema. Para realizar la proyección de demanda de energía eléctrica, la UPME utiliza como variables explicativas:

- Histórico del PIB,
- Proyecciones del PIB,
- Histórico de la demanda de electricidad,
- Histórico de la población del país,
- Histórico de la temperatura,
- Proyecciones de la temperatura.

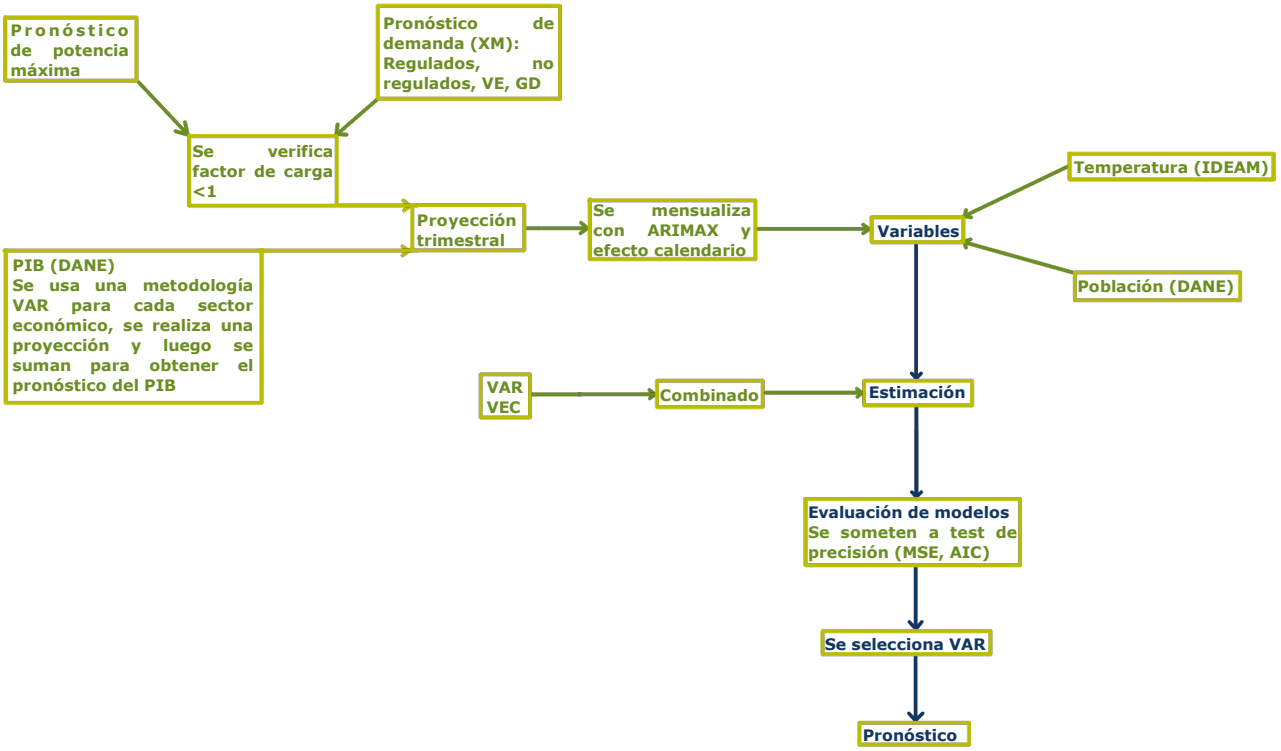

Figura 3.1: Metodología actual para la proyección de la energía eléctrica de la UPME

El proceso de pronóstico de la demanda de energía eléctrica se divide en tres componentes. En primer lugar, se realiza una estimación del consumo del Sistema Interconectado Nacional (SIN); luego se estima el consumo de grandes cargas, y por último, se analizan las proyecciones para el consumo de electricidad de los vehículos eléctricos y la generación distribuida.

Para la proyección del PIB la UPME realiza los siguientes pasos:

(I) Se toman la serie anual y trimestral para cada sector económico con precios constantes.

(II) Para cada uno de los sectores se construye una serie de proyecciones por niveles (largo plazo) y con tasas de crecimiento (corto plazo).

(III) Por un lado, para las estimaciones de corto plazo se estiman vectores autorregresivos-VAR, mientras que para las estimaciones de largo plazo se estiman modelos VEC. En estas estimaciones se utilizan el rezago del PIB de cada sector y los rezagos correspondientes a uno o dos indicadores líderes asociados a cada sector. Cuando no es posible encontrar un indicador asociado a un sector económico las proyecciones se realizan utilizando modelos ARIMA.

(IV) Para las proyecciones se encadenan los últimos cuatro trimestres obtenidos por los modelos VAR, ARIMA y VEC.

(V) Por último, se procede con la suma de cada uno de los trece sectores, para obtener el PIB total desde el enfoque de la oferta.

Descripción de la metodología UPME para proyección de demanda de energía eléctrica

Para realizar el pronóstico de la demanda de energía eléctrica se construye un modelo de combinación de pronósticos basado en los modelos multivariados VAR y VEC. Se propone un sistema de ecuaciones simultáneas, de manera que cada variable es explicada por los retardos de ella misma y por los retardos de las demás variables. Las especificaciones de los modelos son las siguientes:

- Modelo VAR endógeno,
- Modelo VAR exógeno,
- Modelo VEC.

Para cada uno de estos modelos se realiza un análisis de sesgo sistemático que permite calificar la precisión de cada una de las proyecciones que desarrollan los modelos anteriores. Estas medidas analizan el Error Promedio Porcentual (APE), el Error Promedio Absoluto (AAE), el Error Cuadrático Medio (MSE), el Sesgo (B), el Modelo(M) y los Aleatorios (R) (Considine y Clemente, 2007).

A la hora de seleccionar el modelo con mejores proyecciones se emplean los siguientes criterios de selección (Greene, 2000); (UPME, 2014b):

- Logaritmo de la función de máxima verosimilitud, priorizando el modelo con un valor superior.
- Criterio de información de Akaike, priorizando el modelo con un valor inferior.
- Estadístico de Schwarz, siendo preferible el que presente un valor menor R^2 ajustado, priorizando el que presente un valor mayor Estadístico de Hannan-Quinn, y se prefiere el modelo que presente un valor menor de este estadístico.

Con base en los criterios mencionados anteriormente, se concluye que el modelo que cumple con la mayoría de estos es el VAR exógeno, debido a que cumple con los criterios de máxima verosimilitud y el menor valor del estadístico Schwarz.

Con el fin de obtener las proyecciones definitivas para la demanda de energía eléctrica, la UPME emplea una metodología de combinación de pronósticos empleando para tal fin los tres modelos previamente estimados, mediante un sistema de ecuaciones, con el objetivo de obtener proyecciones más exactas que las obtenidas de forma individual por los modelos analizados (UPME, 2014b).

Proyecciones de demandas de consumidores especiales

En este proceso la UPME realiza una proyección de demanda según dos situaciones, primero documentan cuales son los grandes proyectos previstos a desarrollarse en el país y luego se estima su participación dentro de la demanda de energía eléctrica. Por ejemplo, se espera que en Colombia se lleven a cabo la entrada en operación de proyectos como:

- Sociedades portuarias para el año 2021
- Drumond “La Loma” para 2021,
- Ternium Sabanalarga para 2021,
- Minesa para 2021,
- Quebradona para 2025.

Para estos casos, la UPME estima que la participación dentro de la demanda de energía eléctrica se ubica entre un 2 % y un 5 %.

Proyecciones de potencia máxima

La UPME utiliza un modelo de regresión entre los rezagos de la potencia y las proyecciones que se realizaron previamente, respecto a la demanda de energía eléctrica en los diferentes modelos. Adicionalmente, se valida que el factor de carga sea menor a 1; este factor garantiza que la potencia máxima es factible dado el pronóstico de energía. En esencia, se debe satisfacer que

$$\text{Factor de carga} = \frac{p_t^{\max} \Delta t}{E_t} < 1$$

donde $p_t^{\max}$ es la potencia máxima pronosticada para el mes t, Δt corresponde al número de horas del mes t y E_t es la energía pronosticada para el mismo mes.

3.2 Metodología de proyección de demanda de gas natural

La Figura 3.2 ilustra el procedimiento empleado por la UPME para pronosticar la demanda de gas natural en Colombia. En el caso de la proyección del gas natural se emplea un procedimiento similar al usado para proyectar la demanda de energía eléctrica y potencia, con la diferencia de que en el caso del gas se hacen estimaciones para siete sectores. Para cuatro de ellos se utilizan modelos VAR o VECM, mientras que para los demás se usa otro tipo de modelos.

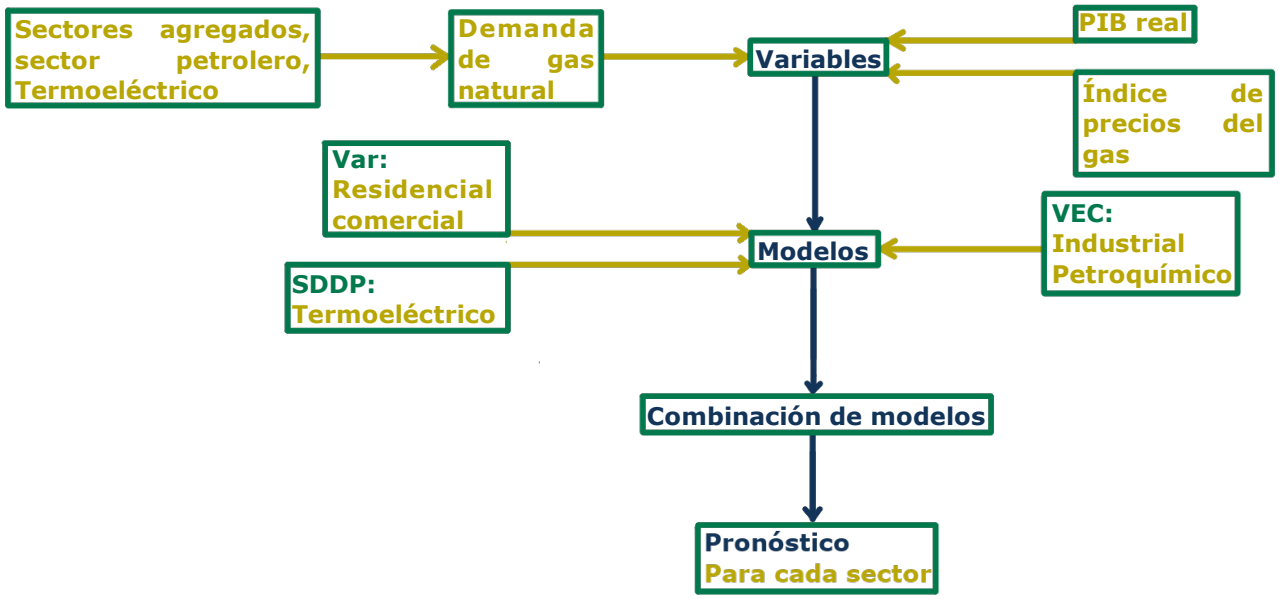

Figura 3.2: Metodología actual para la proyección del gas natural de la UPME

Sector residencial y comercial

Para estos sectores se utilizaron metodologías VAR en primeras diferencias debido a que algunas series son no estacionarias. En el caso de la demanda residencial, se modelaba la relación entre la demanda nacional de gas, con la cobertura y facturación del servicio. La proyección de gas en el sector comercial se obtiene de las estimaciones realizadas en el sector residencial y doméstico. En primer lugar, se modela el sector doméstico en función de la cobertura en el sector residencial y la facturación doméstica, luego a estos resultados se le restan los datos obtenidos con el modelo del sector residencial, para así obtener el consumo en el sector comercial.

Sector industrial y Petroquímico

Para el sector industrial las proyecciones del gas natural utilizaron el modelo VEC, en el que se establecen las relaciones entre la demanda de gas, con el índice de precios del gas; y el valor agregado de la industria manufacturera. Para el caso del sector petroquímico se relaciona el consumo de gas natural con la variable de índice de precios del gas y el índice de producción real del subsector de químicas básicas.

Sector Transporte

Para la proyección de este sector se tomó como punto de partida cifras como:

- número de vehículos a nivel nacional y
- número de viajes y distancias recorridas por vehículos.

3.2.1 Sector Termoeléctrico

El comportamiento de la demanda de este sector está ligado principalmente a condiciones propias del sistema interconectado, así como otras variables que pueden ser los aportes hidrológicos y el volumen útil de los embalses. La demanda de gas para la generación de energía eléctrica se estima teniendo en cuenta estos dos criterios:

- Generación por despacho ideal. Se realiza con el software SDDP, que tiene en cuenta el nivel de las variables mencionadas previamente y los precios de los diferentes energéticos. De acuerdo a su comportamiento, se establecen los costos de las plantas con las cuales se va a suplir la demanda y la cantidad del energético.
- Consumo de generación de seguridad. En este se tienen en cuenta las restricciones de funcionamiento de todo el sistema y las condiciones que se deben presentar para que el sistema opere de manera segura y confiable.

Teniendo en cuenta estos criterios se establece la cantidad de gas natural necesario para cada momento.

Ecopetrol

Los datos de demanda de gas natural por parte del sector petrolero son los consumos que reportan las refinerías de Ecopetrol ubicadas en Cartagena y Barrancabermeja. Las proyecciones para este sector son suministradas por esta Empresa, basadas en las expectativas que tiene de la operación de estas refinerías.

Capítulo 4 Recomendaciones y conclusiones sobre metodología UPME para proyectar energía eléctrica y gas natural

4.1 Introducción

En el informe realizado por la Unidad de Planeación Minero-Energética en Junio de 2021 (UPME, 2021), se presentan las proyecciones de la demanda de energía eléctrica y gas natural que tendrá Colombia en el periodo 2021-2035. En donde el objetivo central del informe es la estimación de la tendencia que tendrá la demanda de estos energéticos a nivel nacional a largo plazo. Para ello, se emplea la información que se conoce sobre la demanda histórica de los energéticos y las expectativas de crecimiento económico esperadas en el país durante los próximos quince años, con el fin de brindar una herramienta que sirva de apoyo a las diferentes autoridades para la toma de decisiones en el área de inversión e infraestructura energética.

Es de anotar que además de usar la información sobre la demanda histórica y las expectativas de crecimiento económico para la realización de sus pronósticos, la UPME (2021) también tiene en cuenta en su informe el efecto generado por la pandemia del COVID-19. Fruto de la pandemia, se ha observado un impacto negativo a nivel mundial y nacional en materia de crecimiento económico, desempleo, niveles de pobreza, y en especial sobre el consumo de los diferentes energéticos. La demanda de los diferentes energéticos ha mostrado contracciones durante este periodo debido en gran medida al confinamiento y cierre de unidades productivas que impuso el gobierno mediante cuarentenas extensas, con el fin de reducir la velocidad de propagación que traía el virus en ese momento. Es de anotar que dichos efectos generados por la Pandemia también se observaron a nivel mundial en donde se evidenciaron tendencias similares (IEA, 2020).

Aunque los efectos de la pandemia sobre la economía fueron desfavorables, las proyecciones realizadas por la UPME (2021) mediante los modelos VAR y VEC son alentadoras, puesto que a partir del 2022 se logra evidenciar una recuperación económica a niveles pre-pandemia, y se proyecta un crecimiento económico anual promedio con rango entre 3 % a 3.6 % para los años 2021-2035.

Así mismo, según las proyecciones presentadas por la UPME (2021), se espera que durante el mismo periodo el crecimiento que registre la demanda de energía eléctrica aumente en un rango anual de 2.28 % y 2.68 %, mientras que, para el caso de la demanda de gas natural, se espera un crecimiento anual promedio en un rango de 0.74 % y 1.60 %.

Ahora bien, con base en el trabajo realizado por la UPME, y con el objetivo de comple-

mentar las metodologías de proyección de demanda de gas natural y energía eléctrica, en este capítulo planteamos una serie de recomendaciones que podrían ser de utilidad para la UPME. Estos complementos, dirigidos tanto para los modelos matemáticos como para variables explicativas, buscan mantener altos niveles de desempeño en términos de proyecciones. Las recomendaciones y observaciones que realizamos son el resultado del análisis de la literatura, del análisis teórico y conceptual de los modelos usados por la UPME, y de la adopción de algunas aplicaciones que buscan mostrar el buen desempeño que pueden mostrar algunos modelos alternativos.

4.2 Variables explicativas

Como se expuso en la revisión de la literatura desarrollada en el Capítulo 2, existe un amplio espectro de variables explicativas que se han usado en la literatura para la modelación de la demanda de los diferentes energéticos, que podrían tal vez, ser de utilidad para obtener buenos pronósticos para el caso colombiano, tales como: tamaño poblacional, temperatura, velocidad del viento, radiación, humedad relativa, entre otras variables climáticas, heating degree days (HDD), cooling degree days (CDD), precio del gas residual, consumo de gas en la industria, precio de la gasolina, ingreso per-cápita, emisiones de CO_2, balanza comercial, tasa de cambio, entre otras.

Sin embargo, es necesario señalar que algunas de estas variables podrían no ser viables para la creación de modelos de pronóstico, debido a la disponibilidad de la información, las características bajo las cuales se usaron dichas variables en los trabajos de referencia y la fiabilidad que se tenga de los pronósticos de las variable exógenas; puesto que las condiciones colombianas no son necesariamente las mismas del entorno internacional en el que se han desarrollado estos trabajos.

Cabe señalar que dada la disponibilidad de información que se tiene en el país y el horizonte de 15 años que requiere la UPME para sus proyecciones, no es posible probar cada unas de las variables observadas en la revisión de la literatura, ya que para poder hacerlo es necesario tener proyecciones para cada una de las variables exógenas durante los próximos 15 años.

Además, si hipotéticamente se tuvieran las proyecciones de dichas variables para los próximos 15 años, tampoco sería posible garantizar que realmente dichas variables sean viables para la estimación de los modelos, ya que no hay garantía de que realmente éstas mejoren el poder predictivo, lo cual puede considerarse como una barrera en la información para los proceso de estimación y pronóstico.

Es de anotar que la necesidad de que las variables exógenas de los modelos posean proyecciones durante los próximos 15 años, parte del marco teórico de los modelos VARX y VEC, en donde el primer paso en la estimación de estos modelos consiste en identificar qué variables son endógenas y qué variables son exógenas, con el fin de plantear correctamente el modelo. Ya que al plantear dos o más variables endógenas estamos diciendo que los movimientos de una

de las variables deberá afectar a las demás variables endógenas (Lütkepohl, 2013).

Por ejemplo, puede ser correcto decir que el PIB y la demanda de energía eléctrica son endógenas, ya que como es de esperarse mayor producción implica mayor demanda de energía y mayor demanda de energía implica que la producción posiblemente va a aumentar. Sin embargo, si decimos que la demanda de energía y la población son endógenas, estamos diciendo que si aumenta la demanda de energía, también aumenta la población, lo cual no necesariamente es correcto.

Caso similar ocurre con variables como la temperatura, en donde la temperatura puede aumentar por causas muy distintas al aumento de la demanda de energía. Desde este punto de vista, la población y la temperatura son variables que deberían ingresar a los modelos VARX y VEC de forma exógena, y por tanto son variables deben tener proyecciones definidas para los próximos 15 años, si se desean ingresar a los modelos.

Es claro que aunque en otros países algunas variables pueden funcionar bien, las realidades de estos países suelen ser muy distintas a la Colombiana, por lo que en nuestro caso incluir dichas variables en los modelos no necesariamente tienen que mejorar el poder predicativo de los modelos. Por último, en el caso de tener proyecciones de tales variables, puede ser difícil en muchos casos garantizar que las predicciones de estas variables serán realmente certeras, sobretodo para aquellas variables de corte financiero tales como las tasas de cambio, precios del oro, entre otros[1].

Por tanto, al no tener garantía alguna con respecto al aporte que pueden hacer estas variables para los modelos de pronóstico de energía y gas natural en Colombia, sólo se recomendará probar algunas de las variables expuestas en la revisión de la literatura, en la medida de la disponibilidad y confiabilidad que se tenga de sus proyecciones.

Dado lo anterior, se ilustra en la Subsección 4.2.1 y la Subsección 4.2.2 algunas de las variables que podrían ser empleadas para el pronóstico de la demanda de energía eléctrica y gas natural en Colombia, basados en el hecho de la disponibilidad de la información con la que cuenta el país y en los pronósticos que podrían realizar las entidades encargadas de la recolección y validación de la información.

4.2.1 Energía eléctrica

Partiendo de lo expuesto en la Sección 4.2, se tiene que los procesos de estimación y pronóstico en el caso colombiano poseen una barrera a la información que se observa de forma considerable cuando se refiere a proyecciones con horizontes a largo plazo, tales como los que plantea la UPME, debido a que para las proyecciones de la demanda de energía eléctrica con un horizonte de 15 años, requieren también que cada una de las variables explicativas que se

[1]Es de anotar, que la construcción de modelos de pronóstico confiables para variables de este tipo a un horizonte de 15 años es sumamente complicado, debido a que el comportamiento de las mismas depende de factores externos no controlables por la economía.

introduzcan en los modelos de forma exógena también posean un pronóstico igual o mayor a 15 años, lo cual es un gran obstáculo debido a que pocas entidades gubernamentales o privadas que cuentan con este tipo de pronósticos.

Por tanto, tomando de referencia las variables explicativas empleadas por la UPME (2021) en la realización de las proyecciones de la demanda de energía eléctrica, y tomando de referencia las variables presentadas la revisión de la literatura realizada en la Sección 2.1, es posible construir un listado de variables que podrían tenerse en cuenta en la estimación de modelos de pronóstico de la demanda de energía eléctrica.

En el Cuadro 4.1 se presentan algunas recomendaciones sobre la inclusión de variables explicativas que pueden encontrarse o construirse para un horizonte de 15 años en Colombia, y que pueden ser incluidas en los modelos de proyección de la demanda de energía eléctrica.

Variable	**Descripción**	**Fuente**	**Usada por:**
Demanda de energía eléctrica	Se refiere a la demanda nacional de energía eléctrica reportada mensualmente por la UPME	**H:** UPME **P:** N/A	UPME/UdeA
Producto interno bruto real precios constantes año base 2015	Se refiere al valor total de los bienes y servicios finales producidos por el país, y reportados trimestralmente por el DANE.	**H:** DANE **P:** UPME	UPME/UdeA
Temperatura promedio nacional	Se refiere a la temperatura promedio registrada a nivel nacional y reportada de forma diaria o mensual por el IDEAM.	**H:** IDEAM **P:** IDEAM	UPME/UdeA
Población	Se refiere al total de individuos que se encuentran habitando en el territorio nacional y reportados de forma anual por el DANE.	**H:** DANE **P:** DANE	UPME/UdeA
Días mes	Variable creada que realiza el conteo de número total de días que posee el mes.	**H:** Calendario **P:** Calendario	UdeA
Días laborales por mes	Variable creada que realiza el conteo de días laborales hábiles en el mes.	**H:** Calendario **P:** Calendario	UdeA
Días domingo por mes	Variable creada que realiza el conteo de días domingo.	**H:** Calendario **P:** Calendario	UdeA

Efecto fenómeno niño y niña	Se refiere a las temperaturas oceánicas del Pacífico tropical central y oriental en las cuales se registran aguas superficiales relativamente más calidas ($\geq 0.5^oC$ - Niño) o más frías ($\leq -0.5^oC$ - Niña) que las temperaturas normalmente registradas.	**H:** Climate Prediction Center[2] **P:** N/A	UdeA
Dummy cierres económicos	Variable creada que busca capturar el efecto mensual que generan los cierres económicos causados por la pandemia, o paros nacionales	**H:** Prensa **P:** N/A[3]	UdeA
Dummy efecto calendario (Tipo de mes)	Variables creadas que busca capturar el comportamiento estacional determinístico que pueden poseer las variables explicativas	**H:** Calendario **P:** Calendario	UdeA
Dummy campañas racionamiento ("Apagar-Paga")	Variables creadas que busca capturar el efecto generado por las campañas de racionamiento impulsadas por el gobierno nacional.	**H:** Prensa **P:** N/A[4]	UdeA
H: Hace referencia a la serie histórica; **P** Hace referencia a las proyecciones de la serie. **N/A:** Hace referencia a datos no disponibles a la fecha o no necesarios para el modelo de proyección de Energía Eléctrica.			

Cuadro 4.1: Variables explicativas sugeridas para el planteamiento de modelos de pronóstico de la demanda de la Energía Eléctrica en Colombia.

4.2.2 Gas Natural

Similar al caso de energía eléctrica y teniendo en cuenta el planteamiento sobre la barrera a la información que se encuentra en Colombia para la construcción de modelos de proyección con horizontes a largo plazo, en el Cuadro 4.2 se plantea un listado de variables explicativas recomendadas que podrían encontrarse o construirse en Colombia para un horizonte de 15

[2] https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php

[3] A pesar de que las proyecciones de esta variable no son conocidas, se podría suponer que no hayan cierres económicos en días futuros dado el efectividad de las vacunas y que los cierres generado por los paros no sean muy prolongados. En caso de conocer alguna novedad sobre futuros cierres será necesario registrar las fechas en la variable de los cierres económicos para mejorar los pronósticos de los modelos.

[4] Caso similar al presentado por la variable cierres económicos, en donde las proyecciones de esta variable no son conocidas pero se espera que no hayan campañas de racionamiento en el corto plazo. En caso de conocer alguna novedad sobre épocas de racionamiento futuros será necesario registrar las fechas en la variable de campañas de racionamiento para mejorar los pronósticos de los modelos.

años, en donde, dichas recomendaciones se realizan basandos en las variables presentadas en el informe de la UPME (2021) y las variables encontradas en la revisión de la literatura realizada en la Sección 2.2

Variable	**Descripción**	**Fuente**	**Usada por:**
Demanda de gas natural	Se refiere a la demanda nacional de gas natural reportada mensualmente por la UPME	**H:** UPME **P:** N/A	UPME/UdeA
Demanda sectorial de gas natural	Se refiere a la demanda por sectores de gas natural reportada mensualmente por la UPME	**H:** UPME **P:** N/A	UPME/UdeA
Índice de precios del gas natural año base 2018.	Se refiere el cambio porcentual que tiene el precio del gas natural respecto al año base 2018.	**H:** UPME **P:** UPME	UPME/UdeA
Producto interno bruto real precios constantes año base 2015	Se refiere al valor total de los bienes y servicios finales producidos por el país, y reportados trimestralmente por el DANE.	**H:** DANE **P:** UPME	UPME/UdeA
Producto interno bruto real por sectores a precios constantes año base 2015	Se refiere al valor total de los bienes y servicios finales producidos por los diferentes sectores del país, y reportados trimestralmente por el DANE.	**H:** DANE **P:** UPME	UdeA
Precios promedio gas natural usuarios regulados	Se refiere al precio promedio que se cobra por m^3 de gas natural a los usuarios regulados, y reportada mensualmente por la SUI.	**H:** SUI **P:** N/A	UdeA
Precios promedio gas natural usuarios no regulados	Se refiere al precio promedio que se cobra por m^3 de gas natural a los usuarios no regulados, y reportada mensualmente por la SUI.	**H:** SUI **P:** N/A	UdeA
Precios promedio GLP usuarios regulados	Se refiere al precio promedio que se cobra por m^3 de GLP a todos los usuarios regulados, y reportada mensualmente por el SUI.	**H:** SUI **P:** N/A	UdeA
Precios promedio carbón	Se refiere al precio promedio que se cobra por tonelada de carbón, y reportada mensualmente por la UPME.	**H:** UPME **P:** N/A	UdeA

Días mes	Variable creada que realiza el conteo de número total de días que posee el mes.	**H:** Calendario **P:** Calendario	UdeA
Días laborales por mes	Variable creada que realiza el conteo de días laborales hábiles en el mes.	**H:** Calendario **P:** Calendario	UdeA
Días domingo por mes	Variable creada que realiza el conteo de días domingo.	**H:** Calendario **P:** Calendario	UdeA
Efecto fenómeno niño	Se refiere a los fenómenos climáticos en las cuales se registran aguas superficiales relativamente más cálidas ($\geq 0.5^{o}$C - Niño) que las temperaturas normales registradas en el Pacífico tropical central y oriental,	**H:** Climate Prediction Center[5] **P:** N/A	UdeA
Dummy cierres económicos	Variable creada que busca capturar el efecto mensual que generan los cierres económicos causados por la pandemia, o paros nacionales	**H:** Prensa **P:** N/A[6]	UdeA
Variable dummy efecto calendario (Tipo de mes)	Variables creadas que busca capturar el comportamiento estacional determinístico que pueden poseer las variables explicativas	**H:** Calendario **P:** Calendario	UdeA
H: Hace referencia a la serie histórica; **P** Hace referencia a las proyecciones de la serie. **N/A:** Hace referencia a datos no disponibles a la fecha o no necesarios para el modelo de proyección de Gas Natural.			

Cuadro 4.2: Variables explicativas sugeridas para el planteamiento de modelos de pronóstico de proyección de Gas Natural en Colombia.

4.3 Modelos alternativos

Un aspecto adicional que se encuentra en la revisión de la literatura, además del amplio espectro de variables explicativas que se mencionaron en la Sección 2.1 y Sección 2.2, es la gran cantidad de modelos alternativos que se suelen emplear para el pronóstico de los diferentes energéticos, en donde se evidencia que la combinación de los modelos presentados en conjunto con una adecuada combinación de variables explicativas, pueden lograr metodologías de pronóstico con resultados certeros para los diferentes energéticos.

[5] https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php

[6] A pesar de que las proyecciones de esta variable no son conocidas, se espera que no hayan cierres económicos en días futuros dado el efectividad de las vacunas y que los cierres generado por los paros no sean muy prolongados.

Aunque muchos de los modelos que se encuentran en la literatura pueden ser de difícil aplicación, debido en gran medida al número de supuestos que deben cumplir para su correcta utilización, también se encuentran modelos que no tienen supuestos altamente restrictivos o que simplemente poseen estructura no paramétrica, lo que facilita significativamente su implementación, permitiendo considerar diferentes alternativas al momento de estimar modelos de proyección.

Por ello, se decide implementar en la Subsección 4.3.1, algunos modelos alternativos de fácil aplicación que podrían considerarse para el pronóstico de la demanda de energía eléctrica y de gas natural, junto a una serie de pequeñas aplicaciones de los mismos, con el fin de ilustrar su desempeño predictivo bajo diferentes escenarios.

Para tal propósito se considera en cada aplicación algunas de las variables presentadas en el Cuadro 4.1 y el Cuadro 4.2 con el objetivo de realizar los aplicación de ejercicios que se presentan en la Subsección 4.3.1, con el fin de mostrar el desempeño predictivo que tienen algunos modelos alternativos que se sugieren emplear para la proyección de la demanda de energía eléctrica y gas natural que realiza la UPME.

4.3.1 Aplicaciones

Aunque hay una gran variedad de modelos que pueden implementarse para realizar el pronóstico de la demanda de los diferentes energéticos, tal como se presenta en el Cuadro 4.1 y el Cuadro 4.2, la aplicación de todos estos modelos se escapa del alcance de este informe, por lo cual solo nos centraremos en la aplicación de cuatro metodologías de mayor uso en la demanda de energía eléctrica y en la demanda de gas natural, a saber, una Regresión lineal múltiple (MLR), un Modelo aditivo generalizado (GAM), una Regresión lineal con shirinkage y operador de selección (LASSO), y una Regresión spline adaptativa multivariante (MARS), en donde la presentación teórica de los modelos aquí mencionados se presenta de forma resumida en el Anexo B.

Con el fin de ejemplificar la aplicación de los modelos alternativos y comparar el desempeño predictivo que estos poseen respecto al que se obtendría con un modelo VARX o VEC, se plantea un total de 9 escenarios, 2 para la demanda energía eléctrica y 7 para la demanda de gas natural.

- **Caso 1:** Pronóstico de la demanda de energía eléctrica bajo un escenario de simulación del COVID-19, con datos para estimación de 2009-1 a 2017-12 y horizonte de pronóstico 2018-1 a 2019-12.
- **Caso 2:** Pronóstico de la demanda de energía eléctrica con datos originales, con datos para estimación de 2009-1 a 2019-8 y horizonte de pronóstico 2019-9 a 2021-8, al incluir campaña “Apagar-Paga” y efecto del fenómeno del niño y niña.
- **Caso 3:** Pronóstico de la demanda de gas natural para el sector industrial, con datos para estimación de 2009-1 a 2019-8 y horizonte de pronóstico 2019-9 a 2019-12.
- **Caso 4:** Pronóstico de la demanda de gas natural agregada, con datos para estimación

de 2009-1 a 2017-12 y horizonte de pronóstico 2018-1 a 2019-12, escenario base.

- **Caso 5:** Pronóstico de la demanda de gas natural agregada, con datos para estimación de 2009-1 a 2017-12 y horizonte de pronóstico 2018-1 a 2019-12, al incluir el efecto del fenómeno del niño.
- **Caso 6:** Pronóstico de la demanda de gas natural agregada, con datos para estimación de 2009-1 a 2017-12 y horizonte de pronóstico 2018-1 a 2019-12, al incluir los precios promedio del gas para usuarios regulados.
- **Caso 7:** Pronóstico de la demanda de gas natural agregada, con datos para estimación de 2009-1 a 2017-12 y horizonte de pronóstico 2018-1 a 2019-12, al incluir los precios promedio del gas para usuarios no regulados.
- **Caso 8:** Pronóstico de la demanda de gas natural agregada, con datos para estimación de 2009-1 a 2017-12 y horizonte de pronóstico 2018-1 a 2019-12, al incluir los precios promedio del GLP para usuarios regulados.
- **Caso 9:** Pronóstico de la demanda de gas natural agregada, con datos para estimación de 2009-1 a 2017-12 y horizonte de pronóstico 2018-1 a 2019-12, al incluir los precios promedio del carbón.

La decisión bajo la cual se seleccionan los periodos de estimación y pronóstico de los casos se hace sujeto a la disponibilidad de información suministrada por la UPME y a los supuestos realizados en cada caso, debido a que el objetivo de los ejercicios aplicados es mostrar el desempeño predictivo de los modelos bajo diferentes condiciones y no realizar pronósticos a largo plazo.

Además se tendrá que solo será factible emplear la información que va hasta el periodo 2021-08, ya que ésta información nos permitirá comparar los resultados obtenidos por los modelos estadísticos respecto a la información real observada.

4.3.1.1 Caso 1: Demanda energía eléctrica bajo escenario de simulación del COVID-19

Como se presentó en la revisión de la literatura de la Sección 2.1, existe una gran variedad de variables que podrían ser usadas para el pronóstico de la demanda de energía eléctrica, pero debido a la poca disponibilidad de información con la que se cuenta en Colombia, se recomienda solo probar la utilidad de las variables en la medida de su fácil consecución. Puesto que no hay garantía respecto a que todas las variables utilizadas en la literatura sean relevantes para el pronóstico de la demanda de energía eléctrica en Colombia; y más aún, cuando estas variables fueron empleadas en aplicaciones para escenarios internacionales que pueden diferir sustancialmente del caso colombiano.

Debido a lo anterior, y con el objetivo de ilustrar el desempeño de los modelos MLR, GAM, LASSO y MARS con respecto al desempeño de un modelo VARX o VEC, en este caso se decide emplear del total de variables listadas en el Cuadro 4.1, las variables demanda de

energía eléctrica, PIB, temperatura promedio nacional, población, días mes, días laborales por mes, días domingos por mes, la dummy de cierres económicos y el efecto calendario, con el fin de realizar el ajuste de los modelos de demanda de energía eléctrica.

Para el planteamiento de este ejercicio, se decide imitar el efecto que tuvo la pandemia del COVID-19, con el fin de ilustrar cómo debería ser el desempeño predictivo de los modelos una vez la economía volviera a la senda de crecimiento que tenía antes de la pandemia. Los lineamientos que se tuvieron en este ejercicio son los siguientes:

- **Datos entre 2009-01 a 2017-12**: Se toman los datos definidos en este periodo con el fin de realizar el ajuste de los modelos.
- **Simulación del comportamiento de la demanda de energía eléctrica y PIB para imitar el efecto COVID-19**: Se reemplaza la información reportada durante el periodo 2016-4 a 2017-8 con la información de la demanda registrada en 2015-12 multiplicada por los factores de crecimiento observados entre el periodo 2020-4 y 2021-8 respecto a 2019-12. Es de anotar que los factores de crecimiento observados se calculan mediante la división entre el periodo objetivo (2020-4 a 2021-8) y se dividen por el periodo base 2019-12.
- **Datos entre 2018-01 a 2019-12**: Se emplean los datos definidos en este periodo únicamente para comparar los pronósticos realizados por los modelos con respecto a la demanda real reportada y poder probar el desempeño de los mismos.

Es de anotar que dado el proceso de simulación que se realizó en el periodo 2016-4 a 2017-8 para imitar el efecto del COVID-19 que se observó en el periodo 2020-4 a 2021-8, se decide no utilizar el tramo de información que va de 2020-1 al 2021-8 para evitar contaminar los resultados obtenidos al momento de probar el desempeño de los modelos alternativos para la predicción de la demanda de energía eléctrica.

En la Figura 4.1 se ilustra la serie original y modificada luego del proceso de simulación para la demanda de energía eléctrica durante el periodo que va de 2009-1 a 2019-12.

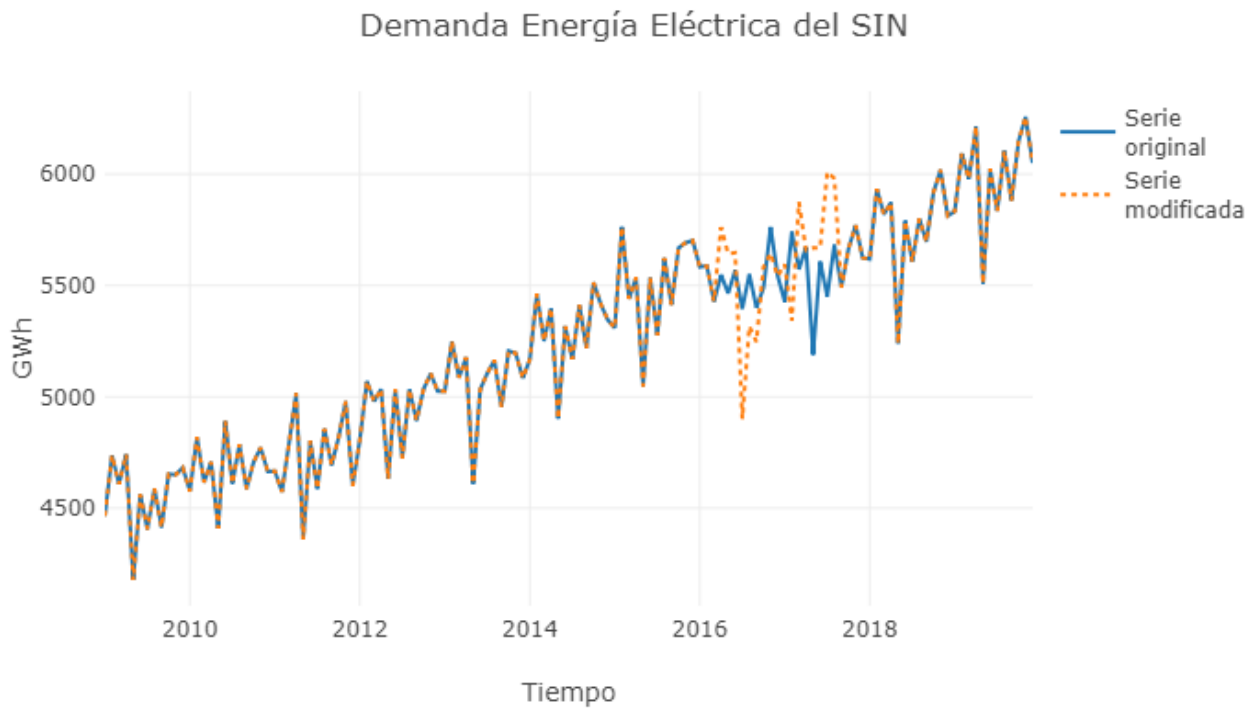

Figura 4.1: Serie original y modificada de la demanda de energía eléctrica para el periodo 2009-1 a 2019-12

Y en la figura Figura 4.2 se ilustra la serie original y modificada luego del proceso de simulación para PIB durante el mismo periodo.

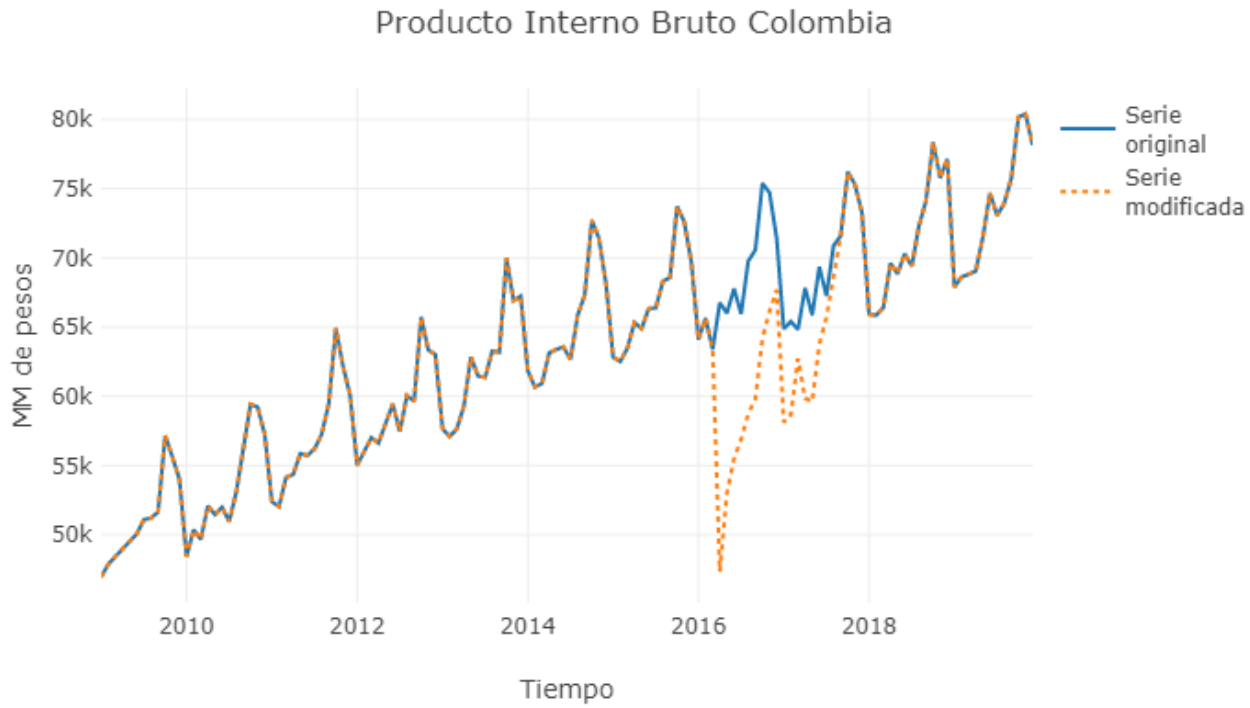

Figura 4.2: Serie original y modificada del PIB para el periodo 2009-1 a 2019-12

Al comparar las series presentadas en la Figura 4.1 y la Figura 4.2 respecto a sus series

originales, se logra observar una caída en la demanda de energía eléctrica y el PIB durante el periodo 2016-4 a 2017-8 para las series modificadas, debido al proceso de simulación que se hizo sobre la serie original durante este periodo, de tal forma que pudiera imitar el comportamiento que tuvo la pandemia sobre las variables.

Ahora, a causa de las modificaciones realizadas sobre las variables de demanda de energía eléctrica y PIB, se crea la variable Dummy de cierres económicos, la cual toma el valor de 1 durante el periodo 2016-4 a 2016-12 y cero en otro caso, buscando imitar el efecto real de los cierres que generó la pandemia entre los meses abril-diciembre de 2020.

Una vez definidas las variables que se van a emplear en el ejercicio de estimación y evaluación de los modelos de pronóstico para la demanda de energía eléctrica, se decide realizar un ejercicio de diagnóstico para las variables endógenas al modelo que se encuentran dentro del tramo que va desde 2009-1 hasta 2017-12, con el fin de observar si alguna de ellas exhibe algún comportamiento estacional.

Para tal propósito, se decide realizar el cálculo de la ACF para la demanda de energía eléctrica y el PIB, con el objetivo de verificar si alguna de ellas presenta un patrón repetitivo que dé indicios sobre el componente estacional que tendrán las series, encontrando que ambas presentan un comportamiento estacional con una periodicidad de 12 rezagos.

En el caso del PIB, se registra un componente estacional claramente marcado, en donde se ve con claridad que la variable registra una ACF con picos cada 12 rezagos, indicando que el PIB es una variable estacional con periodicidad anual. Por su parte, en el caso de la demanda de energía eléctrica, a pesar de que su componente estacional no es tan evidente como en el PIB, también es posible observar un gran pico cada 12 rezagos para su ACF, indicando que para esta serie también debe considerarse un componente estacional no estacionario. La ACF para la demanda de energía eléctrica y PIB se presenta en la Figura 4.3.

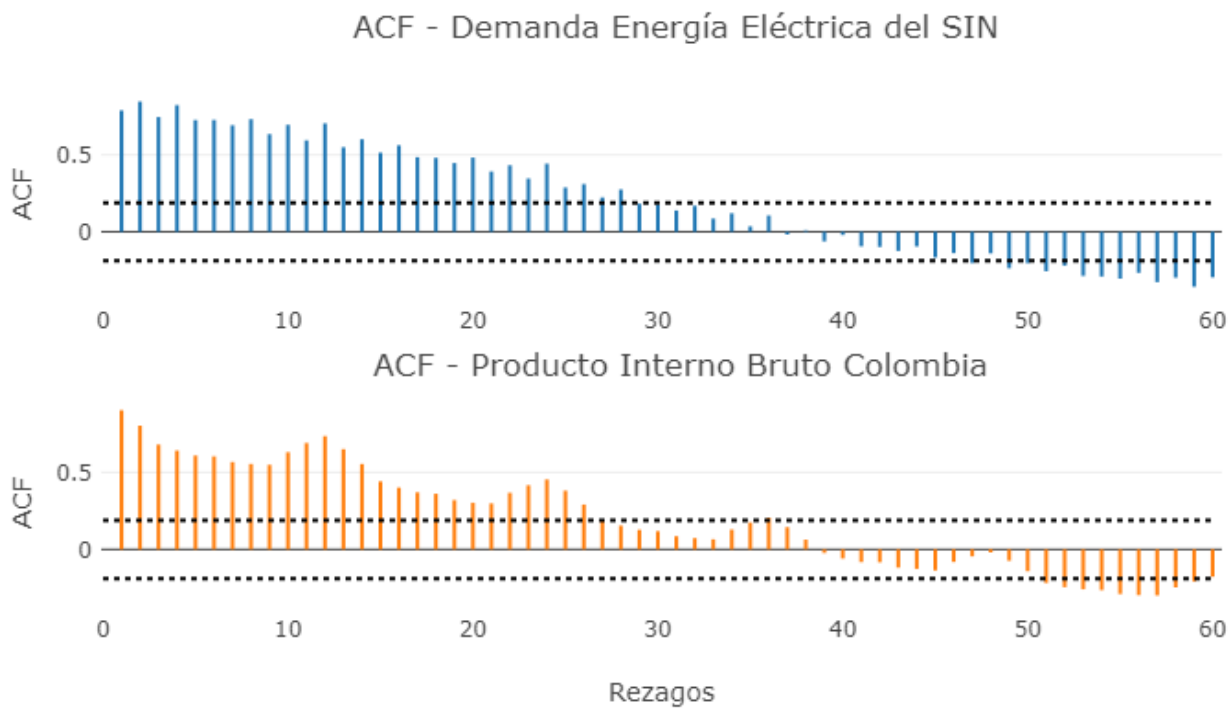

Figura 4.3: Función de autocorrelación para la demanda de energía eléctrica y el PIB, para un total de 60 rezagos.

El comportamiento estacional que exhiben las variables de la demanda de energía eléctrica y el PIB en la Figura 4.3, sumado a la tendencia creciente que se evidencia en la Figura 4.1 y Figura 4.2, hacen necesaria la verificación de la existencia de raíces unitarias estacionales y no estacionales. En caso de no tener en cuenta la existencia de las mismas, podrían generarse sesgos en los procesos de estimación que se verían reflejados en los pronósticos del modelo.

En este sentido, se decide aplicar la prueba Hylleberg, Engle, Granger, and Yoo (HEGY) propuesta por Hylleberg et al. (1990) y extendida a frecuencias mensuales por Franses (1991), con el fin de probar de forma simultanea la existencia de raíces unitarias estacionales y no estacionales que puedan tener las series temporales. El planteamiento teórico de la prueba HEGY se presenta en el Anexo A.

Entonces, dado que se tiene la sospecha sobre la existencia de raíces unitarias estacionales para las variables de demanda de energía eléctrica y el PIB debido a los picos que exhibía su ACF y la tendencia creciente que registran, se decide aplicar la prueba HEGY, con el fin de verificar si dichas sospechas se encuentran sustentadas desde un punto de vista más formal, obteniendo los resultados presentados en el Cuadro 4.3.

Prueba HEGY Energía Eléctrica			
	Estadístico	P-valor	
t_1	−2.5399	0.9778	
t_2	−1.1033	0.3666	
$F_{3:4}$	5.6574	0.0036	**
$F_{5:6}$	3.7317	0.0170	*
$F_{7:8}$	4.2184	0.0097	**
$F_{9:10}$	19.6895	0.0000	***
$F_{11:12}$	0.9767	0.2733	
$F_{2:12}$	7.6034	0.2574	
$F_{1:12}$	7.4158	0.1545	

Prueba HEGY PIB			
	Estadístico	P-valor	
t_1	−2.6617	0.9896	
t_2	−1.7067	0.1678	
$F_{3:4}$	0.2597	0.6535	
$F_{5:6}$	1.3466	0.2299	
$F_{7:8}$	2.6662	0.0674	.
$F_{9:10}$	5.4783	0.0044	**
$F_{11:12}$	2.1236	0.1087	
$F_{2:12}$	2.4172	0.2848	
$F_{1:12}$	2.8413	0.0084	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cuadro 4.3: Prueba HEGY demanda energía eléctrica y PIB

Dado que las pruebas de hipótesis que plantea la prueba HEGY pueden generalizarse para cada periodo de tiempo como[7]

$$H_0 : \text{Existe una raíz unitaria para la frecuencia establecida}$$

$$H_1 : \text{No existe una raíz unitaria para la frecuencia establecida}$$

entonces significa que de los resultados presentados en el Cuadro 4.3, no es posible rechazar la hipótesis nula de los estadístico t_1 y $F_{11:12}$ para las series de la demanda de energía eléctrica y el PIB, concluyendo con ello que ambas variables poseen al menos una raíz unitaria no estacional (a la frecuencia cero) y una raíz unitaria estacional anual.

Una vez detectada la existencia de raíces unitarias estacionales y no estacionales que poseen las variables, el siguiente paso es definir el número de rezagos que se van a incluir en las variables endógenas del modelo VARX(p), y por ello se decide emplear una función de optimización que identifica de forma iterativa cuál es el número óptimo de rezagos que deben incluirse para las variables de demanda de energía eléctrica y PIB, a través de cuatro criterios de información, a saber, el crítерio de informacion de Akaike (AIC), el criterio de información de Hannan-Quinn (HQ), el criterio de información de Schwarz (SC) y el criterio de errores de predicción final (FPE), obteniendo los resultados presentados en el Cuadro 4.4.

Número óptimo de rezagos modelo VARX(p)			
AIC(n)	HQ(n)	SC(n)	FPE(n)
22	17	1	17

Cuadro 4.4: Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p)

[7]Esto es debido a que, como se expone en el Anexo A, cada una de las pruebas que se realizan con HEGY, posee como hipótesis nula la existencia de una raíz unitaria para diferentes frecuencias (no estacionales, mensual, bimensuales, trimestrales, etc), contra la alternativa de que no existe una raíz unitaria para dicha frecuencia.

Por tanto, dado que dos criterios de información (HQ y FPE) sugieren que el número óptimo de rezagos para el modelo VARX(p) es de 17, se decide emplear un total de 17 rezagos para las variables endógenas de demanda de energía eléctrica y PIB durante el proceso de estimación del modelo VARX[8]. Es de anotar que con este número de rezagos se logra capturar además del efecto anual, otros efectos (bimensual, trimestral, etc) que puedan poseer las series.

Como paso final, antes de realiza la estimación del modelo VARX, es necesario verificar si existen relaciones de cointegración entre las variables de demanda de energía eléctrica y PIB, debido a que son las variables que se establecen como endógenas dentro del modelo. Para ello se decide aplicar la prueba de Johansen con traza (razón de verosimilitud), la cual tiene por objetivo verificar si existen o no relaciones de cointegración entre las variables endógenas del modelo VARX[9].

Al aplicar la prueba secuencial de Johansen a las variables endógenas del modelo VARX para saber si existen o no relaciones cointengrantes, se encuentran los resultados planteados en el Cuadro 4.5. Para más información sobre la prueba de Johansen consulte el Anexo B.

Prueba Johansen con traza

Hipótesis	Estadístico	VC-10 %	VC-5 %	VC-1 %
$r \leq 1$	10.70	7.52	9.24	12.97
$r = 0$	105.43	17.85	19.96	24.60

Cuadro 4.5: Prueba de Johansen con traza para las variables de la demanda de energía eléctrica y el PIB

En donde, en el primer paso de la prueba se observa que se rechaza la primera hipótesis nula de que $r = 0$, bajo los tres niveles de significancia del $1\,\%, 5\,\%$, y $10\,\%$, debido a que el estadístico cae para los tres casos dentro de la región de rechazo establecida por los valores críticos, concluyendo que el número de relaciones cointegrantes es superior a 0. En el segundo paso de la prueba se observa que no se rechaza la hipótesis nula de que $r \leq 1$, debido a que el estadístico de prueba cae dentro del área de no rechazo establecida por el valor crítico del

[8]Es de anotar que además de los resultados obtenidos a partir de los criterios de información, es necesario considerar los casos en los cuales se posean variables endógenas que exhiban comportamientos estacionales, ya que en estos casos se hace necesario establecer una longitud mínima de rezagos que se ingresan al modelo, que dependerá de la periodicidad de la serie temporal, en donde si se tiene una periodicidad mensual se tendrán que considerar como mínimo 12 rezagos, mientras que, si la periodicidad es trimestral se tendrán que considerar como mínimo 4 rezagos, ya que es muy importante en los proceso de modelación con datos estacionales considerar al menos el primer rezago estacional que posea la serie.

[9]Es necesario aclarar que cuando se tienen componentes estacionales en las variables endógenas, se debe probar cointegración en las distintas frecuencias estacionales de las variables. Aunque el modelo de cointegración estacional existe y se presenta en Harris y Sollis (2003), todavía no está disponible ni el estadístico de prueba de cointegración en distintas frecuencias estacionales, ni tampoco la estimación de dicho modelo. Otra alternativa sería usar la metodología de modelos SVARMA, los cuales son una extensión de los modelos SARIMA. Sin embargo, hay que dejar claro que la implementación que hay disponibles hasta el momento no admite el uso de variables exógenas. La implementación de los modelos SVARMA se encuentra solo para variables endógenas en la librería MTS del paquete estadístico R.

1 %, concluyendo que existe a lo más una relación cointegrante entre las variables endógenas del modelo VARX.

Dado el resultado obtenidos en el Cuadro 4.5 por la prueba de Johansen, y el número óptimo de rezagos mostrado en el Cuadro 4.4, se tendrá que el modelo a estimar será un VEC con un total de 17 rezagos para las variables endógenas.

Una vez definidas las condiciones óptimas para la estimación del modelo VEC, y dado que el objetivo es comparar el desempeño predictivo de este modelo respecto a los modelos alternativos MRL, GAM, LASS y MARS, se decide estimar todos los modelos bajo las mismas condiciones y se presentan los resultados en la Subsección 4.3.2, con el fin de facilitar la lectura de los resultados obtenidos para éste y los demás casos desarrollados en la Subsección 4.3.1.

4.3.1.2 Caso 2: Demanda energía eléctrica al incluir campaña "Apagar-Paga" y efecto del fenómeno del niño-niña

Al igual que la Subsección 4.3.1.1, este caso de estudio consiste en estimar la demanda de energía eléctrica, pero con la diferencia de que en este caso se decide no realizar ninguna simulación a los datos para imitar el efecto del COVID-19, sino que se decide estimar los cinco modelos mencionados previamente, con los datos reales de la demanda de energía eléctrica con dos variables adicionales a las que se usaron en el caso anterior, a saber, una dummy que representa el efecto de la campaña de racionamiento de energía eléctrica "Apagar-Paga" impulsada por el gobierno nacional entre el periodo 2015-12 a 2016-8 y la temperaturas registradas en el pacífico tropical central y oriental asociadas al fenómeno del niño y de la niña.

Para tal fin se decide emplear la información correspondiente al periodo 2009-1 a 2019-8 para realizar los procesos de estimación, y el periodo 2019-9 a 2021-8 para probar el desempeño productivo de los diferentes modelos, en donde se usará el MAPE como estadístico para calcular los errores de estimación que poseen los modelos.

Para observar cuales son las condiciones óptimas bajo las cuales se debería estimar el modelo VARX, se realiza todo el análisis sobre componentes estacionales, rezagos óptimos y pruebas de cointegración, tal como se realizó en la Subsección 4.3.1.1 para el caso de simulación del efecto del COVID-19.

Con esto en mente, se presenta en la Figura 4.4 el comportamiento de las series de la demanda de energía eléctrica y PIB, para el periodo 2009-1 a 2019-8 con el fin de observar si las series originales de estas variables poseen comportamientos estacionales bien definidos.

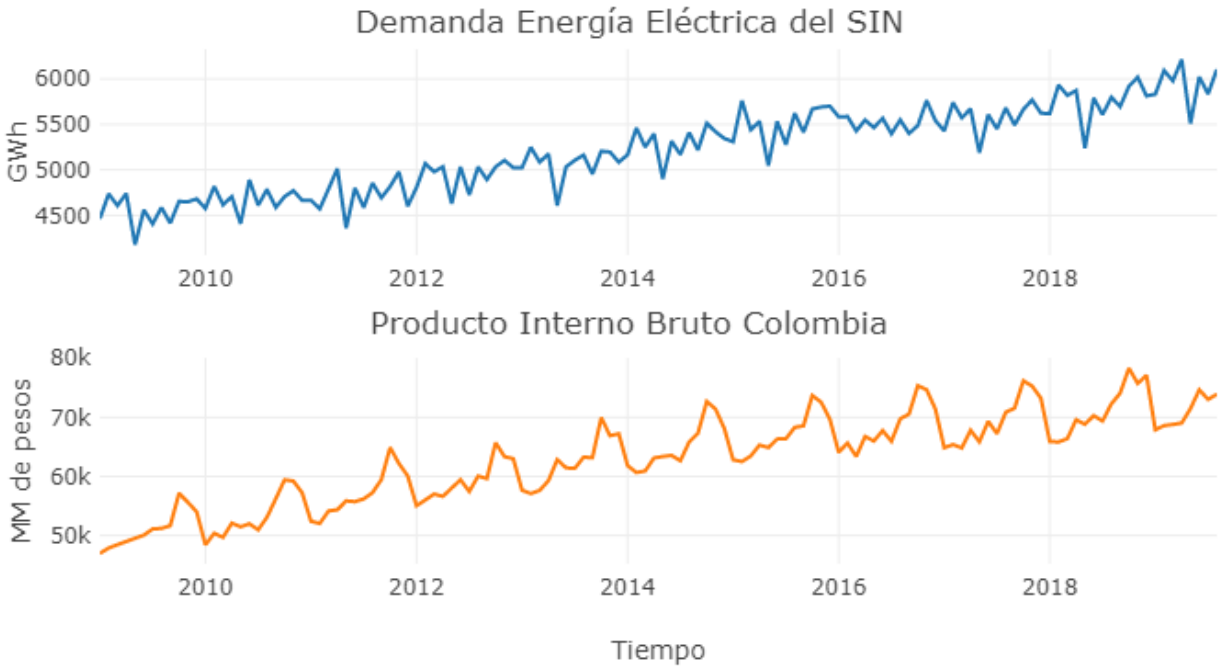

Figura 4.4: Serie original la demanda de energía eléctrica y PIB para el periodo 2009-1 a 2019-8

De la Figura 4.4 se evidencia que a pesar de que no se realiza un proceso de simulación como en la Subsección 4.3.1.1, se observa que tanto al serie de la demanda de energía eléctrica como el PIB exhiben comportamientos estacionales marcados que se repiten de forma anual. Dicho comportamiento puede corroborarse mediante el cálculo de las ACF para las dos variables. La ACF para la demanda de energía eléctrica y el PIB se presentan en la Figura 4.5.

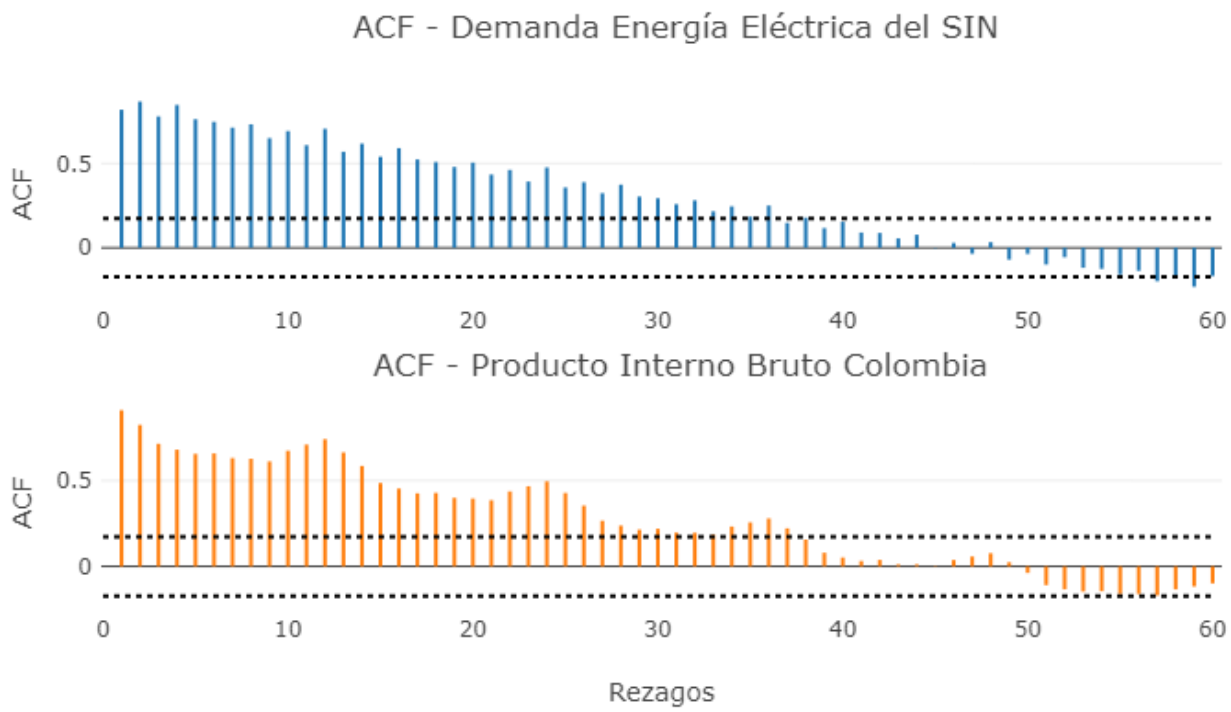

Figura 4.5: Función de autocorrelación para la demanda de energía eléctrica y PIB, para un total de 60 rezagos.

En la parte superior de la Figura 4.5 se presenta la ACF para la demanda de energía eléctrica, en donde observa que a pesar de que cada 2 rezagos se nota un incremento en su ACF, también se evidencia que cada 12 rezagos hay un incremento superior que en rezagos previos, por lo cual se podría pensar sobre la existencia de un componente estacional de frecuencia anual para dicha serie. Por su parte en el caso del PIB, se evidencia de forma más plena la existencia de un componente estacional anual para la serie, pues se observa un comportamiento cíclico que aumenta cada 12 rezagos.

Dada la existencia de componentes estacionales para estas dos series, las cuales se usarían como variables endógenas del modelo VARX, se hace necesario verificar la existencia de raíces unitarias estacionales y no estacionales que poseen las series, por lo cual se realiza la prueba HEGY debido a que permite identificar la existencia de raíces unitarias para diferentes frecuencias, tal como se expone en el Anexo B. Los resultados obtenidos en la prueba HEGY para estas dos series se presenta en el Cuadro 4.6.

Prueba HEGY Energía Eléctrica				Prueba HEGY PIB			
	Estadístico	P-valor			Estadístico	P-valor	
t_1	−2.7642	0.1382		t_1	−1.9055	0.5533	
t_2	−0.7020	0.3252		t_2	−0.4869	0.4045	
$F_{3:4}$	0.7904	0.4611		$F_{3:4}$	0.2916	0.7369	
$F_{5:6}$	2.6319	0.0817	.	$F_{5:6}$	0.2962	0.7336	
$F_{7:8}$	0.4671	0.6270		$F_{7:8}$	2.6883	0.0749	.
$F_{9:10}$	7.7442	0.0001	***	$F_{9:10}$	2.1322	0.1269	
$F_{11:12}$	0.0394	0.9599		$F_{11:12}$	1.8068	0.1749	
$F_{2:12}$	2.3491	0.1674		$F_{2:12}$	1.3852	0.2927	
$F_{1:12}$	2.8185	0.0000	***	$F_{1:12}$	1.6028	0.0000	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cuadro 4.6: Prueba HEGY demanda energía eléctrica y PIB

Del resultado obtenido en el Cuadro 4.6, y los juegos de hipótesis establecidos en el Anexo B, se tendrá que al no rechazar las hipótesis para las pruebas t_1 y $F_{11:12}$ se tendrá que tanto para la demanda de energía eléctrica como para el PIB, se concluye que existen tanto raíces unitarias no estacionales, como raíces unitarias estacionales de frecuencia anual.

Para encontrar el número óptimo de rezagos que se deben emplear para realizar el ajuste del modelo VARX, se usa la función de optimización descrita en la Subsección 4.3.1.1, la cual entrega el número óptimo de rezagos a partir de cuatro criterios de información, a saber, AIC, HQ, SC y FPE. Los resultados obtenidos por la función de optimización se presenta en el Cuadro 4.7.

Número óptimo de rezagos modelo VARX(p)			
AIC(n)	HQ(n)	SC(n)	FPE(n)
4	3	2	4

Cuadro 4.7: Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p)

Del Cuadro 4.7 se observa que de los cuatro criterios de información calculados, el AIC y el FPE sugieren que el número óptimo de rezagos a emplear para las variables endógenas del modelo VARX es de 4 rezagos, por lo cual se asumirá dicho número de rezagos durante los procesos de estimación tanto del modelo VARX como para los modelos alternativos.

Para concluir el análisis del modelo VARX, se aplica la prueba secuencial de Johansen para comprobar si existen relaciones de cointegración entre las variables endógenas del modelo, con el objetivo de verificar si el modelo estimado debe ser VARX o si por el contrario debe ser un VEC. Los resultados obtenidos por la prueba de Johansen con traza se resumen en el Cuadro 4.8.

Prueba Johansen con traza

Hipótesis	Estadístico	VC-10 %	VC-5 %	VC-1 %
$r \leq 1$	4.96	7.52	9.24	12.97
$r = 0$	49.60	17.85	19.96	24.60

Cuadro 4.8: Prueba de Johansen con traza para las variables de la demanda de energía eléctrica y el PIB

De los resultados obtenidos por la prueba secuencial de Johansen se evidencia que para el caso de la hipótesis nula de no existencia de relaciones de cointegración entre las variables, $r = 0$, se encuentra que el estadístico de prueba está por encima de los tres valores críticos del 1 %, 5 % y 10 %, por lo cual se rechaza la hipótesis nula y se concluye que hay al menos una relación de cointegración. Dicho resultado se corrobora cuando se observa la segunda prueba de hipótesis, en donde se quiere probar la existencia de a lo más una relación de cointegración $r \leq 1$, en donde se observa que el estadístico de prueba cae por debajo de los valores críticos 1 %, 5 % y 10 %, lo cual no puede rechazarse la hipótesis nula, y se concluye que existe a lo más una relación de cointegración entre las variables endógenas del modelo.

Dados los resultados obtenidos en el Cuadro 4.7 y Cuadro 4.8 significa que el modelo adecuado para estimar la relación entre las variables endógenas, será realmente un modelo VEC con un total de 4 rezagos.

El desempeño predictivo del modelo aquí definido, junto al obtenido por los modelos alternativos se presenta en la Subsección 4.3.2, con el fin de facilitar la lectura de los resultados, y poder comparar el desempeño predictivo de los modelos para las diferentes aplicaciones de forma más amigable.

4.3.1.3 Caso 3: Demanda gas natural para el sector industrial

Similar a las aplicaciones realizadas en la Subsección 4.3.1.1 y Subsección 4.3.1.2 para la demanda de energía eléctrica, esta sección plantea las condiciones bajo las cuales se estimarán los modelos MLR, GAM, LASSO, MARS y VARX para la variable de demanda de gas natural.

Con el objetivo de realizar una aplicación que ilustre el desempeño predictivo de los modelos para el caso de la demanda gas natural, se decide seleccionar como variables explicativas algunas de las variables presentadas en el Cuadro 4.2, tales como el efecto calendario, días laborales por mes y PIB, debido a que se pretende usar aquellas variables que emplea la UPME en sus ejercicios de pronóstico, junto con algunas variables que son usadas en la literatura y de las cuales se tiene disponibilidad para el caso colombiano.

Es de anotar que la aplicación que se realiza en esta subsección es similar a la realizada en la Subsección 4.3.1.2, con la diferencia de que en este caso el PIB ingresa al modelo VARX como variable exógena para el pronóstico de la demanda de gas natural; y por tanto, se quiere observar si la inclusión de estas variables es suficiente para reflejar la caída que tuvo la demanda

de gas natural del sector industrial causada por el efecto COVID-19.

Dado lo anterior, se emplea la información reportada para las variables en el periodo que va desde 2009-1 hasta 2021-08, para dividirlas en dos grupos. El primer grupo esta conformado por la información reportada entre 2009-1 y 2019-8, y será usada para realizar el ajuste de los modelos. El segundo grupo contiene la información reportada para los últimos 24 meses (2019-9 y 2021-8) y se usará para comparar los pronósticos realizados por los diferentes modelos con respecto a la información real reportada, y así poder observar el desempeño predictivo de estos ante la aparición del COVID-19.

En la Figura 4.6 se presenta el comportamiento que han tenido las series de la demanda de gas natural para el sector industrial y la serie del IPG, en donde se busca verificar si estas series exhiben un comportamiento estacional durante el periodo 2009-1 y 2019-8.

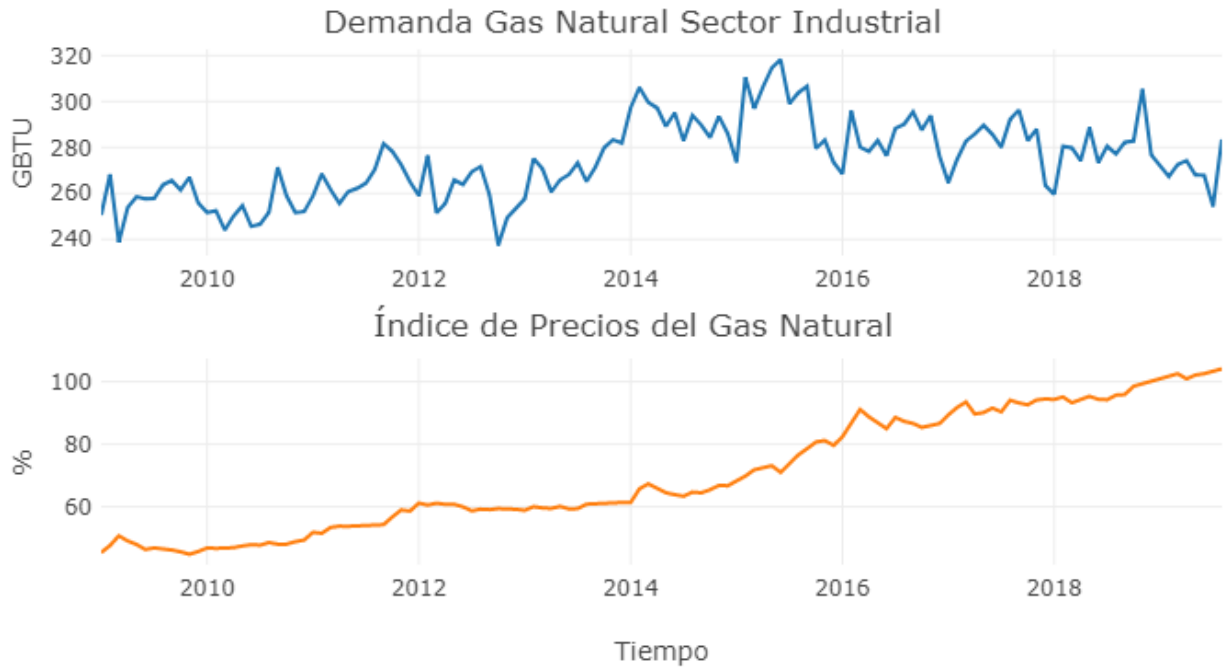

Figura 4.6: Serie de la demanda de gas natural del sector industrial y del IPG para el periodo 2009-1 y 2019-8

De la Figura 4.6 se observa que la demanda de gas natural del sector industrial pareciera tener un componente estacional, debido a que se logra identificar con facilidad un comportamiento que se repite periódicamente en la serie. Mientras que, en el caso del IPG se puede afirmar con certeza que esta serie no posee un componente estacional, debido a que la serie solo tiene un comportamiento de tendencia creciente sin ningún patrón visible.

Dado que no es posible negar la existencia de un patrón estacional en la serie de la demanda de gas natural para el sector industrial, se decide plantear la ACF para las dos series temporales para poder tener un diagnóstico más acertado. La ACF para la demanda de gas natural para el sector industrial y para el IPG se presenta en la Figura 4.7.

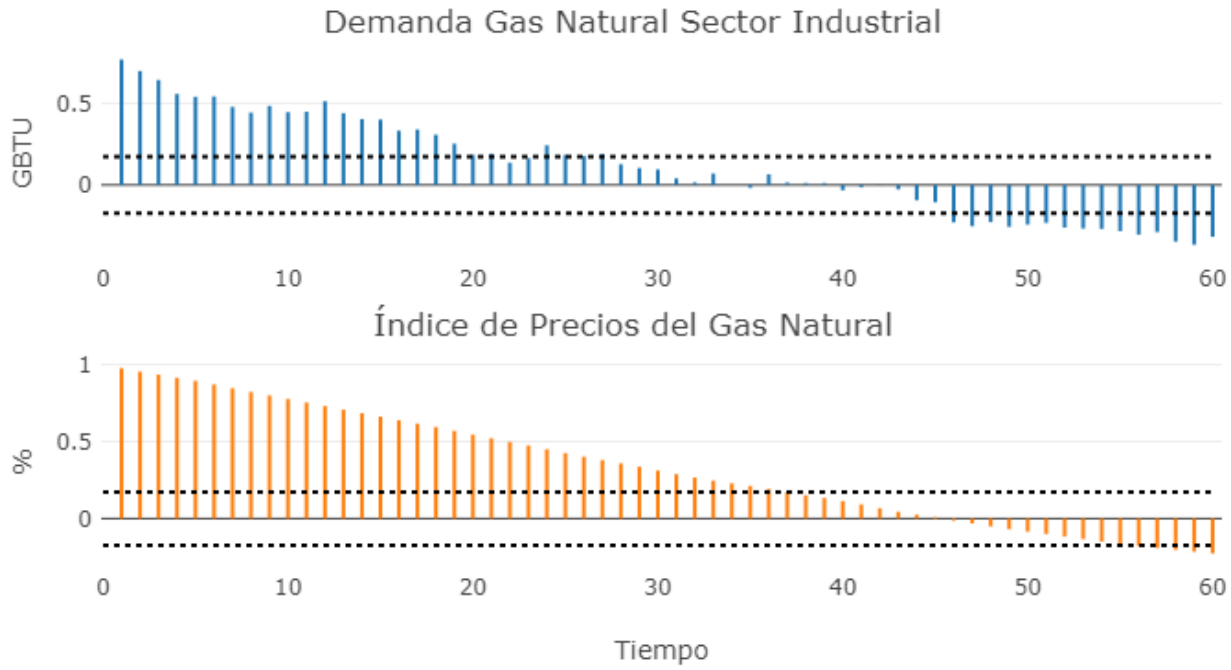

Figura 4.7: Función de autocorrelación para la demanda de gas natural del sector industrial y el IPG, para un total de 60 rezagos.

Aunque se esperaba, en la Figura 4.7 se observa que la ACF para la demanda de gas natural del sector industrial presenta decaimiento seguido por un incremento en el rezago 12 y 24, por lo cual se tendrá que esta serie sí posee un comportamiento estacional con periodicidad anual. Por otro lado, en la misma figura se observa que la ACF para el IPG posee un decaimiento uniforme sin ningún tipo de incremento, por lo cual se concluye que esta serie no posee un componente estacional.

Dado que la demanda de gas natural industrial posee un componente estacional, se hace necesario verificar si dicha variable posee raíces unitarias estacionales y no estacionales. Mientras que, a pesar de que la serie del IPG no posea un componente estacional, se tiene que éste posee una tendencia marcada, por lo cual se requiere probar la existencia de raíces unitarias no estacionales.

Con el fin de probar la existencia de las raíces unitarias estacionales y no estacionales para las dos variables de interés, se decide aplicar la prueba HEGY, obteniendo los resultados registrados en el Cuadro 4.9

Prueba HEGY Demanda Gas Natural del Sector Industrial			
	Estadístico	P-valor	
t_1	−1.2444	0.9257	
t_2	−2.2959	0.0255	*
$F_{3:4}$	5.7049	0.0034	**
$F_{5:6}$	4.2513	0.0138	*
$F_{7:8}$	6.6405	0.0014	**
$F_{9:10}$	2.3372	0.0920	.
$F_{11:12}$	3.4113	0.0319	*
$F_{2:12}$	4.5531	0.1864	
$F_{1:12}$	4.3729	0.0295	*

Prueba HEGY IPG			
	Estadístico	P-valor	
t_1	−2.3035	0.4354	
t_2	−2.2203	0.0304	*
$F_{3:4}$	5.4512	0.0043	**
$F_{5:6}$	5.9221	0.0027	**
$F_{7:8}$	7.0100	0.0010	**
$F_{9:10}$	6.2528	0.0020	**
$F_{11:12}$	8.1052	0.0004	***
$F_{2:12}$	7.3118	0.1189	
$F_{1:12}$	7.7227	0.0000	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cuadro 4.9: Prueba HEGY demanda gas natural agregada e IPG

En la parte izquierda del Cuadro 4.9 se presenta la prueba de HEGY para la demanda de gas natural para el sector industrial, y se observa que la variable no posee raíces unitarias estacionales debido a que se rechaza la hipótesis nula para $F_{11:12}$, pero se tiene que la serie si posee raíces unitarias no estacionales, debido a que no se rechaza la hipótesis nula para t_1. Sin embargo, como el rezago 12 de la ACF de la demanda de gas industrial indica efectos estacionales, y en la prueba HEGY se observa que no existen raíces estacionales anuales, quiere decir que los efectos estacionales de la demanda de gas natural para el sector industrial son estacionarios.

Por su parte, en la parte derecha del Cuadro 4.9, se presenta la prueba HEGY para el IPG, y se evidencia que la serie no posee raíces unitarias estacionales con periodicidad anual debido a que rechaza la hipótesis nula para el estadístico $F_{11:12}$, pero si se puede concluir que posee raíces unitarias no estacionales debido a que el estadístico t_1 resultó ser no significativo.

El siguiente paso que debe desarrollarse para poder realizar el ajuste del modelo VARX, es encontrar el número óptimo de rezagos que deben incluirse para las variables endógenas del modelo. Por ello, se decide aplicar una función de optimización que indica cuál es el número óptimo de rezagos que deben considerarse para las variables endógenas del modelo VARX, de tal forma que la selección de los rezagos se base en criterios de información. El cálculo de la función de optimización para el modelo de gas natural para el sector industrial se presenta en el Cuadro 4.10

Número óptimo de rezagos modelo VARX(p)			
AIC(n)	HQ(n)	SC(n)	FPE(n)
1	1	1	1

Cuadro 4.10: Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p)

Encontrando que el número óptimo de rezagos para el modelo VARX de la demanda de gas del sector industrial, según los cuatro criterios de información es de 1 rezago, y por tanto se realizará el ajuste de los modelos con esta cantidad de rezagos.

Finalmente, el paso restante será probar si existen relaciones de cointegración entre la demanda de gas natural industrial respecto al IPG, y para ello se emplea la prueba de Johansen con traza, obteniendo los resultados registrados en el Cuadro 4.11.

Prueba Johansen con traza				
Sector Industrial				
Hipótesis	Estadístico	VC-10 %	VC-5 %	VC-1 %
$r \leq 1$	3.89	6.50	8.18	11.65
$r = 0$	16.59	15.66	17.95	23.52

Cuadro 4.11: Prueba de Johansen con traza para las variables de la demanda de gas natural para el sector industrial y el IPG

Del Cuadro 4.11 puede concluirse que no existen relaciones de cointegración entre la demanda de gas natural para el sector industrial y el IPG, debido a que el estadístico calculado para la hipótesis $r = 0$, cae por debajo del valor crítico del 5 % y y el 1 %, por lo cual se concluye que no hay relaciones cointegrantes entre las variables endógenas del modelo VARX. Dados los resultados anteriores se tendrá que para el caso de la demanda de gas natural del sector industrial el modelo a estimar será un VARX con un solo rezago para las variables endógenas.

Basados en los resultados obtenidos en el Cuadro 4.10 y el Cuadro 4.11 se plantea en la Subsección 4.3.2, el desempeño predictivo obtenido por el modelo VARX junto al obtenido con los modelos alternativos MLR, GAM, LASSO y MARS, empleando para ello, las mismas variables y los mismos rezagos que se identificaron como óptimos en el proceso de diagnóstico.

4.3.1.4 Caso 4: Demanda gas natural agregada escenario base

Del mismo modo como se realizó la aplicación en la Subsección 4.3.1.3 sobre demanda de gas natural para el sector industrial, se decide realizar una aplicación adicional considerando la demanda de gas natural agregada, en donde el objetivo será encontrar cuales son las condiciones óptimas bajo las cuales deberá estimarlse el modelo VARX, para comparar posteriormente su desempeño predictivo frente a los modelos alternativos MLR, GAM, LASSO y MARS.

La finalidad de la aplicación es estimar los cinco modelos planteados bajo las mismas condiciones óptimas que tendrá el modelo VARX, empleando como variables endógenas la demanda de gas natural agregada, el IPG y como variables exógenas los días del mes, los días domingos, los días laborales, el efecto calendario y el PIB, para tener un punto de comparación al emplear las mismas variables y las mismas condiciones óptimas bajo las cuales se deberían estimar el modelo VARX.

A diferencia del modelo presentado en la Subsección 4.3.1.3, se decide en este caso solo emplear la información contenida en el periodo 2009-1 a 2019-12, para mostrar cuál es el desempeño predictivo de los cinco modelos en una situación en la cual hay ausencia total de los efectos generados por la pandemia del COVID-19.

Por ello, se realiza entonces la división de la base de datos de la siguiente manera: Para el proceso de estimación y prueba del desempeño de los modelos, se toma la información reportada para cada variable entre 2009-1 y 2017-12, mientras que la información reportada entre 2018-1 y 2019-12 se emplea para medir el desempeño predictivo de los modelos estimados en ausencia del COVID-19.

Es de anotar que dado que no se realizó ningún tipo de modificación al comportamiento original de las variables, se decide solo analizar la variable de demanda de gas natural agregada, debido a que los resultados obtenidos por la variable IPG que sería la otra variable endógena del modelo VARX, son similares a los presentados en la Subsección 4.3.1.3, y en consecuencia las conclusiones obtenidas para ésta serán las mismas.

En este sentido, se presenta en la Figura 4.8 el comportamiento de la serie temporal de la demanda agregada de gas natural para el periodo 2009-1 a 2017-12, junto a su ACF con el fin de verificar si dicha variable presenta un comportamiento estacional.

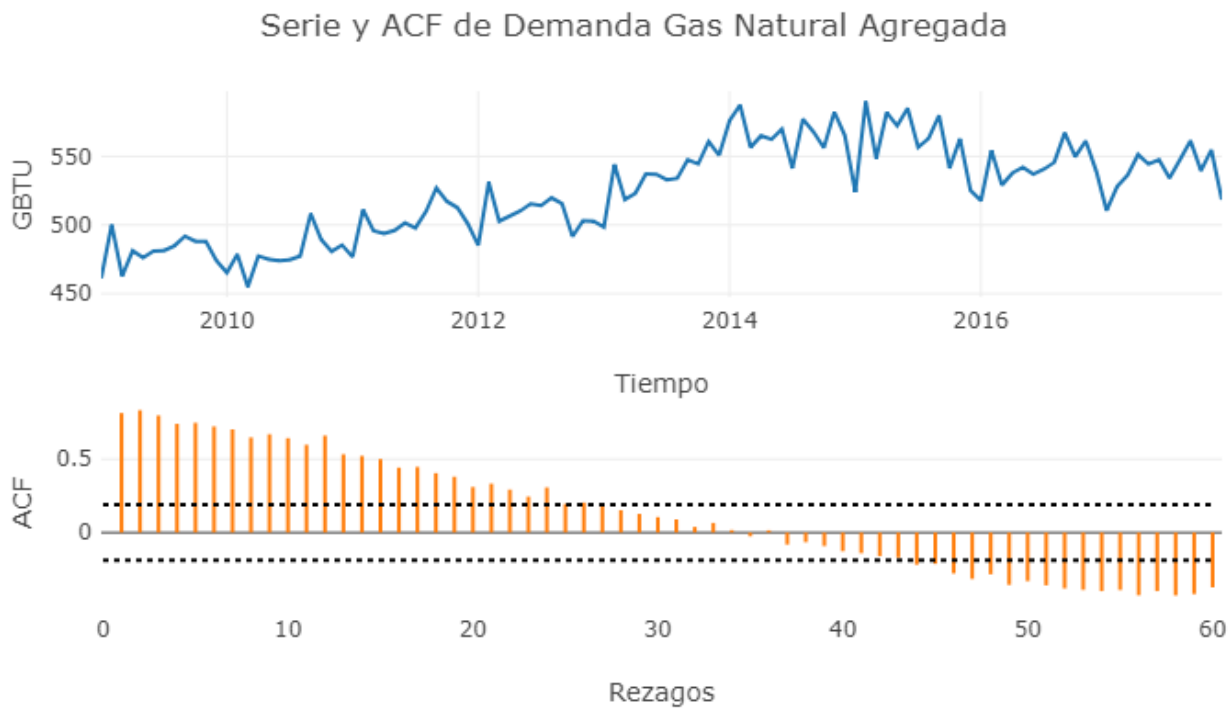

Figura 4.8: Serie y ACF de la demanda de gas natural agregada para el periodo 2009-1 y 2017-12

En la parte superior de la Figura 4.8 se presenta el comportamiento de la la demanda agregada de gas natural, en donde se evidencia que no es posible garantizar que exista o no un comportamiento estacional para la variable, ya que no se identifican patrones repetitivos definidos en la serie.

Por su parte en la parte inferior de la Figura 4.8 se muestra la ACF de la serie, en la cual se observa que si es posible afirmar que la demanda de gas natural agregada posee un comportamiento estacional de frecuencia anual, ya que al observar la serie se evidencia un pico que sobresale cada 12 rezagos.

A pesar de que el comportamiento estacional que presenta la variable no es muy marcado, es necesario verificar si esta variable exhibe raíces unitarias estacionales y no estacionales. Para ello se realiza la prueba HEGY ilustrando los resultados obtenidos en el Cuadro 4.12.

Prueba HEGY Demanda Gas Natural Agregada

	Estadístico	P-valor
t_1	−1.1826	1.0000
t_2	−2.4814	1.0000
$F_{3:4}$	6.5667	0.9183
$F_{5:6}$	6.9009	0.9084
$F_{7:8}$	13.4681	0.3969
$F_{9:10}$	8.4900	0.8328
$F_{11:12}$	8.0267	0.8616
$F_{2:12}$	19.7855	0.3711
$F_{1:12}$	18.5120	0.4678

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cuadro 4.12: Prueba HEGY demanda gas natural agregada

Del resultado presentado en el Cuadro 4.12 se observa que ninguna de las hipótesis establecidas en la prueba HEGY son rechazadas para la demanda de gas natural agregada, por lo cual se puede concluir que esta variable posee tanto raíces unitarias estacionales como raíces unitarias no estacionales.

El siguiente paso es encontrar el número óptimo de rezagos del modelo VARX, y para ello se emplea una función de optimización para identificar por diferentes criterios cuál es el número óptimo de rezagos que deben considerarse para las variables endógenas del modelo. Los resultados obtenidos por la función de optimización se presentan en el Cuadro 4.13

Número óptimo de rezagos modelo VARX(p)

AIC(n)	HQ(n)	SC(n)	FPE(n)
24	1	1	1

Cuadro 4.13: Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p)

En el Cuadro 4.13 se encuentra que tres de los cuatro criterios de información calculados concluyen que el numero óptimo de rezagos para estimar el modelo VARX(p) es de un solo rezago, encontrando que dicho resultado es similar al encontrado en el Caso 3 para la demanda de gas natural industrial.

Ahora, para probar si existen relaciones de cointegración entre la demanda de gas natural agregada y el IPG, se decide realizar la prueba de Johansen con traza, con el fin de decidir si debe estimarse en este caso un modelo VARX o un modelo VEC. Los resultados encontrando por la prueba de Johansen se presentan en el Cuadro 4.14.

Prueba Johansen con traza

Hipótesis	Estadístico	VC-10 %	VC-5 %	VC-1 %
$r \leq 1$	2.17	7.52	9.24	12.97
$r = 0$	33.66	17.85	19.96	24.60

Cuadro 4.14: Prueba de Johansen con traza para las variables de la demanda de gas natural agregada y el IPG

De los resultados encontrados en el Cuadro 4.14 se rechaza la hipótesis de que no hay relaciones de cointegración entre las variables de demanda de gas natural y el IPG, debido a que el estadístico de prueba calculad es mayor a los diferentes valores críticos del 10 %, 5 % y 1 %. Por su parte, para la hipótesis de que existe a lo más una relación de cointegración se encuentra que el estadístico de prueba se encuentra por debajo de los valores críticos, concluyendo que existe una relación de cointegración entre las variables endógenas del modelo, por lo cual se concluye que debe estimarse un modelo VEC en lugar de un modelo VARX.

Similar que los demás casos, se presenta en la Subsección 4.3.2 el reporte del desempeño predictivo del modelo VEC, junto con los modelos alternativos MLR, GAM, LASSO y MARS.

4.3.1.5 Caso 5: Demanda gas natural agregada al incluir efecto fenómeno del niño

En este escenario se decide replicar el Caso 4 presentado en la Subsección 4.3.1.4 pero con la diferencia de que en este caso se introduce como variable exógena del modelo el efecto del fenómeno del Niño, y para ello se introducen las temperaturas registradas en el pacífico tropical central y oriental que sean mayores o iguales a 0.5^oC, con el fin de observar si este fenómeno puede o no mejorar el desempeño predictivo de los modelos.

La razón por la cual se decide incluir en el Caso 5 el fenómeno del Niño dentro de las estimaciones de gas natural se debe a que la ocurrencia de este fenómeno genera escasez de recursos hídricos para las plantas de generación hidroeléctrica, lo cual puede crear la necesidad de generar energía en el país a partir de las plantas de gas natural. En otras palabras, se tendrá que el consumo de gas en el sector termoeléctrico se verá afectado ante la presencia de tal fenómeno meteorológico. Por tal razón, es que se decide en esta aplicación incluir un análisis adicional en donde se considera el ONI (*Oceanic Niño Index*) como variable explicativa para los modelos de la demanda gas natural agregada. Los datos históricos para este ejercicio fueron tomados de NOAA (2021).

Es de anotar que para esta aplicación, se obtuvieron resultados similares a los presentados en el Cuadro 4.12, Cuadro 4.13 y Cuadro 4.14 y la Figura 4.8, por lo cual no se presentan en esta caso dichos análisis, debido a que sería información redundante.

El reporte de los resultados obtenidos para el desempeño predictivo de los diferentes modelos se presenta en la Subsección 4.3.2, con el objetivo de realizar el análisis de los resultados y

poder compararlos respecto a los obtenidos en el Caso 4, en el cual no se considera la inclusión del efecto del niño durante los proceso de estimación.

4.3.1.6 Casos 6 y 7: Demanda gas natural agregada al incluir precios del gas natural usuarios regulados y no regulados

En los casos 6 y 7, se replica el ejercicio realizado en el Caso 4 presentado en la Subsección 4.3.1.4, con la diferencia de que en este caso se introducen las variables precios promedio del gas natural para usuarios regulados (Caso 6) y para usuarios no regulados (Caso 7) entre las variables endógenas del modelo VARX, con el fin de observar si la inclusión de las variables de precio del gas natural mejoran el desempeño predictivo de los modelos.

La razón por la cual se decide incluir en estos casos los precios promedio del gas natural a los cálculos, se debe al principio básico sobre el que se basa la economía de mercado, también conocida como la ley de la oferta y la demanda, la cual establece que manteniendo lo demás constante, la cantidad demandada de un bien disminuye cuando el precio de dicho bien aumenta, mientras que la oferta de un bien aumenta cuando el precio de dicho bien aumenta, lo cual generará que existirá una relación inversa entre la oferta y la demanda, o en este caso entre el precio del gas natural y la demanda del gas natural.

Dado que en estos dos caso se incluye la variable del precio promedio de gas natural para usuarios regulados (Caso 6) y usuarios no regulados (Caso 7), como variable endógena del modelo VARX, será necesario realizar los mismos análisis planteados previamente en la Subsección 4.3.1.4 para estas dos variables.

En la Figura 4.9 el gráfico de la serie de los precios promedio del gas natural regulado y no regulado, con le fin de buscar si éstas poseen comportamientos estacionales.

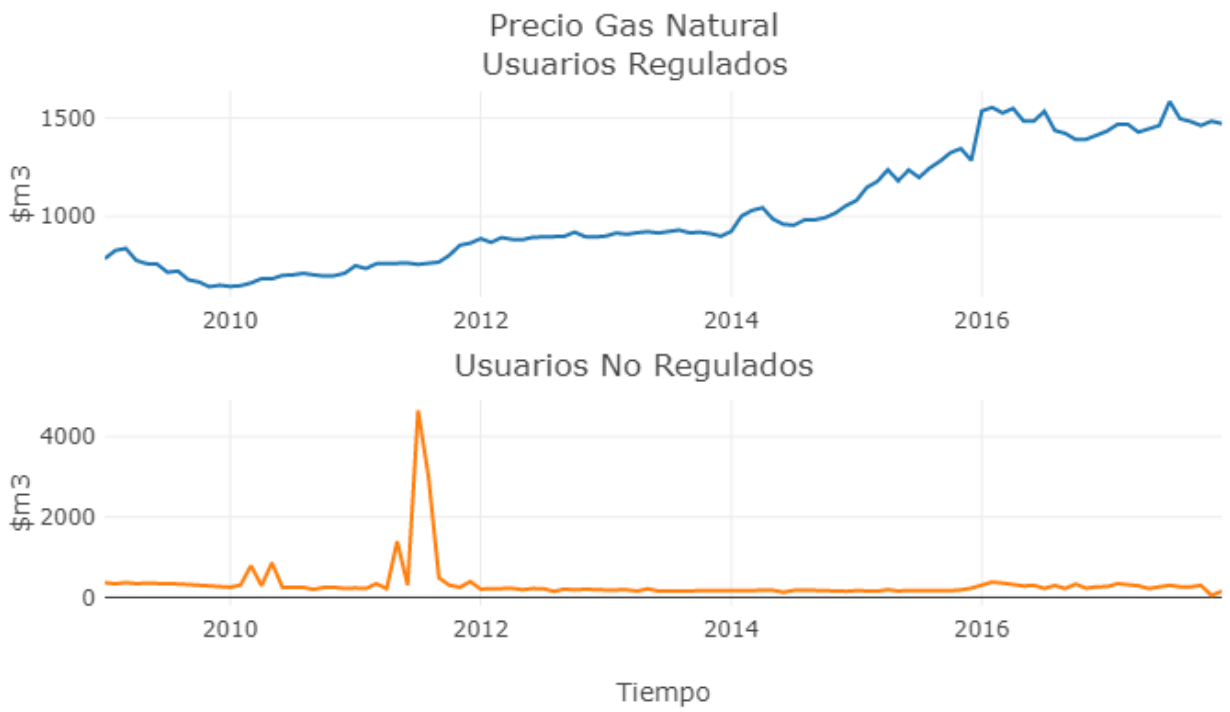

Figura 4.9: Serie precios promedio de gas natural para usuarios regulados y usuarios no regulados para el periodo 2009-1 y 2017-12

De la Figura 4.9 se observa que ninguna de las dos variables de precios promedio, exhibe comportamientos estacionales definidos, ya que ninguna muestra patrones repetitivos. Es de anotar que la variable de precios promedio de gas natural para usuarios regulados muestran una tendencia creciente a través del tiempo, mientras que los precios promedio de gas natural para usuarios no regulados exhiben un comportamiento relativamente constante con picos en los años 2010 y 2011, de los cuales se debería revisar la información reportada durante el año 2011, ya que presenta valores extremos que ascienden hasta $4659 en Julio de 2011.

Para realizar un análisis más formal sobre la no existencia de un comportamiento estacional de los precios promedio del gas natural para usuarios regulados y no regulados, se presenta la ACF de ambas series en la Figura 4.10.

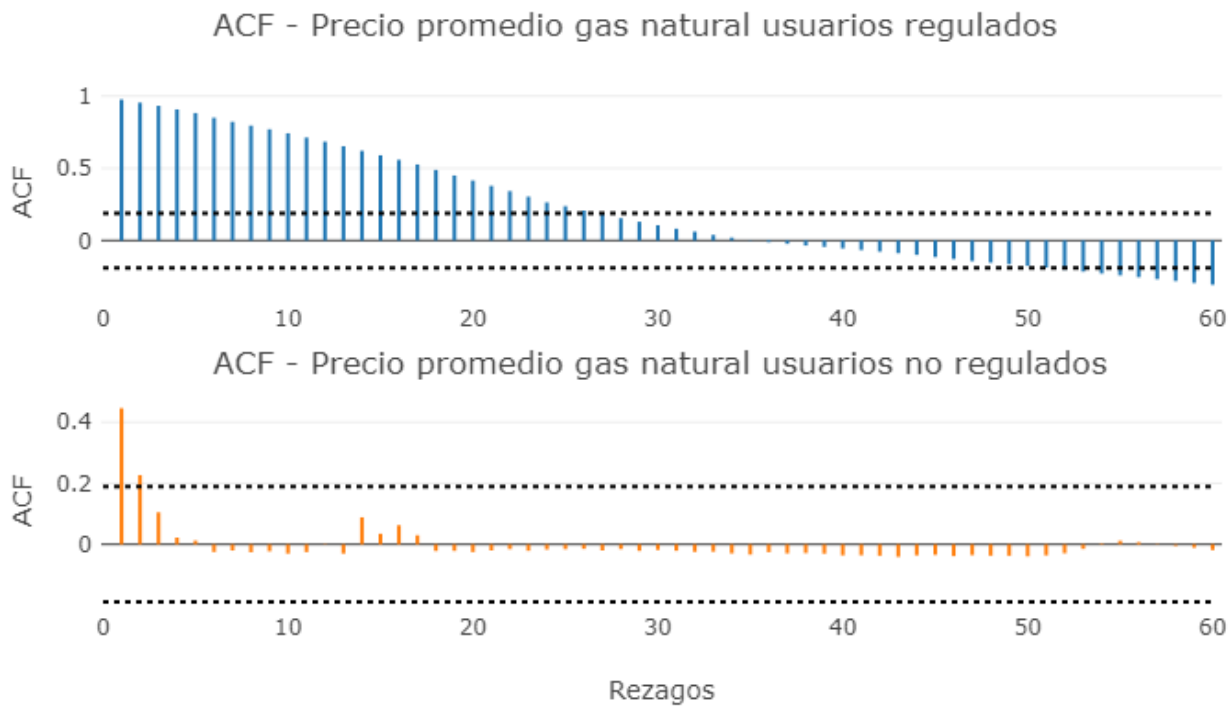

Figura 4.10: ACF precios promedio de gas natural para usuarios regulados y usuarios no regulados para el periodo 2009-1 y 2017-12

Como se esperaba, en la Figura 4.10 se corrobora la creencia sobre la no existencia de patrones estacionales para las variables de precios promedio de gas natural, tanto para usuarios regulados como para usuarios no regulados, dado que no se observa ningún pico en los rezagos estacionales de la ACF. Adicionalmente, puede observarse en la Figura 4.10 que los precios promedio para usuarios regulado presentan una caída uniforme en su ACF cae uniformemente, lo cual puede indicar existencia de raíces unitarias no estacionales, debido a la tendencia creciente que posee la serie temporal.

Para probar la existencia de raíces unitarias no estacionales para las dos variables de precios promedio, se podría realizar la prueba aumentada de Dickey-Fuller (ADF), pero con el fin de ser consistentes con las pruebas presentadas en las otras subsecciones, se decide realiza la prueba HEGY, la cual además de probar la existencia de raíces unitarias estacionales, también prueba la existencia de raíces unitarias no estacionales, por lo que podría considerarse como una prueba mejorada respecto a la ADF. Los resultados obtenidos por la prueba HEGY se presentan en el Cuadro 4.15.

Prueba HEGY precio promedio del Gas Natural Usuarios Regulados			
	Estadístico	P-valor	
t_1	−2.2244	0.9872	
t_2	−1.8503	0.1354	
$F_{3:4}$	2.4472	0.0804	.
$F_{5:6}$	2.8502	0.0574	.
$F_{7:8}$	4.6305	0.0093	**
$F_{9:10}$	4.2427	0.0139	*
$F_{11:12}$	4.8054	0.0076	**
$F_{2:12}$	4.0708	0.2780	
$F_{1:12}$	4.5209	0.0000	***

Prueba HEGY precio promedio del Gas Natural Usuarios No Regulados			
	Estadístico	P-valor	
t_1	−2.5389	0.9924	
t_2	−2.7853	0.0367	*
$F_{3:4}$	6.2478	0.0025	**
$F_{5:6}$	7.2159	0.0013	**
$F_{7:8}$	7.2477	0.0013	**
$F_{9:10}$	7.2001	0.0013	**
$F_{11:12}$	6.8943	0.0016	**
$F_{2:12}$	11.0441	0.2108	
$F_{1:12}$	10.4293	0.1155	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cuadro 4.15: Prueba HEGY precio promedio gas natural usuarios regulados y no regulados

Del Cuadro 4.15 se evidencia que las dos series de precios promedio no rechazan la hipótesis nula de la prueba t_1, por lo cual se concluye que existen raíces unitarias no estacionales tanto para el caso de usuarios regulados, como para usuarios no regulados. Adicionalmente, se observa que para el caso de usuarios regulados no se rechaza la hipótesis nula de la prueba t_2, por lo que se puede concluir que la serie de precio promedio para usuarios regulados posee raíces unitarias estacionales de periodicidad bimensual.

Para encontrar el número de rezagos que se deben usar en el cálculo de los modelos VARX, al agregar los de precios promedio del gas natural como variable endógena, se emplea la función de optimización para identificar por diferentes criterios cuál es el número óptimo de rezagos que se deben emplear en el ajuste del modelo VARX, obteniendo los resultados presentados en el Cuadro 4.16.

Número óptimo de rezagos modelo VARX(p)

Caso Estudio	AIC(n)	HQ(n)	SC(n)	FPE(n)
Caso 6	12	2	1	2
Caso 7	1	1	1	1

Cuadro 4.16: Criterios de información para identificar el número de rezagos óptimo para el modelo VARX(p)

De los resultados presentados en el Cuadro 4.16 se encuentra que el número óptimo de rezagos que sugieren los criterios de información presentados que el caso de usuarios regulados (Caso 6) es de 2, debido a que dos de los cuatro criterios calculados sugiere emplear 2 reazagos, mientras que en el caso de usuarios no regulados (Caso 7) los cuatro criterios de información sugieren emplear 1 rezago como el óptimo para el cálculo del modelo VARX.

Finalmente, se realiza las respectivas pruebas de cointegración para las variables endógenas usadas en cada caso, con el fin de observar cuál es el número de relaciones de cointegración existentes entre las tres variables endógenas del modelo VARX. Para tal propósito se decide emplear la prueba secuencial de Johansen, tal como se hizo en casos anteriores, obteniendo los resultados presentados en el Cuadro 4.17.

Prueba Johansen con traza Caso 6

Hipótesis	Estadístico	VC-10 %	VC-5 %	VC-1 %
$r \leq 2$	4.34	7.52	9.24	12.97
$r \leq 1$	17.64	17.85	19.96	24.60
$r = 0$	51.95	32.00	34.91	41.07

Prueba Johansen con traza Caso 7

Hipótesis	Estadístico	VC-10 %	VC-5 %	VC-1 %
$r \leq 2$	2.07	7.52	9.24	12.97
$r \leq 1$	29.07	17.85	19.96	24.60
$r = 0$	62.02	32.00	34.91	41.07

Cuadro 4.17: Prueba de Johansen con traza para las variables de la demanda de gas natural agregada, el IPG y el precio promedio gas natural, para usuarios regulados (Caso 6) y usuarios no regulados (Caso 7)

En la parte superior del Cuadro 4.17, se observa que la prueba secuencial de Johansen rechaza la hipótesis de que existen cero relaciones cointegrantes $r = 0$ debido a que el estadístico de prueba es superior a los valores críticos del 1 %, 5 % y 10 %, pero no logra rechazar la hipótesis sobre que existe a lo más una relación cointegrante $r \leq 1$, entre las variables de demanda de gas natural agregada, IPG y precio promedio gas natural para usuarios regulados, debido a que el estadístico de prueba es inferior a los tres valores críticos, por lo cual se concluye que el modelo adecuado para calcular en el Caso 6, será un modelo VEC con una relación de cointegración[10].

Por su parte, en la parte inferior del Cuadro 4.17, se evidencia que se rechaza tanto la hipótesis de que existen cero relaciones cointegrantes $r = 0$, y la hipótesis sobre que existe a lo más una relación cointegrante $r \leq 1$, ya que en ambos casos el estadístico de prueba es superior a los valores críticos establecidos para el 1 %, 5 % y 10 %, pero se observa que no se rechaza la hipótesis de que existen a lo más dos relaciones cointegrantes $r \leq 2$, entre las variables de demanda de gas natural agregada, IPG y precio promedio gas natural para usuarios no

[10]Es de anotar de que a pesar de que la prueba secuencial de Johansen planteada en la parte superior del Cuadro 4.17 el resultado para la hipótesis de que existen a lo más dos relaciones cointegrantes $r \leq 2$, este resultado se ignora debido a que ya se rechazó la hipótesis para $r \leq 1$, y al ser una prueba secuencia, la conclusión que debe seleccionarse será en la primera prueba en que no se rechaza la hipótesis nula.

regulados, por lo cual se concluye que el modelo adecuado para el Caso 7, será un modelo VEC con dos relaciones de cointegrantes.

Definidos los rezagos y las relaciones de cointegración para los modelos VEC para los casos 6 y 7, se realiza el cálculo de estos modelos junto con el de los modelos alternativos, y se presentan los resultados obtenidos en el Subsección 4.3.2, con el objetivo de poder comparar los resultados obtenidos para cada caso respecto a los obtenidos en el escenario base (Caso 4), en el cual no se considera la inclusión de las variables de precio del gas natural.

4.3.1.7 Casos 8 y 9: Demanda gas natural agregada al incluir precios de sustitutos, Carbón y GLP

Para los casos 8 y 9 se replican los procedimientos realizados en el Caso 4 presentado en la Subsección 4.3.1.4 pero con la diferencia de que en éstos se decide introducir el precio promedio del carbón (Caso 8) y el precio promedio del GLP para usuarios regulados (Caso 9) dentro de las variables exógenas del modelo, con el fin de observar si al introducir el precio de bienes sustitutos se puede mejorar el desempeño predictivo que tienen los modelos.

La razón de introducir dentro de los modelos como variable exógena el precio promedio del carbón y el precio promedio del GLP para usuarios regulados, se debe a que estos éstos son sustitutos en muchas situaciones del gas natural. Por ejemplo, el GLP es un compuesto químico orgánico cuya composición química se conforma mayormente por butano y propano, que se usa comúnmente en cocción de alimentos, calefacción, generación de energía, entre otros, mientras que el carbón, es un combustible fósil que se obtiene luego de realizar de una serie de transformaciones sobre restos vegetales, y también se usa para cocción de alimentos, calefacción, generación de energía, entre otros, por lo cual, al coincidir con los usos más frecuentes del gas natural, se tendrá que tanto el carbón como el GLP pueden ser usados como un remplazo del gas natural, y por tanto su precio podría influir en si se consume uno u otro.

Es de anotar, que en la aplicación del Caso 8 y 9 se obtuvieron resultados similares a los presentados en las pruebas de raíces unitarias estacionales, número óptimo de rezagos y relaciones cointegrantes que en la Subsección 4.3.1.4, por lo cual no se presentan en esta subsección dichos análisis debido a presentar estos resultados serían redundantes con los presentados en la Subsección 4.3.1.4.

Por ello, solo se decide presentar en la Subsección 4.3.2 los resultados obtenidos para el desempeño predictivo de los modelos aplicados a los casos 8 y 9, buscando comparar los resultado obtenidos para éstos respecto a los que se encontraron en el escenario base (Caso 4), en donde no se consideraban las variables de precios de sustitutos en los análisis.

4.3.2 Resultados casos de estudio

Esta subsección presenta de forma resumida los resultados obtenidos en los nueve casos de estudio presentados en la Subsección 4.3.1, para hacer más fácil la visualización del desempeño predictivo obtenido por el modelo VARX/VEC, respecto al obtenido por los modelos alternativos MRL, GAM, LASSO y MARS, además de poder visibilizar el desempeño predictivo en algunos casos donde se incluyeron variables explicativas adicionales.

Como se mencionó en subsecciones anteriores, el cálculo tanto del modelo VARX/VEC y como de los modelos alternativos MLR, GAM, LASSO y MARS, se realiza bajo las mismas condiciones, con el objetivo de hacer comparables los resultados obtenidos. Debido a esto, se realiza la estimación de los cinco modelos con las mismas variables y los mismos rezagos óptimos encontrados en cada caso de estudio, teniendo en cuenta los horizontes de tiempo de estimación y pronóstico que se establecieron en cada uno de las aplicaciones presentadas en la Subsección 4.3.1.

Es de anotar que no necesariamente los modelos alternativos deben usar los mismos rezagos óptimos encontrados para el modelo VARX/VEC, pues se tiene que en muchos casos eliminar los rezagos de las variables pueden conllevar a mejores resultados para los modelos alternativos, ya que evitan problemas asociados al sobre ajuste del modelo, lo cual puede afectar al desempeño predictivo de los modelos.

Para ilustrar el desempeño predictivo de los diferentes modelos de pronóstico de la demanda de energía eléctrica y gas natural, se realiza en cada caso el cálculo del Error Porcentual Absoluto Medio (MAPE), debido a que este indicador nos permite comparar el resultado obtenido por los pronósticos de los diferentes modelos, respecto a la información realmente reportada para la demanda del energético de interés, con el fin de evaluar lo acertado que fueron los pronósticos generados por los modelos.

Los MAPE obtenidos para los casos de estudio presentados en la Subsección 4.3.1, se resumen en el Cuadro 4.18.

MAPE(%) de modelos estimados bajo casos de estudio

Caso Estudio	p	r	Horizonte Pronóstico	Modelo				
				MRL	GAM	LASSO	MARS	VARX/VEC
Caso 1	17	1	2018-1/2019-12	8.2303	**1.5967**	2.7233	3.9344	2.5046
Caso 2	4	1	2019-9/2021-8	**2.5994**	2.6020	2.6934	3.7332	4.8241
Caso 3	1	0	2019-9/2021-8	12.7024	**11.1672**	14.8550	14.7584	13.685
Caso 4	1	1	2018-1/2019-12	2.4567	**2.0632**	2.5578	2.9398	2.0725
Caso 5	1	1	2018-1/2019-12	2.4588	**1.8236**	2.7307	1.9471	1.9180
Caso 6	2	1	2018-1/2019-12	1.8626	1.8955	**1.8614**	1.9629	2.6031
Caso 7	1	2	2018-1/2019-12	2.4690	**1.9376**	2.5579	1.9471	2.0455
Caso 8	1	1	2018-1/2019-12	2.1486	2.0507	**2.0135**	2.9395	2.7314
Caso 9	1	1	2018-1/2019-12	2.2704	**1.9742**	2.5740	2.3576	2.3258

p: Número de rezagos óptimos encontrados para la estimación del modelo VARX/VEC
r: Número de relaciones de cointegración.
Caso 1: Demanda energía eléctrica bajo escenario de simulación del COVID-19
Caso 2: Demanda energía eléctrica al incluir campaña "Apagar-Paga" y efecto del fenómeno del niño-niña
Caso 3: Demanda gas natural para el sector industrial
Caso 4: Demanda gas natural agregada escenario base
Caso 5: Demanda gas natural agregada al incluir el efecto del fenómeno del niño
Caso 6: Demanda gas natural agregada al incluir el precio promedio del gas natural para usuarios regulados
Caso 7: Demanda gas natural agregada al incluir el precio promedio del gas natural para usuarios no regulados
Caso 8: Demanda gas natural agregada al incluir el precios promedio del carbón
Caso 9: Demanda gas natural agregada al incluir el precios promedio del GLP

Cuadro 4.18: MAPE modelos de demanda de energéticos para los diferentes casos de estudio

En el Cuadro 4.18 se presenta para los casos de estudio planteados en la Subsección 4.3.1, el número óptimo de rezagos para el modelo VARX/VEC con la cual se estimaron los cinco modelos, el número de relaciones de cointegración que poseen las variables endógenas del modelo VARX/VEC, el horizonte de tiempo empleado como periodo de pronóstico y el MAPE calculado para los cinco modelos estimados.

De los resultados encontrados en el Cuadro 4.18, se destaca inicialmente el hecho de que el número óptimo de rezagos sugeridos por los criterios de información para el ajuste de los modelos VARX para la demanda de gas natural, tanto para el sector industrial como para el agregado de los sectores, fue de solo 1 o 2 rezagos.

La razón por la cual se sugieren tan pocos rezagos como óptimos para la estimación de los modelos de demanda de gas natural, puede ser debido a que las variables exógenas introducidas

en los modelos logran capturan en este caso el efecto estacional estacionario de las variables endógenas del modelo. Posiblemente las variables explicativas de efecto calendario y el PIB, serán las variables con la capacidad de recoger el efecto estacional estacionario que posee la demanda de gas natural.

La variable de efecto calendario tiene por objetivo capturar el efecto estacional estacionario que poseen las series temporales, por lo cual su inclusión puede en este caso, ser suficiente para capturar el efecto estacional que poseen las series temporales. Por su parte, como se muestra en la Subsección 4.3.1.2 la variable PIB es una variable que exhibe un comportamiento estacional claramente definido para una periodicidad anual, lo cual podría contribuir en la captura del comportamiento estacional que posee la demanda de gas natural.

Adicionalmente, en el Cuadro 4.18 se observa que el único caso en el cual no se detectó relaciones cointegrantes entre las variables endógenas del modelo VARX, fue en el Caso 3 para la demanda de gas natural industrial, debido a que al realizar la prueba secuencial de Johansen (Cuadro 4.11) se concluye que no hay relaciones cointegrantes entre la demanda de gas natural para el sector industrial y el IPG, debido a que el estadístico de prueba para $r = 0$ cae por debajo de los valores críticos de la prueba, llevando al no rechazo de la hipótesis sobre la no existencia de relaciones cointegrantes.

Por su parte, el único caso donde se observa más de una relación de cointegración es en el Caso 7, en el cual se incluye además de la demanda del gas natural agregado y el IPG, la variable de precios del gas natural para usuarios no regulados como variable endógena del modelo, y en la cual se encuentra que la prueba secuencial de Johansen (Cuadro 4.17) rechaza las hipótesis de no existencia de relaciones de cointegración $r = 0$, y de a lo más una relación de cointegración $r \leq 1$, pero que no logra rechaza la hipótesis de la existencia de a lo más dos relaciones de cointegración $r \leq 2$.

Otro resultado encontrado en el Cuadro 4.18 es el desempeño predictivo que presenta el modelo GAM en cada uno de los casos de estudio, en donde se observa que este modelo registra los mejores desempeños predictivo en casi todos los casos planteados, siendo superado en el Caso 2 por el modelo MRL y en los Casos 6 y 8 por la regresión LASSO, pero en los cuales se observa que los MAPE registrados por el GAM respecto a sus competidores son muy similares, y por tanto podría pensarse en el GAM como una alternativa viable para realizar proyecciones de la demanda de los diferentes energéticos.

Como resultado final del Cuadro 4.18 se realizan los análisis comparativos de los resultados obtenidos en el Caso 4, respecto a los Casos 5, 6, 7, 8 y 9, debido a que estos escenarios fueron estimados bajo las mismas condiciones, con la única diferencia de que en el Caso 4 se usan las variables explicativas de IPG, PIB, días del mes, días domingos, días laborales, efecto calendario, mientras que en los Casos 5, 6, 7, 8 y 9, se integra una variable adicional para observar si la inclusión de esta variable dentro del modelo de demanda de gas natural, contribuye a la mejora del desempeño predictivo de los modelos.

Es de anotar que en comparación con el escenario base, se integra en el Caso 5 el efecto del

fenómeno del niño, en el Caso 6 el precio del gas natural para usuarios regulados, en el Caso 7 el precio del gas natural para usuarios no regulados, en el Caso 8 el precio promedio del Carbón y en el Caso 9 el precio promedio del GLP para usuarios regulados.

De los resultados registrados en el Cuadro 4.18, al comparar los Casos 4 y 5 se destaca el hecho de que a pesar de que la inclusión del efecto del fenómeno del niño deteriora el desempeño predictivo de los modelos LASSO y MRL, se observa que esta variable mejora marginalmente las predicciones realizadas por los modelos GAM y VEC, en donde ambos modelos reducen su MAPE en 0.23 % y 0.15 %, respectivamente, mientras que en el caso del modelo MARS, el incluir dicha variable mejora notablemente el desempeño predictivo del modelo, ya que se logra reducir su MAPE en 0.99 %, pero que en donde dicha reducción no logra superar el MAPE registrado por los modelos GAM y VEC.

Por su parte, al comparar los Casos 4 y 6, se evidencia que a pesar de que el desempeño predictivo de los cinco modelos alternativos mejora, se logra observar un deterioro para el modelo VEC, ya que registra un incremento de su MAPE en 5.306 %. La comparación de los Casos 4 y 7, no muestra realmente cambios significativos en los modelos MRL, LASSO, MARS y VEC, debido a que los MAPE registrados en ambos casos son prácticamente los mismos, mostrando solo una leve mejora en el MAPE obtenido por el modelo GAM, con reducciones de 0.13 %.

De los casos 4 y 8, se evidencia un deterioro significativo de 0.66 % en el desempeño predictivo del modelo VEC, una mejora de 0.54 % para el LASSO y de 0.3081 % para el MRL, y valores similares para los modelos GAM y MARS. Finalmente en la comparación de los casos 4 y 9, se evidencia que todos los modelos tienen una leve mejora a excepción de los modelos LASSO y VEC que incrementa sus MAPE en 0.02 % y 0.25 %, respectivamente.

De lo resultados anteriores obtenidos por la comparación de resultados, se destaca el hecho de que la inclusión de la variable del fenómeno del niño o de los diferentes precios mejora el desempeño predictivo de uno u otro modelo, siendo la variable del fenómeno del niño la que logra registrar el MAPE más pequeño de todos los modelos, para el caso del modelo GAM, seguido por la inclusión del precio promedio del gas natural para usuarios regulados, para el caso del modelo LASSO.

En consecuencia, se podría pensar que la inclusión de estas variables dentro de las estimación de los modelos de proyección de gas natural podrían ser acertadas para mejorar un poco el desempeño predictivo de los modelos. Es de anotar que la inclusión conjunta de una o más de estas variables podría mejorar o no el desempeño predictivo de los modelos, pero dado el objetivo principal de esta sección, dicho análisis se escapa del alcance de nuestro planteamiento, debido a que la finalidad de la sección era mostrar el desempeño predictivo de los modelos alternativos bajo diferentes escenarios.

Finalmente, para concluir esta subsección, se decide presentar en la Figura 4.11, Figura 4.12, Figura 4.13, Figura 4.14, Figura 4.15, Figura 4.16, Figura 4.17, Figura 4.18 y Figura 4.19, los pronósticos obtenidos por los diferentes modelos para cada caso de estudio, en donde

el objetivo será mostrar de forma gráfica el ajuste que tiene tanto el modelo VARX/VEC como los modelos alternativos respecto a la serie original reportada para el energético de interés, durante los horizontes de pronóstico presentados en el Cuadro 4.18.

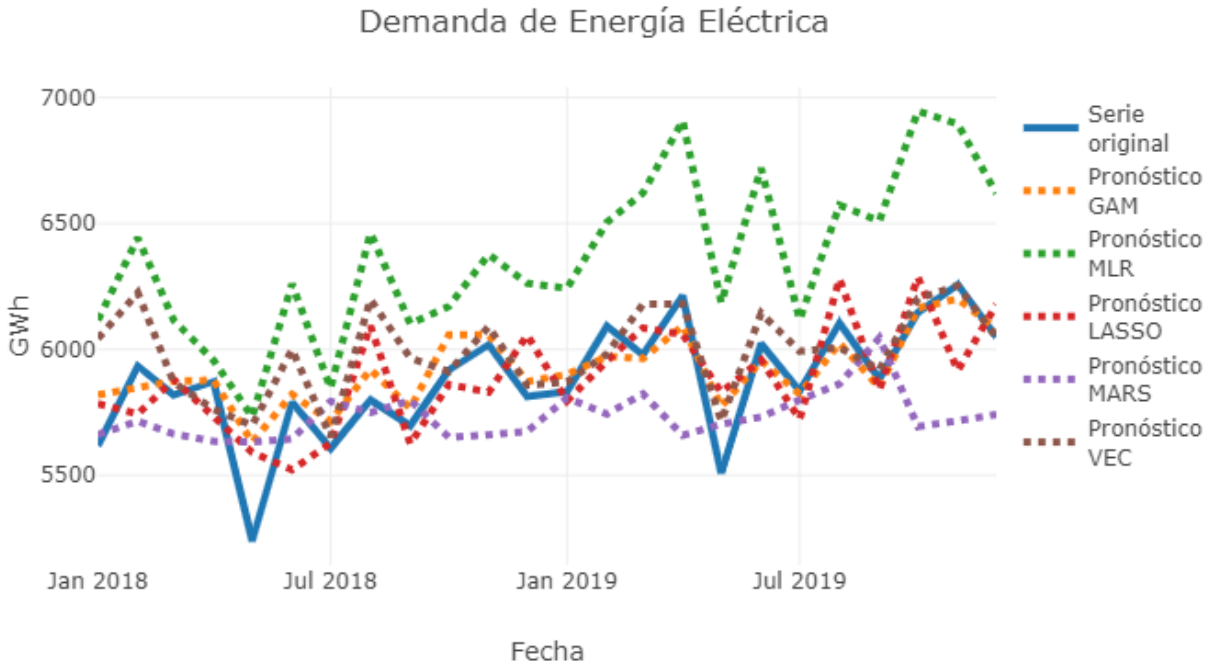

Figura 4.11: Ajuste pronóstico modelos Caso 1: Demanda energía eléctrica bajo escenario desimulación del COVID-19.

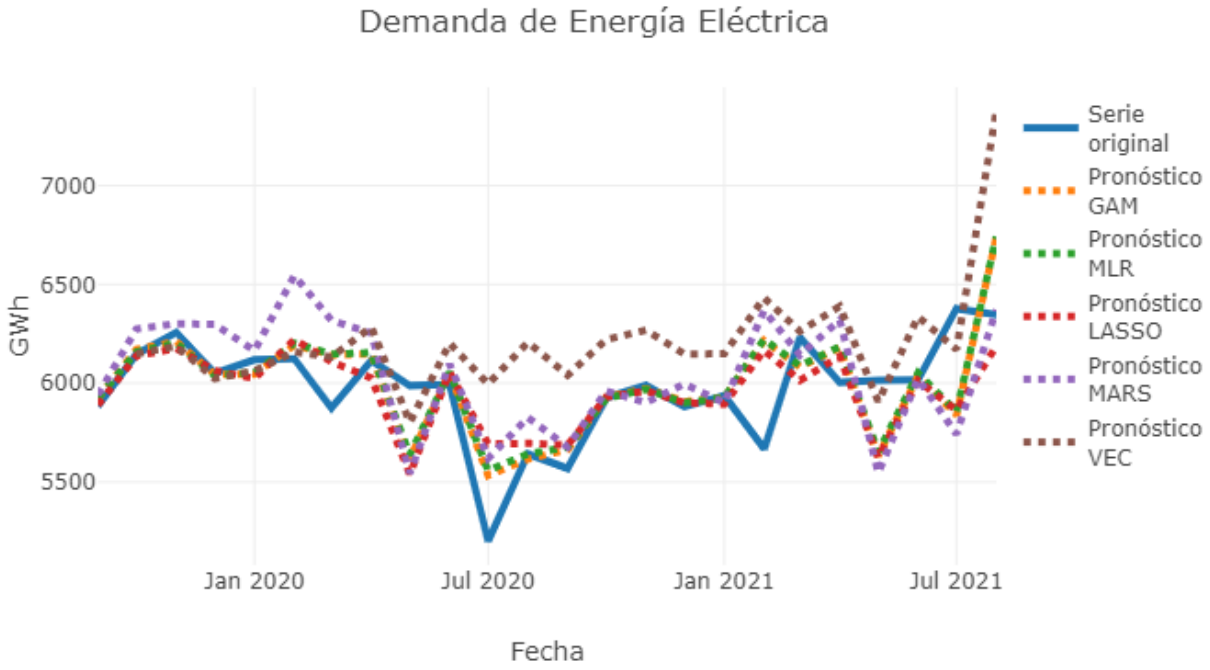

Figura 4.12: Ajuste pronóstico modelos Caso 2: Demanda energía eléctrica al incluir campaña "Apagar-Paga" y efecto delfenómeno del niño-niña.

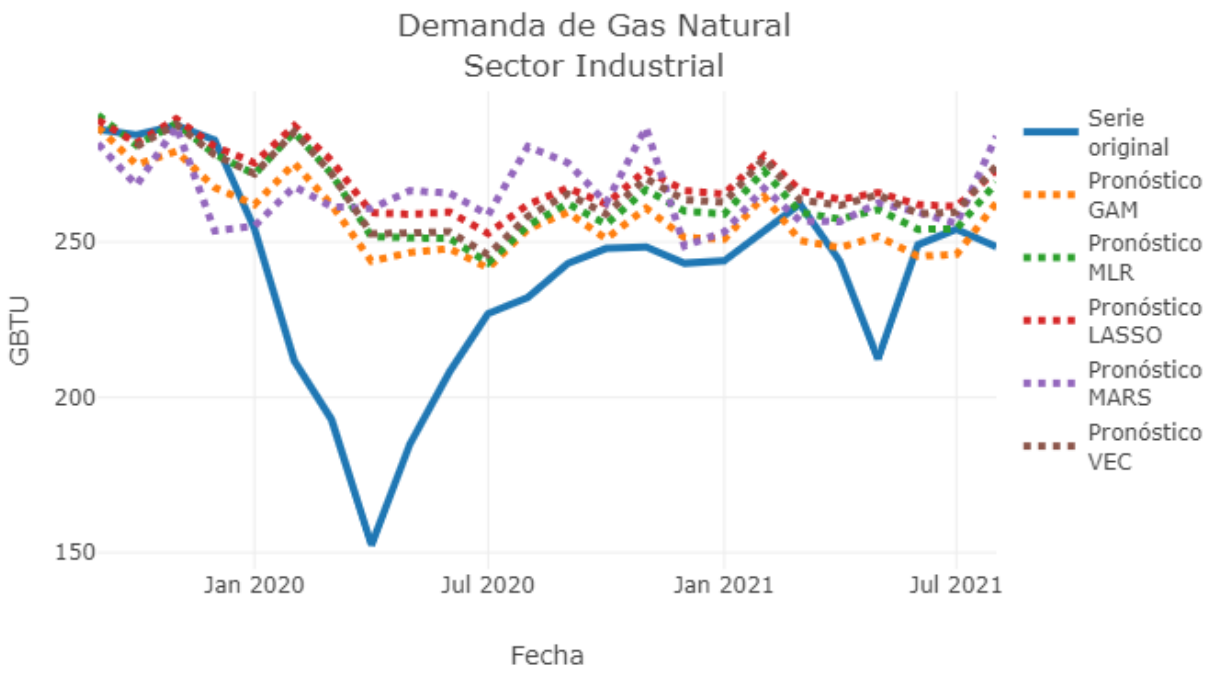

Figura 4.13: Ajuste pronóstico modelos Caso 3: Demanda gas natural para el sector industrial.

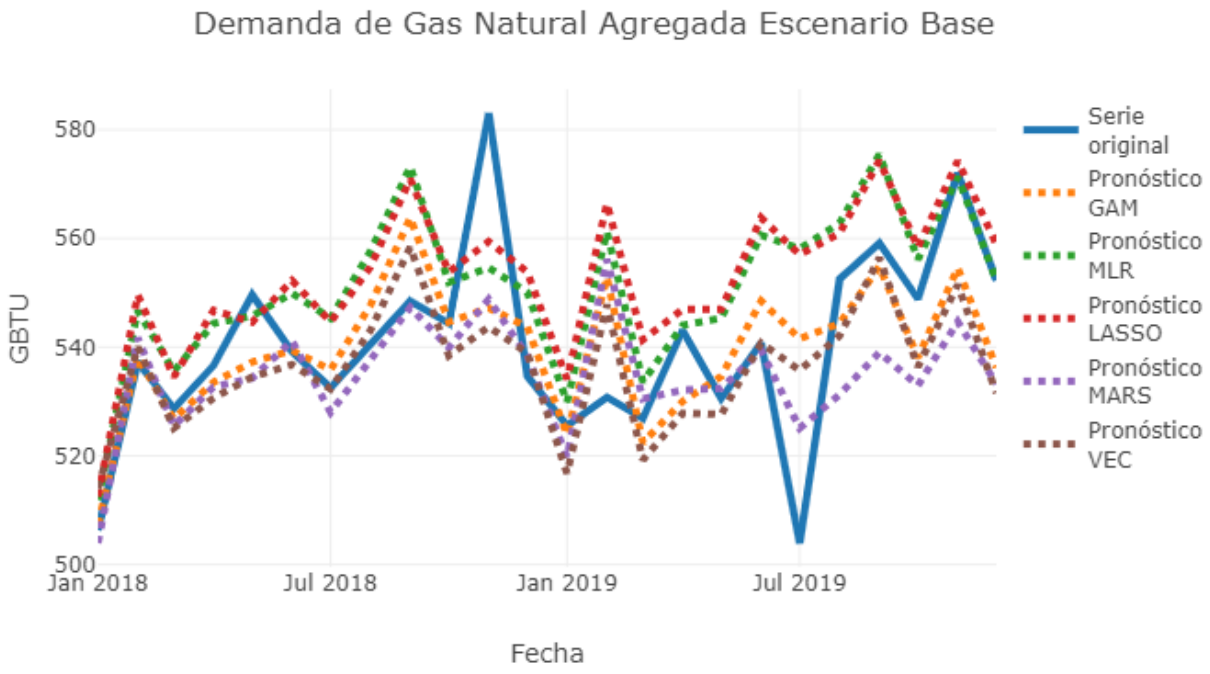

Figura 4.14: Ajuste pronóstico modelos Caso 4: Demanda gas natural agregada escenario base.

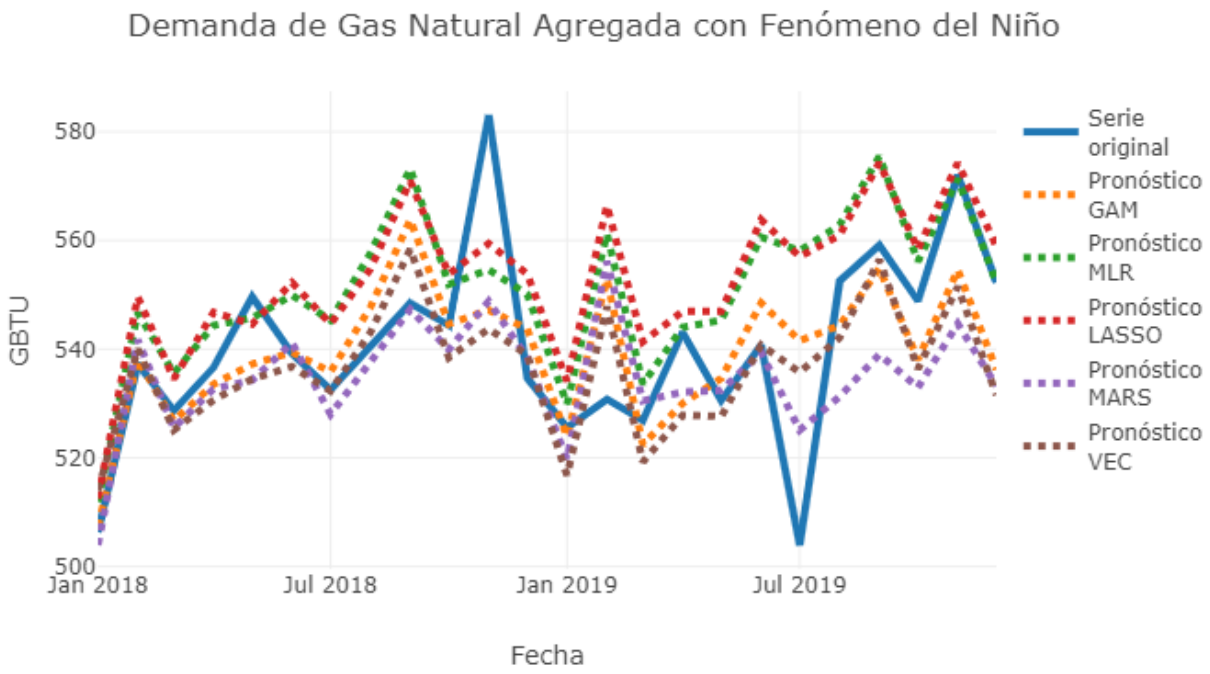

Figura 4.15: Ajuste pronóstico modelos Caso 5: Demanda gas natural agregada al incluir el efecto del fenómeno del niño.

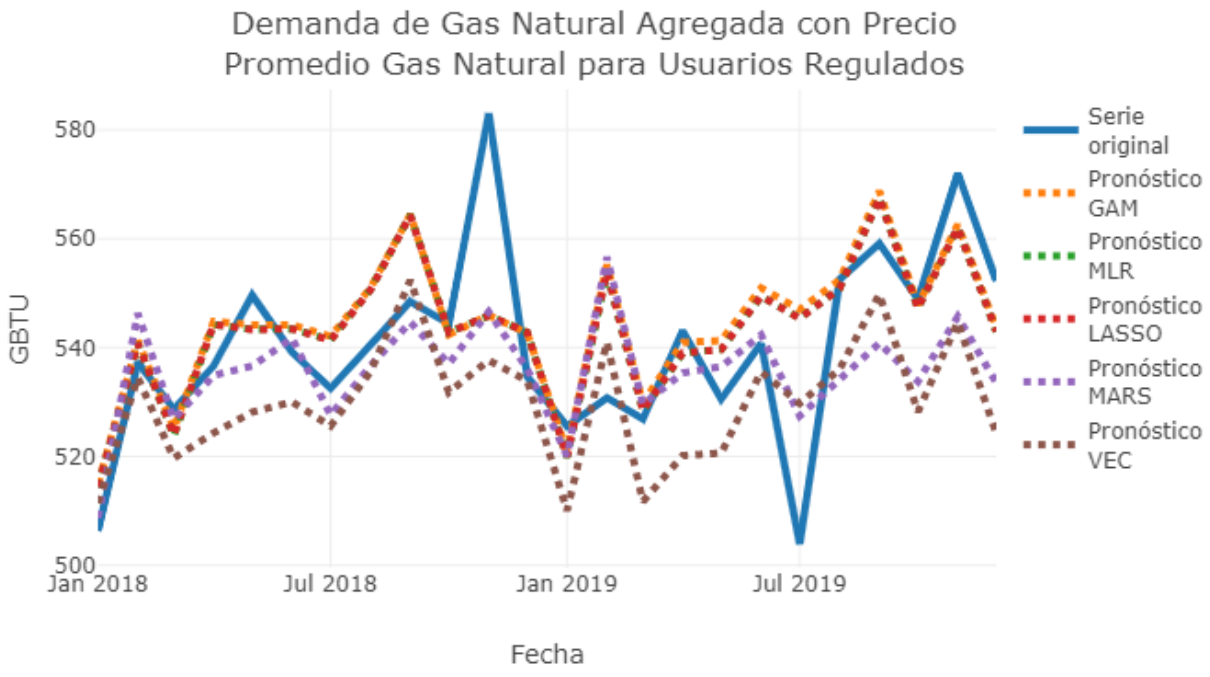

Figura 4.16: Ajuste pronóstico modelos Caso 6: Demanda gas natural agregada al incluir el precio promedio del gas natural para usuarios regulado.

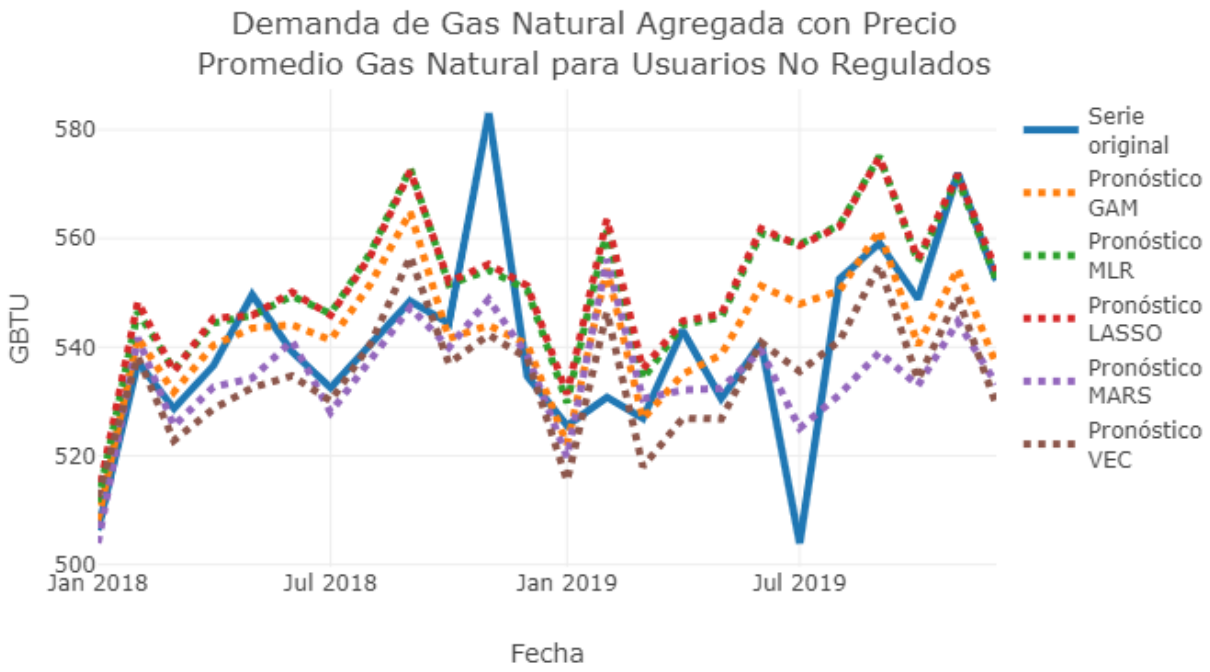

Figura 4.17: Ajuste pronóstico modelos Caso 7: Demanda gas natural agregada al incluir el precio promedio del gas natural para usuarios no regulados.

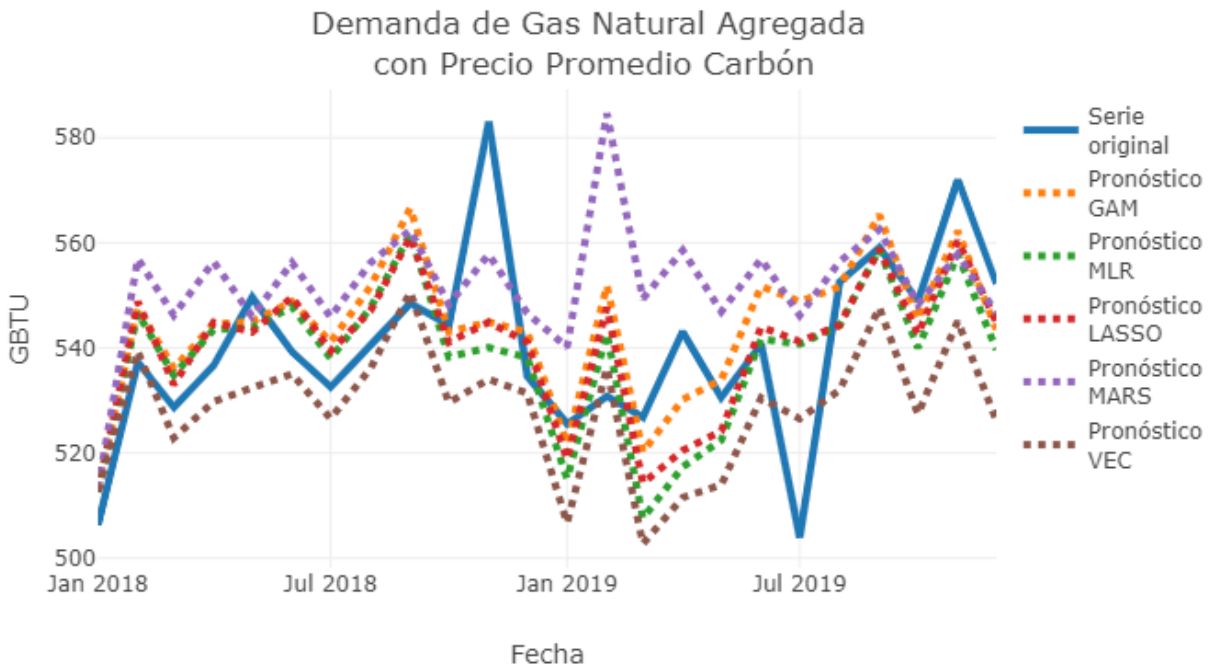

Figura 4.18: Ajuste pronóstico modelos Caso 8: Demanda gas natural agregada al incluir el precios promedio del carbón.

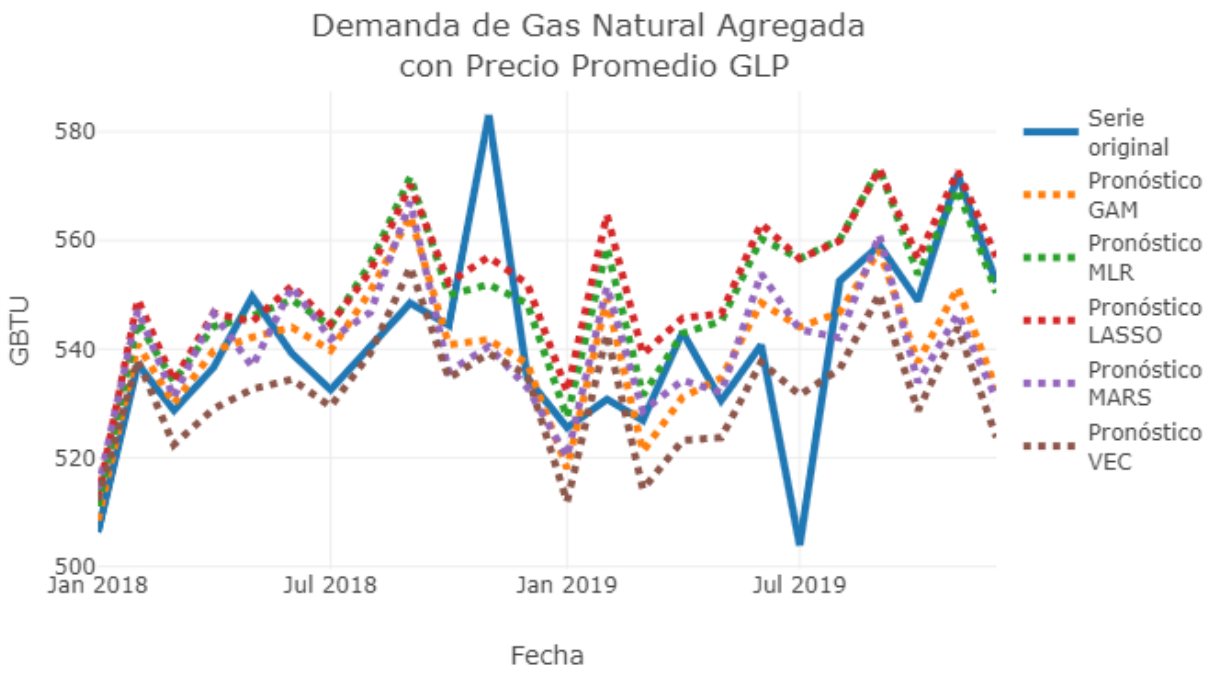

Figura 4.19: Ajuste pronóstico modelos Caso 9: Demanda gas natural agregada al incluir el precios promedio del GLP.

La Figura 4.11 correspondiente al Caso 1, muestra que aún cuando se simula el efecto de la pandemia en el conjunto de datos usados para la estimación de los modelos, se encuentra que las proyecciones realizadas para la demanda de energía eléctrica durante los dos años posteriores presentan un buen ajuste respecto al comportamiento original que se registra para la demanda energía eléctrica, a excepción del modelo MRL, el cual captura durante el proceso de estimación la tendencia que traían los datos ante la caída de la demanda causada por el COVID-19, y por ello, sus pronóstico se encuentre por encima de los valores reportados.

En la Figura 4.12 se presenta el ajuste de los modelos estimados para el Caso 2, en donde se evidencia que en general todos los modelos estimados presentan un buen ajuste al comportamiento que tuvo la demanda de energía eléctrica durante el periodo 2019-9 a 2021-8, destacando el buen ajuste que ofrecen los modelos alternativos aún cuando durante este periodo se registró una caída en la demanda causada por la pandemia del COVID-19.

Por su parte, en la Figura 4.13 se presentan los pronósticos de los modelos ajustados para el Caso 3, en donde se observa que todos los modelos ajustados presentaron problemas al momento de pronosticar el comportamiento de la serie original, debido en gran parte por los efectos que tuvo la pandemia del COVID-19 sobre el sector industrial durante el segundo y tercer trimestre de 2020, ya que los cierres económicos impuestos por el gobierno nacional para controlar la propagación de la pandemia, afectó significativamente el crecimiento económico del sector industrial, lo cual conllevó a su vez a una reducción del 40 % aproximadamente para la demanda de gas natural realizada por el sector.

En la Figura 4.14 se evidencia que ante la ausencia del periodo COVID-19 en el horizonte

de pronóstico, el ajuste presentado por los modelos es relativamente bueno, ya que logran capturar de forma adecuada el comportamiento que posee la serie original. Lo cual indica que las variables explicativas usadas para el modelo son adecuadas para la implementación de modelos de pronóstico de la demanda agregada de gas natural.

Por su parte, al comparar el ajuste obtenido en la Figura 4.15, Figura 4.16, Figura 4.17, Figura 4.18 y Figura 4.19 respecto al presentado en la Figura 4.14, se evidencia que a pesar de que la inclusión de las variables adicionales, logra mejorar el ajuste de algunos modelos, destacando la mejora del modelo MARS para el Caso 5, los modelos LASSO, MRL y MARS en el caso 6, el modelo MARS en el caso 7, los modelos LASSO y MRL en el caso 8 y el modelo MARS en el caso 9, debido a que fueron las mejora más notorias visualmente hablando. Sin embargo, también se destaca el nivel de ajuste del modelo GAM, debido a que presenta una leve mejora en cada uno de los casos, y es en términos generales el modelo que sobre salen por su nivel de ajuste en todos los escenarios planteados.

4.4 Observaciones adicionales

4.4.1 Pronósticos demanda sector termoeléctrico y refinerías

Actualmente, la UPME toma las proyecciones de demanda de gas natural de las termoeléctricas y refinerías para proyectar el consumo del gas natural. En el caso de tener la información real de los consumos mensuales que se presenta en el sector, se recomienda que la UPME, paralelamente, emplee modelos estadísticos que permitan hacer proyecciones alternativas y así comparar y evaluar la calidad de las proyecciones entregadas por estos dos sectores. También es muy importante mantener información sobre los proyectos productivos futuros de estos dos sectores ya que estos proyectos pueden influenciar significativa la demanda de gas.

4.4.2 Realización de pronósticos bajo los diferentes escenarios del PIB

La versión tres de las proyecciones del PIB entregadas por la UPME, en mayo de 2020 plantea cuatro escenarios a saber: base, bajo, alto y medio, siendo el último escenario el promedio de los tres primeros. Entonces, dado que la variable del PIB se usa como variable endógena o exógena en los modelos de proyección de la demanda de energía eléctrica y gas natural, significa que se podrían también presentar diferentes escenarios para las proyecciones de los energéticos, basados estas proyecciones en los diferentes escenarios mencionados para el PIB.

4.4.3 Uso de los pronósticos del PIB en la demanda de la energía eléctrica

Dado que los modelos VARX y VEC pronostican simultáneamente la demanda de energía eléctrica y el PIB, significa que estos modelos no harán uso de los pronósticos oficiales que reporta la UPME para el PIB, en la obtención de los pronósticos de la demanda de energía eléctrica.

Es decir, que cada vez que el modelo VARX o VEC realice un pronóstico dentro del modelo, se tendrá que éste generará de forma interna un nuevo valor para la demanda de energía eléctrica y otro para el PIB, en donde ese nuevo valor generado dentro del modelo para el PIB, no será necesariamente igual al valor oficial reportado para la misma fecha.

Por tanto, la recomendación ante esta situación será extraer la ecuación ajustada para la demanda de energía eléctrica, una vez se estime el modelo, y simplemente reemplazar en ésta los pronósticos oficiales que se tienen para el PIB, para encontrar con ello los pronósticos para la demanda de energía eléctrica.

4.4.4 Demanda de energía eléctrica de grandes consumidores

Actualmente, la UPME consulta a las empresas que consumen grandes cantidades de energía eléctrica (por ejemplo, empresas dedicadas a la minería) sobre sus proyecciones de consumo. Consideramos que este procedimiento es el adecuado puesto que tales empresas poseen toda la información necesaria sobre su cadena productiva, por lo cual éstos pueden estimar sus consumos futuros de una manera más precisa. Sin embargo, una vez se tenga el consumo real de estas empresas, se pueden aplicar los modelos planteados en este documento u otros que la ciencia vaya generando, con el fin de comparar si las proyecciones que los grandes consumidores entregan son similares a las obtenidas por estos modelos.

4.4.5 Proyecciones de vehículos eléctricos

El comportamiento del número de vehículos eléctricos matriculados en Colombia no tiene un patrón regular que permita emplear modelos estadísticos para proyectar el número de vehículos de manera confiable. Por tanto, lo más recomendable sería emplear tasas de crecimiento para los vehículos matriculados desde 2018, donde se observa un mayor número de estos. Al aplicar las tasas de crecimiento se tendría una proyección del número de vehículos eléctricos, que a su vez se deben multiplicar por el factor de consumo individual para así obtener un estimado del consumo de energía.

4.4.6 Escenarios de generación distribuida

La penetración de generación distribuida en Colombia podría alcanzar niveles significativos que lleven a representar posibles cambios en los patrones de consumo en los diferentes sectores (residencial, comercial e industrial). Por tanto, construir diferentes escenarios de participación de la generación distribuida en el mercado, obtenidos a partir de estimaciones y/o estudios de crecimiento de entrada de estas tecnologías en el sector eléctrico en Colombia, podría ser un ejercicio con el cual se construyan escenarios de demanda de energía en el país.

4.5 Conclusiones

- En este trabajo encontramos modelos alternativos al VARX y VEC que también dan buenos resultados en términos de poder predictivo. De hecho, como puede observarse en la revisión de la literatura, estos modelos alternativos son usados más frecuentemente para el pronóstico de demanda de gas natural y demanda de energía eléctrica que los modelos VARX y VEC.
- La ventaja de los modelos alternativos con respecto a los modelos VARX y VEC en el escenario de demanda de energía eléctrica es que pueden darle uso a las proyecciones del PIB de una forma más directa y simple.
- Una ventaja de los modelos MARS, GAM y RNN es que al ser modelos semiparamétricos o no paramétricos no dependen de supuestos restrictivos que en un momento dado impidan su aplicación. Por otro lado, los modelos VARX y VEC a pesar de que den buenas predicciones, en muchos casos puede ser difícil que los supuestos bajo los cuales se construyen se cumplan, lo cual limita su aplicación.
- A pesar de que en la literatura internacional han mostrado un gran uso de variables explicativas, tales como: tamaño poblacional, temperatura, velocidad del viento, radiación, humedad relativa, entre otras variables climáticas, heating degree days (HDD), cooling degree days (CDD), precio del gas residual, consumo de gas en la industria, precio de la gasolina, ingreso per-cápita, emisiones de CO_2, balanza comercial, tasa de cambio, entre otras; en el caso colombiano, el uso de variables está limitado debido a que la disponibilidad de variables pronosticadas a 15 años no se encuentran disponibles de forma confiable. Sin embargo, de este listado de variables, en este trabajo rescatamos las variables de efecto calendario, días laborales, días domingo en el mes, y la variable de cierre de la economía.

Capítulo 5 Revisión de literatura de combustibles líquidos y GLP

Para la Unidad de Planeación Minero Energética (UPME) es esencial tener un conocimiento contextualizado del sector minero-energético con el fin de proveer señales informadas, de manera periódica, de las posibles condiciones futuras de demanda de combustibles líquidos y GLP en el país. Múltiples sectores del país analizan las proyecciones de la Unidad con el fin de ajustar sus estrategias de producción, o de transporte, o de comercialización de los combustibles.

Para realizar un ejercicio contextualizado de proyecciones a 15 años de las demandas de combustibles líquidos y GLP, en este informe se presenta una revisión de literatura en la que se han revisado más de 30 publicaciones. La revisión se ha enfocado en las estrategias, variables, modelos matemáticos y metodologías empleadas en la proyección de demandas de combustibles líquidos. Se ha evidenciado que es crucial contar con datos que reflejen la posible evolución de variables demográficas, financieras, económicas, sociales, meteorológicas, entre otras.

Este capítulo se ha subdividido en diferentes secciones según el combustible a analizar. Se inicia ilustrando la experiencia que Ecopetrol ha empleado para analizar los escenarios de proyección de combustibles a largo plazo.

5.1 Ecopetrol

En esta revisión de literatura también se incluye un análisis de los procesos que emplea la principal empresa colombiana en el sector de hidrocarburos, Ecopetrol. En particular, el equipo de la Universidad de Antioquia agradece a la UPME por haber intermediado para llevar a cabo conversaciones con Ecopetrol.

En el largo plazo, Ecopetrol realiza proyecciones de demanda de combustibles líquidos con horizontes hasta 2050. La última proyección se realizó en 2018 y se espera actualizar en este 2021. Para esto, se apoyan de consultorías externas y encuestas a expertos para definición de escenarios. Emplean un modelo balance-ENPEP[1] junto a variables como la población (tomada del banco mundial), precios nacionales de combustibles (proyectados internamente por Ecopetrol), PIB (tomado de la UPME), precio internacionales de combustibles (consultados en agencias internacionales), precios de carbón (de agencias internacionales considerando ajustes locales), precio gas (proyectados internamente por la vicepresidencia de gas en Ecopetrol) y el precio de la electricidad (a partir del modelo SDDP a través de un consultor).

[1] El ENPEP es un modelo de equilibrio en el que se tienen en cuenta los objetivos de diferentes sectores energéticos como eléctrico, industrial y residencial. Este tipo de modelos se diferencia de los modelos de optimización convencionales que únicamente modelan el objetivo de un planeador.

En las encuestas que realiza Ecopetrol al grupo de expertos, se consulta sobre el análisis de variables como el PIB, población, ruralidad, precios de importación y transporte, formación precio GNV, precio autogas y gas natural licuado, parque automotor carretero, grado de penetración de vehículos eléctricos particulares, movilidad activa (caminar, usar bicicleta, etc.), hábitos de conducción, uso hidrógeno, impacto trabajo virtual, uso de diésel para carga urbana, transición hacia otros energéticos, modo férreo o fluvial para transporte intermunicipal, demanda de energía eléctrica, precio del kWh, hidrología, energías renovables, restricciones al uso de energéticos contaminantes, producción de carbón, eficiencias y re-cambio de tecnologías en la industria, entro otros.

A partir de estas encuestas, Ecopetrol construye algunos escenarios que tienen como punto de partida la Transición Energética (TE) y el Crecimiento Económico (CE), puesto que, en conjunto, definen la geopolítica del análisis. Algunos de los aspectos más relevantes a destacar de la más reciente encuesta realizada por Ecopetrol son mostrados a continuación:

- El grupo de expertos considera que la TE dependerá de los valores históricos de los consumos y del CE que tenga el país, teniendo un cuenta que el escenario de crecimiento esperado es de 3 %, tal y como lo prevé la UPME.
- El hidrógeno sería un energético fundamental para participar en la atención de la demanda a 2030, podría comenzar a reemplazar al GNV.
- El GLP es considerado como el recurso energético que comenzaría a desaparecer de la matriz.
- El escenario de crecimiento de los precios del petróleo sería medio teniendo en cuenta la situación mundial.
- En los escenarios optimistas se esperan precios del Gas entre 3 y 6 dólares por cada MBTU. Se destaca que los precios del GNV, el GLP, GNL y el hidrógeno estarán determinados, en los escenarios más optimistas, como variables independientes de la gasolina o de los productos refinados.

También se ha hecho un análisis por diferentes sectores de consumo energético: transporte, industria, generación de electricidad.

Transporte

Para el sector transporte, con respecto a las necesidades de movilidad urbana, según las encuestas realizadas por Ecopetrol (Ecopetrol, 2021), se cree que el teletrabajo podría afectar en un 5 % la movilidad y por ende el consumo de combustibles para el transporte. Asimismo, se espera un aumento en la movilidad para el transporte colectivo cercano, mientras se construyen los sistemas de transporte masivo; no obstante, es posible que comiencen a popularizarse otros medios de transporte como las bicicletas eléctricas o las convencionales. Aunque en los escenarios donde se plantea un CE no tan favorable, los expertos consideran que la movilidad estaría dada por motos en primer lugar. En este mismo sentido, se espera en los escenarios optimistas la inclusión de vehículos eléctricos, mientras que los de gas aparecen como el medio de transporte

en los escenarios conservadores.

La movilidad en las ciudades, para los expertos, continuará desmejorando, lo cual puede explicarse a la luz de la entrada constante de nuevo parque automotor y la baja tasa de recambio o chatarrización de los vehículos existentes. Estos problemas de movilidad generan mayor consumo de combustible; y por ello los expertos esperan, en el escenario más optimista, que se adopten los criterios de Eco-Driving[2]. Pero en los escenarios pesimistas, se espera que la tendencia de consumo continúe igual.

Con respecto a la atención de la demanda de pasajeros interurbanos, para distancias inferiores a los 150 km, los escenarios optimistas plantean la posible penetración de energías renovables. También, se espera que entre dos y cinco años se reestablezca el crecimiento en la cantidad de vuelos, lo cual dinamizaría el transporte aéreo e incrementaría el consumo de estos combustibles. Es de anotar que basados en CORSIA (Carbon Offsetting and Reduction Scheme for International Aviation), se espera que la reducción de emisiones se haga mediante el diseño de equipos que mejoren la eficiencia, pero que se continúe con la compensación de emisiones. Por otra parte, se espera que para la carga urbana se llegue a una desestimulación del uso de diésel como energético mediante incentivos para evitar su uso al interior de las ciudades.

Con respecto a la carga interurbana, en los escenarios optimistas de CE, los expertos esperan que se transportaría en modos tanto férreo como fluvial. Aunque en los escenarios no optimistas, se presume que no habrá la superación de la meta (10 mil vehículos chatarrizados). Además, la penetración de energéticos como hidrógeno, electricidad y GNV podrían estar en la cima del uso.

Industria

Con respecto a las restricciones para el uso de energéticos que se asocian con gases de efecto invernadero, en el escenario optimista se espera que se restrinja el uso de estos dentro de la matriz energética, lo cual impactará industrias como la del carbón. Sin embargo, el mejorar la eficiencia de los procesos podría ser el principal frente de trabajo en las industrias que tengan estos consumos.

Generación Eléctrica

En cuanto a la generación de energía, los escenarios optimistas, para los expertos, auguran la entrada de proyectos de energías renovables como solares y eólicos, lo cual disminuiría el precio del kwh, y que respondería a un escenario en el que aumenta la demanda de energía eléctrica.

[2]El Eco-Driving es una técnica de conducción que se basa en el control óptimo de las variables que controla el conductor como la velocidad, la aceleración, la desaceleración y la marcha, buscando minimizar la pérdida de energía para lograr un consumo reducido de combustible.

A continuación se presenta la revisión de literatura para los diferentes combustibles líquidos y GLP analizados en este proyecto. Adicionalmente, se han incluido cuadros que resumen la literatura revisada de cada combustible. En estos se presentan las variables y modelos matemáticos empleados en tales publicaciones.

5.2 Diésel

El combustible diésel, también conocido como gasóleo o gasoil, es un producto derivado del petróleo que se obtiene mediante un proceso de destilación y purificación. Este producto es generalmente utilizado en automóviles que tengan su motor adecuado para este combustible. La característica principal del diésel es la elevada temperatura en el proceso de combustión, lo que le hace alcanzar una gran eficiencia termodinámica en comparación a los motores de gasolina.

En primer lugar, se presentan los trabajos más relevantes en cuanto a inferencia; y posteriormente, los artículos que tienen como objetivo desarrollar de modelos de pronóstico.

Iniciando con el caso colombiano, López Valderrama et al. (2015) realizan estimaciones de las elasticidades precio de la demanda, cruzada de la demanda y de ingreso a través de series de tiempo. En su trabajo, los autores encuentran que muestran que la demanda de diésel no es muy sensible a los cambios que sufre su respectivo precio, si no más bien, que dicha demanda es explicada principalmente por el ingreso per-cápita de las personas. Es de anotar que, en este trabajo se toma como variable proxy de los ingresos de las personas, el variable del PIB per-cápita en logaritmo, encontrando con ello que en general la capacidad adquisitiva de las personas es un determinante de la demanda de gasolina y diésel en Colombia.

De la misma manera que en el trabajo de López Valderrama et al. (2015), la investigación presentada por García et al. (2016) calcula las elasticidades precio de la demanda para el diésel en Colombia. Emplea el modelo casi ideal de demanda en donde de manera inicial se estiman modelos de ecuaciones aparentemente no relacionadas con toda la industria para luego estimar cada uno de los parámetros. Los resultados muestran que el diésel sigue siendo insensible al precio, además con la elasticidad cruzada se valida la sustituibilidad que tiene el gasolina respecto al diésel. En relación al ingreso, se ve como el diesel tiene un comportamiento normal; es decir, ante aumentos en el ingreso de las personas, la cantidad demandada también aumenta.

En el trabajo de Adom et al. (2016) se realiza una estimación de las elasticidades precio de gasolina y diésel en el sector de transporte de Ghana, con datos desde 1971 hasta 2011 de frecuencia anual. A través de un modelo VAR, que emplea como variables la demanda de diésel, precio del diésel, precio de la gasolina y el PIB, se encuentra que las variaciones del consumo del diésel respecto a los movimientos de su precio resultan ser mínimas en el corto plazo. Sin embargo, a través de la incorporación de rezagos se observa cómo esta respuesta, ante cambios en los precios, se incrementa ligeramente en el largo plazo.

Con la intención de tener un panorama amplio de los cálculos que se han hecho a nivel

internacional de las elasticidades precio e ingreso de la demanda de diésel. Para el caso de los Estados Unidos, el trabajo de Uri y Herbert (1992) tiene como objetivo principal analizar la relación entre la demanda de diésel y su respectivo precio, el ingreso de las personas y precipitaciones. A partir de un modelo regresión lineal múltiple, se encuentra que ante un aumento de 1 % en el precio del diésel, la demanda cae entre 1.15 % y 0.26 %. Es decir, para este grupo de referencia, la cantidad demanda de diésel resulta ser sensible al precio. Sin embargo, es importante mencionar que este trabajo tomó como referencia datos de demanda el sector agrícola en diferentes estados; pero no se tuvo en cuenta uno de los sectores más importantes como lo es el automotriz.

El trabajo de Dahl (2012) resume los principales estudios relacionados con la elasticidad precio y demanda de diésel. Este autor emplea datos de aproximadamente 100 países con diferentes tamaños e ingresos para los cálculos de las elasticidades. Se encuentra que en aproximadamente el 95 % de los países, la demanda de diésel es inelástica o insensible al precio. Esto quiere decir que las políticas de consumo vía precio no serían eficientes si se busca una transición energética.

En el caso donde el objetivo principal es proyectar la demanda de diésel, el trabajo realizado por Rao y Parikh (1996) presenta una aplicación para el caso de la India. A través del modelo tradicional de mínimos cuadrados y con variables como el PIB, precios de los combustibles, índices de producción y población en los centros urbanos, se busca tener estimaciones de los valores futuros del consumo diésel. Las estimaciones muestran que la demanda de diésel experimentará fuertes tasas de crecimiento anual. Por esta razón, los autores recomiendan buscar alternativas en eficiencia energética que reduzcan los impactos sobre el medio ambiente.

En Ecuador, Rivera-González et al. (2020) estiman diferentes escenarios de demanda de combustibles para el sector del transporte. Para la proyección, se consideran variables relacionadas con los diferentes tipos de vehículos y el consumo medio de estos. Con esta información se analiza la participación de cada uno sobre el total de la demanda de diésel. Los autores han empleado un modelo de simulación LEAP. Cuando se simulan proyecciones con aumentos en la eficiencia energética de los automóviles, se observan disminuciones significativas en el consumo de combustibles líquidos.

El trabajo de Ertuğrul et al. (2020) muestra el comportamiento en el mercado de diésel teniendo en cuenta los shocks generados por la pandemia de la COVID-19. Para este caso se recogen datos diarios del consumo de diésel. Con un análisis descriptivo inicial se evidencia cómo la volatilidad del mercado del diésel aumentó significativamente en los primeros meses del año 2020, sobre todo por las restricciones en la movilidad impuestas por el gobierno nacional. Un modelo SARMA, con parámetros (7,7)(1,1), y al ser datos con resolución diaria, la variabilidad es alta. Por esta razón, los autores también han empleado la familia de modelos ARCH con el fin de explicar, en cierta medida, las disminuciones de la demanda durante el confinamiento.

Finalmente, el Cuadro 5.1 presenta un resumen de los modelos y variables empleadas en las publicaciones analizadas para la proyección del consumo de diésel.

Cuadro 5.1: Resumen revisión de literatura sobre diésel

Autores	Frecuencia	Modelo	Variables
Rao y Parikh (1996)	1974-2000 anual	Mínimos cuadrados	PIB, precios de los productos petroleros, índice industrial de producción, población urbana
Lee y Cho (2009)	Microdatos 2003	Preferencia revelada con regresión lineal	
Chai et al. (2012)	1985-2009 Anual	Modelo de regresión Bayesiana y ARIMA para el pronóstico	Nivel de urbanización, PIB per cápita, Rotación de pasajeros, Número de carros por persona
Dahl (2012)		Análisis Exploratorio	Demanda de gasolina y diésel, precios de gasolina y diésel, PIB
Kim et al. (2006)	Microdatos	Logit	Variables sociodemograficas (sexo, educación, ocupación,ingreso, tipo de carro, intención de compra de automóvil)
Uri y Herbert (1992)	1971-1989 Anual	Mínimos cuadrados	Precios, precipitaciones, ingresos personas
López Valderrama et al. (2015)	2001-2014 Mensual	Modelo de Koyck	Demanda de gasolina y de ACPM, PIB per cápita, respectivos precios
Adom et al. (2016)	1971-2011 Anual	VAR y ARDL	Demanda de diesel, Demanda de gasolina, precio del diesel, precio de la gasolina, PIB
García et al. (2016)	2003-2012 mensual	Modelo casi ideal de demanda de combustibles	Participación de la gasolina, GNV y diésel en el gasto total del mercado de combustibles, precios promedios de cada uno de los combustibles, gasto total en los combustibles, Dummy que recoge el efecto estacional.

Ertuğrul et al. (2020)	Diario	ARIMA ,GARCH, ARCH, EGARCH	Consumo diario de diesel
Rivera-González et al. (2020)		Software LEAP	Ventas de automóviles, tipo de automovil, número de pasajeros, optimización de energía

5.3 Fuel Oil

A nivel internacional se han presentado algunos aportes en cuanto a metodologías de pronóstico de fuel oil. En el trabajo de Gómez-Villalva y Ramos (2004) describen la fuerte correlación lineal que existe entre los precios del Brent y el precio del fuel oil en España. Inicialmente, mediante un modelo de optimización estocástico, el cual usa como insumo el precio de los futuros del Brent se generan varios escenarios de precios del Brent para finalmente usar un modelo de regresión lineal para el pronostico de precios del fuel oil con un horizonte de tiempo de un año en una frecuencia mensual.

Los autores en Shekarchian et al. (2012) reportan el pronostico de consumo de fuel oil usado para aire acondicionado. La cantidad de combustible se estima a partir de la energía suministrada por la red a sistemas de aire acondicionado y el porcentaje de cada combustible primario usado para la generación de esta energía. Con base en 2 tipos de tecnología, aire acondicionado convencional y enfriadores de absorción, se generan 4 escenarios de consumo de acuerdo a al porcentaje de penetración de cada tipo de tecnología. Finalmente, teniendo en cuenta el histórico de consumo de cada tipo de combustible y la tecnología de aire acondicionado, se pronostica el consumo de combustibles mediante un ajuste polinomial a una resolución anual, 17 años adelante.

En el informe de IEA (2021c) reportan, de manera general, los efectos del COVID-19 sobre la demanda de fuel oil y como las políticas de reducción de contaminación afectan el consumo de este combustible. Debido a cierres temporales de todo tipo de puertos y al trabajo remoto, el consumo de fuel oil para el transporte disminuyó considerablemente. La organización internacional del transporte marítimo, a su vez, tiene la meta de incrementar la eficiencia del transporte. Esto llevará a que el aumento en el consumo de fuel oil sea relativamente lento. Respecto a la generación de electricidad, con el uso de energías renovables para cumplir las metas de emisiones, se espera una disminución en el consumo de fuel oil en el sector de generación.

Finalmente, el Cuadro 5.2 presenta un resumen de los modelos y variables empleadas en las publicaciones analizadas para la proyección del consumo de fuel oil.

Cuadro 5.2: Resumen revisión de literatura sobre fuel oil

Autores	Frecuencia	Modelo	Variables
Gómez-Villalva y Ramos (2004)	2002-2003 mensual	Modelo de regresión estocástico y regresión lineal	Precio de los futuros del Brent, Brent spot price
Shekarchian et al. (2012)	2009-2025 Anual	Ajuste polinomial	Histórico de consumo de combustibles, tipo de tecnología de calefacción

5.4 GLP

El Gas Licuado del Petróleo (GLP) es un compuesto químico orgánico que surge de diversas mezclas de hidrocarburos, en cuya composición química predomina el butano y el propano, y en menor medida, el propileno, el butileno, el etano, el pentano, isobutano o una mezcla de éstos.

El GLP se obtiene regularmente a través del proceso de destilación o refinamiento del petróleo al separar los componentes de la gasolina, la nafta, el queroseno, entre otros. Adicionalmente, puede ser obtenido también de forma natural durante la extracción del gas natural, en donde al someter éste a un proceso de enfriamiento, el butano y el propano se condensan en la parte inferior del mismo, facilitando así su extracción.

Aunque durante los últimos años el GLP ha venido cobrando cada vez más importancia en diferentes sectores, debido a que entre sus características se destaca que es un compuesto incoloro, inoloro, no es tóxico y es de rápida combustión. Su aparición y uso data desde el siglo XX, en donde dicho compuestos se utilizaba como un sustituto de la leña para calefacción o cocción.

Entre los usos que tiene el GLP en la actualidad se remarca el empleo que tiene en el hogar, la industria, el transporte y la agricultura. En el caso del hogar, el GLP se emplea para la cocción de alimentos, calentadores de agua y fuente de energía para equipos de calefacción, neveras, chimeneas, entre otros. En la industria, se emplea para la generación de energía en zonas apartadas, para el funcionamiento de maquinaria especializada y como fuente de energía en restaurantes, hoteles y otros sectores.

Por su parte, en el caso del transporte, el GLP se emplea como combustible alternativo en vehículos terrestres y fluviales que posean motores de combustión interna, ya que éste otorga un excelente rendimiento, menor contaminación, menor costo y mejor combustión que el diésel, mientras que en la agricultura se emplea para el manejo climático y de temperatura de los cultivos, para el secado de semillas, frutas y tabaco, y para el control de insectos, plagas, maleza, entre otros.

Debido a la gran variedad de usos que tiene el GLP, es que la determinación de su demanda se ha convertido un tema de especial interés para los productores, distribuidores y comerciantes del compuesto. Esto ha generado que se desarrollen trabajos que buscan pronosticar el comportamiento que tendrá la demanda del GLP para los próximos años, con el fin de poder idear planes de expansión que permitan un crecimiento sostenible a largo plazo y evitar problemas derivados al suministro del producto.

Entre los trabajos que se han realizado para pronosticar la demanda del GLP, se presenta inicialmente el trabajo de Rodríguez y Da Silva (2010), en donde las autoras proponen un modelo SARIMA para obtener pronósticos de la demanda diaria y mensual del GLP, y un modelo de VEC para obtener los pronósticos mensuales de la demanda del GLP. Las autoras establecen que el GLP posee un comportamiento estacional asociado a la temporada del año, debido a que este compuesto se usa principalmente para cocción y calefacción en los hogares durante los meses de frío. Además del término estacional empleado en los modelos univariados y multivariados, las autoras incorporan el efecto calendario[3], la temperatura media máxima, el precio real del GLP y para el caso multivariado adicionaron el consumo de energía eléctrica residencial, las variaciones en el ingreso sobre la demanda del producto y el precio relativo de la energía eléctrica.

Luego de estimar los diferentes modelos, las autoras concluyen que a pesar de que los modelos univariados presentan un buen desempeño predictivo cuando se incorpora la temperatura mediante tres variables regresoras, a saber, una para invierno (junio, julio y agosto), otra para otoño (abril y mayo) y otra para primavera (septiembre y octubre), los resultados predictivos obtenidos en el modelo multivariado son mucho mejores ya que se observa con claridad que a nivel acumulado de error de predicción, los modelos multivariados reducen notoriamente dicho error respecto a los univariados.

Otro trabajo que se ha realizado en el tema es el presentado por Sánchez y Reyes (2016), en el que proponen tres Modelos VAR para estimar el consumo de gasolina, GLP y electricidad, respectivamente, mediante la aplicación de tres rezagos para los modelos de gasolinas y electricidad, dos rezagos para el modelo de GLP y variables dummy para el dato de la gasolina reportado en el año 1999 y para el dato de electricidad reportado en 2004. Adicionalmente, emplean como variables exógenas al modelo el precio de la gasolina, del GLP y electricidad, además de incluir el ingreso que es representado por el Producto Interno Bruto (PIB). A pesar de que en el trabajo de Sánchez y Reyes (2016) no se realizan pronósticos usando los modelos estimados, los autores demuestran que tanto las variables de consumo y precios de la gasolina, GLP y electricidad, junto con la variable del ingreso representado por el PIB, son series no estacionarias con orden de integración I(1). Además, demuestran que los modelos VAR calculados para el consumo de gasolina, GLP y electricidad cumplen las pruebas de normalidad, auto-

[3]Las ventas de GLP se realizan fundamentalmente de lunes a sábados y sólo en casos excepcionales se vende el producto los domingos o los festivos, lo cual implica considerar en términos de demanda que no todos los días del mes son iguales, lo cual significa que entre mayor sea la cantidad de domingos y festivos que haya en un mes, menor es la demanda acumulada.

correlación y heterocedasticidad. Finalmente, señalan que los estadísticos de la raíz del error cuadrático medio para los modelos indican un buen ajuste, y que por tanto, dichos modelos podrían realizar buenos pronósticos para las variables de consumo.

Un trabajo adicional realizado sobre la estimación de la demanda del GLP, es el planteado por Gesto (2016), el cual deriva de su tesis de grado Gesto (2015). El trabajo emplea un modelo VEC con un total de dos y tres rezagos. Para la estimación de su modelo, la autora emplea como covariables la demanda de GLP, la demanda de energía eléctrica, el índice de salario real con base en 2012, el precio del GLP relativo al índice de precios al consumidor (IPC) base 2012 y el precio de la energía eléctrica relativo al IPC con base 2012. Además de variables dummies estacionales y variables de intervención[4]. Una vez realizadas las estimaciones de los modelos VAR(2) y VAR(3), Gesto (2015) encuentra que estos modelos no son estacionarios y que cuentan con 3 raíces características próximas al círculo unitario, lo cual significa que el modelo cuenta con hasta 2 relaciones de cointegración, por lo cual reemplaza los modelos VAR estimados por modelos VEC para adicionar la cointegración entre las variables.

Finalmente, Gesto (2015) realiza pruebas de pronóstico con horizonte a 12 meses para probar la capacidad de ajuste de los modelos, encontrando que los modelos de predicción poseen un error de predicción menor al 1 % al considerar los datos agregados anualmente, pero que al considerar las predicciones puntuales mes a mes se encuentra que el modelo no logra predecir bien la demanda de ninguno de los dos energéticos.

Por su parte, Correia et al. (2020) plantea una metodología de modelos múltiples con el fin de realizar un pronóstico para la demanda de cilindros de GLP, en donde el objetivo de los autores es realizar los pronósticos de la variable de interés mediante la combinación de tres técnicas que se emplean usualmente para pronosticar la demanda y las ventas de combustibles. Los tres modelos de pronóstico empleados por los autores fueron: un modelo de coeficientes estacionales y suavización exponencial de Holt's para series de tiempo (TS), un modelo de regresión lineal múltiple (MLR) y un modelo de redes neuronales artificiales (ANN). Para la estimación de los modelos calculados, Correia et al. (2020) emplean como variables explicativas, la demanda de cilindros de GLP, la temperatura, la implementación de campañas promocionales, objetivos de ventas y las expectativas de los precios.

Una vez realizadas las estimaciones y los pronósticos con cada una de las tres metodologías propuestas, los autores plantean una combinación lineal ponderada para calcular el pronóstico final que tendrá la variable de interés durante el mes m, tal que

$$Y(m) = \alpha_{TS} Y_{TS}(m) + \alpha_{MLR} Y_{MLR}(m) + \alpha_{ANN} Y_{ANN}(m)$$

En donde las variables Y_{TS}, Y_{MLR} y Y_{ANN} representan los valores pronosticados para el mes m con cada una de las metodologías, mientras que, los valores α_{TS}, α_{MLR} y α_{ANN} representan los coeficientes de ponderación que se le dará a cada uno de los pronósticos calculados con cada metodología con el fin de obtener el mejor pronóstico final. Estos coeficientes se obtienen

[4]Las intervenciones corresponden a determinados eventos (shocks) que se dan a lo largo del período bajo estudio y serán incluidas como variables determinísticas exógenas a la relación de cointegración (Gesto, 2015)

a través de un proceso de simulación de Monte Carlo, en el cual los autores suponen que los valores pronosticados por cada una de las metodologías poseen una distribución normal con media igual al valor pronosticado y varianza igual al error cuadrático medio.

La metodología empleada por Correia et al. (2020) se describe a través del diagrama de flujo presentado en la Figura 5.1, el cual es desarrollado por los mismos autores

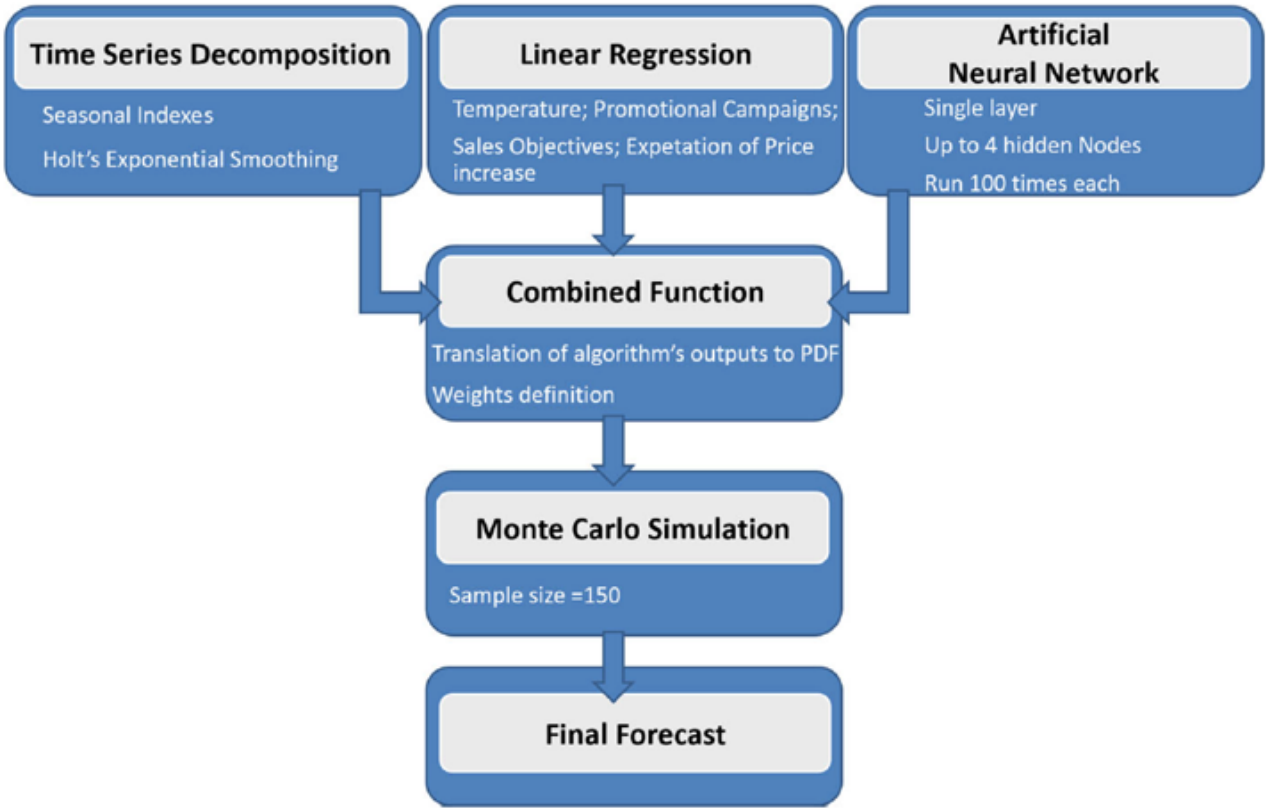

Figura 5.1: Resumen de la metodología propuesta por Correira et al. (2020)

De los resultados encontrados, los autores concluyen que la metodología de ponderación de pronósticos permite eliminar los problemas de sobreajuste que pueden tener las metodologías individuales, además de que permiten conducir a pronósticos más sólidos que los que se realizan normalmente con las otras metodologías, ya que la estructura propuesta de ponderación permite lidiar con los inconvenientes que pueden presentarse por la no linealidad y la estacionalidad.

En el caso de Rehman et al. (2017), los autores presentan un trabajo que tiene por objetivo pronosticar para diferentes sectores[5] la demanda del petróleo, el gas natural, el carbón, la electricidad y el GLP a largo plazo en Pakistán. Para ello, deciden estimar tres modelos con el fin de observar cuál presenta mejores resultados. Los tres modelos implementados por Rehman et al. (2017) para cada uno de los seis energéticos fueron: un modelo autorregresivo integrado de media móvil (ARIMA), un modelo de suavización exponencial de Holt-Winter y un modelo

[5]En los pronósticos realizados por los autores, éstos dividen las estimaciones y pronósticos de la demanda de petróleo en un total de en seis sectores, para el gas natural en siete sectores, para el carbón en dos sectores, para la electricidad en cinco sectores y para el GLP en tres sectores.

de planificación de alternativas energéticas a largo plazo (LEAP). Es de anotar que para la estimación de los diferentes modelos, los autores no introducen ninguna variable endógena o exógena en sus estimaciones; sino que realizan los pronósticos de las diferentes demandas por sector, uno a uno, empleando solamente la variable de demanda para el energético de interés.

Una vez realizadas todas las estimaciones y las proyecciones de la demanda de los energéticos a 2035, Rehman et al. (2017) concluyen que el modelo ARIMA es el que presenta las proyecciones más apropiadas. Pero, señalan que hay ocasiones en las cuales la demanda de los energéticos es negativa[6], y por ello sugieren comparar resultados de esta naturaleza con los que se obtienen mediante otras metodologías, haciendo alusión al modelo LEAP y Holt-Winter.

Finalmente, el Cuadro 5.3 presenta un resumen de los modelos y variables empleadas en las publicaciones analizadas para la proyección del consumo de GLP.

Cuadro 5.3: Resumen revisión de literatura sobre GLP

Autores	**Frecuencia**	**Modelo**	**Variables**
Koshala et al. (1999a)	Anual	Ajuste de curva	Histórico de la demanda, precio del queroseno, ingreso per-cápita
Bhattacharyya y Blake (2009)	Anual	Regresión lineal	Precio real de los productos petroliferos
Rodríguez y Da Silva (2010)	Mensual, diario	Modelos ARIMA y VECM	Demanda GLP, demanda de energía residencia, el IMSR, logaritmo del precio relativo entre GLP y energía eléctrica, efecto calendario e intervenciones y la temperatura
Iwayemi et al. (2010)	Anual	Cointegración multivariada	Histórico de la demanda
Abdullahi (2014)	Anual	Structural time series	PIB, precio real, tendencia de la demanda de energía
Gesto (2015)	Mensual	VAR	Precio del gas, temperatura
Sánchez y Reyes (2016)	Anual	Modelo de demanda casi ideal (AIDS)	Cantidad demanda de GLP, nivel de gasto de cada hogar.

[6]Los autores encuentran que para el caso de la demanda de gasolina, los pronósticos obtenidos por el modelo ARIMA arrojan un resultado negativo para los sectores doméstico y agricultura, debido a la tendencia histórica a la baja. Por consiguiente, prefieren omitir los resultados obtenidos para estos dos sectores.

Rehman et al. (2017)	Anual	ARIMA, software LEAP, Holt-Winter	Demanda de GLP.
Rasouli (2018)	Anual	Regresión lineal	PIB, número de vehículos, población.
Correia et al. (2020)	Mensual	Series de tiempo, regresión lineal, red neuronal artificial, combinación	ventas nacionales de cilindros, capacidad de los cilindros, temperatura, promociones, incremento de ventas esperado, fecha, objetivos de venta

5.5 Gasolina

De acuerdo con el Ministerio de Minas y Energía, la gasolina se obtiene a partir de una combinación de diferentes hidrocarburos líquidos, volátiles e inflamables como lo son el carbono e hidrógeno. Este combustible se obtiene por medio de la destilación fraccionada del crudo, a la que se le añaden una serie de aditivos para que sus características naturales sean optimizadas. El principal uso que recibe la gasolina es como combustible en motores de combustión interna.

De los diferentes combustibles derivados del petróleo, uno de los más importantes y más utilizados es la gasolina. La razón es que la mayoría de los vehículos y maquinaria utilizan motores de combustión. Por esta razón, es importante, en primer lugar, conocer los principales determinantes de la demanda de gasolina para, posteriormente, proyectar su consumo en los próximos años. De esta manera, las instituciones gubernamentales pueden diseñar políticas que garanticen el abastecimiento continuo de gasolina.

Para la estimación de la demanda de gasolina, la literatura ha tomado generalmente dos rutas metodológicas: inferencia y pronóstico. En la primera, se ha tenido interés por observar la relación y fuerza entre la demanda de gasolina y diferentes variables que podrían tener influencia sobre ésta. En general, la prioridad ha sido calcular las elasticidades precio e ingreso de la demanda, y así, determinar la sensibilidad que tiene la gasolina frente a los movimientos de los precios e ingresos de los consumidores. Para esta tarea los modelos más comunes son la regresión lineal, vectores autorregresivo(VAR) y modelos de corrección de errores (VEC)((Páez Martínez, 2009); (Moran Rugel et al., 2009); (Villamarin Lafaurie et al., 2007); (Galindo et al., 2015); (Rodrıguez, 2012)).

Cuando el enfoque del estudio es el pronóstico, el interés ya no está en explicar relaciones entre la demanda de gasolina y las variables regresoras. Más bien, el objetivo principal

está en encontrar los mejores modelos que puedan proyectar el consumo de gasolina en el futuro(Waheed Bhutto et al. (2017);Nasr et al. (2002);Mardiana et al. (2020);Kazemi et al. (2009);Hsing (1990)). Para esto, se utilizan diferentes modelos como ARIMA, SARIMA y redes neuronales. Estas estimaciones se hacen de manera mixta, primero se estima la demanda futura de gasolina explicada por sus valores pasados; y luego, se incluyen variables exógenas que puedan mejorar las predicciones.

El punto de partida en la modelación de la demanda de gasolina es considerar que se puede explicar como cualquier otro tipo de bien (Galindo et al., 2015). Siguiendo a la teoría económica el consumo esta determinado principalmente por los precios y los ingresos. Dentro de los precios se encuentra el relacionado al bien mismo, y adicionalmente se incluyen los precios de los bienes sustitutos y complementarios de la gasolina (Galindo et al., 2015). Los bienes sustitutos son entendidos como aquellos productos que cumplen con una función muy parecida. Son bienes que se pueden utilizar como reemplazo de los demás debido a que sus características son muy similares (Varian (1992), Nicholson (2005)). En el mercado hay algunos productos que pueden cumplir con la función de la gasolina como son el gas licuado vehicular, el oxifuel, hidrógeno y la energía eléctrica. Desde la teoría económica, se espera que la disminución en los precios de los sustitutos generen al mismo tiempo una disminución en demanda de gasolina, ya que esto incentiva a los consumidores a seguir realizando sus actividades a un menor coste.

Dos bienes son complementario cuando es necesario que ambos sean usados de manera simultánea (Varian, 1992); (Nicholson, 2005). Un complemento para la gasolina sería, por ejemplo, las ventas de automóviles. Se esperaría que, si se producen fuertes aumentos en la ventas de automóviles, la demanda de gasolina también aumentará en una proporción similar. En la siguiente tabla se presentan las variables utilizadas dentro de la literatura.

Sustitutos	Precio del gas natural vehicular, stock de autos híbridos y eléctricos, costo de transporte, precio diésel
Complementarios	Precio del petróleo,Precio relativo de los combustibles, cantidad de vehículos a motor de combustión.
Otros	Precio de la gasolina, impuesto a la gasolina, Tasa de cambio, PIB, tasa de desempleo, población, número de días de descanso, gasto en educación

Un análisis de inferencia en el que se estiman las elasticidades precio e ingreso de la demanda de gasolina corriente y acpm en Colombia es el presentado por Villamarin Lafaurie et al. (2007). En este trabajo se utiliza como variables explicativas el precio de la gasolina, la Tasa representativa del mercado (TRM), el precio del ACPM y el PIB como proxy de los ingresos de los individuos. A través de una regresión lineal múltiple se obtienen las elasticidades precio e ingreso de la demanda de gasolina. Se encuentra que la demanda de gasolina es un

bien inelástico, es decir que su demanda no cambia mucho cuando se producen cambios en los precios. Sin embargo, se observa cómo los cambios relativos en los precios de la gasolina frente al diésel si produce cambios significativos en la demanda de gasolina. Es decir, ante incrementos en los precios relativos de gasolina frente al diésel, el consumo de diésel aumenta.

También el trabajo realizado por Sapnken (2018) tiene como objetivo objetivo estimar las elasticidades precio e ingreso de la gasolina en Camerún a través de modelos de regresión Lineal. El valor agregado de este trabajo se encuentra en que con el mismo modelo que estima las elasticidades también es utilizado para realizar sus respectivos pronósticos. El autor utiliza como variables regresoras el PIB, precios de los combustibles y el costo del transporte. Los resultados son muy similares a los de Villamarin Lafaurie et al. (2007), en donde la demanda de gasolina resulta ser muy poco sensible ante los cambios en su respectivo precio. Sin embargo, es un poco más variable ante las variaciones del ingreso de los consumidores. Con base en estas estimaciones se espera que la demanda de gasolina en este país pueda crecer anualmente alrededor de 7 %.

Del mismo modo, Páez Martínez (2009) busca encontrar cuáles son las variables más importantes para explicar el consumo de gasolina en México. Utilizando como modelo la regresión lineal múltiple, establece una interacción entre la demanda de gasolina con el precio promedio del barril de petróleo, los impuestos a la gasolina, PIB per cápita, el tipo de cambio y el precio de la gasolina. Luego de un proceso de simulación, en donde el autor propone múltiples cambios en las variables regresoras, se encuentra que solo es significativa a nivel individual el tipo de cambio. Es decir, según este estudio, en México, la demanda de gasolina solo puede ser explicada con las fluctuaciones que tiene su moneda frente al dólar, situación que lleva a concluir que toda política interna en relación a la gasolina podría no ser significativa.

En general cuando se trabaja la demanda de gasolina por la vía de inferencia, el algoritmo más utilizado ha sido el de regresión lineal múltiple (Galindo et al., 2015). Esto se debe a la facilidad de obtener las elasticidades, ya que solo es suficiente con transformar los valores originales con el logaritmo natural y los coeficientes serán interpretados como el cambio porcentual que experimenta la demanda de gasolina ante cambios porcentuales en las variables explicativas. Sin embargo, en la literatura se han encontrado diferentes alternativas para observar la relación entre la demanda de gasolina y sus respectivas variables explicativas.

Por ejemplo, Moran Rugel et al. (2009) utilizan Mínimos Cuadrados Dinámicos (MCD) para estimar la elasticidad precio de la demanda y la elasticidad ingreso de la gasolina en Ecuador. Los MCD se basan en la misma estructura funcional que la regresión lineal múltiple. Sin embargo, se incluyen interacciones y rezagos de las primeras diferencias de las variables explicativas. De esta manera, los índices temporales ya no son los mismos, sino que permiten analizar las relaciones dinámicas entre variables. Los resultados de este trabajo son muy similares a los anteriores. En relación con el precio, se encuentra que la demanda de gasolina no cambia significativamente ante sus cambios. A pesar de esto, los autores resaltan el hecho de que el aumento del ingreso per cápita produce un aumento en la demanda de gasolina en el

largo plazo.

De acuerdo con Melikoglu (2014), el PIB, la población, el número de vehículos, la eficiencia de los combustibles y los precios de los combustibles determinan el comportamiento de los combustibles líquidos en Turquía. Para el pronóstico, el autor emplea un modelo de regresión lineal, cuyo MAPE resulta ser menor al 1 %. También plantea modelos cuadráticos y exponenciales donde la variable independiente es el tiempo. Estos últimos tienen MAPEs entre el 10 % y el 14 %. Para el diésel también se emplearon modelos lineales, cuadráticos y exponenciales. El modelo lineal en este último combustible tiene un MAPE menor de 11.8 %. Para el GLP, el mejor modelo tiene MAPE del 19.6 %. GLP representa el combustible de mayor uso en las carreteras de Turquía.

El trabajo de Al-Fattah (2020) también presenta un caso de pronóstico de gasolina en Arabia Saudita. Allí, la demanda de gasolina aumentó entre 1975 y 2015; sin embargo, dadas las nuevas reformas en los precios de los combustibles, el aumento en la eficiencia de los motores de los vehículos de combustión, la disminución de la tasa de crecimiento de la población y los cambios en el comportamiento del consumidor, la tasa de crecimiento de la demanda de gasolina ha venido disminuyendo desde 2016. Los autores plantean una red neuronal cuyas variables de entrada son el PIB, los precios de la gasolina y el diesel, la población, el número de vehículos de pasajeros, el número de vehículos de carga y una variable proxy de la eficiencia energética (tendencia del modelo). La salida de la red neuronal es la demanda de gasolina. Un algoritmo genético fue empleado para determinar la mejor combinación de variables del modelo. Los autores reportan que iniciaron con 48 variables candidatas. Al final, después de la selección de variables, el modelo emplea las siete variables descritas previamente. Se puede resaltar que el modelo no plantea una estructura autorregresiva de la demanda de combustible. Además, los mejores cuatro modelos obtenidos se promediaron para formar lo que se conoce como un modelo "ensamble." Se encontró que la demanda de gasolina en Arabia Saudita es elástica con respecto al ingreso (PIB) pero inelástica con respecto al precio. Un hecho importante, según el autor, es que el modelo tuvo un error de 0.046 % cuando se pronosticó que la demanda de gasolina de 2017 caería 2.5 % con respecto a la de 2016.

Por ejemplo, Mardiana et al. (2020) utiliza modelo como ARIMA, suavizamiento exponencial y redes neuronales para realizar el pronóstico mensual de la demanda de gasolina en Indonesia. En este trabajo se utiliza como variables la demanda mensual de gasolina con su respectivo precio. El autor compara los modelos para encontrar las mejores especificaciones econométricas de estos. Los resultados muestran que a través del suavizamiento exponencial y las redes neuronales las estimaciones son las más consistentes; mientras que el modelo ARIMA es el de menor rendimiento.

Sin embargo; para estimar la demanda de gasolina en Pakistán, Waheed Bhutto et al. (2017) ajustan dos modelos, a saber un modelo ARIMA (1,0,1) y un modelo polinómico de segundo orden. Para el caso del modelo ARIMA, los autores no emplean ninguna variable explicativa ni consideran estacionalidad debido a que los autores emplean datos anuales. Para

el caso del modelo polinómico de segundo orden, solo se considera la variable tiempo. De los resultados, el modelo que ofrece un mejor ajuste es el polinómico de segundo orden, cuyo $R^2 = 0.8779$.

Del mismo modo, Cervero (1985) utiliza el modelo ARIMA con datos mensuales para pronosticar la demanda de gasolina en los Estados Unidos. A diferencia del trabajo de Waheed Bhutto et al. (2017), el trabajo de Cervero (1985) considera datos mensuales de demanda de gasolina, con las cuales logra obtener estimaciones con una Raíz del Error Cuadrático Médio (RMSE) de 0.384.

El análisis con series de tiempo (ARIMA) también permite incluir variables exógenas adicionales a los valores rezagados de la variable dependiente. Por ejemplo, Rodríguez y Da Silva (2010) utiliza un modelo VEC para el pronóstico de la demanda de gasolina en Puerto Rico. Como variables exógenas, se utiliza el precio de la gasolina, el ingreso de las personas, la cantidad de vehículos a motor y la tasa de desempleo. Los resultados muestran que todas las variables explicativas propuestas tienen una relación directa con la demanda de gasolina.

Para el caso de Colombia, el trabajo realizado por Alonso Cifuentes et al. (2019) se plantea como objetivo encontrar el mejor modelo para el pronóstico de la demanda de gasolina. Para esto, el autor utiliza los datos de galones consumidos de gasolina corriente de Bogotá. El autor analiza los desempeños de pronósticos de el método de suavizamiento exponencial y los modelos autorregresivos. Concluye que el suavizamiento exponencial explica mejor la demanda de gasolina en Bogotá. A pesar de haber presentado un análisis univariado, los autores proponen emplear, adicionalmente, variables como flota de vehículos eléctricos y precios del gas natural vehicular.

Las redes neuronales artificiales para para el pronóstico de la demanda de gasolina han sido altamente empleadas. Nasr et al. (2002) utiliza en el proceso de entrenamiento del algoritmo la función de entropía cruzada, una especificación útil cuando se dispone de un horizonte temporal corto. Después de analizar diferentes configuraciones de variables exógenas, los autores concluyen que el precio de la gasolina y el número de vehículos registrados son las más adecuadas para el pronóstico de la demanda de gasolina.

Kazemi et al. (2009) utiliza redes neuronales artificiales con variables explicativas para explicar el comportamiento de la demanda de gasolina. Esta vez son incluidas variables que den cuenta de las características socio-económicas como la población, el PIB y el número de vehículos registrados. De esta manera, las estimaciones son consistentes con los valores tomados como prueba y la senda de pronóstico es coherente con los valores observados. La estimaciones de este trabajo comprende el periodo de 2007 a 2030 con frecuencia anual.

El trabajo propuesto por Azadeh et al. (2010) pone de manifiesto la superioridad de las redes neuronales frente a los modelos paramétricos más utilizados en el pronóstico de la demanda de gasolina. Esta investigación emplea datos de países como USA, Canadá, Japón, Kuwait e Irán desde el año 1992 hasta el 2005. De la misma manera que los trabajos anteriores, se incluye en la modelación variables explicativas como la población, número de vehículos registrados, PIB

y el precio real de la gasolina. Comparando los pronósticos realizados de los modelos clásicos con la red neuronal a través de sus respectivos MAPES se concluye que las mejores estimaciones son realizadas por las redes neuronales artificiales.

Más recientemente, (Güngör et al., 2021) demuestran a partir de la implementanción de modelos ARIMA, ARCH y GARCH que el brote de Covid-19 afectó de forma abrupta los pronósticos del consumo de gasolina en Turquía. Antes de pandemia la diferencia entre el valor observado y el pronosticado es del 0.8 % y después de pandemia esta diferencia se sitúa en 30 %. De manera que, para mejorar el desempeño predictivo de los modelos tras la crisis sanitaria es necesario agregar un factor de volatilidad para controlar la heterocedasticidad desatada en el consumo de gasolina e incluir como variable exógena una dummy que capture el cambio estructural de la serie de tiempo como consecuencia de la emergencia sanitaria.

(Afkhami et al., 2021) argumentan que el índice de volumen de búsqueda de Google (GSVI) de bus y tren puede utilizarse como una variable proxy para medir la tendencia de los consumidores a utilizar el transporte público. La inclusión del GSVI en el modelo de regresión lineal de la demanda de consumo de gasolina en el mercado estadounidense arroja coeficientes estadísticamente significativos y mejora el poder explicativo del modelo. Los signos de los parámetros estimados sugieren que una búsqueda masiva en Google sobre el transporte público puede coincidir con una menor demanda de gasolina. Al analizar de forma gráfica el comportamiento histórico de la demanda de gasolina, GSVI por bus y GSVI por tren se evidencia de una posible relación cointegrante entre las series de tiempo, lo cual es consistente con los resultados obtenidos de los autores.

(Al-Fattah, 2020) estima un modelo predictivo para analizar y pronosticar la demanda de gasolina de Arabia Saudita a partir de un algoritmo de inteligencia artificial denominado GANNATS. Está fundamentado en algoritmo genético (GA), redes neuronales artificiales (ANN), data mining (DM) y series de tiempo (TS), empleando como covariables el precio del diesel, PIB, población, transporte de vehículos ligeros y pesados (total vehículos) y variable proxy de eficiencia energética y tecnológica. Con base en un análisis de impacto de variables de los factores impulsores de la demanda de gasolina, se demostró que los vehículos ligeros y la población son los impulsores más influyentes de la demanda de gasolina, seguido de vehículos pesados y la variable proxy del avance tecnológico y la eficiencia energética. El desempeño predictivo del modelo GANNATS fue evaluado a partir de diferentes indicadores como MRE, MAE, MSE, RMSE cuyos resultados fueron -0.0118 %, 0.0792 %, 0.8392 % y 0.0916 %, respectivamente. El modelo predictivo exhibió un coeficiente de determinación del 0.98, lo que indica que el 98 % de los datos de entrenamiento son explicados mediante el modelo propuesto.

Finalmente, el Cuadro 5.4 presenta un resumen de los modelos y variables empleadas en las publicaciones analizadas para la proyección del consumo de gasolina.

Cuadro 5.4: Resumen revisión de literatura sobre gasolina

Autores	**Frecuencia**	**Modelo**	**Variables**

Cervero (1985)	1977-1983	ARIMA	Demanda de gasolina
Hsing (1990)	1960-1981	Box-Cox extended autoregres-sive	Demanda de gasolina, precio de la gasolina
Moreno Guerrero (2000)	2000	Muestreo y estimación del volumen	Demanda de gasolina total
Kayser (2000)	1981 Microdatos	Regresión lineal múltiple	Precio de la gasolina, ingreso, interacciones entre las variables
Nasr et al. (2002)	1993-1998 Mensual	Redes neu-ronales ar-tificiales	Demanda de gasolina,Precio y número de carros registrados
Villamarin Lafaurie et al. (2007)	Anual y Trimestral (1980-2006)	Regresión Lineal Múltiple	Precio de la gasolina corriente en promedios aritméticos, PIB anual,TRM
Kazemi et al. (2009)	1968-2030	Redes neuronales multinivel	Población,PIB, número de vehiculos, demanda de gasolina
Rao y Rao (2009)	1970-2005	Engle Granger dinámico, Mínimos cuadrados modificados, GETS Bound test, JML	Demanda de la gasolina, precio,ingreso
Azadeh et al. (2010)	1992-2005	Redes neu-ronales	PIB,Población, Número de vehículos, precio real de la gasolina

Rodríguez y Da Silva (2010)	1999-2006	Optimización por restricciones y luego análisis de cointegración con ECM	Precio gasolina,ingreso, cantidad de vehículos de motor, tasa de desempleo
Carbonell y Semerena (2014)	Anual 1980-2012	Modelo de regresión lineal	PIB,precio relativo del combustible
Melikoglu (2014)	Anual	Modelo de regresión lineal	PIB,población lineal,población, cantidad de vehículos y precios
Ackah y Frank (2013)	1971-2010	Distribución autorregresiva	Demanda de gasolina, índice de precios a los consumidores de gasolina, pib, gasto en educación
Galindo et al. (2015)	Anual 1960-2013	Panel de datos	Precio de la gasolina, ingreso de los consumidores, precios de bienes sustitutos y complementarios
Reyes Müller (2015)	2009-2013	Modelo de Grey-Markov	Demanda de gasolina, precio de gasolina , PIB, precio del petróleo, importación y exportación de gasolina
Azadeh et al. (2015)	2009-2011	Support vector machine	Demanda de gasolina, número de días de descanso por semana, pasajeros por kilómetro
Waheed Bhutto et al. (2017)	1991-2014	ARIMA	Demanda de Gasolina
Alonso Cifuentes et al. (2019)	Mensual (2006-2017)	ARIMA, Suavizamiento exponencial	Demanda de Gasolina

Sapnken (2018)	1994-2010 anual	Regresión lineal, cointegra-ción	Demanda de gasolina, PIB, precios , costo de transporte
Atalla et al. (2018)	1975-2015 Anual	Modelo autorregre-sivo	Ingreso per cápita, precio real de la gasolina y la tendencia subyacente de la demanda de energía
Algunaibet y Matar (2018)	Microdatos, encuesta	Modelo de elección de transporte	Tipo de transporte, precios de los combustibles e ingreso
Páez Martínez (2009)	Anual	Regresión Lineal Múltiple	Precios promedio del barril del petróleo, PIB per cápita deflactados, Impuestos a la gasolina, tasa de cambio
Mikayilov et al. (2020)	1980-2017 anual	Modelo de variación en el tiempo con coeficiente de cointegración Park Zhao	Precio real de la gasolina e ingreso per cápita
Al-Fattah (2020)	Anual anual	Redes Neuronales	PIB,Precios de la gasolina, población, número de vehículos, vehículos de carga y eficiencia energética
Mardiana et al. (2020)	2015-2019 mensual	ARIMA, Holt-Winters, Regresión lineal y redes neuronales	Demanda de gasolina, Precio de la gasolina

Moran Rugel et al. (2009)	Mensual	Mínimos cuadrados dinámicos, MCE,VAR	Ingreso de la personas, precio de la gasolina y precio del bien complementario
Güngör et al. (2021)	Mensual	ARIMA ARCH GARCH	Demanda de gasolina
Afkhami et al. (2021)	Tiempo real	Regresión lineal	Demanda de gasolina GSUS

5.6 Jet Fuel

El jet fuel o combustible de aviación es un tipo de combustible derivado del queroseno usado generalmente en motores de reacción o turbohélices. Es refinado y mezclado con pequeñas cantidades de otros aditivos con el fin de impedir el crecimiento de organismos dentro del motor, evitar la congelación del combustible al transitar en grandes alturas, controlar la manera como arde el combustible, evitar que el combustible se cargue eléctricamente, entre otros.

Dada las condiciones tan específicas que debe tener el jet fuel y las exigencias que deben cumplir los motores de las aeronaves para su adecuado funcionamiento, la estimación de la demanda del combustible se ha convertido en un insumo fundamental para mejorar la eficiencia de la cadena de suministros y medir el poder competitivo de las empresas dentro de la industria, ya que la reducción de los costos de funcionamiento estará relacionada con el mejor consumo del combustible.

Con esto en mente, en este informe también se ha revisado la bibliografía sobre proyecciones de la demanda del combustible a nivel mundial para identificar variables y modelos empleados. Entre los trabajos que se ha realizado para pronosticar la demanda de jet fuel o combustible de aviación, se presenta inicialmente el realizado por Chai et al. (2014), en donde el autor propone un análisis estructural de descomposición de la eficiencia del combustible de aviación[7], la facturación total del transporte aéreo[8] y del costo de combustible, para identificar cuales son las variables que poseen fuertes relaciones con las variables a descomponer, y poder emplear estos resultados en la predicción de la demanda de combustible de aviación en China. Una vez realizado el análisis estructural de descomposición para saber qué variables afectan la eficiencia del combustible y la facturación del transporte, los autores deciden plantear dos modelos univariados de series de tiempo para pronosticar estas dos variables, junto

[7]La eficiencia del combustible de aviación es la distancia máxima de vuelo de un avión por litro de combustible medido en km/tonelada

[8]La facturación total del transporte aéreo es la suma de la facturación del transporte de mercancías y del transporte de pasajeros.

con la demanda de combustible, y para ellos emplean un modelo de suavización exponencial con intercepto, tendencia y componente estacional y un modelo ARIMA(0,1,0).

Posteriormente, Chai et al. (2014) deciden aplicar un modelo de regresión lineal multivariante bayesiano (BMRL) con el fin realizar pronosticar las tres variables de interés a un horizonte de ocho años. Finalmente, combinan los pronósticos obtenidos por las metodologías anteriormente empleadas, a saber, el modelo de suavización exponencial, el modelo ARIMA(0,1,0) y el BMRL. De los resultados obtenidos por los modelos, los autores concluyen que a pesar de que el modelo ARIMA(0,1,0) presenta un desempeño predictivo superior al modelo ETS en el corto plazo, el modelo ETS ofrece resultados más cercanos a los datos históricos y a las expectativas teóricas.

Otro modelo de pronóstico implementado para la demanda de jet fuel, es el presentado por Chèze et al. (2011), el cual presenta un modelo de pronóstico a largo plazo con un horizonte a 2025 para un total de ocho regiones geográficas[9] y a nivel mundial[10], empleando para ello un modelo econométrico que busca pronosticar el tráfico aéreo para luego convertirlo en la demanda de jet fuel a través de los supuestos sobre mejoras en la eficiencia del tráfico aéreo, que dependerán del factor de carga y la eficiencia energética.

Para el caso de la estimación del tráfico aéreo, los autores emplean un modelo econométrico de datos de panel con variables como PIB, precio del jet fuel, choques exógenos y madurez del mercado para cada una de las ocho regiones bajo ciertos escenarios de PIB. Además de los pronósticos para el tráfico aéreo, los autores plantean un análisis de sensibilidad en donde muestran que el tráfico aéreo difiere dependiendo del grado de madurez del mercado que se está considerando, y señalan el papel relevante que tiene el PIB y el precio del jet fuel en el crecimiento del tráfico aéreo mundial. Por su parte, la estimación del factor de carga se calcula directamente como una medida de ocupación de las aerolíneas, en donde existe una relación directa entre altos factores de carga y una mejor eficiencia en el transporte, tal que mejorar los factores de carga de las aerolíneas disminuyen hace que disminuya significativamente el consumo de jet fuel sin recurrir a progresos tecnológicos (Chèze et al., 2011).

Finalmente, los autores concluyen que la demanda de combustible tendrá un aumento anual promedio de 1.9 % hasta el 2025, impulsado mayormente por el crecimiento que tendrá el PIB de las regiones, el incremento del tráfico aéreo y la mejora de la eficiencia energética. También señalan que a pesar de la contribución del tráfico aéreo para impulsar la demanda del jet fuel, este incremento se ve mitigado por los avances tecnológicos, los cuales indican que son variables que deben considerarse en modelos de pronóstico.

En el trabajo de Baumann y Klingauf (2020), los autores modelan el consumo de combustible de aviación mediante el empleo de un modelo *feed-forward network* (FFN) y un árbol de decisión, con el fin de brindar una herramienta que permita monitorear el combustible reque-

[9]Centro América y Norte América, Latinoamérica, Europa, Rusia y CIS (Comunidad de estados independientes), África, Medio Oriente, Asia y Oceanía (sin China) y China como la octava región.

[10]La suma de las ocho regiones.

rido para una misión de vuelo específica, para aumentar la eficiencia de las operaciones de la aeronave y disminuir los costos asociados a los vuelos. Los autores emplean datos de perfiles de combustible basados en el flujo de combustible para diferentes fases del vuelo, además de integrar en sus análisis la ruta del vuelo, el tipo de aeronave y las características ambientales para más de 180 mil vuelos que ocurrieron entre 2001 y 2003, con el fin de crear diagnósticos y pronósticos individuales.

Es de anotar que del total de 186 posibles variables encontradas en la base de datos divididas entre datos sobre el vuelo, parámetros del motor, entradas de control, datos de navegación, datos del entorno y parámetros del sistema, Baumann y Klingauf (2020) deciden consideran para sus modelos 30 parámetros que presentaban altas correlaciones. Una vez entrenados, validados y probados los dos modelos no paramétricos, los autores concluyen que el modelo de redes neuronales es el que presenta la menores métricas de validación para modelar el flujo de combustible durante las misiones de vuelo.

A partir de la investigación de (Atems, 2021), se pueden considerar para el estudio de la demanda de jet Fuel variables como: tarifas aéreas, tráfico aéreo, capacidad de las aeronaves, precios promedio de los boletos de avión, índice de precios de las tarifas aéreas (CPI Airface), total de aviones para vuelos regulares de pasajeros nacionales e internacionales de compañías aéreas, toneladas de ingresos aéreos de carga y correo, millas de pasajeros de ingresos aéreos, total de millas de asientos disponibles para los vuelos regulares de pasajeros nacionales e internacionales de las compañías aéreas, millas recorridas por los vehículos, pasajeros en transporte público, millas de pasajeros por tren.

De acuerdo con Melikoglu (2017) existe una necesidad general por el jet fuel para la aviación, cuya demanda se encuentra muy relacionada con el tráfico aéreo y el crecimiento económico de la región. El PIB, el número y eficiencia del combustible de los vehículos, la población y el precio del combustible son las variables independientes que, generalmente, se emplean para modelar la demanda de jet fuel de un país a una escala macro. Sin embargo, una alternativa es usar modelos de regresión lineal y no lineal basados en indicadores de crecimiento económico, los cuales proporcionan un medio sencillo y preciso para calcular previsiones. Aplicar un modelo lineal, cuadrático y exponencial para predecir el consumo anual de jet fuel de Turquía obtienen un MAPE de 16.6 %, 15.9 % y 29.9 % y un R-cuadrado de 0.9535, 0.9536 y 0.9068, respectivamente. Los autores argumentan que estos modelos pueden disponerse para pronosticar la demanda de jet fuel a mediano plazo en países en desarrollo con sistemas económicos sólidos y altas tasas de crecimiento del PIB.

Chèze et al. (2011) estima a partir de econometría dinámica de datos de panel que la demanda de jet fuel aumentará a una tasa de crecimiento media de 1.9 % al menos hasta 2025 y el tráfico aéreo incrementará a una tasa de crecimiento promedio de 4.7 %. En virtud de ello, se considera que las mejoras en la eficiencia energética permiten reducir el efecto del aumento del tráfico aéreo en el incremento de la demanda de jet fuel; por tanto, es poco probable que la demanda de combustible jet fuel disminuya durante el periodo 2008–2025 a menos que haya un

cambio tecnológico radical o se restrinjan los viajes aéreos. Sin embargo, dada la emergencia sanitaria en 2020, de todos los productos petrolíferos, el jet fuel fue proporcionalmente el más afectado según (IEA, 2021c) por la pandemia cuando los gobiernos realizaron el cierre de las fronteras internacionales, los consumidores cancelaron sus vacaciones y las conferencias de negocios fueron llevados a cabo de forma virtual. Por tanto, se espera que la demanda de jet fuel y queroseno retornen a sus niveles pre-pandémicos a partir de 2024, cuya tasa de crecimiento estimada 2019–2026 es de 3.8 %.

Finalmente, el Cuadro 5.5 presenta un resumen de los modelos y variables empleadas en las publicaciones analizadas para la proyección del consumo de jet fuel.

Cuadro 5.5: Resumen revisión de literatura sobre jet fuel

Autor	Frecuencia	Modelo	Variables
Chèze et al. (2011)	Anual	Modelos de datos de panel, Macro-Level Methodology.	Tráfico y capacidad de las líneas aéreas regualres internacionales y nacionales, Ingresos toneladas-kilómetro, Ingresos pasajeros-kilómetro, Factor de Carga de Peso (Ingresos toneladas-kilómetro/ Toneladas-kilómetro disponible)
Melikoglu (2017)	Anual	LR, NLR	PIB, cantidad y eficiencia del combustible de los vehículos, población y precio del combustible
Atems (2021)	Mensual	VAR	Producción mundial del crudo de petróleo, consumo de Jet Fuel, tarifas aéreas, tráfico aéreo, capacidad de las aeronaves, precios promedio de los boletos de avión, índice de precios de las tarifas aéreas (CPI Airface), total de aviones para vuelos regulares de pasajeros nacionales e internacionales, toneladas de ingresos aéreos de carga y correo, millas de pasajeros de ingresos aéreos, total de millas de asientos disponibles para los vuelos regulares de pasajeros nacionales e internacionales, millas recorridas por los vehículos, pasajeros en transporte público, millas de pasajeros por tren.

5.7 Queroseno

El queroseno es un tipo de combustible derivado del petroleo usado principalmente en el sector industrial y comercial. Dependiendo del nivel de refinación, se encuentran distintos tipos de combustible que pueden entrar en la categoría de queroseno; y a su vez dependiendo del tipo, se le destina un uso. La categoría general del queroseno aplica a hidrocarburos que hierven a temperaturas de alrededor de 150°C o 290°C entrando en la clasificación de destilados intermedios. A temperatura ambiente se encuentra en estado líquido y tiene una baja presión de vapor. Entre sus usos están la aviación civil y militar, combustible para generadores, calentadores e iluminación. Dependiendo de la aplicación se emplean combustibles con bajo punto de congelación o niveles de azufre.

Koshala et al. (1999b) desarrollan un ajuste de curva con base en los históricos de consumo en el país (entre 1957 y 1992), además de otras variables descriptivas. El modelo arroja que el kerosene no es un bien complementario de la electricidad y por ende no debería tener subvenciones del Estado. Por otra parte, se concluye que en la medida en que se extiendan los abonados de energía eléctrica irán disminuyendo los consumidores de kerosene.

Iwayemi et al. (2010) parece una actualización del modelo presentado por Koshala, ya que se desarrolla un modelo que parte de determinar la curva que representa el consumo del país (con datos entre 1977 y 2006). Se concluye casi lo mismo; es decir, el kerosene no es un bien complementario de la electricidad y por ende no debería tener subvenciones del Estado. Por otra parte, se concluye que en la medida en que se extiendan los abonados de energía eléctrica irán disminuyendo los consumidores de kerosene.

Rasouli (2018) pronostica la demanda anual de Irán de queroseno, GLP, gasoil, fuel oil y gasolina para el periodo 2017–2036 empleando los modelos de regresión lineal, la función de producción Cobb-Douglas y regresión difusa. De acuerdo con los resultados obtenidos a partir del MAPE, el test ANOVA y Tukey, el modelo más apropiado para realizar previsiones de la demanda de los productos petrolíferos de Irán es la regresión difusa; se estima que la demanda de gasolina y gasoil aumentará considerablemente para los próximos veinte años lo cual implica que estos productos son estratégicos y su demanda requiere una planificación integral y de largo plazo. Adicionalmente, el autor argumenta que los precios de los productos petrolíferos fueron factores sustanciales en las previsiones realizadas y sugiere para futuras investigaciones considerar el efecto de los combustibles líquidos alternativos en la construcción de los modelos.

Abdullahi (2014) estima a partir de modelos de series de tiempo estructurales (STSM) la demanda de gasolina, diésel, queroseno, fuel oil y LPG, los cuales son los principales productos derivados del petróleo de Nigeria. Este modelo permite incluir la estimación de una tendencia estocástica, la cual resulta fundamental para estimar la sensibilidad de la demanda de los productos petrolíferos ante cambios porcentuales del precio. Según el autor, los modelos enfocados en la cointegración no son apropiados cuando se requiere generar estimaciones sólidas de las elasticidades, puesto que la excesiva dependencia de la cointegración sin la debida consideración de los cambios estructurales conduce a un potencial sesgo significativo en la elasticidad precio e ingreso. Esta investigación sustenta que todas las demandas de los productos petrolíferos son inelásticas al precio y al ingreso, dado que las elasticidades estimadas usando STSM son (0.11 y -0.23), (0.17 y -0.30) y (0.10 y -0.20) para la gasolina, el diésel y el queroseno respectivamente, mientras que (0.27 y -0.18) y (0.64 y -0.58) para el fuel oil y GLP, respectivamente. Dadas la bajas elasticidades al precio, estos productos representan una base impositiva importante que podría ser explotada por el gobierno en el futuro. Adicionalmente, este estudio encuentra que la tendencia subyacente de la demanda de gasolina y su nivel es de carácter estocástico, la demanda de GLP tiene una tendencia suave, un nivel fijo pero con pendiente estocástica. Mientras que el diésel, queroseno y fuel oil exhiben una tendencia a nivel local que es estocástica pero con pendiente fija.

Bhattacharyya y Blake (2009) realizan una estimación de la demanda de los productos petrolíferos del Medio Oriente y África del Norte a partir del modelo de regresión lineal múltiple, empleando como variables exógenas el PIB per cápita, el precio real y el rezago del consumo per cápita de los productos petrolíferos. Los resultados obtenidos demuestran que la mayoría de los coeficientes del precio estimados son no estadísticamente significativos, ello en virtud de que los países objeto de estudio han mantenido precios relativamente bajos lo cual conduce a pequeñas elasticidades. Asimismo, se argumenta que en el periodo de análisis existe una falta de variación de los precios que no permite que el modelo capture de forma adecuada las elasticidades; también, puede justificarse por razones estructurales dentro de las economías de esos países que amortiguan los efectos de los ingresos sobre la demanda de los productos petrolíferos. Otras de las razones es que, por ejemplo, para el caso del queroseno este se comporta, para niveles bajos de ingresos, como un bien normal para cocinar e iluminar dado que es un sustituto de combustibles más baratos. Sin embargo, para niveles altos de ingreso, el queroseno se comporta como un bien inferior en comparación con otras fuentes comerciales como GLP, gas natural y electricidad. Además, en vista de que existe el efecto de precios bajos y controlados del queroseno, el modelo puede fallar al capturar la elasticidad precio de la demanda de forma significativa. Por tanto, este trabajo sugiere, para los pronósticos de la demanda de petrolíferos, utilizar estrategias alternativas como enfoques económicos de ingeniería y considerar variables socioeconómicas y políticas no relacionadas con los precios de los gobiernos de la región puesto que pueden jugar un papel importante.

Finalmente, el Cuadro 5.6 presenta un resumen de los modelos y variables empleadas en las publicaciones analizadas para la proyección del consumo de queroseno.

Cuadro 5.6: Resumen revisión de literatura sobre queroseno

Autor	Frecuencia	Modelo	Variables
Koshala et al. (1999b)	Anual	Ajuste de curva	Precio del queroseno, Ingreso per cápita.
Kazemi et al. (2009)	Anual	LR	PIB per cápita, Precio Real del Queroseno, Rezago del consumo del Queroseno.
Iwayemi et al. (2010)	Anual	Cointegración Multivariada	Histórico de la demanda del queroseno.
Abdullahi (2014)	Anual	STSM, ARDL	PIB, Precio Real del Queroseno, Tendencia subyacente de la demanda de energía (UEDT)
Rasouli (2018)	Anual	LR, Fuzzy LR, Cobb-Douglas	PIB, población

5.8 Conclusiones

A partir de la revisión exhaustiva de literatura, se ha encontrado que para el planteamiento de los modelos de cada combustible líquido es importante considerar entre las variables explicativas de los modelos, las variables macroeconómicas del PIB, la población, tasa de desempleo, el precio de los combustibles líquidos, demanda de energía eléctrica, demanda de gas natural, flota de vehículos (de combustión interna y eléctricos) y motocicletas; y demanda de energía eléctrica, las variables de efecto calendario y las variables meteorológicas.

Es de anotar que las variables relacionadas con la eficiencia energética y tecnológica, eficiencia de los combustibles líquidos, flota aérea, número de pasajeros, tarifas aéreas, entre otras, son consideradas como una alternativa promisoria para modelar algunos combustibles líquidos.

Es importante que en Colombia se pueda contar con la mayor cantidad de datos posibles de estas variables con el fin de experimentar y analizar los casos de demanda de combustibles a nivel local de manera detallada. No sólo es importante contar con datos históricos, sino también con proyecciones oficiales con el fin de realizar proyecciones de largo plazo confiables que contribuyan de la mejor manera en la construcción de los modelos.

Capítulo 6 Metodología propuesta de proyección de combustibles líquidos

6.1 Introducción

En este trabajo, se han construido metodologías de proyección de combustibles líquido y GLP basadas en herramientas de aprendizaje de máquina e inteligencia artificial por los resultados satisfactorios reportados en la literatura y su facilidad en la implementación.

Algunos trabajos que sugieren emplear modelos VAR han planteado *endogeneidad* entre el PIB, la tasa de desempleo y la demanda de combustibles como la gasolina o el diésel. Sin embargo, estos supuestos, para el caso de Colombia, son difíciles de argumentar.

Los modelos VEC utilizados en los trabajos mencionados en la revisión de literatura, a pesar de trabajar con datos mensuales, únicamente consideran la cointegración a la frecuencia cero. No consideran la cointegración a las frecuencias estacionales como debería ser ante series estacionales como las de demandas de combustibles. Esto, probablemente, generaría sesgos en sus resultados.

Para el caso de los modelos SARIMA, la mayoría de las aplicaciones no consideran variables exógenas en el modelo, excepto uno de los trabajos mencionados, que considera la variable de temperatura como variable exógena. Los modelos SARIMA con variables exógenas presentan problemas de convergencia en los algoritmos numéricos que buscan maximizar la función de verosimilitud cuando hay una cantidad considerable de variables exógenas. En general, en múltiples ocasiones no se alcanza la convergencia. Por tal razón, no es recomendable depender de modelos cuyos procesos de estimación de parámetros no siempre convergen. Al final de este capítulo se presenta un análisis de tipo numérico de este modelo.

Este trabajo presenta la modelación predictiva de varios combustibles líquidos utilizados en Colombia y para esto se recurre entre otras, a metodologías de aprendizaje de máquinas o equivalentemente a modelos de inteligencia artificial, que según Hastie et al. (2001), estas metodologías predictivas son probablemente las de mayor uso en la actualidad. La razón se debe a que estas metodologías predictivas no solo son útiles para pronosticar demandas de combustibles líquidos o energía eléctrica, sino para permiten pronosticar tipo de variables, independientemente de cuál sea su tipo de escala. En particular, en este trabajo, se ha planteado proponer metodologías de proyección de combustibles líquidos como:

- Regresión lineal: es un modelo muy transparente a la hora de ser interpretado por su simple formulación. Además, si no hay multicolinealidad perfecta, es fácil estimar el modelo por el método de mínimos cuadrados. Según Greene (2000), cuando este modelo se

plantea de forma adecuada, los resultados de inferencia y predicción son útiles.

- Redes neuronales recurrentes de corta y larga memoria (LSTM): son modelos no paramétricas. A pesar de que son modelos no interpretables, la revisión de literatura muestra un muy buen poder predictivo.
- Modelos aditivos generalizados (GAM): son también no paramétricos, y según Gareth et al. (2013), son modelos que además de ser útiles para la inferencia, también lo son para los pronósticos.
- Regresión spline adaptativa multivariante (MARS): modelos no paramétricos que, a pesar de no ofrecer interpretabilidad, tienen capacidades similares a las redes neuronales en cuanto a pronóstico.
- Regresión con y sin penalización de parámetros (LASSO): son modelos de que evitan el problema del sobre ajuste (bajos errores de entrenamiento y altos errores de predicción) o el problema de alta varianza. También son útiles para inferencia y pronóstico.

En general, entre las ventajas que poseen estos modelos de aprendizaje de máquina es que no requieren verificar supuestos sobre el conjunto de datos, que en ocasiones son difíciles de cumplir con datos de la vida real. Por ejemplo, la normalidad de la variable dependiente no siempre se tiene. De hecho, autores como Pathak et al. (2018), utilizan el aprendizaje de máquinas para hacer predicciones en combustibles líquidos. Para el pronóstico de la demanda de energía eléctrica, estos modelos también han sido ampliamente empleados (Sigauke, 2017).

El modelado planteado en este trabajo es esencialmente univariada. Las variables explicativas determinan el comportamiento de los combustibles líquidos; pero el consumo de un combustible líquido no determina el comportamiento de las variables explicativas. Es decir, el tamaño de la población puede determinar el consumo de combustibles líquidos, pero el consumo de los combustibles líquidos no determina el tamaño de la población.

En las siguientes secciones se presenta una breve discusión de los modelos predictivos utilizados en este trabajo. También se presenta una metodología de combinación de pronósticos que permite mejorar (en algunos casos) los pronósticos entregados por los modelos individuales. Posteriormente se presenta la estrategia de *bootstrap* empleada para la construcción de los intervalos de confianza de las proyecciones. Luego, se describen todas las variables explicativas empleadas en los ejercicios de proyección de combustibles. Posteriormente, se presentan las limitaciones de las metodologías clásicas de pronósticos VAR, VECM y SARIMAX en el caso de pronósticos de combustibles líquidos. Al final de este capítulo se ilustran diferentes medidas de bondad de ajuste que se tienen en cuenta en las metodologías planteadas en este trabajo.

6.2 Modelo de Regresión Lineal Múltiple (MLR)

Un modelo de regresión lineal es una metodología que busca establecer la relación existente entre una variable dependiente o respuesta Y respecto a una o más variables independientes X_j, con $j = 1, 2, 3, \ldots, k$.

Entonces, si suponemos una muestra de tamaño n para las variables Y y X_j, para $j = 1, 2, \ldots, k$, entonces tendremos que la relación lineal entre estas variables podrá plantearse de la forma:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \ldots + \beta_k x_{k,t} + \varepsilon_t$$

Donde en este caso y_t representa las t-ésima observación de la variable dependiente, $x_{j,t}$ representa la t-ésima observación para la j-ésima variable, los β_j representan los parámetros de ajuste que relacionan a y_t con $x_{j,t}$ y μ_t representa el término de error aleatorio, el cual explica por qué el modelo de regresión lineal no se ajusta exactamente los datos.

Finalmente, para realizar el proceso de estimación, se emplea el método de mínimos cuadrados con el fin de calcular el valor de los parámetros β_k.

6.3 Modelo Aditivo Generalizado (GAM)

Para superar el problema de relaciones lineales entre las variables dependiente y explicativas, Hastie y Tibshirani (2017) proponen reemplazar en el modelo MLR, el componente lineal $\beta_0 + \sum_{j=1}^{k} \beta_j x_{j,t}$ por una función de suavización no lineal. La función propuesta por estos autores se define como:

$$y_t = \beta_0 + \sum_{j=1}^{k} f_j(x_{j,t}) + \mu_t$$

donde

$$f_j(x_{j,t}) = \sum_{j=1}^{q} b_j\left(x_{j,t}\right)\beta_j$$

se conoce como función de suavización, $b_j\left(x_{j,t}\right)$ la función base de $x_{j,t}$ y q el número de nodos de la función de suavización.

Es de anotar que el modelo aquí planteado se conoce como modelo aditivo debido a que está dado por la suma de funciones de suavización no lineales desconocidas que lo hacen semiparamétrico. En lugar de que cada variable explicativa sea incorporada al modelo de una forma lineal, en este caso las variables explicativas ingresan al modelo como una función no lineal a través de la función base respectiva.

Como lo señala Wood (2017) el modelo GAM en comparación con el modelo MLR, permite una relación más flexible entre la variable de respuesta y sus variables explicativas, ya que le permite recoger de esta forma, posibles relaciones no lineales entre las variables. Sin embargo, esta flexibilización tiene un costo; si queremos una relación muy flexible, y por tanto, un ajuste más preciso a los datos, la función base tendrá generalmente una gran dimensión, lo que nos puede conllevar aun posible sobreajuste. De acuerdo con Wood (2017), si se asume que $y_t \sim EF(\mu_t, \phi)$ donde EF indica familia de distribuciones exponenciales, el sobre ajuste se puede

combatir recurriendo a la función de verosimilitud penalizada dada por:

$$l*(\beta) = l(\beta) - \frac{1}{2}\sum_{j=2}^{k} \lambda_j \beta_j' S_j \beta_j$$

donde $l(\beta)$ es la log-verosimilitud del modelo no restringido, mientras que los términos S_j son matrices de coeficientes conocidos que ayudan en el proceso de penalización de las funciones de suavización. Los λ_j son los hiperparámetros de penalización que controlan el trade-off entre ajuste y suavización de $f_j(x_{j,t})$ y se seleccionan de forma que minimicen el score de doble validación cruzada dado por:

$$v_d = \frac{n\|y - \hat{\mu}\|}{[n - 1.5tr(A))]^2}$$

donde

$$A = X\left(X'X + \sum_{j=2}^{k} \lambda_j S_j\right) X'$$

Para más detalles sobre GAM véase Wood (2017).

6.4 Regresión lineal con Shrinkage (LASSO)

De acuerdo con Hastie y Tibshirani (2017) los modelos de pronóstico que sufren del problema de la multicolinealidad o que tienen un gran número de variables explicativas, tienden a desempeñarse muy bien en la muestra de entrenamiento, pero no tan bien en la muestra de prueba, aún cuando éstos exhiban altos R^2 en sus estimaciones.

Según Hastie y Tibshirani (2017), una de las razones por las cuales se presenta esta situación se debe al gran numero de variables, en donde al momento del proceso de entrenamiento del modelo, se aprenden patrones falsos que no pertenecen al proceso generador de los datos, y por lo tanto a la hora de hacer predicciones el modelo intentará pronosticar los patrones falsos, logrando de esta forma que su poder de predicción se vea comprometido.

Esta situación se conoce como sobreajuste o alta varianza, y se da cuando el modelo se ajusta muy bien a los datos de entrenamiento, pero no los datos de prueba, debido a que en lugar de trazar una función suave que pase cerca a los valores de la variable respuesta, el modelo traza una función que pasa sobre todos las observaciones, de forma que se genera una alta variabilidad de los valores ajustados por el modelo.

Para analizar el problema del sobre ajuste, definamos inicialmente el error cuadrático medio que se obtiene para los datos al realizar el cálculo del modelo de entrenamiento como (Hastie y Tibshirani, 2017)

$$MSE = \frac{\sum_{i=1}^{n}(y_t - \hat{y}_t)^2}{n}$$

Como lo muestra Greene (2000), entre más variables se incluyan en la estimación de un modelo, es decir, entre más se sobre ajuste el modelo, probablemente más bajo será su MSE de entrenamiento, de hecho, Greene (2000) plantea que el MSE de entrenamiento es una función decreciente del numero de variables explicativas. Por su parte, Hastie y Tibshirani (2017) muestra algo contrario, y es que a medida de que nos movemos de muy pocas variables a muchas variables en un modelo, pasamos del sub ajuste del modelo, donde el poder predictivo es bajo (alto MSE), al sobre ajuste del modelo donde el poder predictivo del modelo también es bajo (alto MSE).

Del ejercicio realizado por Hastie y Tibshirani (2017), los autores encuentran que en la trayectoria de camino que hay entre el sub ajuste al sobre ajuste, hay un punto donde el MSE de prueba es el más bajo posible. Dicho punto corresponderá a la situación óptimo bajo la cual deberá ser estimado el modelo, debido a que se tendrá la suficiente información como para poder obtener adecuados pronósticos.

Existen diferentes métodos para solucionar el problema del sobre ajuste y la multicolinealidad en un modelo de regresión, entre los cuales aparece el método regularizado conocido como Regressión LASSO.

Si centramos la variable dependiente respecto a su media, es decir, si partimos del modelo de regresión lineal

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + e_i$$

y usamos en lugar de la variable $y_{j,i}$, a $y_{j,i} - \bar{Y}_j$, tendremos el modelo de regresión en desviaciones dado por[1]

$$(y_t - \bar{Y}) = \beta_1(x_{1,t} - \bar{X}_1) + \beta_2(x_{2,t} - \bar{X}_2) + \ldots + (x_{k,t} - \bar{X}_k) + e_t$$

Entonces el método de estimación LASSO obtendrá el estimador de mínimos cuadrados como el vector de coeficientes que minimiza la suma de residuales al cuadrado penalizada por la suma del valor absoluto de los coeficientes, conocida como penalización L_1, tal que

$$RSS^* = \sum_{i=1}^{n} \left[y_t - \beta_1(x_{1,t} - \bar{X}_1) - \beta_2(x_{2,t} - \bar{X}_2) - \ldots - \beta_1(x_{k,t} - \bar{X}_k) \right]^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

En donde el parámetro λ será el encargado de eliminar las variables o términos que generen el sobre ajuste del modelo, en donde se le dará un valor de cero a las variables redundantes dentro de la estimación del modelo de regresión LASSO. En otras palabras, el parámetro λ será el encargado de llevar el modelo al punto óptimo, y su estimación se logrará a través de un proceso de validación cruzada tal como se establece en Hastie y Tibshirani (2017).

[1] Greene (2000) demuestra que la estimaciones de los coeficientes de las pendientes parciales para este modelo, serán las mismas estimaciones de mínimos cuadrados; mientras que, la estimación del intercepto se podrá hacer mediante la ecuación de mínimos cuadrados para intercepto, de la forma

$$\beta_0 = \bar{Y} - \sum_{j=1}^{k} \beta_j \bar{X}_j$$

6.5 Regresión Spline Adaptativa Multivariante (MARS)

El modelo MARS es un modelo muy similar al GAM, debido a que ambos parten de una estructura de modelación similar dada por:

$$y_t = \beta_0 + \sum_{j=1}^{L} \beta_j b_j(X) + \mu_t$$

donde en el caso del modelo MARS, el término $b_j(X)$ está dado por

$$b_j(X) = \prod_{m=1}^{M_j} \Phi_{j,m}(X_{q(j,m)})$$

es el producto de M_j funciones multivariadas $\Phi_{j,m}(X)$, donde M_j es un número finito y $q(j,m)$ es un índice que depende de la m-ésima función base y la m-ésima función spline. De esta forma para cada j, $b_j(X)$ puede consistir en una sola función spline o un producto de dos o más funciones spline, en las cuales, ninguna variable explicativa puede aparecer más de una vez dentro del producto. Estas funciones spline (para j impares) a menudo se toman como lineales de la forma $\Phi_{j,m}(X) = X - t_{l,m}$ y $\Phi_{l+1,m}(X) = (t_{l,m} - X)_+$ con:

$$(x-t)_+ = \begin{cases} x - t & si\, x > t \\ 0 & \text{en otro caso} \end{cases}$$

$$(t-x)_+ = \begin{cases} t - x & si\, x < t \\ 0 & \text{en otro caso} \end{cases}$$

donde $t_{l,m}$ es el nudo de $\phi_{l,m}(X)$ ocurriendo en uno de los valores de $X_{q(l,m)}$ con $m = 1, 2, 3, \ldots, M_t$ y $l = 1, 2, \ldots, L$.

El proceso de estimación del modelo MARS, es similar al que realiza el modelo de regresión lineal hacia adelante, con la diferencia de que en lugar de usar las variables explicativas originales, éste utiliza productos de sus funciones spline $\Phi_{l,m}(X) = (x - t_{l,m})_+$, en donde el spline de cada variable puede aparecer solo una vez.

Es de anotar que, a pesar de que el spline de cada variable pueda aparecer solo una vez, las variables puede estar dentro del producto de un, dos o más spline, lo cual puede generar problemas de sobre ajuste. Por esta razón en este modelo se busca minimizar los criterios de validación cruzada generalizada similares al v_d en el caso del modelo GAM. Similar al caso del modelo de regresión lineal, en MARS minimizamos la suma de residuales al cuadrado, pero penalizando esta última función, también como en el caso de los modelos GAM, la penalización busca combatir el sobre ajuste.

Para más detalles de esta metodología véase (Hastie y Tibshirani, 2017).

6.6 Modelo de Redes Neuronales Recurrentes de Corta y Larga Memoria (LSTM)

La arquitectura de la RNN de corta y larga memoria (LSTM) fue propuesta por Hochreiter y Schmidhuber (1997). Como lo señalan estos autores LSTM es una arquitectura en la cual la información transmitida a una capa oculta es procesada iterativamente por una celda de estado c_t y dos componentes adicionales conocidos como puerta de entrada y puerta de olvido. Dadas las variables de entrada en x_t y el estado oculto en el periodo anterior h_{t-1}, la puerta de olvido se calcula como:

$$f_t = \sigma\left(w_{fx}x_t + w_{fh}h_{t-1} + b_f\right)$$

Donde $\sigma(\cdot)$ es generalmente una función sigmoidal, aunque hay más posibilidades para esta función. f_t es una matriz de elementos con uno y cero; si uno de los elementos de f_t está cerca de uno, este será un elemento que se mantendrá para los cálculos futuros de la celda de estado y si está cerca de cero, este será descartado. La puerta de entrada se calcula como:

$$i_t = \sigma\left(w_{it}x_t + w_{ih}h_{t-1} + b_i\right)$$

$$g_t = tanh(w_{gx}x_t + w_{gh}h_{t-1} + b_f)$$

De nuevo esta es una matriz con elementos entre cero y uno en el caso de i_t y entre $(-1, 1)$, en el caso de g_t. Esta decide que valores positivos o negativos de la entrada deberían ser utilizados para actualizar la celda de estado, si un elemento en i_t está cerca de uno se utilizará la mayoría de la información en g_t, y si está cerca de cero, no se utilizará la información en g_t para actualizar la celda de estado.

Con c_{t-1}, f_t, i_t, y g_t la nueva celda de estado es calculada como:

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t.$$

El estado oculto se calcula recurriendo a la otra puerta, conocida como puerta de salida dada por:

$$o_t = \sigma\left(w_{ox}x_t + w_{oh}h_{t-1} + b_o\right)$$

la cual de nuevo entrega valores entre cero y uno. Si o_t está cerca de uno, indica que la información de la celda de estado debería ser utilizada para calcular la capa oculta y si está cercana a cero indica que tal información debería desecharse. La nueva capa oculta se calcula como

$$h_t = o_t \otimes tanh(c_t)$$

La información de esta capa oculta es enviada a la capa de salida mediante al siguiente

transformación:

$$y_t = g(w_{yh}h_t + b_y)$$

donde $g(\cdot)$ aplicará una función que depende de la escala de la variable dependiente. Por ejemplo, si la variable es continua, será transformación lineal tipo:

$$y_t = b_y + w_{yh}h_t + \varepsilon_t$$

y si estamos es en un escenario de clasificación multinomial, $g(\cdot)$ será una transformación $softmax(b_y + w_{yh}h_t)$.

Como lo señala Goodfellow et al. (2016), LSTM puede ser implementada utilizando varias capas en cada una de las puertas de entrada, olvido y salida como también en la celda de estado. Sin embargo, como lo señalan estos autores, esto último complica el proceso de aprendizaje (estimación) ya que la optimización de la función de pérdida (función objetivo a minimizar) se convierte en un proceso difícil. Una salida a este problema, lo sugiere Graves et al. (2013) quienes sugieren usar estructuras profundas en cada uno de los pasos iterativos de la RNN LSTM y sus estados de la forma:

$$h_t^n = g\left(w_{h^{n-1}h^n}h^{n-1} + w_{h^nh^n}h^n + b_n^n\right)$$

En realidad esta última ecuación corresponde al proceso de aplicar varias capas ocultas en la red LSTM, que es lo que en realidad hoy en día llamamos aprendizaje profundo (deep learning).

Para el caso de la capa de salida se tiene lo siguiente:

$$y_t = g\left(w_h^N h_t^N + b_y\right)$$

donde de nuevo $g(\cdot)$ es una función lineal para una variable dependiente con escala continua. La red LSTM, sobre todo cuando se tienen varias capas ocultas son modelos sobre parametrizados, por tal razón generalmente para evitar un poco el sobre ajuste se utilizan procedimientos de estimación penalizada tipo LASSO, los detalles del procedimiento de estimación se encuentran en Goodfellow et al. (2016), uno de los textos clásicos hoy en día en el tema de redes neuronales.

Por último, a modo de conclusión sobre la presentación de los modelos expuestos en la Sección 6.2, Sección 6.3, Sección 6.4, Sección 6.5 y Sección 6.6, se tiene que en el modelo de regresión lineal múltiple, los coeficientes representan pendientes parciales y son susceptibles de ser interpretados. En los modelos GAM los coeficientes no son interpretables, pero si sus funciones de suavización, de hecho, $f_j(x_{jt})$ es la relación entre y_t y x_{jt} manteniendo lo demás constante, la cual es en realidad una función y no un solo valor como en LR por lo que los modelos GAM son mucho más ricos desde el punto de vista de la inferencia estadística, con la ventaja de que ajustan las relaciones no lineales entre y_t y x_{jt}. Por otro lado, los modelos

LSTM y MARS son modelos de caja negra donde ni las capas ocultas ni los productos spline y sus coeficientes correspondientes tienen una interpretación útil, sin embargo, estos modelos se utilizan y han dado buenos resultados en la predicción demanda de energía eléctrica y demanda de gas natural, véase Liu et.al (2021).

6.7 Metodología de combinación de pronósticos

Con el fin de ofrecer alternativas a la UPME sobre estrategias de combinación de pronósticos, se ha adoptado e implementado una estrategia de combinación que se enfoca en la minimización de la varianza del error de pronóstico, y posiblemente, también mejorar el ajuste obtenido tanto dentro como fuera del conjunto de datos de entrenamiento.

Considere que se tiene un conjunto de modelos $\mathcal{M}$ de pronósticos. El i-ésimo modelo de pronóstico para un mes t se denota por $f_{i,t}$. Cada modelo, tiene un error aleatorio con respecto a la verdadera demanda de combustible y_t^{real}. Es decir,

$$f_{i,t} = y_t^{\text{real}} + e_{i,t}, \ \forall i \in \mathcal{M}.$$

Y sea $\sigma_i^2 = \text{Var}\left(e_{i,t}\right)$ la varianza del error de pronóstico del modelo $i \in \mathcal{M}$.

El pronóstico combinado para el mes t, denotado por $\hat{f}_t$, sería construido como la combinación lineal de los pronósticos en el conjunto $\mathcal{M}$, dado por:

$$\hat{f}_t = \sum_{i \in \mathcal{M}} w_i f_{i,t} \tag{6.1}$$

donde w_i representa el peso asignado al i-ésimo modelo. Estos pesos se pueden determinar a través de diferentes metodologías como se ilustra en Bates y Granger (1969) y Bichpuriya et al. (2016). Para este trabajo, como se mencionó previamente, se ha empleado la estrategia de combinación que minimiza la varianza del error de pronóstico, donde el error del pronóstico es determinado como la diferencia $\hat{f}_t - y_t$. Basados en lo anterior, el problema de optimización a resolver que permite obtener los pesos que minimizan la varianza del error es dado por:

$$\begin{aligned} \underset{w}{\text{minimizar}} \quad & \text{Var}\left(\hat{f}_t - y_t^{\text{real}}\right) = \text{Var}\left(\sum_{i \in \mathcal{M}} w_i f_{i,t} - y_t^{\text{real}}\right) = \text{Var}\left(\sum_{i \in \mathcal{M}} w_i e_{i,t}\right) = \sum_{i \in \mathcal{M}} w_i^2 \sigma_i^2 \\ \text{sujeto a:} \quad & \sum_{i \in \mathcal{M}} w_i = 1 \\ & w_i \geq 0, \ \forall i \in \mathcal{M}. \end{aligned} \tag{6.2}$$

Asumiendo que los errores $e_{i,t}$ son independientes, la solución a este problema de optimización cuadrática lleva a que los pesos óptimos de la combinación sean

$$w_i = \frac{1/\sigma_i^2}{\sum_{j\in\mathcal{M}} 1/\sigma_j^2},\ \forall i \in \mathcal{M}. \tag{6.3}$$

La Ecuación 6.3 establece que a medida que la varianza de un modelo sea más pequeña, el peso asignado a dicho modelo será significativo. Y por el contrario, cuando la varianza es grande, el error del modelo será pequeño.

Finalmente, los pesos dados por la Ecuación 6.3 se emplean en la combinación de pronósticos presentada en la Ecuación 6.1 para obtener finalmente mezcla adecuada de los modelos individuales.

6.8 *Bootstrap* en Modelos de Regresión

Toda la sección de bootstrap mostrada aquí se basa en Chernick y LaBudde (2014) y Lahiri y Lahiri (2003). El bootstrap es un enfoque general de inferencia estadística basada en la construcción de una distribución de muestreo para una estadística de interés mediante el muestreo repetido de los datos disponibles. Booststrap ofrece las siguientes ventajas:

- Debido a que no requiere supuestos de distribución (como errores distribuidos normalmente), el bootstrap puede proporcionar inferencias más precisas cuando los datos no se comportan bien en el sentido de aproximar una distribución paramétrica o cuando el tamaño de la muestra es pequeño.
- Es posible aplicar el bootstrap a estadísticas con distribuciones de muestreo que son difíciles de obtener, incluso asintóticamente, como por ejemplo las distribuciones de los percentiles muestrales. Por esta razón, el Bootstrap que es una técnica no paramétrica es adecuada en métodos no paramétricos que no ofrecen formulas analíticas de estimación.
- En muchos casos donde los errores estándar son asintóticos y posiblemente muy amplios, los errores estándar Bootstrap, que son errores de muestras finitas, ofrecen una mejor aproximación.

En su versión más básica, el bootstrap para modelos no paramétricos de regresión lo definimos a partir del siguiente modelo:

$$y_t = g\left(x_t\right) + u_t \tag{6.4}$$

El Bootstrap tiene los siguientes pasos:

(I) Estime la Ecuación 6.4 y obtenga los valores ajustados $\widehat{g}\left(x_t\right)$ y los residuales ajustados del modelo e_t.

(II) Sea T la longitud de los residuales ajustados del modelo. Tome muestras aleatorias de tamaño T con reemplazo de estos residuales y denotelas como $e_{m,t}$, para $m = 1, 2, ..., M$ y $t = 1, 2, ..., T$. En este caso M es el número de replicas bootstrap.

(III) Para cada una de las replicas forme los nuevos pseudo valores $y_{m,t}$ como

$$y_{m,t} = \hat{g}(x_t) + e_{m,t}$$

y obtenga la estimación de la ecuación (6.4) con los nuevos pseudo valores $y_{m,t}$ y los mismos valores x_t. Realice las $l = 1, 2, ..., L$ predicciones adelante de T.Para cada punto l de pronóstico guarde las M predicciones.

(IV) Como es costumbre en el bootstrap y basado en Chernick y LaBudde (2014), se mantiene la predicción con los datos originales del modelo y se calcula el error estándar de cada una de las l predicciones a partir del error estándar de las M correspondientes estimaciones bootstrap. El intervalo de confianza para el punto de pronóstico l es:

$$\left[\widehat{y}_{M,T+l} - 1.96\,\text{s.e}\left(\widehat{y}_{M,T+l}\right), \widehat{y}_{M,T+l} + 1.96\,\text{s.e}\left(\widehat{y}_{M,T+l}\right)\right].$$

En este caso M denota las estimaciones a partir de las M replicas bootstrap.

6.9 El *Wild Bootstrap*

El bootstrap basado en muestras aleatorias de los errores estimados del modelo no es válido si los términos de error no son homoscedásticos. Para el caso de modelos de regresión no paramétricos con errores heterocedásticos el Bootstrap aplicado es conocido como "Wild Bootstrap" y fue propuesto por Wu (1986) para modelos de regresión. El Wild Bootstrap parte del siguiente modelo

$$y_t = g(x_t) + h(u_t)\, v_t \tag{6.5}$$

donde $h(u_t)$ es una función de los residuales u_t y v_t es una variable aleatoria de media cero y varianza uno. La selección más utilizada para v_t es

$$v_t = \begin{cases} \frac{-\sqrt{5}+1}{2}, & \text{con probabilidad } \frac{\sqrt{5}+1}{2\sqrt{5}} \\ \frac{\sqrt{5}+1}{2}, & \text{con probabilidad } \frac{\sqrt{5}-1}{2\sqrt{5}} \end{cases}$$

Los pasos en este modelo son ahora:

(I) Estime la Ecuación 6.4 y obtenga los valores ajustados $\hat{g}(x_t)$ y los residuales ajustados del modelo e_t.

(II) Sea T la longitud de los residuales ajustados del modelo. Tome muestras aleatorias de tamaño T con reemplazo de estos residuales y denotelas como e_{mt}, para $m = 1, 2, ..., M$ y $t = 1, 2, ..., T$. En este caso M es el número de replicas bootstrap.

(III) Genere de la distribución de v_t muestras aleatorias de tamaño T y denotelas como v_{mt}.

(IV) Para cada una de las M replicas calcule los nuevos pseudo valores y_{vmt} como

$$y_{v,m,t} = \hat{g}(x_t) + h(e_{m,t})\ v_{m,t}$$

y obtenga la estimación de la Ecuación 6.4 con los nuevos pseudo valores $y_{v,m,t}$ y los mismos valores x_t. En cada uno de los M modelos estimados haga las $l = 1, 2, ..., L$ predicciones y guárdelas.

(v) Dado que para cada una de las l predicciones se tiene M estimaciones bootstrap promedie las predicciones y obtenga sus errores estándar. El Intervalo de confianza esta dado de nuevo por

$$\left[\widehat{y}_{M,T+l} - 1.96 \text{ s.e}\left(\widehat{y}_{v,M,T+l}\right), \widehat{y}_{M,T+l} + 1.96 \text{ s.e}\left(\widehat{y}_{v,M,T+l}\right)\right].$$

En este caso v denota la intervención de la variable v_t en el wild bootstrap.

6.10 Variables explicativas

Dado que el interés de este trabajo se centra en el pronóstico anual de la demanda de los diferentes combustibles líquidos, a saber, ACPM/Diesel, Fuel Oil, GLP, GM y Jet Fuel, fue necesario construir una base de datos en la cual además de reportar la demanda de los diferentes combustibles para el periodo 2010-1 a 2021-9, se agrega también una serie de variables explicativas que poseen información histórica y proyectada, desde el periodo 2010-1 hasta el periodo 2035-12, lo cual permite el planteamiento de los diferentes escenarios que se presentan en el Capítulo 7.

El total de variables contenidas en la base de datos se muestra en el Cuadro 6.1, en donde se presenta el nombre clave que se usó para cada variable dentro de la base de datos, la fuente a partir de la cual se obtuvo la información sobre la variable, y la descripción de la misma.

Variable	**Fuente**	**Descripción**
fecha	Creación Propia	Hace referencia a la fecha en la cual es registrada cada observación
d_glp	Observaciones 2010/01 - 2021/07 documento GLP_dem_021121 (UPME)	Demanda Gas Licuado del Petroleo en $GBTUD$
d_acpm	Observaciones 2010/01 - 2021/09 documento Consumos combustibles líquidos 2010 - 2021 JUL mensuales (UPME)	Demanda ACPM, Diesel Marino y Electrocombustible en gal/mes
d_gm	Observaciones 2010/01 - 2021/09 documento Consumos combustibles líquidos 2010 - 2021 JUL mensuales (UPME)	Demanda Gasolina Motor Corriente, Extra y Avigas en gal/mes

d_jetfuel	Observaciones 2010/01 - 2021/09 documento Consumos combustibles líquidos 2010 - 2021 JUL mensuales (UPME)	Demanda Jet Fuel y Queroseno en gal/mes
d_fueloil	Observaciones 2010/01 - 2021/09 documento Consumos combustibles líquidos 2010 - 2021 JUL mensuales (UPME)	Demanda Fuel Oil y Combustoleo en gal/mes
febrero	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de febrero
marzo	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de marzo
abril	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de abril
mayo	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de mayo
junio	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de junio
julio	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de julio
agosto	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de agosto
septiembre	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de septiembre
octubre	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de octubre
noviembre	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de noviembre

diciembre	Creación Propia	Dummy que busca capturar el efecto determinístico de las series en los meses de diciembre
pib	Observaciones 2010/01 - 2021/08 y proyecciones 2021/09 - 2035/12 documento Datos GN C-075-2021 (UPME)	PIB Real Precios Constantes Año base 2015. Observaciones 2010/01 - 2021/08 y proyecciones 2021/09 - 2035/12.
prop_lab	Creación Propia	Proporción de día laborales en el mes
poblacion	Desagregación serie anual 2010 a 2035 proyecciones población con base en los resultados del Censo Nacional de Población y Vivienda - CNPV- 2018 (DANE)	Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
covid	Creación Propia	Dummy que busca capturar el efecto negativo en la demanda de los combustibles causado por la pandemia del COVID-19 entre el 2020/04 - 2020/11
covidjet	Creación Propia	Dummy que busca capturar el efecto negativo en la demanda de Jet Fuel causado por la pandemia del COVID-19 entre el 2020/04 - 2021/03
d_gnresi	Observaciones 2010/01 - 2021/08 documento Datos GN C-075-2021 y proyecciones 2021/09 - 2035/12 documento Anexo_Proyeccion_Demanda_EE_GN_2021-2035 (UPME)	Demanda Gas Natural Residencial en $GBTUD$
d_gntran	Observaciones 2010/01 - 2021/08 documento Datos GN C-075-2021 y proyecciones 2021/09 - 2035/12 documento Anexo_Proyeccion_Demanda_EE_GN_2021-2035 (UPME)	Demanda Gas Natural Transporte en $GBTUD$

d_gnindu	Observaciones 2010/01 - 2021/08 documento Datos GN C-075-2021 y proyecciones 2021/09 - 2035/12 documento Anexo_Proyeccion_Demanda_EE_GN_2021-2035 (UPME)	Demanda Gas Natural Industrial y otros en $GBTUD$
d_gnterc	Observaciones 2010/01 - 2021/08 documento Datos GN C-075-2021 y proyecciones 2021/09 - 2035/12 documento Anexo_Proyeccion_Demanda_EE_GN_2021-2035 (UPME)	Demanda Gas Natural Terciario en $GBTUD$
d_gntot	Observaciones 2010/01 - 2021/08 documento Datos GN C-075-2021 y proyecciones 2021/09 - 2035/12 documento Anexo_Proyeccion_Demanda_EE_GN_2021-2035 (UPME)	Demanda Gas Natural Total en $GBTUD$
d_energ	Observaciones 2010/01 - 2021/08 documento Demanda Energía SIN (XM) y proyecciones 2021/09 - 2035/12 documento Anexo_Proyeccion_Demanda_EE_GN_2021-2035 (UPME)	Demanda Comercial de Energía Eléctrica en $GWh - mes$
temp	Observaciones y proyecciones 2010/01 - 2035/12 documento Temperaturas IDEAM (IDEAM)	Temperatura promedio en grados centígrados
p_glp	Observaciones 2010/01 - 2021/09 documento precios históricos, y proyecciones de tendencia 2021/10 - 2035/12 documento precios térmicas (UPME)	Precio promedio del GLP en $\$/m3$
p_carbon	Observaciones 2010/01 - 2021/11 documento precios historicos, y proyecciones de tendencia 2021/12 - 2035/12 documento precios térmicas (UPME)	Precio promedio del Carbón en $\$/kg$

p_gm	Observaciones 2010/01 - 2021/02 documento precios historicos, y proyecciones de tendencia 2021/03 - 2035/12 documento precios térmicas (UPME)	Precio promedio de gasolima motor en $\$/gal$
p_acpm	Observaciones 2010/01 - 2021/02 documento precios historicos, y proyecciones de tendencia 2021/03 - 2035/12 documento precios térmicas (UPME)	Precio promedio de acpm en $\$/gal$
p_jetfuel	Observaciones 2010/01 - 2021/02 documento precios historicos, y proyecciones de tendencia 2021/03 - 2035/12 documento precios térmicas (UPME)	Precio promedio de jet fuel en $\$/gal$
p_fueloil	Observaciones 2010/01 - 2021/02 documento precios historicos, y proyecciones de tendencia 2021/03 - 2035/12 documento precios térmicas (UPME)	Precio promedio fuel oil en $USD/MBTU$
p_crudo	Observaciones 2010/01 - 2021/02 documento precios historicos, y proyecciones de tendencia 2021/03 - 2035/12 documento precios térmicas (UPME)	Precio del crudo en USD/BI
tot_aut_gm	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos de gasolina motor proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
tot_aut_diesel	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos de Diesel proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12

tot_aut_elect	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos electricos proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
tot_aut_gnv	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos de GNV proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
tot_aut_glp	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos de GLP proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
tot_aut_gnl	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos de GNL proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
tot_mot_gm	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de motos de gasolina motor proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
tot_mot_diesel	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de motos de diesel proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
tot_mot_elect	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de motos electricas motor proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12

tot_mot_gnv	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de motos de GNV proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
aut_liv_diesel	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos livianos de pasajeros (automóviles, camperos y taxis) de Diesel proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
aut_carga_diesel	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos livianos de carga (camionetas) de Diesel proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
aut_pesados_diesel	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos pesados (Microbus, bus, camión y tractocamión) de Diesel proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
aut_liv_gm	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos livianos de pasajeros (automóviles, camperos y taxis) de Gasolina Motor proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
aut_carga_gm	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos livianos de carga (camionetas) de Gasolina Motor proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12

aut_pesados_gm	Desagregación serie anual 2010 - 2035 documento Flota proyectada PEN - LEAP Actualización 2010-2050 (UPME)	Stock anual de vehículos pesados (Microbus, bus, camión y tractocamión) de Gasolina Motor proyectada entre 2010 - 2050. Se realiza la desagregación de serie anual a mensual para el periodo 2010/01 a 2035/12
rel_aut_elect/gm	Creación Propia	Cociente entre el stock mensual de vehículos electricos y vehículos de gasolina motor
rel_mot_elect/gm	Creación Propia	Cociente entre el stock mensual de motos electricas y motos de gasolina motor
niñoniña	Observaciones 2010/01 - 2021/09 national weather service climate prediction center	Índice Oceánico Niño (ONI) en ºC
niño	Observaciones 2010/01 - 2021/09 national weather service climate prediction center	Índice Oceánico Niño (ONI) mayores a 0.5ºC
dum_niño	Creación Propia	Dummy que busca capturar los periodos en donde se registran ONI mayores a 0.5ºC

Cuadro 6.1: Variables explicativas empleadas en los modelos de demanda de combustibles líquidos.

Es de anotar que en el Cuadro 6.1 se plantea un total de 18 variables que fueron creadas y un total 17 variables que fueron desagregadas temporalmente a una periodicidad mensual, con el fin de poder emplearlas dentro de los modelos de pronósticos propuestos. Entre las variables creadas se parte de la variable `fecha`, la cual se crea con la estructura $mmm - yy$, con el fin de poder identificar el mes y el año asociado a cada uno de los registros de las variables de demanda de combustibles líquidos u otras variables explicativas que posee la base de datos.

De las variables creadas restante, se destacan las 11 variables dummy asociadas al efecto calendario, a saber, `febrero`, `marzo`, `abril`, `mayo`, `junio`, `julio`, `agosto`, `septiembre`, `octubre`, `noviembre` y `diciembre`, que buscan capturar el efecto determinístico que tiene el i-ésimo mes sobre la demanda del combustible de interés, en donde dichas dummy son creadas bajo la

estructura

$$\texttt{mes}_i = \begin{cases} 1 & \text{Si el mes asociado a la fecha en que se registró la observación es igual al i-ésimo mes de interés.} \\ 0 & \text{Si el mes asociado a la fecha en que se registró la observación es diferente al i-ésimo mes de interés.} \end{cases}$$

En el caso de la variable de `prop_lab`, se realizó el conteo del número de días totales que tiene cada mes, junto al número de días laborales que tiene cada uno de los meses, omitiendo los días sábados, domingos y festivos, para posteriormente crear la variable de proporción de días laborales mediante el cociente

$$\texttt{prop\_lab} = \frac{\text{número de días laborales que posee el mes}}{\text{número total de días que posee el mes}}$$

Las variables dummy `covid` y `covidjet`, fueron creadas de tal forma que trataran de capturar el efecto que tuvo la pandemia del COVID-19 sobre la demanda de los combustibles. Por ello se crean dos variables dummy cuyo valor será 1 durante las fechas en las cuales se presentaron cierre económico que afectaron la demanda del demanda energético de interés, y se le da el valor de 0 en otro caso. La asignación que se le da a cada variable de `covid` se presenta a continuación

$$\texttt{covid} = \begin{cases} 1 & \text{si la fecha corresponde al periodo 2020/04 - 2020/11} \\ 0 & \text{en otro caso} \end{cases}$$

mientras que, la asignación que se le da a la variable `covidjet` es de la forma

$$\texttt{covidjet} = \begin{cases} 1 & \text{si la fecha corresponde al periodo 2020/04 - 2021/03} \\ 0 & \text{en otro caso} \end{cases}$$

Es de anotar que la diferencia entre la variable `covid` y la variable `covidjet` radica en el intervalo de tiempo para el cual se considera que hubo un cierre en la economía que afectó la demanda del energético de interés. La diferencia anterior entre la especificación de las dos variables se debe al impacto más prolongado que sufrió la demanda del Jet Fuel respecto a los demás combustibles líquidos, debido al cierre prolongado que tuvieron los aeropuertos a nivel nacional e internacional a causa de la pandemia del COVID-19, con el fin de minimizar la propagación del virus, generando que la recuperación en la demanda del Jet Fuel tardara más tiempo en retornar a los niveles pre-pandemia, y por lo cual se decide establecer para este combustible una variable "`covid`", diferente a la usada para los otros combustibles.

En el caso de la variable `rel_aut_elect/gm`, se crean a partir de las variables `tot_aut_gm` y `tot_aut_elect`, mediante el cociente

$$\texttt{rel\_aut\_elect/gm} = \frac{\texttt{tot\_aut\_elect}}{\texttt{tot\_aut\_gm}}$$

con el objetivo de capturar el efecto del cambio tecnológico de los autos de gasolina por autos eléctricos. Proceso similar se realiza para la variable `rel_mot_elect/gm`, la cual se crea a partir

del cociente entre las variables `tot_mot_elect` y `tot_mot_gm`, tal que

$$\texttt{rel\_mot\_elect/gm} = \frac{\texttt{tot\_mot\_elect}}{\texttt{tot\_mot\_gm}}$$

para capturar el efecto del cambio tecnológico que tendrán las motocicletas que emplean gasolina hacia motores eléctricos.

Finalmente se crea la variable dummy `dum_niño`, la cual tiene por objetivo capturar aquellos periodos en los cuales se registra en el Índice Oceánico Niño (ONI), temperaturas que sean superiores a 0.5ºC. Por ello, se parte de la variable `niñoniña`, que contiene los registros ONI, y se crea la siguiente variable dummy

$$\texttt{dum\_niño} = \begin{cases} 1 & \text{si valor registrado en variable } \texttt{niñoniña} > 0.5 \\ 0 & \text{en otro caso} \end{cases}$$

En el caso de las variables desagregadas temporalmetne a una frecuencia mensual, se tiene inicialmente la variable `poblacion`, la cual es obtenida de las observaciones y proyecciones anuales realizadas por el DANE entre el periodo 1993 - 2050. Para realizar la desagregación temporal para la `poblacion`, se parte del hecho de que esta variable no es un variable que presente comportamientos estacionales, si no que es una variable que posee solo un comportamiento tendencial, por lo cual luego de limitar el periodo entre 2010 a 2035 dado que es nuestro periodo objetivo, se decide aplicar el método de desagregación temporal propuesto por Chow y Lin (1971), el cual plantea la estimación por mínimos cuadrados generalizados, de un modelo lineal que relaciona la serie temporal de interés a desagregar Y, con una variable indicadora X, cuya estructura en este caso particular es lineal, con el fin de recoger consideraciones a priori sobre la evolución de la tendencia de la serie, y con ella llevar a cabo el proceso de desagregación temporal para la población `poblacion`.

Por su parte, para la desagregación de las 16 variables asociadas al parque motor que se presentan en la base de datos, a saber, `tot_aut_gm`, `tot_aut_diesel`, `tot_aut_elect`, `tot_aut_gnv`, `tot_aut_glp`, `tot_aut_gnl`, `tot_mot_gm`, `tot_mot_diesel`, `tot_mot_elect`, `tot_mot_gnv`, `aut_liv_diesel`, `aut_carga_diesel`, `aut_pesados_diesel`, `aut_liv_gm`, `aut_carga_gm` y `aut_pesados_gm`, se aplica también la metodología de desagregación temporal de variables de Chow y Lin (1971), pero con la diferencia de que en este caso no se emplea una variable indicadora lineal para su desagregación, si no que se construye una variable secundaria basada en las ventas de vehículos.

Para realizar el proceso de desagregación temporal de este grupo de variables, se siguieron los siguientes pasos:

1. Se tomaron las ventas mensuales de vehículos en Colombia desde el 2010 hasta el 2021, junto al `pib` proyectado por la UPME, las variables dummy de efecto mensual calendario descritas en el Cuadro 6.1, la variable desagregada de la `poblacion` reportada por el DANE y la variable `covid` para reflejar el efecto de la pandemia del COVID-19 en la venta de vehículos. Con estas variables se realiza la estimación de un modelo de regresión lineal, y se proyectaron las ventas de vehículos al 2035.

2. Se toman las flotas proyectadas reportadas por la UPME en el PEN, y al dato de cada año de las flotas se le suman las ventas mensuales acumuladas reales y proyectadas correspondientes al año. Esta última suma efectivamente no era el dato real de las flotas, pero dado que lo que se agregaba mes a mes, eran automóviles nuevos, es un valor muy aproximado al real, logrando así una aproximación de la flota de vehículos para cada combustible. La variable resultante en este procedimiento es la variable que se usará como índice para el proceso de desagregación.
3. Una vez calculada la variable índice que se usará para la desagregación de cada una de las flotas de vehículos presentas en el Cuadro 6.1, se emplea la metodología de desagregación de Chow y Lin (1971) basada en índices junto a las flotas proyectadas reportadas por la UPME en el PEN, para obtener de esta manera las flotas mensuales entre el periodo 2010/01 - 2035/12, las cuales se usaran para realizar las proyecciones de la demanda de los combustibles.

Es de anotar que, a pesar de haber empleado esta metodología para realizar la desagregación de las flotas de vehículos a una periodicidad mensual, sería preferible obtener esta información mensual directamente desde una fuente oficial tal como lo es el Ministerio de Transporte, u otra fuente fidedigna que permitiera asegurar que tanto los valores obtenidos como el comportamiento observado de las series a nivel mensual en la desagregación si son consecuentes con el comportamiento real que se esperaría tuviera este tipo de variables.

6.11 Metodologías Clásicas de Predicción y sus Restricciones en la Modelación de Combustibles Líquidos

6.11.1 Modelos de Vectores Autorregresivos y Modelos de Corrección de Error

Decimos que las variables $y_{1,t}$ y $y_{2,t}$ son endógenas dentro de un proceso generador de datos, si el aumento de $y_{1,t}$ genera aumentos en $y_{2,t}$ y de la misma manera los aumentos en $y_{2,t}$ generan aumentos en $y_{1,t}$. Si el proceso cuenta con tres o mas variables endógenas lo anterior también es cierto para todas las variables. Sea

$$y_t = \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ . \\ . \\ . \\ y_{k,t} \end{bmatrix} \tag{6.6}$$

un vector de variables endógenas. Entre los modelos mas utilizados para analizar el comportamiento temporal de este vector tenemos el modelo VAR(p) estructural dado por

$$Ay_t = \Gamma_0{}' + \Gamma_1{}' y_{t-1} + \Gamma_2{}' y_{t-2} + \cdots + \Gamma_p{}' y_{t-p} + B' x_t + u_t \tag{6.7}$$

donde x_t es el vector de variables exógenas, $u_t \sim NM\left(0_{k,k}, \Sigma\right)$ y cuya forma reducida puede ser presentada en una de las siguientes dos formas:

(I) El modelo VAR(p) dado por

$$y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + B x_t + e_t \tag{6.8}$$

$A_i = A^{-1}\Gamma_i{}'$, $e_i = A^{-1}u_i$ y $B = A^{-1}B'$

La Ecuación 6.8 corresponde al caso en que las variables están correlacionadas pero no siguen una tendencia común, es decir, no existe cointegración entre las variables.

(II) El modelo VECM(p-1) dado por

$$\Delta y_t = A_0 + \Gamma_1 \Delta y_{t-1} + \Gamma_2 \Delta y_{t-2} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + B x_t + e_t \tag{6.9}$$

La Ecuación 6.9 corresponde al caso cuando las variables del vector siguen una tendencia común, es decir, existe cointegración entre las variables.

Para nuestro trabajo, aunque se puede sospechar que la demanda de combustibles (como el ACPM/Diésel), se ve afectada por el comportamiento de la economía medido a través del PIB, los movimientos del PIB no necesariamente se ven directamente afectados por el consumo del diésel. Aunque el consumo de este combustible puede afectar el comportamiento de la economía, el consumo de diésel, es en realidad un pequeño componente del PIB. De hecho, otras variables como la inversión, el sistema financiero y en general el comportamiento de otros sectores de la economía, pueden determinar el comportamiento del PIB de forma más efectiva. Por esta razón, es inadecuado plantear una relación endógena entre el PIB y el consumo del ACPM/Diésel en Colombia.

Por tanto, el uso de modelos VAR y VEC implica la existencia de una relación endógena. De manera que si la endogeneidad no está presente, estos modelos no satisfacen sus principios teóricos y su uso no es justificable. Luego, lo más conveniente es recurrir a modelos de una sola ecuación que, al igual que los modelos VAR y VEC, capturen la estructura de autocorrelación y la dependencia con variables exógenas como el PIB.

Por otro lado, la existencia simultánea de raíces unitarias a la frecuencia cero y estacionales, implica que los modelos VARX(p) o VECM(p-1) no son adecuados. A la fecha, estos modelos no están diseñados para capturar los efectos estacionales de forma directa. En particular, para datos trimestrales, Lee (1992) propone el siguiente modelo en el caso de la existencia de cointegración

$$\begin{aligned}\Delta_4 y_t = \Gamma_1 \Delta_4 z_{t-1} + \cdots + \Gamma_{k-4} \Delta_4 z_{t-k+4} + \\ \pi_1 z_{1,t-1} + \pi_2 z_{2,t-1} + \pi_3 z_{3,t-2} + \pi_4 z_{3,t-1} + \\ \Upsilon_1 D_{1t} + \Upsilon_2 D_{2t} + \Upsilon_3 D_{3t} + \Upsilon_4 D_{4t} + e_t\end{aligned}$$

con $\pi = \alpha_i \beta_i{}'$.

En este caso, la variable de frecuencia cero es $z_{1,t} = \left(1 + L + L^2 + L^3\right) y_t$, la variable de

frecuencia bianual es $z_{2,t} = (1 - L + L^2 - L^3)\, y_t$ y la variable de frecuencia anual es $z_{2,t} = -(1 - L^2)\, y_t$.

Note que en el caso de datos estacionales trimestrales podemos tener cuatro frecuencias de cointegración. Por lo tanto, se tienen cuatro distintos vectores cointegrantes. Como lo señala Martin et al. (2013), el caso de datos mensuales es aún más complejo, pues se tendrían más vectores cointegrantes en las distintas frecuencias estacionales que se presentan para este tipo de datos. Esto crea la dificultad de interpretar las relaciones cointegrantes, que según Martin et al. (2013), representa una de las mayores críticas al enfoque de cointegración estacional multivariante. Otro de los inconvenientes que se tiene con este enfoque es que aún no hay implementaciones, ni teóricas ni computacionales, que permitan detectar y estimar el número de relaciones cointegrantes con variables estacionales. Claramente el PIB en Colombia y el consumo de diésel tienen patrones estacionales que deben ser tenidos en cuenta. De hecho, en un caso hipotético donde se pudiera plantear una relación endógena entre el PIB y el consumo de ACPM/Diésel, todavía no se tienen las herramientas teóricas ni computacionales suficientes para realizar la modelación correcta, aparte de la complejidad para interpretar las distintas relaciones cointegrantes a analizar.

Otra alternativa en el caso de variables endógenas sería usar la metodología de modelos SVARMA, los cuales son una extensión de los modelos SARIMA a un vector de variables endógenas. Sin embargo, las implementaciones que hay disponibles a la fecha no admiten el uso de variables exógenas. La implementación de los modelos SVARMA se encuentra solo para variables endógenas en la librería MTS del paquete estadístico `R`. Para el caso multivariado, aún no se tienen paquetes estadísticos ni un desarrollo teórico adecuado para el modelo SVARMA con variables exógenas.

6.11.2 Modelos Estacionales Autorregresivos de Media Móvil con Variables Externas el Modelo SARIMAX

Siguiendo a Brockwell et al. (2016), el modelo $ARIMA(p,d,q)(P,D,Q)_s$ está dado por

$$\phi(B)\,\Phi\left(B^s\right)(1-B)^d\,(1-B^s)^D\,y_t = \theta(B)\,\Theta\left(B^s\right)a_t$$

donde

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots \theta_q B^q$$

$$\Phi(B) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots \Phi_Q B^{QS}$$

$$\Theta(B) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \cdots \Theta_Q B^{QS}$$

Si agregamos variables explicativas en el vector x_t en el modelo, tenemos lo que conocemos como modelos de regresión con errores autocorrelacionados o también funciones de transferencia. De nuevo basados en Brockwell et al. (2016), el modelo ahora es

$$y_t = \beta^\top x_t + w_t$$

con

$$\phi(B)\,\Phi\left(B^s\right)(1-B)^d\,(1-B^s)^D\,w_t = \theta(B)\,\Theta\left(B^s\right)a_t.$$

En forma matricial el modelo de regresión en forma matricial es

$$Y = XB + W$$

Siguiendo a Brockwell et al. (2016) la función de verosimilitud para este modelo es

$$L\left(B;\phi,\theta,\Phi,\Theta,\sigma^2\right) = (2\pi)^{-n/2}\left|\Gamma_n\right|^{-1/2} exp\left\{\frac{1}{2}(Y-XB)^\top\Gamma_n^{-1}(Y-XB)\right\}$$

sin embargo, cuando el modelo $SARIMA(p,d,q)(P,D,Q)_s$ contiene componentes estacionales, así sean de bajo orden, y además en el modelo se tienen variables explicativas, se convierte en un $SARIMAX(p,d,q)(P,D,Q)_s$. Además, el determinante de Γ_n, denotado por $\left|\Gamma_n\right|$, puede tender a cero fácilmente en el proceso de estimación. Esto dificulta que la función de verosimilitud alcance un máximo y no se puedan obtener las estimaciones de los parámetros del modelo. De hecho, algunos programas toman bastante tiempo iterando el algoritmo de búsqueda y finalmente reportan la incapacidad de alcanzar el óptimo; otros simplemente reportan un error.

A continuación se presenta un ejemplo donde se estima un modelo $SARIMAX(p,d,q)(P,D,Q)_s$ con y sin covariables. El objetivo es demostrar que, cuando en estos modelos se incluyen variables explicativas, la convergencia de la verosimilitud se convierte en una tarea compleja para los algoritmos actuales.

6.11.3 Modelo SARIMAX para la demanda de gasolina

Para realizar este ejercicio, se decide crear un modelo SARIMAX para la demanda de Gasolina Motor, empleando para ellos las variable explicativas `pib`, `poblacion`, `prop_lab` y `covid`, tal que:

$$x_t = \begin{bmatrix} \texttt{pib}_t \\ \texttt{poblacion}_t \\ \texttt{prop\_lab}_t \\ \texttt{covid}_t \end{bmatrix}$$

La Figura 6.1 y la Figura 6.2 muestran los errores reportados por los paquetes `R` y `Python` al tratar de maximizar una verosimilitud del modelo ARIMAX con únicamente cuatro variables explicativas. En el caso de R, se reporta de una vez un error; y en el caso de Python se reporta que la matriz de covarianzas de los errores es singular, como era de esperarse de

acuerdo a lo reportado arriba con respecto a la maximización de la verosimilitud en el modelo $SARIMAX(p,d,q)(P,D,Q)_s$.

```
> modelo=Arima(gasolina,order=c(p=0,d=1,q=1),
+ seasonal=list(order=c(P=0,D=1,Q=1),period=12),
+ xreg=xreg,method='ML')
Error in optim(init[mask], armafn, method = optim.method, hessian = TRUE,   :
  non-finite value supplied by optim
```

Figura 6.1: Salidas del Paquete R para un Modelo ARIMAX

```
Dep. Variable:                           gasolina   No. Observations:
139
Model:             SARIMAX(0, 1, 1)x(0, 1, 1, 12)   Log Likelihood
-1822.690
Date:                            Sat, 04 Dec 2021   AIC
3659.380
Time:                                    11:12:06   BIC
3679.234
Sample:                                         0   HQIC
3667.446
                                              - 139
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
PIB              46.5061     12.100      3.844      0.000      22.791      70.221
diaslaborables  9.714e+05   5950.173    163.253      0.000     9.6e+05    9.83e+05
poblacion       -1.14e+06   3176.353   -358.848      0.000   -1.15e+06   -1.13e+06
Covid          -1.066e+06   1.83e+05     -5.816      0.000   -1.42e+06   -7.07e+05
ma.L1             -0.0109      0.030     -0.361      0.718      -0.070       0.048
ma.S.L12          -0.0062      0.017     -0.363      0.716      -0.039       0.027
sigma2          1.921e+11      1.713   1.12e+11      0.000    1.92e+11    1.92e+11
===================================================================================
Ljung-Box (L1) (Q):                   6.38   Jarque-Bera (JB):                 443.52
Prob(Q):                              0.01   Prob(JB):                           0.00
Heteroskedasticity (H):               6.31   Skew:                               0.55
Prob(H) (two-sided):                  0.00   Kurtosis:                          12.13
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 4.43e+26.
Standard errors may be unstable.
```

Figura 6.2: Salidas del Paquete Python para un Modelo ARIMAX

6.12 Medidas de bondad de ajuste y comparación de modelos

Un aspecto de suma importancia en el proceso de modelación, es la evaluación de la precisión o eficiencia del modelo estimado respecto a un conjunto de datos, y por ello, es necesario

emplear un conjunto de diferentes indicadores que permiten identificar qué tan acertado es el nivel de ajuste que ofrece un modelo particular a un conjunto de observaciones reales.

Las medidas de bondad de ajuste de modelos, son un grupo de estadísticos que buscan describir el nivel de ajuste que ofrece un modelo estimado respecto a un conjunto de observaciones, realizando para ello, una comparación entre los valores observados y los valores esperados que se generan dentro del modelo de estudio.

Aunque existe una gran cantidad de medidas de bondad de ajuste, el objetivo principal de todas es proporcionar información sobre qué modelo es el más adecuado para explicar el comportamiento de una variable dependiente, cuando se emplean diferente número de variables explicativas. Así mismo, algunas de estas medidas se extienden para brindar información sobre la comparación de modelos, en donde luego de algunas modificación, permiten comparar cuál modelo estadístico tiene mejor ajuste, cuando se emplean las mismas variables explicativas.

Partiendo de las medidas de error y los criterios de información presentados en Greene (2000), en esta sección se decide presentar algunas de las medidas de bondad de ajustes existentes que pueden ser empleadas para medir el nivel de ajuste que tienen los modelos a un conjunto de datos.

6.12.1 Error Medio (ME)

El Error Medio (ME) de estimación de un modelo estadístico, es una medida de la desviación promedio que tienen los valores estimados respecto a los valores reales, y sirve para observar la dirección que posee el error de estimación, y en consecuencia, el sesgo que poseen las estimaciones. Si definimos el error de estimación como

$$e_i = y_i - \hat{y}_i$$

donde y_i es la i-ésima observación real, y $\hat{y}_i$ es la i-ésima estimación obtenida por el modelo ajustado. Por tanto, basado en el error de estimación, se tendrá que la ecuación para el cálculo del ME estará dada por

$$ME = \frac{1}{n}\sum_{i=1}^{n} e_i$$

donde e_i es el error de estimación de la i-ésima observación y n el número total de observaciones evaluadas.

Es de anotar que en el ME, los resultados obtenidos dependen de la escala de medición, y por tanto sus valores también son afectados por transformaciones de los datos. Además, dado que en esta medida los efectos de los errores positivos y negativos que tenga el e_i se cancelan, no hay una regla clara para saber a partir de cuál valor se tendrá un buen ajuste de parte del modelo, por tanto, solo se tendrá que un valor desea para el ME será aquel donde dicha medida se encuentre lo más cercana a cero como sea posible.

6.12.2 Error Porcentual Medio (MPE)

El Error Porcentual Medio (MPE) es una medida porcentual del error promedio ocurrido en una estimación, en el que, se muestra la dirección del error y, en consecuencia, el porcentaje de sesgo que poseen las estimaciones. La ecuación para el cálculo del MPE estará dada por

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{y_i} \times 100$$

donde y_i es la i-ésima observación real, e_i es el error de estimación de la i-ésima observación y n el número total de observaciones evaluadas.

Similar al ME, los efectos porcentuales de los errores positivos y negativos se cancelan, y por tanto la regla de decisión que debe tomarse para la interpretación de esta medida de error, será elegir aquellos modelos que ofrezcan un MPE lo más cercano a cero como sea posible.

6.12.3 Error Absoluto Medio (MAE)

El Error Absoluto Medio (MAE) o también denominado como Desviación Absoluta Media (MAD), mide la desviación absoluta promedio de los valores estimados respecto a los valores reales, y por tanto, muestra la magnitud del error general, ocurrido dentro del proceso de estimación. Dado que, en a diferencia del ME, en el MAE, los efectos de los errores positivos y negativos no se anulan, entonces no es posible observar el sesgo que tendrán las estimaciones realizadas. La ecuación para el cálculo del MAE está dada por

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

donde e_i es el error de estimación de la i-ésima observación y n el número total de observaciones evaluadas.

Como lo señala Hastie y Tibshirani (2017) esta medida disminuye con el número de variables del modelo en la muestra de entrenamiento del modelo, más no en la muestra de validación. Por lo tanto, no es una buena medida para seleccionar entre modelos donde la variable dependiente es la misma, pero el número de variables explicativas es distinto, ya que siempre ganará el modelo con más variables o términos. Sin embargo, para evaluar el desempeño predictivo de un modelo, es una buena medida, ya que el poder predictivo no depende del número de variables o términos en un modelo. De esta forma, entre más pequeño sea el MAE mejor poder predictivo tiene el modelo.

6.12.4 Error Absoluto Porcentual Medio (MAPE)

El Error Absoluto Porcentual Medio (MAPE) es una medida porcentual del error absoluto promedio que ocurre en una estimación o pronóstico, y por tanto, permite evidenciar el porcentaje de error general que ocurre durante el proceso de estimación. La ecuación del $MAPE$

está dada por

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \times 100$$

donde y_i es la i-ésima observación real, e_i es el error de estimación de la i-ésima observación y n el número total de observaciones evaluadas.

Igual que el MAE, el $MAPE$ no depende del número de variables o términos en el modelo, por lo que no es muy útil para hacer selección de modelos. Sin embargo, si es una medida muy útil para medir el desempeño predictivo del modelo ya que entre más cerca estén los valores predichos fuera de muestra por el modelo de los valores reales, mejor poder predictivo tiene el modelo, en cuyo caso, seleccionamos como el modelo con mayor desempeño predictivo el de menor $MAPE$.

6.12.5 Suma de Cuadrados del Error (SSE)

La Suma de Cuadrados del Error (SSE) es una medida que permite observar la desviación al cuadrado de los valores estimados o pronosticados por un modelo respecto a los valores reales. Adicionalmente, a diferencia de otras medidas de error como el ME, MAE, el SSE penaliza errores extremos ocurridos durante el proceso de estimación, puesto que, al ser una medida cuadrática, los errores grandes de estimación obtienen un peso mucho mayor que los errores pequeños. La ecuación del SSE está dada por

$$SSE = \sum_{i=1}^{n}(e_i)^2$$

donde e_i es el error de estimación de la i-ésima observación y n el número total de observaciones evaluadas.

Es de anotar que la SSE, es una medida altamente sensible a los cambios en la escala de medición y transformaciones que sufran los datos originales, y su interpretación no es tan intuitiva como lo es con el $MAPE$. A pesar de lo anterior, esta medida también servirá para el mismo propósito de las otras medidas de error, en donde al comparar modelos se seleccionará aquel modelo que minimice el SSE.

6.12.6 Error Cuadrático Medio (MSE)

El Error Cuadrático Medio (MSE) es una medida similar al SSE, con sus mismas propiedades, entre las que destaca su sensibilidad a los cambios de escala y transformaciones, y el enfoque que hace sobre los grandes valores que puede tener esta medida a causa de la penalización que le realiza a aquellos cuando hay grandes errores, los cuales generan un mayor impacto sobre la medida que cuando los errores de estimación son pequeños. La ecuación del MSE está

dada por

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2$$

donde e_i es el error de estimación de la i-ésima observación y n el número total de observaciones evaluadas.

Como en el caso de la SSE, se tendrá que el MSE de un modelo, será mejor entre más pequeño sea su valor, por tanto, se buscará aquel modelo que minimice el MSE, en donde entre más cercano a cero se encuentre esta medida indicará que es mejor el modelo planteado.

6.12.7 Raíz del Error Cuadrático Medio (RMSE)

Como su nombre lo indica, la Raíz del Error Cuadrático Medio (RMSE), es simplemente la raíz cuadrada del MSE, y por tanto, todas las propiedades y análisis que posee el MSE también son aplicables para el $RMSE$, con la diferencia de que en el caso del $RMSE$, la unidad de medida bajo la cual son calculados los datos retorna a su medida original, y por tanto su interpretación es un poco más intuitiva que la del MSE. La ecuación del $RMSE$ está dada por

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i)^2}$$

donde e_i es el error de estimación de la i-ésima observación y n el número total de observaciones evaluadas.

Dado que las propiedades del MSE también son aplicables para el $RMSE$, se tendrá que entre menor sea el $RMSE$ de un modelo, mejor será el ajuste de las estimaciones obtenidas por el modelo al conjunto de datos.

6.12.8 Criterio de Información de Akaike (AIC)

El Criterio de Información de Akaike (AIC), es una de los criterio de información o medidas de bondad de ajuste más utilizadas en la práctica, debido a que permite medir la calidad relativa del ajuste que tiene un modelo estadístico a un conjunto de datos. La ecuación del $AIC(p)$ está dada por

$$AIC(p) = n \times \ln\left(\frac{SSE}{n}\right) + 2p$$

donde SSE es la suma de cuadrado del error, n en número total de observaciones y p en número de variables explicativas usadas en el modelo.

Es de anotar que, la SSE de un modelo tiende a disminuir a medida que aumenta el número de variables explicativas que ingresan dentro del modelo, tal como se presenta en Greene (2000), y debido a esto, con el fin de penalizar la reducción que toma la SSE por el número de variables

explicativas que se ingresan en el modelo, el AIC agrega un término de penalización dado por $2p$, con el fin de compensar la reducción que tiene la SSE con el aumento que se tendrá en el número de variable p.

Por esta razón, si una variable realmente aporta en la explicación de la variabilidad de la variable dependiente y_t, entonces al ingresarla, disminuirá el $AIC(p)$. De esta manera, en un modelo con distinto número de variables explicativas y la misma variable dependiente, lo recomendable es seleccionar el modelo de menor $AIC(p)$. Caso similar ocurre en la comparación de distintos modelos, en donde al comparar diferentes metodologías con las mismas variables explicativas, se seleccionará aquel modelo que brinde un menor $AIC(p)$.

6.12.9 Criterio de Información de Akaike Corregido (AICc)

Una alternativa al AIC que se suele emplear cuando se tiene que el cociente entre el número de observaciones n y el número de parámetros estimados p es menor a 40 (Burnham y Anderson, 2002), es el conocido Criterio de Información de Akaike corregido $AICc$, que tiene por objetivo hacer una corrección al AIC cuando se poseen muestras finitas, reemplazando el término $2p$ de penalización que posee el AIC por $\frac{n(n+p)}{n-p-2}$. El $AICc(p)$ está definido como

$$AICc(p) = n \times \ln\left(\frac{SSE}{n}\right) + \frac{n(n+p)}{n-p-2} \quad \text{para } n \neq p+2$$

donde SSE es la suma de cuadrado del error, n en número total de observaciones y p en número de variables explicativas usadas en el modelo.

Es de anotar que la interpretación que se realiza para el $AICc$ será la misma que para el AIC, en donde el mejor ajuste para un modelo será aquel que ofrezca valores más pequeños, para este estadístico.

6.12.10 Criterio de Información Bayesiano (BIC)

Otra medida que se suele emplear con frecuencia para la selección de modelos, es el Criterio de Información Bayesiano (BIC), el cual posee una ecuación e interpretación similares a los presenta en la Subsección 6.12.8 para el criterio AIC debido a que su ecuación y estructura son similares, con la diferencia de que en lugar de penalizar el número de variables explicativas ingresadas en el modelo con $2p$, éste realiza una penalización de la forma $p \times ln(n)$, en donde además de incluir en la estimación del estadístico el número de parámetros, también tiene en cuenta el número total de observaciones. La ecuación para el $BIC(p)$ está dada por

$$BIC(p) = n \ln\left(\frac{SSE}{n}\right) + p \times \ln(n)$$

donde SSE es la suma de cuadrado del error, n en número total de observaciones y p en número de variables explicativas usadas en el modelo.

La interpretación del BIC es similar a la que se realiza para el AIC y por tanto, se tiene

que entre más pequeño sea el criterio de información, mejor será la bondad de ajuste del modelo evaluado.

6.12.11 Criterio de Información Hannan-Quinn (HQC)

Otro criterio de información popular para la selección de modelos y como alternativa al AIC y al BIC es el Críterio de Información de Hannan-Quinn (HQC), el cual usa como penalización para el número de variables explicativas adicionales que se ingresan al modelo, una combinación de la penalización realizada por el AIC y el BIC, empleando para ello término $2p \times \ln(\ln(n))$, en donde a diferencia del BIC, este criterio de información reduce el peso que se le da al número de observaciones n, aplicándole un doble logaritmo natural. La ecuación para el $HQC(p)$ está dada por

$$HQC(p) = n \ln\left(\frac{SSE}{n}\right) + 2p \times \ln(\ln(n))$$

donde SSE es la suma de cuadrado del error, n en número total de observaciones y p en número de variables explicativas usadas en el modelo.

En este caso, al igual que los criterio de información evaluados en la Subsección 6.12.8, Subsección 6.12.9 y Subsección 6.12.10, el $HQC(p)$ se empleará como un criterio para seleccionar entre modelo, en donde, el modelo estimado que arroje un menor $HQC(p)$ será el modelo que ofrezca el mejor ajuste al conjunto de datos.

6.12.12 Error de Predicción Final (FPE)

A pesar de no contener en su nombre "criterio de información", como los expuestos en la Subsección 6.12.8, Subsección 6.12.9, Subsección 6.12.10 y Subsección 6.12.11, el Error de Predicción Final (FPE) es un estadístico altamente utilizado en la comparación de la bondad de ajuste de los modelos, debido a que también penaliza el número de parámetros que se utilizan en el modelo, pero a diferencia de los otros criterios, éste criterio no lo hace a través de la suma de un valor adicional, si no que crea un cociente de la forma $\frac{n+p}{n-p}$ que se multiplica por la suma de cuadrados del error. La ecuación para el cálculo del $FPE(p)$ está dado por

$$FPE(p) = \frac{SSE}{n} \times \frac{n+p}{n-p} \quad \text{para } n \neq p$$

donde SSE es la suma de cuadrado del error, n en número total de observaciones y p en número de variables explicativas usadas en el modelo.

La interpretación del $FPE(p)$ es la misma de los demás criterios de información expuestos en la Sección 6.12, en donde la especificación del modelo estimado será mejor entre menor sea el valor obtenido en su $FPE(p)$.

Capítulo 7 Resultados combustibles líquidos

En este capítulo se presentan algunos de los escenarios que se tuvieron en cuenta al momento de realizar las proyecciones de la demanda de cada uno de los combustibles líquidos, a saber, ACPM/Diésel, Fuel Oil, GLP, GM y Jet Fuel, en donde se presenta para cada caso, el escenario que logró los resultados más significativos en términos de validación del modelo, siendo la variabilidad resultante de las proyecciones, las variables explicativas usadas en cada escenario y la trayectoria de las proyecciones resultantes las determinantes para seleccionar cuál de todos los escenarios calculados en cada caso fue el que ofrece mejor ajuste, en donde se buscaba que la variabilidad de las proyecciones fuese similar a la que poseía la serie original, que las variables explicativas fuesen consecuentes con la lógica económica y que la trayectoria de las proyecciones resultantes fuese similar a la planteada en el PEN.

Para realizar el proceso de estimación de los diferentes escenarios se decide partir el conjunto de observaciones en dos tramos. El primer tramo compuesto por la información correspondiente al periodo 2010/01 - 2020/09 (aproximadamente 90 % del total de las observaciones totales) se emplea en el proceso de entrenamiento de los modelos planteados en la Sección 6.2, Sección 6.3, Sección 6.4, Sección 6.5 y Sección 6.6; mientras que, el segundo tramo compuesto por la información correspondiente al periodo 2020/10 - 2021/09 (aproximadamente 10 % del total de las observaciones totales) se emplea en el proceso de validación de los modelos estimados y la metodología de combinación de pronósticos presentada en la Sección 6.7, con el objetivo de observar el desempeño predictivo de los modelos [1].

Adicionalmente, para realizar el proceso de estimación de los diferentes modelos se parte de un escenario base para todos los combustibles líquidos, en el cual se emplean variables dummy de efecto calendario (`febrero`, `marzo`, `abril`, `mayo`, `junio`, `julio`, `agosto`, `septiembre`, `octubre`, `noviembre` y `diciembre`), variables macroeconómicas (`pib`, `prob_lab`, y `poblacion`) y una variable dummy asociada a cierres económicos causados por el COVID-19 (`covid` o `covidjet`).

Es de anotar que para todos combustibles se emplea la variable `covid` descrita en el Cuadro 6.1, dentro del escenario base para representar el efecto que tuvo la pandemia del COVID-19 sobre la demanda de cada combustible, a excepción del caso del jet fuel, en donde como se expuso previamente en la Cuadro 6.1, se reemplaza la variable `covid` por la variable `covidjet`, debido a que en el sector de aviación el efecto negativo generado por la pandemia fue más prolongado que para otros sectores, y por tanto se extendió el número de periodos en los cuales la variable dummy toma el valor de 1.

Una vez identificado el modelo que ofrece el mejor desempeño predictivo en términos de

[1] Es de anotar que se decide solo emplear el último año en el proceso de validación del desempeño predictivo de los modelos, debido a que se buscaba capturar en el proceso de estimación de los modelos el efecto que tuvo el COVID-19, sobre la economía, con el fin de poder usar dichos modelo en futuras proyecciones en las cuales se puedan imponer nuevas cuarentenas o restricciones a la economía a causa de las nuevas variantes del COVID-19 que se registran actualmente en el mundo.

validación, para cada combustible se emplea la metodología Bootstrap descrita en la Sección 6.8 para construir los intervalos de predicción del 95 % de confianza. Para tal propósito, se realiza luego de identificar el mejor modelo, un total de 1000 replicas bootstrap en las que se re-estima el modelo ajustado aplicando un error de estimación a cada una de las observaciones. Una vez realizadas las 1000 replicas bootstrap, se calcula, para cada i-ésima observación proyectada, la desviación estándar correspondiente. El intervalo de confianza se calcula a través de la ecuación:

$$\left[\widehat{y}_{M,T+l} - 1.96\,\text{s.e}\left(\widehat{y}_{M,T+l}\right), \widehat{y}_{M,T+l} + 1.96\,\text{s.e}\left(\widehat{y}_{M,T+l}\right)\right].$$

previamente descrita en la Sección 6.8.

En la Sección 7.1, Sección 7.2, Sección 7.3, Sección 7.4 y Sección 7.5 se presenta, para cada combustible, líquido los resultados obtenidos para el escenario que presentó mejor modelo evaluado en cada uno de los casos de estudio, junto con el ajuste del mejor modelo dentro de muestra, en validación y las correspondientes proyecciones obtenidas por el modelo.

7.1 ACPM/Diésel

Para el pronóstico de la demanda de ACPM/Diésel, además de plantear el escenario base compuesto por las variables de efecto calendario, variables macroeconómicas y la variable de cierres económicos, descritas en la Capítulo 7, se presenta un total de 18 casos diferentes al escenario base, probando con ello un total de 264 escenarios en los cuales se probaron diferentes combinaciones entre las variables, `p_acpm`, `d_energ`, `d_gntran`, `d_gntot`, `p_crudo`, `tot_aut_diesel`, `tot_mot_diesel`, `tot_aut_elect`, `aut_carga_diesel` y `aut_pesados_diesel`; junto a un número diferente de rezagos para las variables `d_acpm`, `pib`, `d_gntran` y `d_gntot`, todas escritas en el Cuadro 6.1.

De los 264 escenarios estimados, se seleccionan 3 de ellos para exponen en esta sección, debido a que son los tres escenarios que mejor se adaptan a las expectativas que tiene la UPME, en términos de variabilidad, variables explicativas y trayectoria de proyección.

7.1.1 ACPM/Diésel: Escenario 1

El primer escenario que se presenta para la demanda de ACPM/Diésel, se realiza mediante la metodología de combinación de pronósticos expuesta en la Sección 6.7, a partir de la combinación entre el modelo MARS y el modelo LSTM, empleando para ello solo las variables presentadas en el escenario base, a saber, las variables de efecto calendario, las variables macroeconómicas y la variable del COVID-19, junto a los dos primeros rezagos de la demanda de ACPM/Diésel y el PIB, debido a que fue uno los modelos que ofrece mejores ajustes y resultados más consistentes con lo esperado en el PEN.

Ahora bien, con el objetivo de observar el nivel de ajuste dentro de muestra, el desempeño predictivo, y los pronósticos obtenido por la metodología de combinación de pronósticos, se presentan a continuación 4 figuras, a saber, en la Figura 7.1 se presenta una perspectiva

general del ajuste y las proyecciones obtenidas por el modelo de pronóstico, al mostrar todo el comportamiento del modelo desde el periodo 2010/01 hasta el periodo 2035/12.

En la Figura 7.2 se muestra el ajuste del modelo desarrollado respecto a los datos de entrenamiento que van desde el 2010/01 hasta el 2020/09. En la Figura 7.3 se exhibe el desempeño predictivo de la metodología de combinación de pronósticos respecto a los datos que se dejaron por fuera de muestra que van desde 2020/10 hasta el 2021/09. Finalmente, en la Figura 7.4 se presentan las proyecciones obtenidos por la metodología de combinación de pronósticos para el periodo 2021/10 a 2035/12, junto a sus correspondientes intervalos de confianza del 95 %, los cuales fueron calculados vía Bootstrap, tal como se explica en la Sección 6.8.

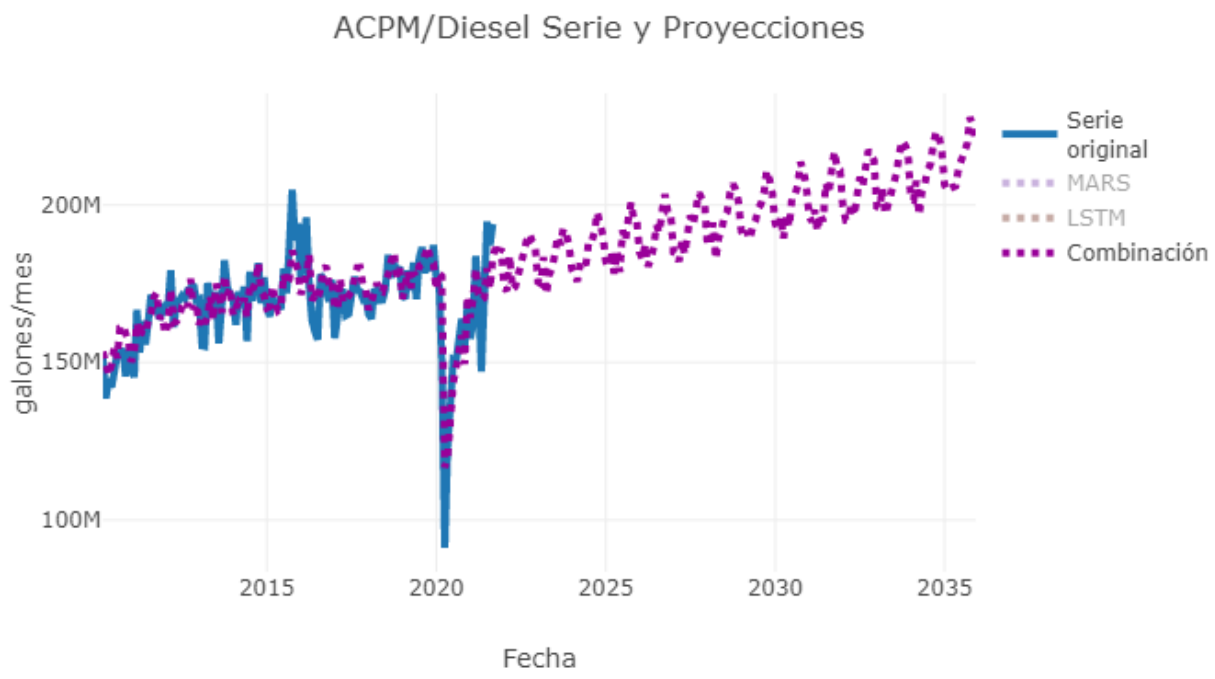

Figura 7.1: Ajuste del modelo de combinación de pronósticos a la demanda de ACPM/Diésel para el periodo 2010/01 - 2035/12.

En la Figura 7.1 se expone el comportamiento de la serie original, junto a las estimaciones y proyecciones realizadas por la metodología de combinación de pronósticos, en donde, se evidencia que el ajuste obtenido por la metodología tanto para datos de entrenamiento como para datos de validación, es relativamente bueno, debido a que sigue la tendencia que presenta la serie original, captura la caída que se registra en la demanda del ACPM/Diésel durante el año 2020 a causa de la pandemia del COVID-19, y exhibe una tendencia de proyección similar a la que traía la serie original antes de la pandemia del COVID-19.

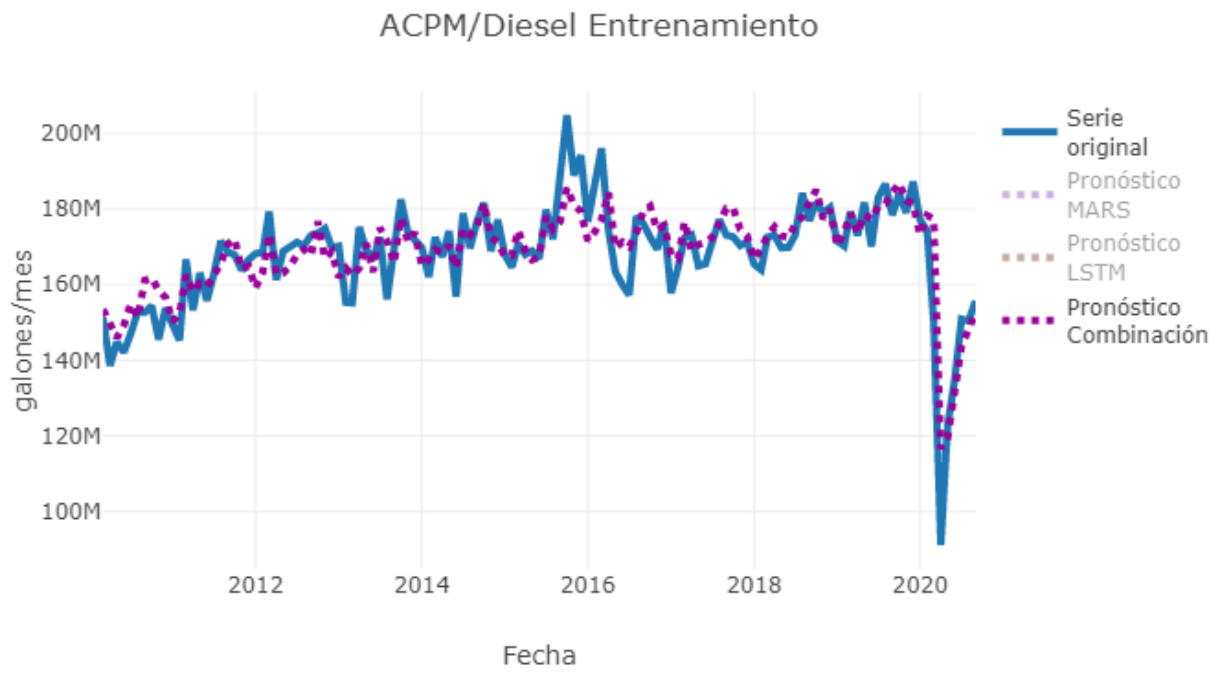

Figura 7.2: Ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda ACPM/Diésel para el periodo 2010/01 - 2020/09.

En la Figura 7.2 se observa que a pesar de que la metodología de combinación de pronósticos no logra tener un ajuste perfecto a los datos de la demanda de ACPM/Diésel, ésta logra capturar relativamente bien el crecimiento de la demanda registrado entre el 2015 y 2017, posiblemente causado por el fenómeno del niño, en donde, pese a que no alcanza los picos observados en la serie original, se logra evidenciar dos pequeños picos que imitan el comportamiento de la serie original. Adicionalmente se observa que la metodología de pronóstico lograr capturar la caída y recuperación que tuvo la demanda de ACPM/Diésel en el 2020 a causa por la pandemia del COVID-19. El buen ajuste del modelo puede corroborarse con el resultado presentado en el Cuadro 7.1, en donde se registra el MAPE, el AIC y el BIC del ajuste.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
3.13611	4035.68287	4089.72242

Cuadro 7.1: Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda de ACPM/Diésel

En el Cuadro 7.1 se presenta el MAPE, el AIC y el BIC obtenido mediante la metodología de combinación de pronósticos, en donde se evidencia que el MAPE obtenido por dicha metodología para los datos de entrenamiento fue de 3.13611 %, valor que según la escala de juicio del criterio MAPE desarrollado por Lewis (1982), para medir la precisión de las proyecciones que realiza un modelo, significa que el ajuste resultante es muy preciso, ya que, según la escala

de juicio del criterio MAPE, se tiene que valores inferiores a 10 % significa una alta precisión de las proyecciones, entre 10 % y 20 % buenas proyecciones, entre el 20 % y 50 % proyecciones razonables, mientras que valores superiores al 50 % significa proyecciones inadecuadas. Lo aquí planteado se presenta en el Cuadro 7.2.

MAPE (%)	Evaluación
MAPE $\leq$ 10 %	Pronósticos Muy Precisos
10 % $<$ MAPE $\leq$ 20 %	Pronósticos Buenos
20 % $<$ MAPE $\leq$ 50 %	Pronósticos Razonables
MAPE $>$ 50 %	Pronósticos Inadecuados

Cuadro 7.2: Escala de precisión de los pronósticos MAPE (Lewis, 1982), (Melikoglu, 2017).

Es de anotar que, pese a que la escala de juicio planteada por Lewis (1982) fue propuesta para medir la bondad de ajuste de un modelo respecto a su desempeño predictivo, ésta también puede ser usada como referencia para medir el nivel del ajuste de un modelo dentro de muestra.

Adicionalmente, en la Cuadro 7.1 se observa que el AIC y el BIC de la metodología de combinación de pronóstico dieron 4035.68287 y 4089.72242, respectivamente; pero dado que dichas medidas se plantean originalmente para la comparación de modelos, y no son conocidos valores de referencia que indiquen a partir de qué niveles el AIC o el BIC son buenos para datos de entrenamiento; el uso que se le dará a estas medidas, será en este caso, para comparar el desempeño dentro de muestra que tienen los tres escenarios presentados en la Sección 7.1, en donde, el criterio para la selección del mejor escenario será el minimice el valor de los criterios de información, tal como se expone en la Sección 6.12.

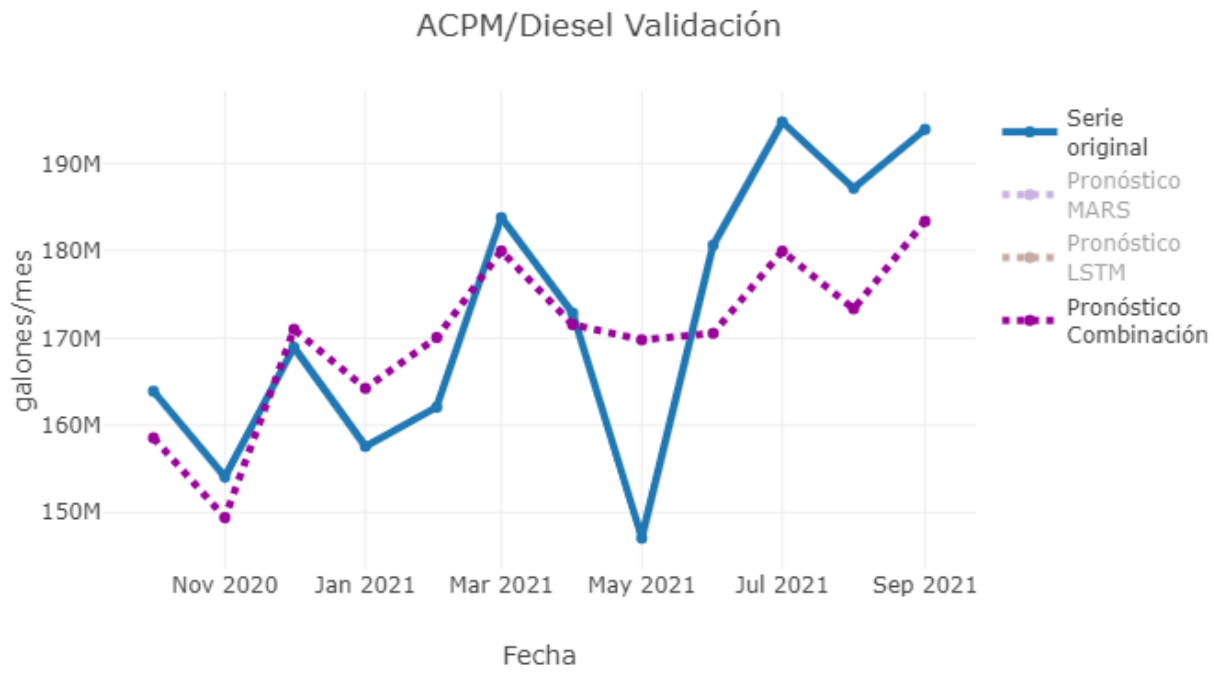

Figura 7.3: Ajuste del modelo de combinación de pronósticos para datos de validación de la demanda ACPM/Diésel para el periodo 2020/10 - 2021/09.

Por su parte, en la Figura 7.3 se evidencia que tanto la demanda de ACPM/Diésel, como el ajuste de la metodología de combinación de pronósticos muestran un comportamiento creciente a través del tiempo debido a la fase de recuperación que presenta la economía post-pandemia luego de la caída que se observa en la Figura 7.2 a partir de los primeros meses del 2020 causada de las cuarentenas estrictas que impuso el gobierno que buscaban controlar el crecimiento que se presentaba en el número de contagios del virus del COVID-19.

Adicionalmente, se observa que los datos estimados por la metodología de combinación de pronósticos presentan valores muy cercanos a las observaciones originales registradas para los últimos meses del 2020, y los primeros meses del 2021, observándose la diferencia más significativa en Mayo de 2021, en donde la demanda real del combustible cae hasta 147.057 mil galones/mes, mientras que el valor estimado por el modelo toma un valor de 166.224 mil galones/mes.

Con el fin de cuantificar el nivel de ajuste de la metodología de combinación de pronósticos respecto a las observaciones que se dejaron para medir la validación del modelo, se presenta el Cuadro 7.3, en el cual se registra el MAPE obtenido por el modelo, junto a los criterios de información AIC y BIC.

Validación		
MAPE (%)	**AIC**	**BIC**
5.08164	425.97241	435.18564

Cuadro 7.3: Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de validación de la demanda de ACPM/Diésel

En la Cuadro 7.3, se observa que el desempeño predictivo medido por el MAPE, para la metodología de combinación de pronósticos fue del 5.08164 %, en donde, tal como se expone en la Cuadro 7.2, al registrarse un MAPE inferior al 10 %, se tendrá que las proyecciones realizadas por el modelo de combinación de pronósticos son muy precisas. Además, se tendrá que el AIC y el BIC de la metodología de combinación de pronósticos fueron respectivamente 425.97241 y 435.18564, y por tanto, se tomarán dichos valores de referencia al momento de concluir cuál de los tres escenarios planteados en la Sección 7.1 es el más preciso en términos de validación de resultados.

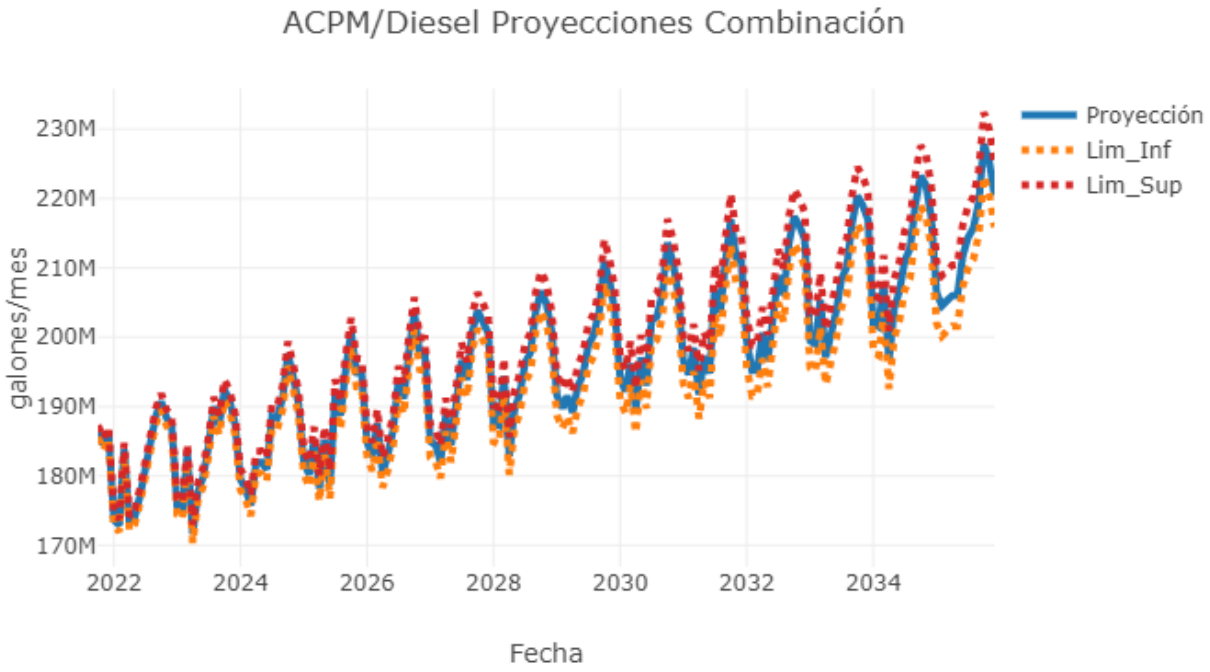

Figura 7.4: Proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésel para el periodo 2021/10 - 2035/12.

Finalmente, en la figura Figura 7.4 se presentan las proyecciones obtenidas por la metodología de combinación de pronósticos, en donde se observa que la demanda de ACPM/Diésel tendrá un comportamiento relativamente constante entre 2022 y 2024, debido a la fase de recuperación de la economía post-pandemia. Adicionalmente se evidencia que la demanda de ACPM/Diésel presenta un comportamiento estacional constante a lo largo del tiempo junto con un leve incremento en su tendencia hasta el año 2035, en donde se proyecta durante este

periodo una demanda mínima de 171.652 millones de galones/mes en Abril de 2023, y una demanda máxima de 227.826 millones de galones/mes en Octubre de 2035.

Con el fin de dar una vista rápida sobre las proyecciones obtenidas por la metodología de combinación de pronósticos

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	158334274.95922	158560321.99317	158786369.02712
2020-11-01	149124817.02283	149425860.03594	149726903.04906
2020-12-01	170549400.55932	170995525.26479	171441649.97026
2021-01-01	163677397.60381	164232794.58998	164788191.57615
2021-02-01	169448235.58528	170054559.39053	170660883.19577
...	...	...	...
2035-08-01	211217356.84818	215727569.00645	220237781.16471
2035-09-01	214848056.64914	219399377.96455	223950699.27996
2035-10-01	223226102.86893	227826306.97653	232426511.08412
2035-11-01	220720573.04156	225334471.97762	229948370.91368
2035-12-01	215849132.44328	220467355.10876	225085577.77425

Cuadro 7.4: Encabezado de validación y proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésel

7.1.2 ACPM/Diésel: Escenario 2

En el segundo escenario presentado para ACPM/Diésel, se encuentra que la metodología de combinación de pronósticos entre el modelo MARS y el modelo LSTM es el más adecuado para ajustar el comportamiento de la demanda de ACPM/Diésel, empleando para ello como variables explicativas adicionales a las variables del escenario base, el `p_acpm` y el `p_crudo`, debido a que desde una perspectiva económica, existe una relación inversa entre el precio de un bien y su demanda, y por tanto se decide incluir `p_acpm` como una de las variables que permite explicar el comportamiento de la demanda de ACPM/Diésel.

De forma similar se realiza el análisis de la inclusión del `p_crudo` dentro del modelo, en donde el ACPM/Diésel también conocido como petrodiésel es un hidrocarburo líquido derivado del petróleo, el cual se obtiene principalmente en los procesos de destilación del petróleo en altas temperaturas, y por consiguiente, precios altos del crudo pueden repercutir de forma directa en el precio del ACPM y de forma indirecta en la demanda del ACPM/Diésel.

Adicionalmente, en el escenario aquí planteado se incluyen tres variables adicionales, a saber, el primer rezago de la variable `d_acpm`, y los dos primeros rezagos de la variable `pib`. Lo anterior se hace con el fin de capturar el comportamiento autoregresivo no estacional que poseen las variables, puesto que, el comportamiento estacional estaría ya siendo capturado por

las variables de efecto calendario, la cual tiene por objetivo capturar el componente estacional determinístico que poseen las series de tiempo.

En la Figura 7.5 se presenta en términos generales el ajuste de la metodología combinación de pronósticos ajustada a partir de los modelos MARS y LSTM, donde el objetivo será observar el nivel de ajuste que tiene el modelo dentro de las observaciones de entrenamiento, el nivel de ajuste para las observaciones de validación y las proyecciones finales obtenidas por el modelo, con el fin de evaluar desde una perspectiva global el grado de ajuste, la variabilidad, y la trayectoria de pronóstico que sigue el modelo.

Por otro lado, en la Figura 7.6 se presenta de forma individual el ajuste de la metodología de combinación de pronósticos para los datos de entrenamiento, en la Figura 7.7 el ajuste del modelo para los datos de validación y finalmente, en la Figura 7.8 las proyecciones realizadas junto con sus correspondientes intervalos de confianza, con el fin dar una perspectiva más concreta del desempeño del modelo predictivo en cada una de sus fases.

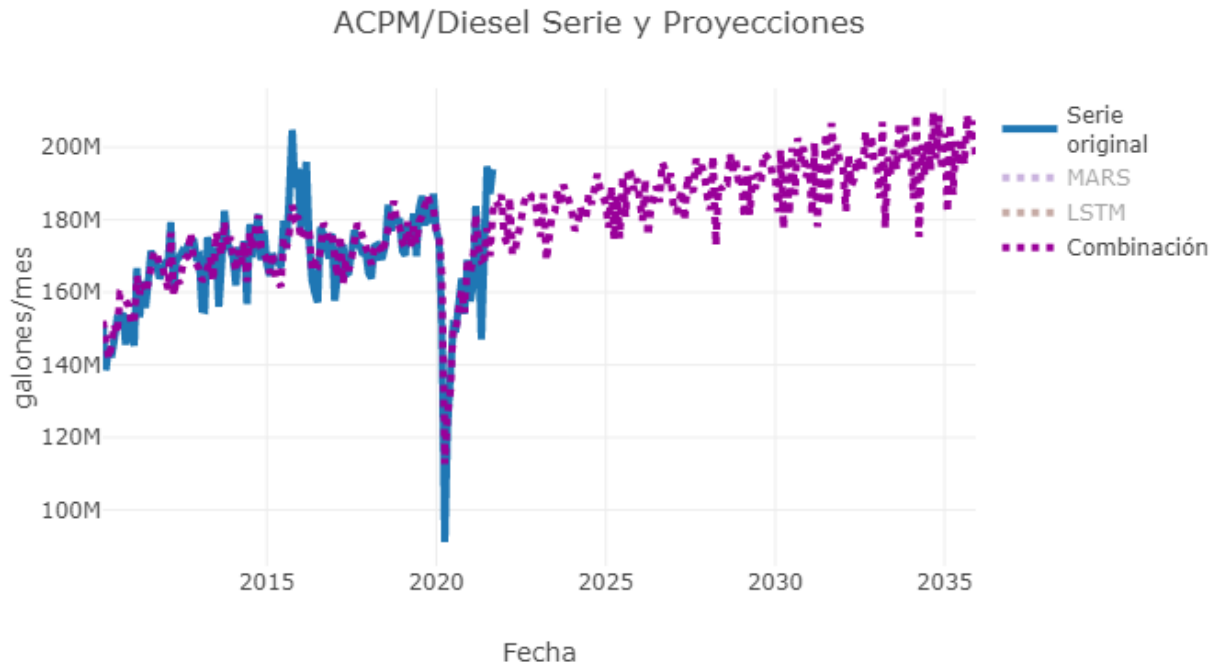

Figura 7.5: Ajuste del modelo de combinación de pronósticos a la demanda de ACPM/Diésel para el periodo 2010/01 - 2035/12.

De la Figura 7.5 se observa que el ajuste del modelo de combinación de pronósticos presenta un buen ajuste para los datos de entrenamiento, puesto que este logra capturar en su mayoría, el comportamiento de toda la serie original, incluyendo la caída que tuvo la demanda de ACPM/Diésel durante el 2020, a causa del efecto de la pandemia del COVID-19 la pandemia. En el caso de las proyecciones, se evidencia que la serie no posee una variabilidad constante, si no que ésta va aumentando un poco a medida que transcurre el tiempo, sin embargo su tendencia sigue el mismo comportamiento de la serie original.

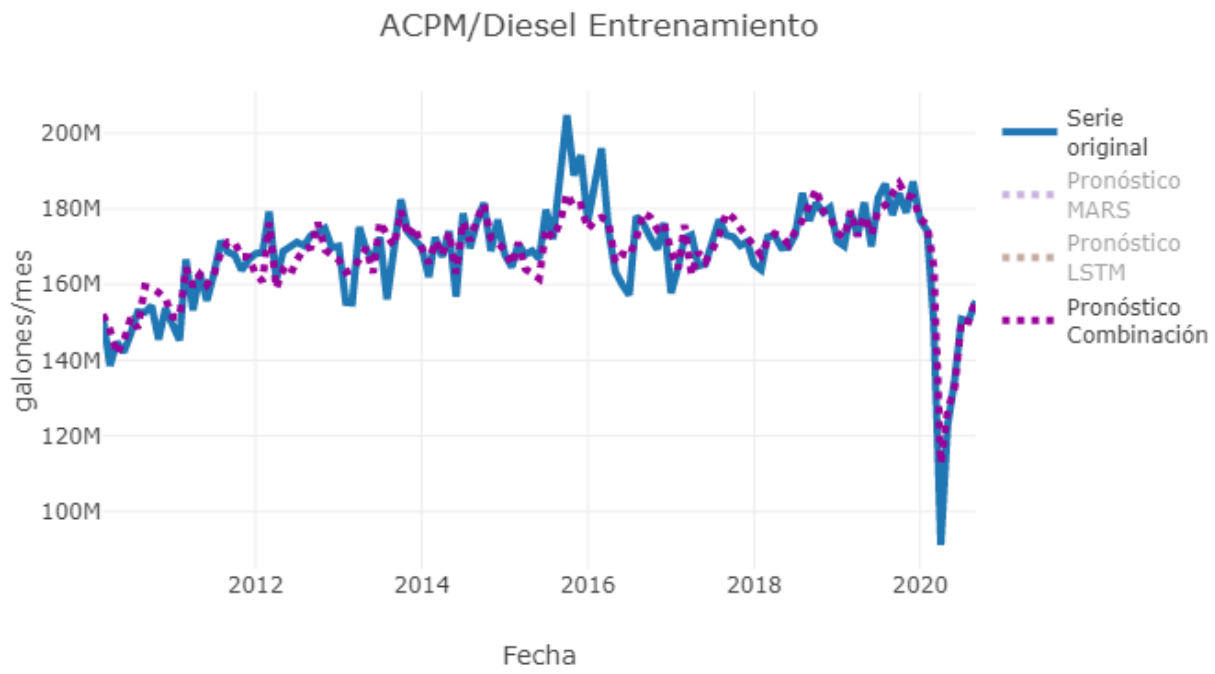

Figura 7.6: Ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda ACPM/Diésel para el periodo 2010/01 - 2020/09

Con el objetivo de apreciar de mejor forma el ajuste de la metodología de combinación de pronósticos, se presenta en la Figura 7.6 el comportamiento de la serie original junto a las estimaciones obtenidas por la metodología de combinación, en donde se evidencia con más detalle lo planteado en la Figura 7.5, puesto que se logra evidenciar como el modelo ajustado logra capturar en su mayoría los picos y valles que posee la serie original, inclusive la caída registrada por la pandemia del COVID-19, sin embargo, se observa que este modelo no logra capturar con precisión los picos de demanda registrados en Octubre-Diciembre de 2015, y Marzo de 2016, cuyos aumentos pueden ser tal vez explicados por el aumento de la demanda de combustibles a causa del fenómeno del niño que se registró por esas época.

Con el propósito de observar si la discrepancia entre el ajuste y la estimación en las fechas señaladas afecta significativamente el nivel de ajuste de la metodología de combinación de pronósticos respecto a los datos de la serie original usados para entrenamiento, se presenta en el Cuadro 7.5 tres medidas de bondad de ajuste ampliamente usadas en la literatura, a saber, el MAPE, el AIC y el BIC del ajuste.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
2.61680	3999.93303	4056.81678

Cuadro 7.5: Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda de ACPM/Diésel

En el Cuadro 7.5 se presenta el MAPE de entrenamiento obtenido por el modelo ajustado, en donde se evidencia que dicho valor es demasiado bajo, puesto que éste registra un valor de tan solo el 2.6158 %, y por tanto, si se parte de la escala de precisión del MAPE propuesta por Lewis (1982), se tendrá que MAPE por debajo del 10 % esta asociado a estimaciones muy precisas. Adicionalmente, en el Cuadro 7.5 se presenta el AIC y el BIC obtenido para el ajuste de entrenamiento, en donde, al compararlo los valores obtenidos en este escenario con respecto a los obtenido en el Cuadro 7.5 del escenario 1 presentado la Subsección 7.1.1, se tiene que la inclusión del `p_acpm` y el `p_crudo` mejoran el nivel de ajuste del modelo dentro de la muestra de entrenamiento.

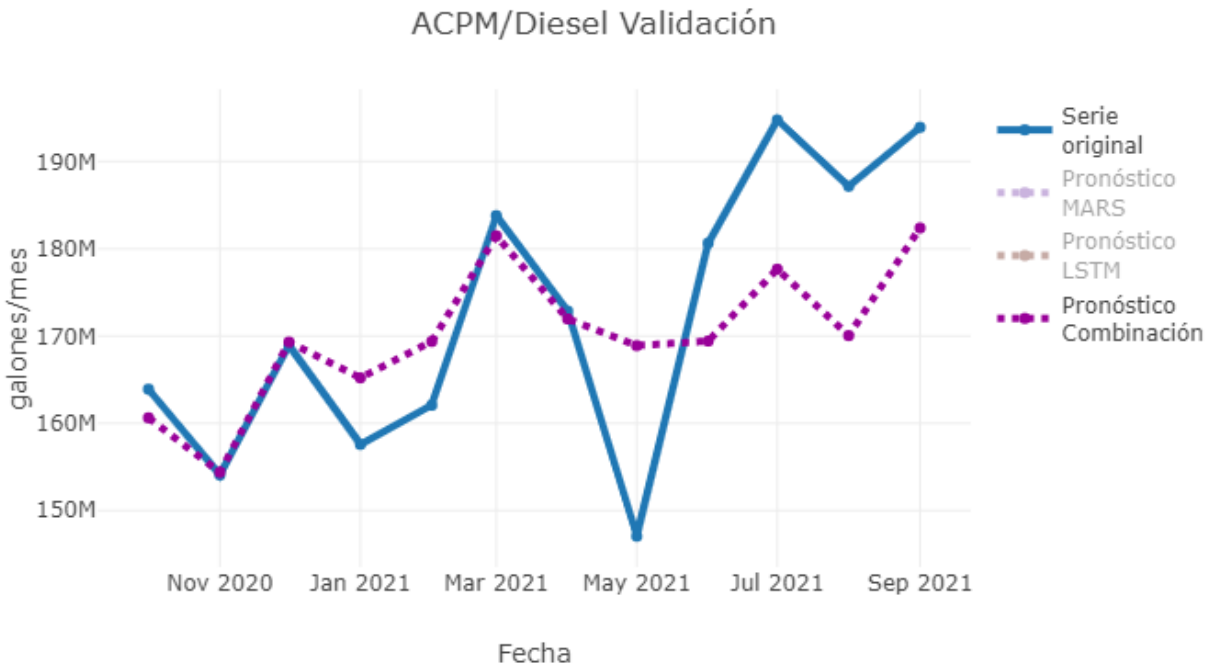

Figura 7.7: Ajuste del modelo de combinación de pronósticos para datos de validación de la demanda ACPM/Diésel para el periodo 2020/10 - 2021/09.

Por su parte, en la Figura 7.7 se presenta el ajuste que tuvo la metodología de combinación de pronósticos respecto a los datos de validación, encontrando que de las 12 observaciones que se dejaron por fuera de muestra, el modelo logra capturar de forma muy precisa el valor de 4 de ella, y 3 observaciones adicionales de forma cercana. Sin embargo, se evidencia que para los meses de Mayo, Julio, Agosto y Septiembre de 2021, el modelo no logra acertar los valores observados con precisión. Sin embargo, con el objetivo de medir de forma estadística el desempeño predictivo de la metodología de combinación de pronósticos a los datos de validación, se registra en el Cuadro 7.6 el MAPE, el AIC y el BIC obtenido por el modelo de pronóstico.

Validación		
MAPE (%)	**AIC**	**BIC**
4.87440	429.11105	438.80918

Cuadro 7.6: Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de validación de la demanda de ACPM/Diésel

Del Cuadro 7.6 se evidencia que el MAPE de validación obtenido por la metodología de combinación de pronósticos es del 4.8744 %, valor que es menor al expuesto en la Cuadro 7.3, del escenario 1 cuyo valor es del 5.0854, lo cual podría hacer pensar que la inclusión de las variables `p_acpm` y el `p_crudo`, mejoran el ajuste del modelo por fuera de muestra, sin embargo, al evaluar los criterios de información del AIC y el BIC, se evidencia que hay una desmejora en estos valores, puesto que, en este escenario se registra un AIC de 429.11105, y un BIC de 438.80918, mientras que en el escenario 1, se registra un AIC de 425.97241 y un BIC del 435.18564.

De lo anterior se tendrá que la selección del modelo que ofrece un mejor ajuste dependerá del criterio de bondad de ajuste que se desee seleccionar, en donde si se selecciona el AIC o el BIC sobre el MAPE, se tendrá que el escenario 1 será quien presente un mejor desempeño predictivo, mientras que, si se selecciona el criterio del MAPE sobre el AIC o el BIC, se tendrá que el escenario 2 será quien presente un mejor desempeño predictivo.

Ahora, dado que el valor obtenido por los tres criterios de información es tan similar entre los dos escenarios evaluados, se tendrá que la selección del modelo más adecuado, tendrá que ser seleccionado a partir de otros criterios, tales como, la variabilidad, la tendencia y el comportamiento estacional de las proyecciones.

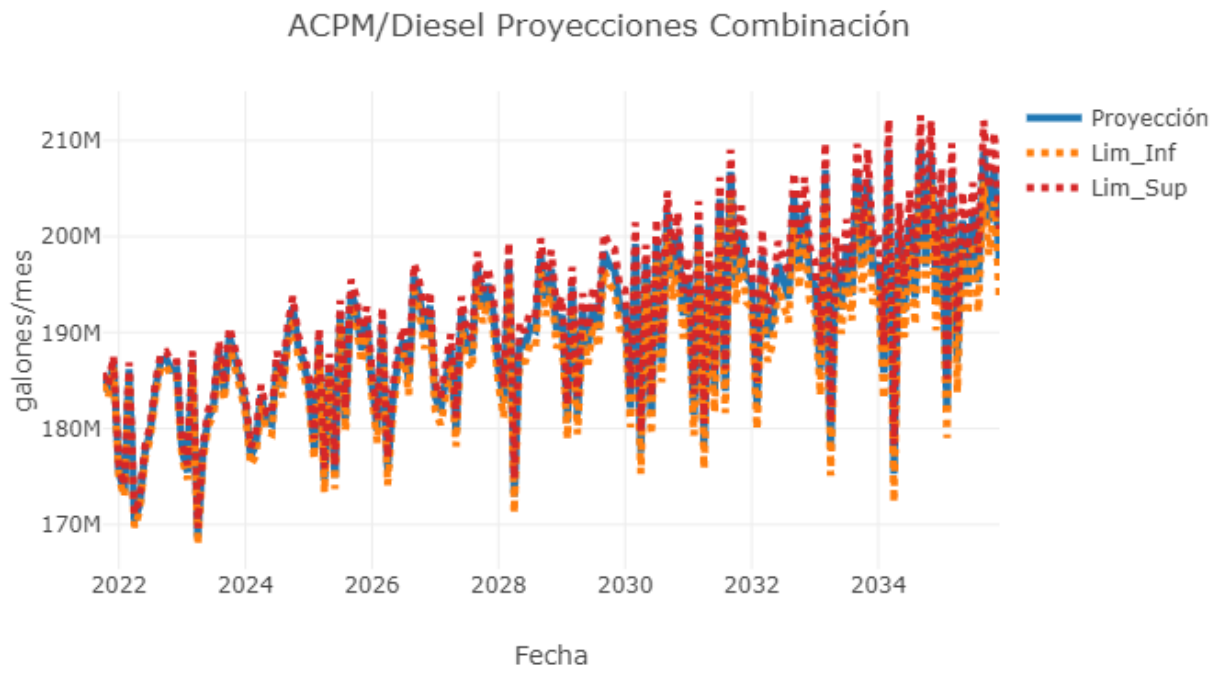

Figura 7.8: Proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésel para el periodo 2021/10 - 2035/12

En la Figura 7.8 se evidencia que la serie proyectada por la combinación de pronósticos no exhibe un claro comportamiento estacional para la variable de demanda de ACPM/Díesel que sea consecuente con el comportamiento observado en la serie original, tal como el observado en la Figura 7.4. Adicionalmente se observa que la variabilidad de las proyecciones obtenidas en este escenario no son constante, y debido a lo anterior los picos y valles exhibidos suelen variar con el tiempo.

Por tanto, a pesar de que el MAPE obtenido por la metodología de combinación de pronósticos al incluir el `p_acpm` y el `p_crudo` es menor al presentado en el escenario 1, se tendrá que las proyecciones obtenidas en el escenario 1 son más cercanas y más consecuentes con lo que se esperaría para la demanda de ACPM/Diésel, y por tanto se concluirá que los resultados obtenidos en el escenario 1 presentado en la Subsección 7.1.1 son de mejor calidad que las presentadas en el escenario 2.

Finalmente, con el fin de exponer el cambio en la variabilidad que poseen las proyecciones en este escenario y soportar el análisis que se realizó en el párrrafo anterior, se presenta en la Cuadro 7.7 el encabezado de los datos de validación y las proyecciones obtenidas por la metodología de combinación de pronósticos planteadas en este escenario, con el objetivo de evidenciar como el intervalo de confianza bootstrap construido pasa de una diferencia aproximada de 200 mil galones/mes en el periodo 2020/10, a una diferencia aproximada de 4 millones de galones/mes en el periodo 2035/12.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	160411629.52937	160626711.58345	160841793.63752
2020-11-01	153925276.60309	154375372.52533	154825468.44758
2020-12-01	168900773.57978	169305862.80458	169710952.02938
2021-01-01	164689735.02059	165207993.79834	165726252.57609
2021-02-01	168633602.17842	169378672.55508	170123742.93174
...	...	...	...
2035-08-01	192077552.26707	195199035.80188	198320519.33669
2035-09-01	205667212.33167	208868471.21399	212069730.09631
2035-10-01	198067507.55017	201489058.39089	204910609.23160
2035-11-01	203640875.28651	207267574.72362	210894274.16073
2035-12-01	193362449.45469	197071140.25636	200779831.05803

Cuadro 7.7: Encabezado de validación y proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésell

7.1.3 ACPM/Diésel: Escenario 3

En el tercer y último escenario presentado para la demanda de ACPM/Diésel se plantea un modelo LSTM, en el cual se usan como variables explicativas adicionales al escenario base, el `p_acpm`, la `d_gntran` y el `tot_aut_diesel`, junto a dos rezagos para la variable `d_acpm`.

La razón de incluir el `p_acpm` se debe a la relación ya expuesta sobre el precio y la demanda de un bien, por su parte, la inclusión de la `d_gntran` se debe a que el gas natural usado en el sector transporte se usa como un bien sustituto del ACPM/Diésel, y por tanto, su inclusión podría contribuir a explicar cuál será la demanda de dicho combustible. Finalmente, la inclusión de `tot_aut_diesel` se debe a la contribución que tiene el sector transporte en la demanda del energético, siendo los camiones, microbuses, buses, camiones y tractocamiones, los que impulsan fuertemente la demanda del ACPM/Diésel en este sector.

Similar a los escenarios anteriores, se presentan cuatro figuras, a saber, en la Figura 7.9 se expone el comportamiento de la serie original junto al ajuste y pronósticos obtenidos por el modelo LSTM, en la Figura 7.10 se muestra el ajuste que tiene el modelo LSTM a los datos usados para el entrenamiento del modelo, en la Figura 7.11 el ajuste del modelo en validación, mientras que, en la Figura 7.12 se presentan las proyecciones de la demanda de ACPM/Diésel, junto a sus correspondientes intervalos de confianza.

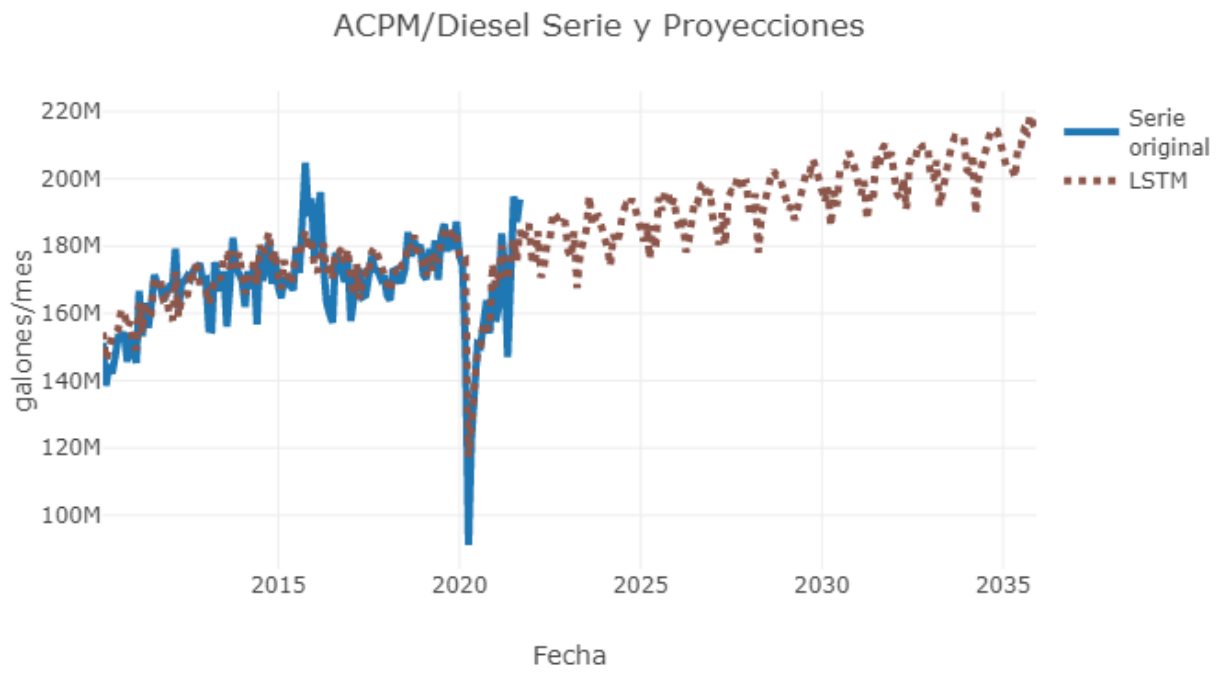

Figura 7.9: Ajuste del modelo LSTM a la demanda de ACPM/Diésel para el periodo 2010/01 - 2035/12

En la Figura 7.9, se evidencia que el ajuste del modelo LSTM a la serie original tanto en entrenamiento como en validación, parece ser relativamente bueno, puesto que sigue la trayectoria que posee la serie y logra capturar la caída en la demanda que se evidencia en el año 2020, a causa de la pandemia del COVID-19. Además se observa que la variabilidad que presentan los pronósticos son similares a los que trae la serie original entre el periodo 2010 - 2019, además, de que la trayectoria de proyección sigue la tendencia que traía la serie original antes de la época de pandemia.

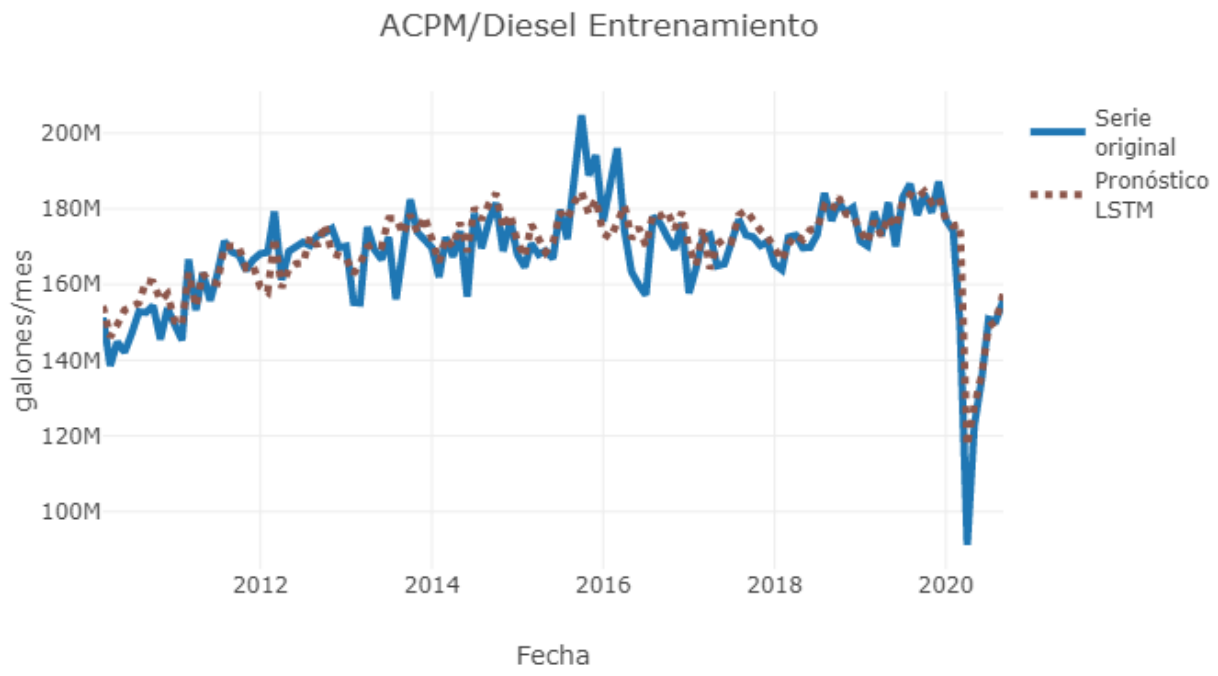

Figura 7.10: Ajuste del modelo LSTM para datos de entrenamiento de la demanda ACPM/-Diésel para el periodo 2010/01 - 2020/09

En la Figura 7.10 se observa de forma más precisa el ajuste del modelo, a los datos de entrenamiento, en donde se evidencia que a pesar de que el modelo no logra capturar del todo bien, algunos picos y valles que presenta la serie original, en general el ajuste dado por el modelo es relativamente bueno, pues captura la pendiente y la estructura que posee la serie original, además de que logra capturar la caída que se registra durante el año 2020.

Con el fin de cuantificar el nivel de ajuste otorgado por el modelo LSTM al conjunto de datos de entrenamiento, en el Cuadro 7.8 se reportan las medidas de bondad de ajuste del MAPE, AIC y BIC.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
3.00367	4038.09250	4094.97624

Cuadro 7.8: Medidas de bondad de ajuste del modelo LSTM para datos de entrenamiento de la demanda de ACPM/Diésel

Del Cuadro 7.8, se evidencia que el MAPE obtenido por el modelo LSTM dentro de entrenamiento es tan solo del 3.00367 %, por lo cual se tendrá que el ajuste ofrecido por el modelo al conjunto de datos de entrenamiento es muy buena, debido a que, como se expuso en el Cuadro 7.2, obtener un MAPE por debajo del 10 %, significa que la estimación obtenida por el modelo de interés es muy preciso. Adicionalmente, al comparar el MAPE obtenido en este escenario respecto a los presentados en los escenarios 1 y 2, se tiene que se encuentra en

un punto medio, es decir, es menor al presentado en el escenario 1 (3.13611 %), pero mayor al presentado en el escenario 2 (2.61680 %).

Por su parte, se tiene que el AIC y BIC obtenidos en el ajuste del modelo LSTM para ajustar la demanda de ACPM/Diésel, es fue respectivamente de 4038.09250 y 4094.97624, en donde se observa que ambos valores son mayores que los reportados en los escenarios 1 y 2, por lo cual podría concluirse que el ajuste dentro de entrenamiento para este escenario es un poco menor a los evidenciados en los escenarios 1 y 2.

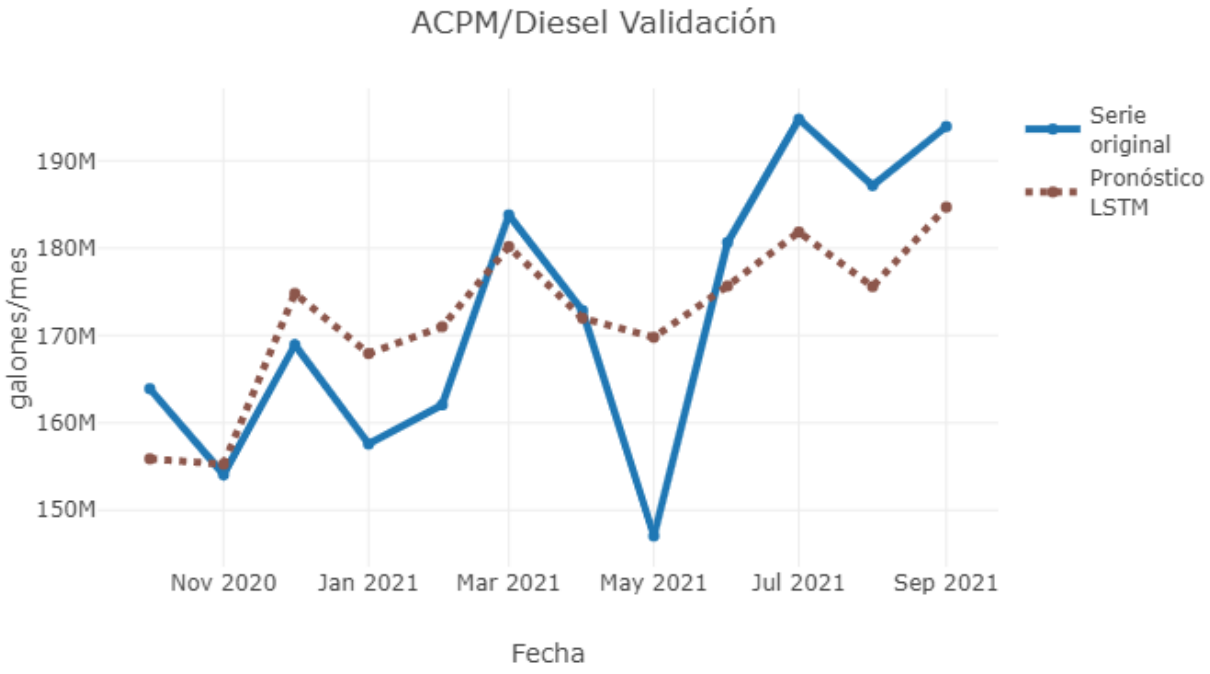

Figura 7.11: Ajuste del modelo LSTM para datos de validación de la demanda ACPM/Diésel para el periodo 2020/10 - 2021/09

En la Figura 7.11 se observa que, a pesar de que el modelo LSTM no logra capturar de forma precisa todos los valles y los picos que se evidencian en las observaciones de validación, el modelo logra capturar la tendencia que posee la demanda, además de capturar de forma muy aproximada cinco de las doce observaciones de validación. Ahora bien, con el objetivo de cuantificar el nivel de ajuste que tiene el modelo a partir de medidas de bondad de ajuste, en el Cuadro 7.9 se presenta el MAPE, el AIC y el BIC del ajuste del modelo.

Validación		
MAPE (%)	**AIC**	**BIC**
4.95919	427.12317	436.82130

Cuadro 7.9: Medidas de bondad de ajuste del modelo LSTM para datos de validación de la demanda de ACPM/Diésel

En el Cuadro 7.9 se observa que el MAPE obtenido por el modelo LSTM para el ajuste de las observaciones de la demanda de ACPM/Diésel que se dejaron por fuera de muestra es del 4.95919 %, siendo dicho valor menor al expuesto en el escenario 1 (5.08164 %), pero mayor al del escenario 2 (4.87440 %), es decir, para validación se conserva al comportamiento del MAPE evidenciado para los datos de entrenamiento.

Por su parte, el AIC y BIC registrado por el modelo LSTM arroja valores de 427.12317 y 436.82130, respectivamente, los cuales son un poco mayores a los presentados en el Cuadro 7.3 del escenario 1, pero menores a los presentados en el Cuadro 7.6 del escenario 2. Por tanto, como se expuso en el escenario 2, al presentarse valores tan similares entre uno u otro escenario, la selección del mejor ajuste se debería basar en otros criterios, tales como las variables utilizadas, la variabilidad, la tendencia, que el comportamiento estacional de las proyecciones sean consecuentes con los de la serie original.

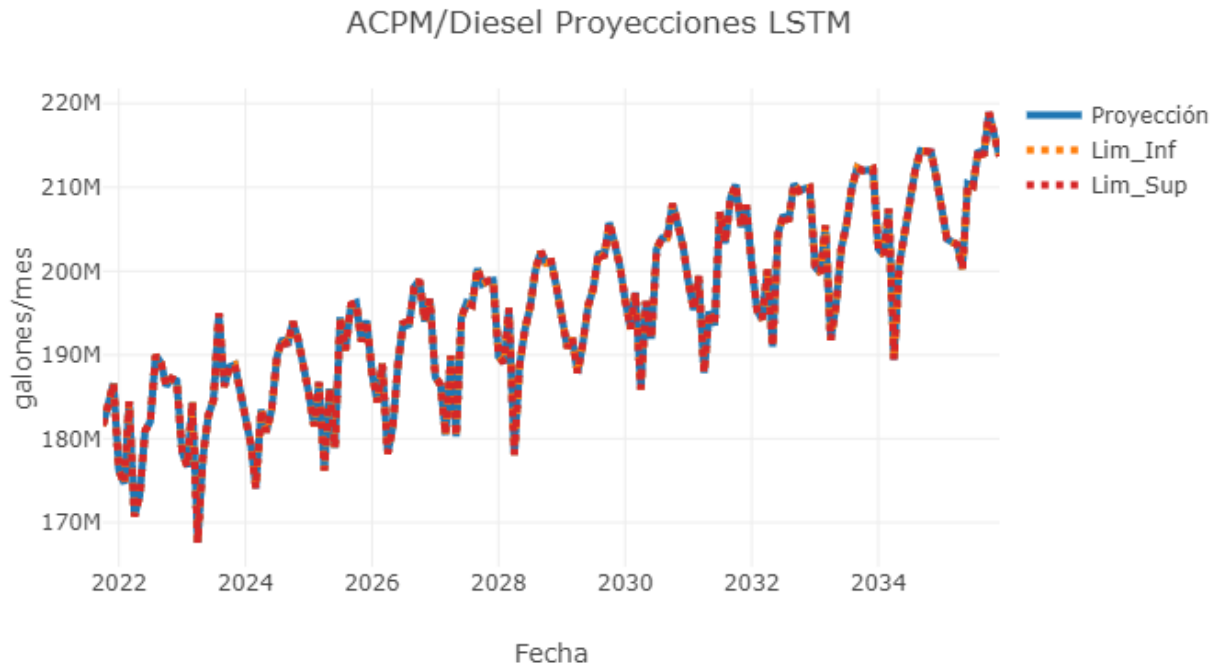

Figura 7.12: Proyecciones del modelo LSTM para la demanda de ACPM/Diésel para el periodo 2021/10 - 2035/12

Dado lo anterior, se realiza el análisis de la Figura 7.12, la cual presenta las proyecciones obtenidas por el modelo LSTM para la demanda de ACPM/Diésel junto a sus intervalos de confianza bootstrap, al realizar un total de 1000 replicas.

En las proyecciones obtenidas por el modelo LSTM se observa que la demanda de ACPM/Diésel registra un comportamiento estacional marcado con picos en los meses de Septiembre y Octubre, y valles en los meses de Febrero, lo cual es consecuente con el comportamiento de la serie original, y donde dicho comportamiento es más consistente que el presentado en la Figura 7.4 del escenario 1. Adicionalmente, se evidencia que tanto la pendiente como la variabilidad

de las proyecciones presentados en la Figura 7.12 es más estable y consistente con el que traía la serie original, que las expuestas en los otros escenarios.

Finalmente se observa que a diferencia de los otros escenarios, no existe mucha diferencia entre los intervalos de confianza bootstrap y el valor efectivamente proyectado, lo cual puede explicarse debido a la robustez que tienen las estimaciones generadas por el modelo LSTM, en donde, la introducción de pequeños errores de estimación no afecta de forma significativa las proyecciones realizadas por el modelo, lo cual hace que la variabilidad generada para cada observación en las 1000 replicas sea pequeña y en consecuencia se obtengan intervalos de confianza muy precisos.

Con el fin de ilustrar la diferencia entre las estimaciones de los intervalos de confianza y las proyecciones se presenta el Cuadro 7.10, en la cual se presentan los encabezados obtenidos por las proyecciones aquí presentadas.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	155872603.23996	155884549.44564	155896495.65132
2020-11-01	155269778.58477	155272152.78856	155274526.99235
2020-12-01	174805549.20592	174809317.03464	174813084.86335
2021-01-01	167933953.01697	167934579.36302	167935205.70906
2021-02-01	171002174.46915	171009946.29506	171017718.12097
...	...	...	...
2035-08-01	214413546.19333	214425520.66121	214437495.12909
2035-09-01	213748182.97660	213763301.07667	213778419.17674
2035-10-01	218899521.74362	218928237.25172	218956952.75983
2035-11-01	216216900.13732	216231107.34541	216245314.55349
2035-12-01	213706907.91322	213715537.17176	213724166.43031

Cuadro 7.10: Encabezado proyecciones modelo LSTM para la demanda de ACPM/Diésel

Dados los resultados y los análisis presentados para los tres escenarios aquí planteados, se tiene que a pesar de no haber diferencias significativas entre los MAPE o los criterios de información descritos, se concluye que las proyecciones más acertadas para la demanda de ACPM/Diésel están dadas por el escenario 3, en el cual se incluyen las variables explicativas del `p_acpm`, la `d_gntran` y el `tot_aut_diesel`, puesto que, tanto su variabilidad, su tendencia y su comportamiento estacional se encuentra más aproximado al comportamiento que se evidencia en la serie original, que lo presentado en los escenarios 1 y 2.

7.2 Fuel oil

Para el pronóstico de la demanda de fuel oil, además de plantear el escenario base compuesto por las variables de efecto calendario, variables macroeconómicas y la variable de cierres económicos, descritas en la Capítulo 7, se presentan 9 casos adicionales al escenario base. Por lo tanto, se alcanzaron a analizar 162 escenarios en los cuales se probaron diferentes combinaciones entre las variables, `p_fueloil`, `p_crudo`, `niñoniña`, `dum_niño`, `temp`; junto a un número diferente de rezagos para las variables `d_fueloil` y `pib`, todas descritas en el Cuadro 6.1.

De los 162 escenarios estimados, se han seleccionado dos de ellos con el soporte de la UPME para ilustrar en este reporte. Estos escenarios satisfacen las expectativas de la UPME en términos de variabilidad en resultados, las variables explicativas y la trayectoria de proyección resultante a 15 años.

7.2.1 Fuel Oil: Escenario 1

En este escenario de proyección de fuel oil se ha obtenido que el mejor modelo corresponde al de redes neuronales LSTM. En este escenario, las variables explicativas adicionales a las del escenario base son el `p_fueloil` y el `p_crudo`. Desde el contexto económico, se justifica tener una ecuación de demanda entre la demanda del bien y su precio, por lo que el precio del fuel oil ha resultado ser relevante en este escenario. Similar al caso del diésel, la variable `p_crudo` también ha mostrado ser útil para explicar el comportamiento de la demanda de fuel oil.

Las especificaciones del mejor modelo en este escenario consideran incluir los dos primeros rezagos de la variable `d_fueloil`. Esta estrategia, como en todos los casos ilustrados en este informe, se logra capturar el comportamiento autoregresivo no estacional del consumo de fuel oil. Como se ha comentado previamente, el comportamiento estacional se captura empleando las variables de efecto calendario. La red LSTM consiste de dos capas ocultas con 40 neuronas por capa. Y la mejor función de activación encontrada fue ReLu.

En la Figura 7.13 se presenta a nivel general el ajuste del modelo de LSTM. El objetivo es observar el ajuste que tiene el modelo dentro de las observaciones de entrenamiento, el ajuste para las observaciones de validación y las proyecciones obtenidas.

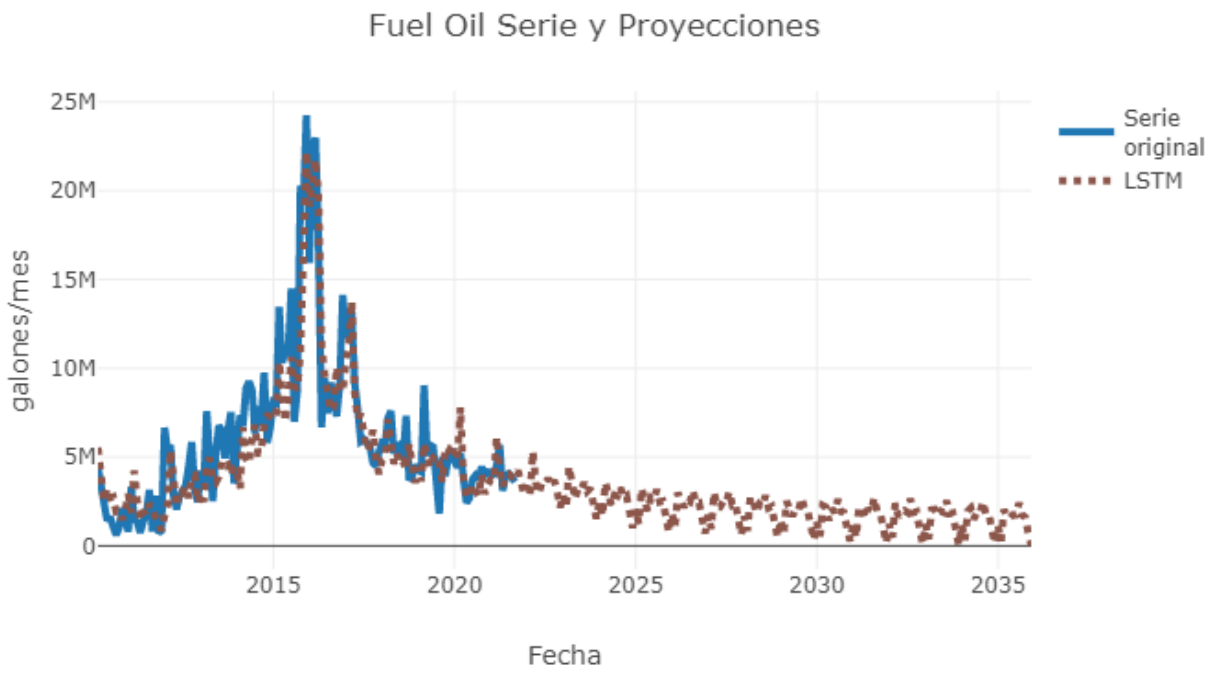

Figura 7.13: Ajuste del modelo LSTM a la demanda de Fuel Oil para el periodo 2010/01 - 2035/12

En la Figura 7.14 se ilustra la serie original y proyección obtenida a través del modelo LSTM de pronósticos para los datos de entrenamiento. En la Figura 7.15 el ajuste del modelo para los datos de validación. Y en la Figura 7.16, se ilustran las proyecciones realizadas junto con sus correspondientes intervalos de confianza.

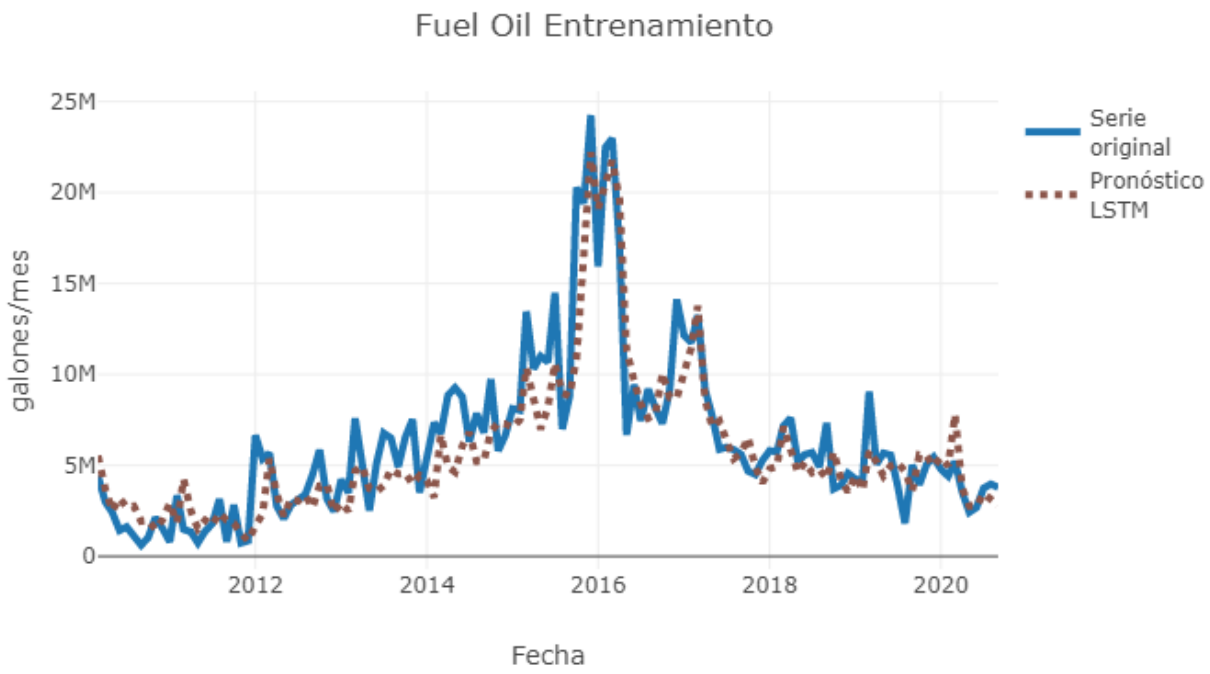

Figura 7.14: Ajuste del modelo LSTM para datos de entrenamiento de la demanda Fuel Oil para el periodo 2010/01 - 2020/09

De la Figura 7.14 se observa un ajuste satisfactorio con la red LSTM. Como se observa, entre 2015 y 2016 se ha presentado un aumento inesperado del consumo de fuel oil dada la ocurrencia del fenómeno del Niño en ese momento. Este fenómeno implicó que algunas plantas de generación aumentaran significativamente su producción con fuel oil. El modelo LSTM, a pesar de todo, captura tales cambios en el consumo de fuel oil. En el caso de las proyecciones, también se evidencia que la serie no posee una variabilidad constante. Esta tiende a aumentar a medida que transcurre el tiempo; sin embargo, su tendencia refleja el comportamiento de la serie original.

Para brindar una perspectiva estadística sobre el nivel de ajuste de la red LSTM, en el Cuadro 7.11 se presenta el MAPE, el AIC y el BIC. El MAPE de entrenamiento obtenido es del 32.4599 %. Sin embargo, a pesar de que el ajuste no es el más preciso según los indicadores, el desempeño con datos fuera de muestra es satisfactorio.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
32.45993	3723.65808	3777.69763

Cuadro 7.11: Medidas de bondad de ajuste del modelo LSTM para datos de entrenamiento de la demanda de Fuel Oil

La Figura 7.15 presenta el desempeño del modelo red LSTM por fuera de muestra. De las 12 observaciones no conocidas por el modelo, el modelo estima con alta precisión 6 observaciones.

En marzo y abril de 2021 se presentan los errores más significativos. Típicamente, los máximos de consumo se han presentado en marzo y en algunos meses de agosto, según datos históricos. Este comportamiento fue aprendido por el modelo de red LSTM. Sin embargo, en 2021, el máximo se presentó en abril, lo cual generó un error adicional fuera muestra. De hecho, el MAPE de validación cae hasta el 12.0022 % como se observa en el Cuadro 7.12. Es decir, el desempeño del modelo por fuera de muestra es más satisfactorio que dentro de muestra, lo que se desea en estos ejercicios de proyección.

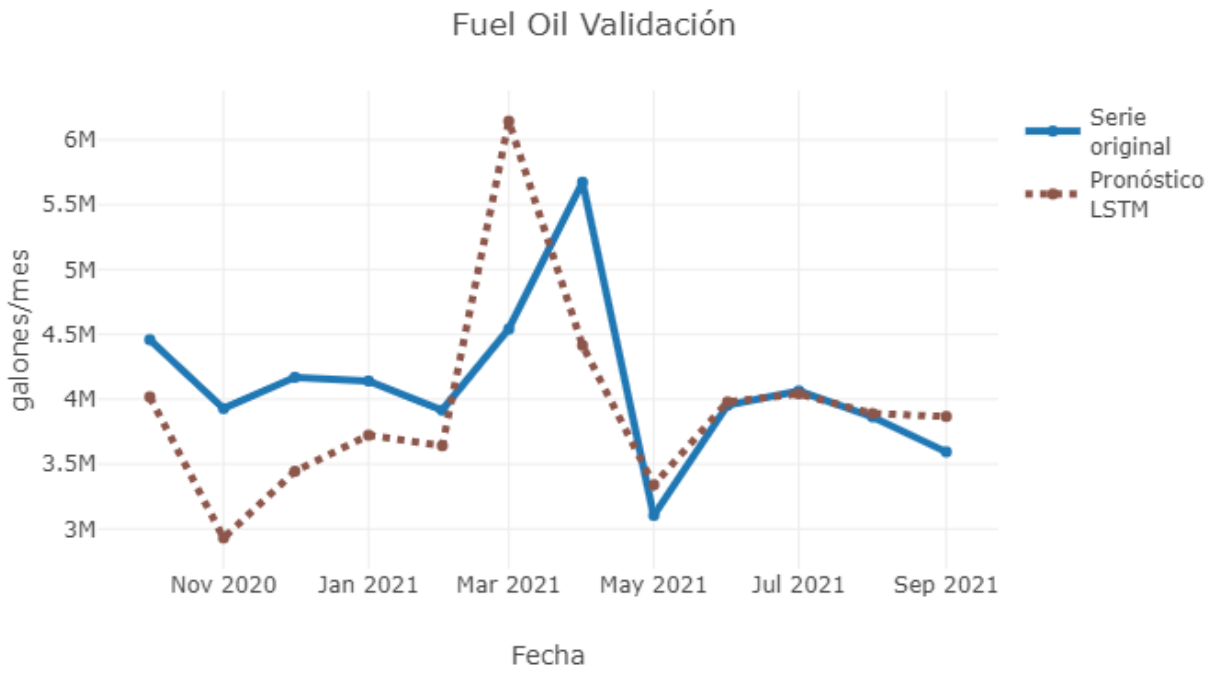

Figura 7.15: Ajuste del modelo LSTM para datos de validación de la demanda Fuel Oil para el periodo 2020/10 - 2021/09

Validación		
MAPE (%)	**AIC**	**BIC**
12.02239	361.74229	370.95551

Cuadro 7.12: Medidas de bondad de ajuste del modelo LSTM para datos de validación de la demanda de Fuel Oil

En la Figura 7.16 se presentan las proyecciones obtenidas por el modelo LSTM para la demanda de fuel oil junto a sus intervalos de confianza bootstrap con 1000 replicas. En las proyecciones obtenidas por el modelo LSTM se observa que la demanda de fuel oil registra un comportamiento estacional marcado con picos en los meses de marzo y agosto, como se ha presentado en la historia de la variable. Adicionalmente, se evidencia que la variabilidad de las proyecciones es estable y similar a la que traía la serie original como se observó en la figura Figura 7.13.

Finalmente se observa que no existe gran diferencia entre los intervalos de confianza bootstrap y el valor efectivamente proyectado tal como se observó en el caso de las proyecciones de diésel. Esto lo explica la robustez de las estimaciones generadas por el modelo LSTM. La introducción de pequeños errores de estimación no afecta de forma significativa las proyecciones realizadas por el modelo.

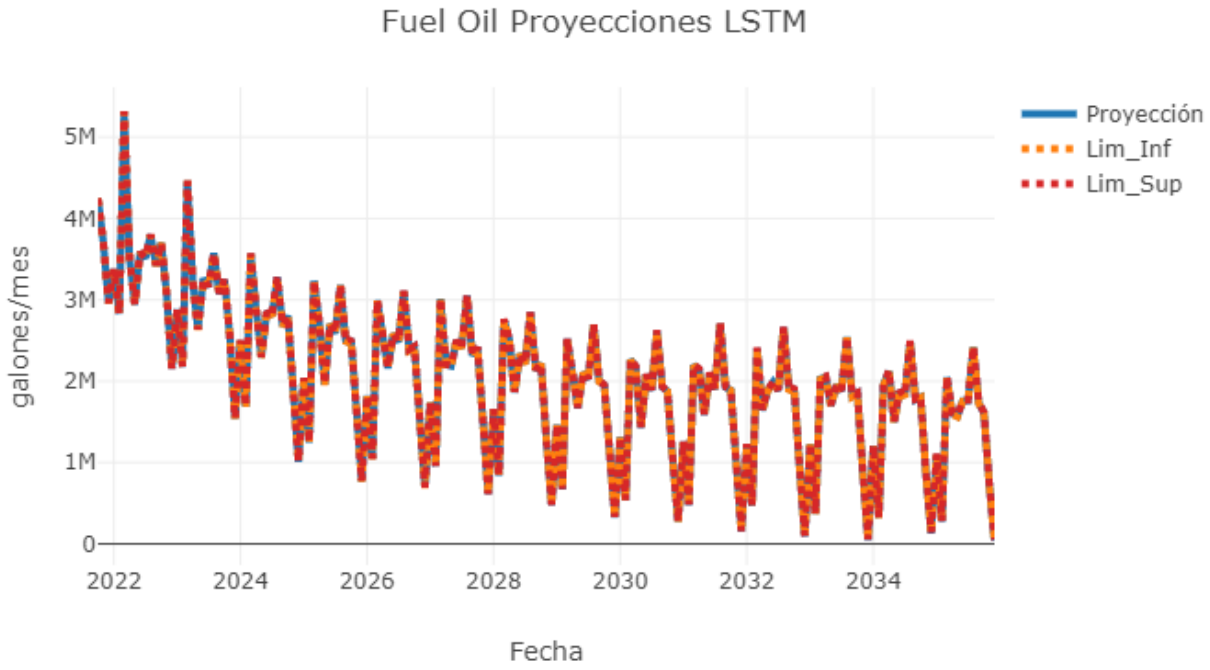

Figura 7.16: Proyecciones del modelo LSTM para la demanda de Fuel Oil para el periodo 2021/10 - 2035/12

Con el fin de ilustrar la diferencia entre las estimaciones de los intervalos de confianza y las proyecciones, se presenta el Cuadro 7.13 donde se presentan los encabezados obtenidos por las proyecciones de fuel oil en este escenario.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	4015369.85142	4017993.29423	4020616.73705
2020-11-01	2927650.75861	2930388.39916	2933126.03971
2020-12-01	3439930.68438	3444178.76592	3448426.84745
2021-01-01	3721454.90526	3722526.84923	3723598.79321
2021-02-01	3638973.80186	3641472.46334	3643971.12482
...	...	...	...
2035-08-01	2412631.56743	2419877.64280	2427123.71818
2035-09-01	1722388.04102	1730840.77806	1739293.51511
2035-10-01	1607705.76864	1612600.71399	1617495.65934
2035-11-01	729201.08808	736505.68170	743810.27532
2035-12-01	34331.71951	42102.33385	49872.94819

Cuadro 7.13: Encabezado proyecciones modelo LSTM para la demanda de Fuel Oil

7.2.2 Fuel Oil: Escenario 2

En este segundo escenario, después de un análisis experimental exhaustivo, se ha eliminado la variable `p_crudo` del escenario 1 y se han incluido variables con asociadas al fenómeno del Niño como `ninonina` (que representa el ONI) y `dum_nino` (que representa el evento de ONI ≥ 0.5). También se debe tener en cuenta que todas las variables asociadas al caso base siguen estando presentes en este escenario. La variable `p_fueloil` se ha mantenido. El mejor modelo para proyectar la demanda de fuel oil en este caso fue la combinación de MARS y la red neuronal LSTM.

Las especificaciones del mejor modelo en este escenario consideran incluir los primeros cinco rezagos de la variable `d_fueloil`. Esta estrategia, como en todos los casos ilustrados en este informe, se logra capturar el comportamiento autoregresivo no estacional del consumo de fuel oil. La red LSTM consiste de una capa oculta con 40 neuronas con función de activación encontrada tangente hiperbólica.

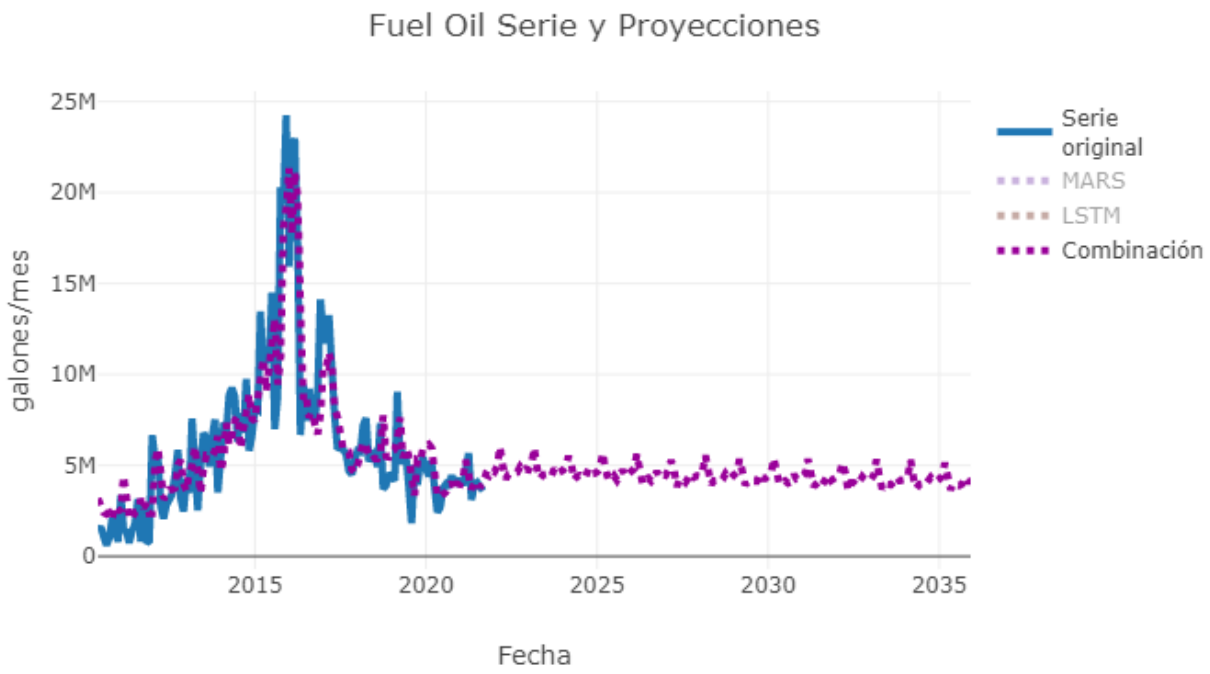

Figura 7.17: Ajuste del modelo de combinación de pronósticos de pronósticos a la demanda de Fuel Oil para el periodo 2010/01 - 2035/12

En la Figura 7.17 se presenta a nivel general el ajuste del modelo de combinación seleccionado. La diferencia con respecto a los resultados del escenario 1 radican en que las proyecciones de fuel oil futuras tienden a ser más constantes.

En la Figura 7.18 se ilustra la serie original y proyección obtenida a través del modelo de combinación entre MARS y LSTM para los datos dentro de muestra. De la Figura 7.18 también se observa un ajuste satisfactorio con este modelo de combinación entre MARS y red LSTM. Como se mencionó previamente, entre 2015 y 2016 se ha presentado un aumento inesperado del consumo de fuel oil dada la ocurrencia del fenómeno del Niño en ese momento. Esto motivó a que en este escenario se consideraran las variables asociadas a este fenómeno.

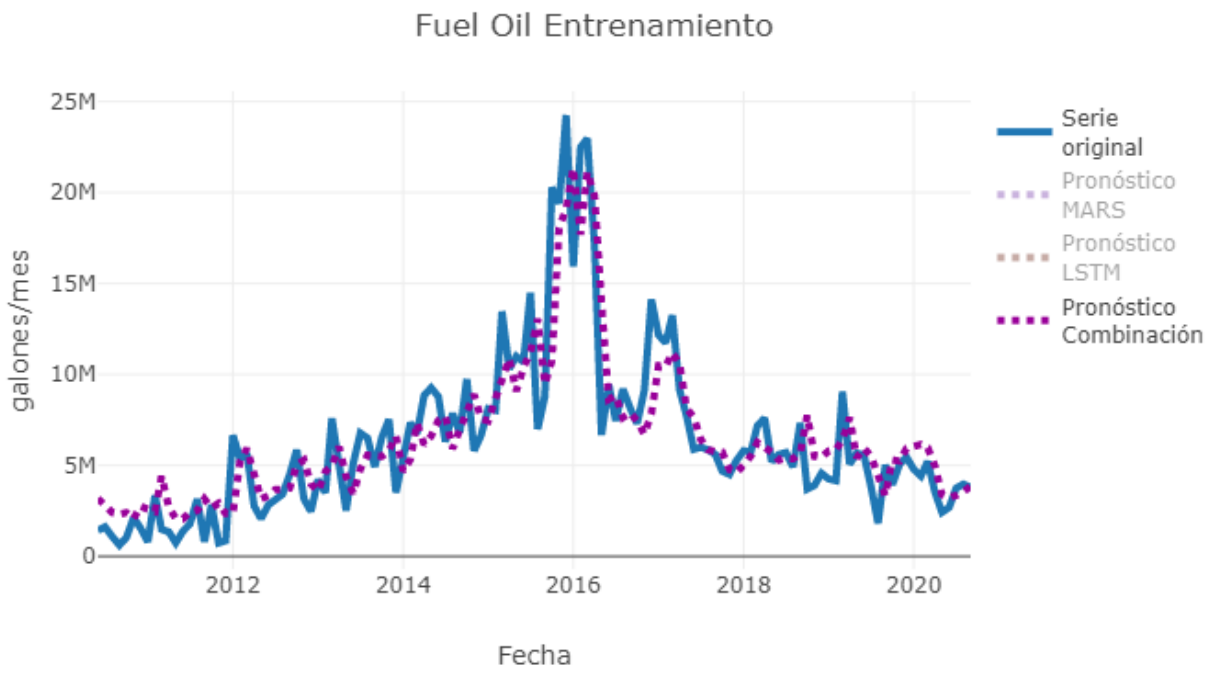

Figura 7.18: Ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda Fuel Oil para el periodo 2010/01 - 2020/09

Para brindar una perspectiva estadística sobre el nivel de ajuste del modelo de combinación usado en este escenario, en el Cuadro 7.14 se presenta el MAPE, el AIC y el BIC. El MAPE de entrenamiento obtenido es del 40.98892 %. A pesar de que las métricas dentro de muestra no son las menores (similar al escenario 1), se notó que el modelo captura las tendencias en la demanda de fuel oil.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
40.98892	3665.37666	3730.24313

Cuadro 7.14: Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda de Fuel Oil

La Figura 7.19 presenta el desempeño del modelo de combinación entre MARS y la red LSTM por fuera de muestra. El MAPE de validación en este escenario se pudo reducir hasta 9.18646 % como se observa en el Cuadro 7.15. Nuevamente, el desempeño predictivo del modelo es bastante satisfactorio.

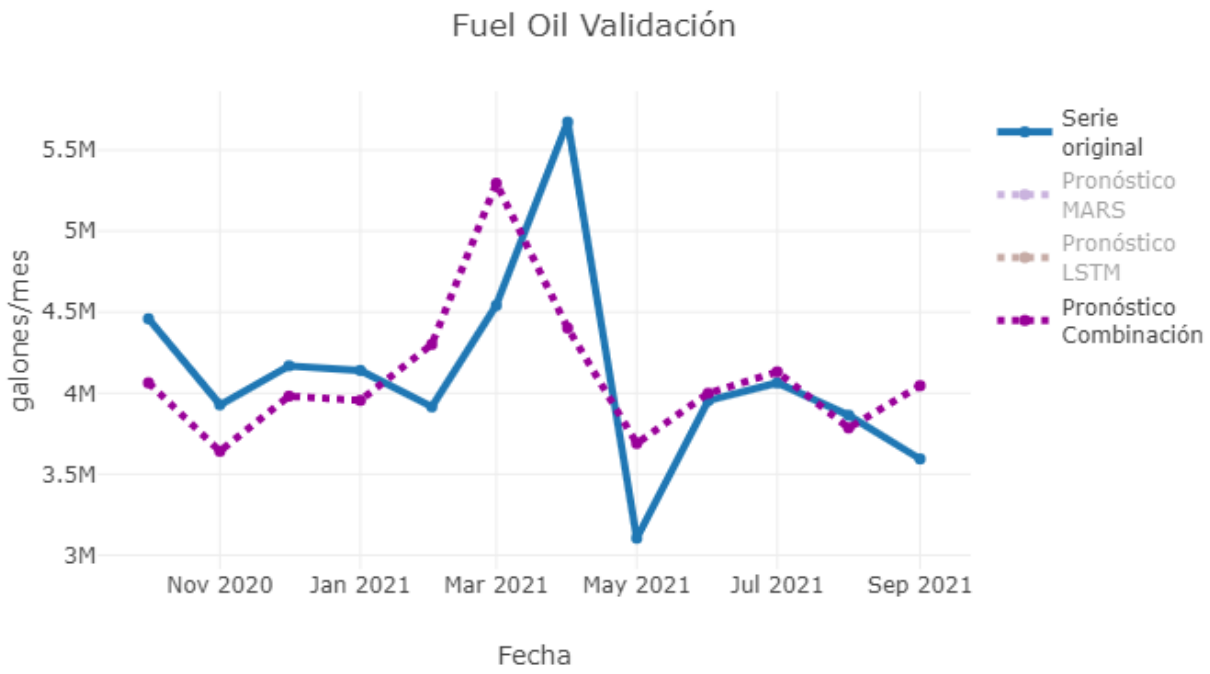

Figura 7.19: Ajuste del modelo de combinación de pronósticos para datos de validación de la demanda Fuel Oil para el periodo 2020/10 - 2021/09

Validación		
MAPE (%)	**AIC**	**BIC**
9.18646	361.71357	372.86642

Cuadro 7.15: Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de validación de la demanda de Fuel Oil

En la Figura 7.20 se presentan las proyecciones obtenidas por el modelo de combinación entre MARS y LSTM para la demanda de fuel oil junto a sus intervalos de confianza bootstrap con 1000 replicas. En las proyecciones obtenidas, como en el escenario 1, se observa que la demanda de fuel oil registra un comportamiento estacional con demandas máximas en marzo como se ha presentado en la historia de la variable. En este escenario, también se presenta un segundo máximo, aunque de menor nivel, en los meses de septiembre. Este comportamiento es más cercano al comportamiento histórico del consumo de fuel oil en el país. Adicionalmente, también se sigue evidenciando que la variabilidad de las proyecciones es estable y similar a la que traía la serie original como se observó en la figura Figura 7.17.

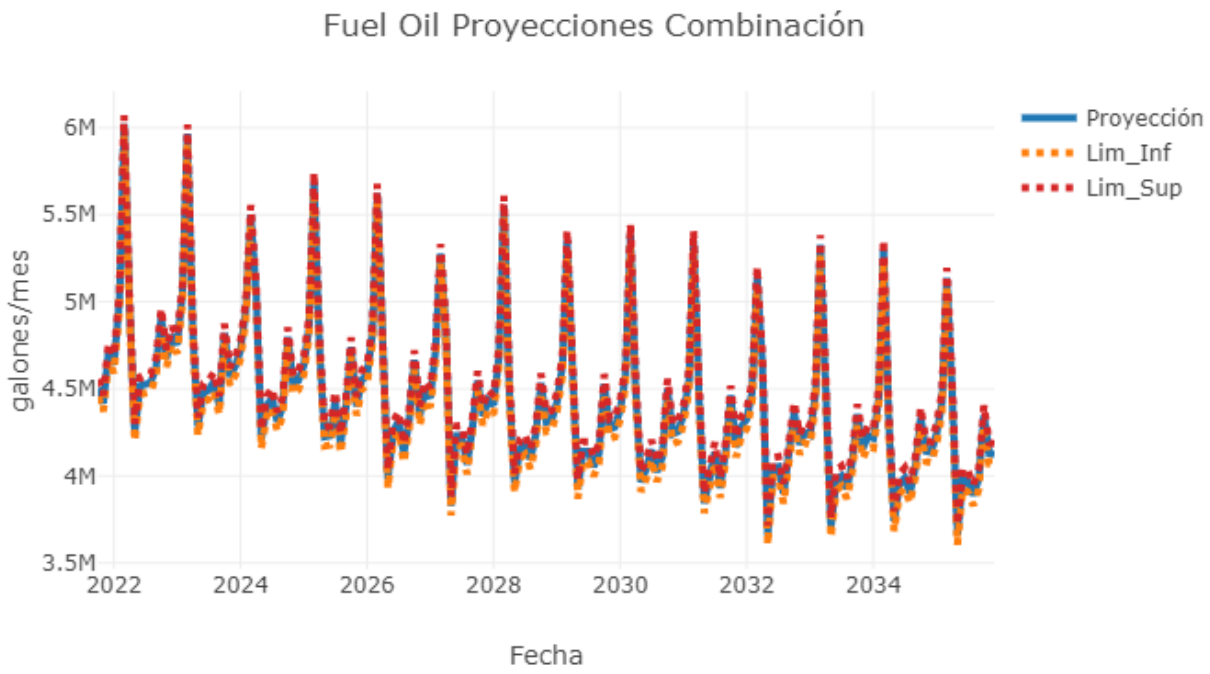

Figura 7.20: Proyecciones del modelo de combinación de pronósticos para la demanda de Fuel Oil para el periodo 2021/10 - 2035/12

Por otro lado, la diferencia entre los intervalos de confianza bootstrap y el valor proyectado es pequeña como se presenta en el Cuadro 7.16.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	4029060.46641	4063870.80883	4098681.15126
2020-11-01	3599845.66359	3641644.07178	3683442.47996
2020-12-01	3936687.23070	3981266.41722	4025845.60374
2021-01-01	3913873.65918	3955391.00923	3996908.35927
2021-02-01	4254932.76915	4300901.11444	4346869.45972
...	...	...	...
2035-08-01	3830217.45039	3889286.32451	3948355.19864
2035-09-01	3922667.52925	3981359.65940	4040051.78956
2035-10-01	4288454.72030	4347410.74052	4406366.76075
2035-11-01	4068699.23794	4129259.32374	4189819.40953
2035-12-01	4061343.41924	4123171.43230	4184999.44537

Cuadro 7.16: Encabezado proyecciones modelo de combinación de pronósticos para la demanda de Fuel Oil

7.3 Gas Licuado de Petróleo (GLP)

En el caso de los modelos de demanda del Gas Licuado de Petróleo (GLP), se realiza la estimación de un total de 100 escenarios, divididos en 13 casos adicionales al escenario base, en los cuales se las variables explicativas de *p_glp*, *p_crudo*, *d_gnresi*, *d_gnterc*, *d_gntot*, *temp* y *tot_aut_glp*. Es de anotar que las variables explicativas adicionales al escenario base son descritas en el Cuadro 6.1.

Es de anotar que, para el caso de la demanda de GLP (d_gpl), no se tienen en cuenta rezagos de la variable dependiente en el planteamiento de los escenarios que se evaluaron en cada caso. Al ingresar rezagos para la demanda de GLP se evidenciaba que las proyecciones y los intervalos de confianza se veían gravemente afectados. Esto podría generar que los valores proyectados y sus intervalos de confianza tomaran valores con tendencia al infinito.

Una vez estimados los 100 escenarios, se preseleccionó un grupo de 21 escenarios distintos, para ser evaluados de la mano del juicio experto de la UPME, concluyendo que de éstos, 2 escenarios en particular son los que satisfacen las expectativas de la UPME en materia de variables explicativas, tasa de crecimiento y variabilidad, en su trayectoria de proyección a 2035. Dichos escenarios se presentan a continuación junto a un análisis detallado y comparativo de sus resultados.

7.3.1 GLP: Escenario 1

El primer escenario seleccionado para la demanda de GLP, se calcula mediante el empleo de un modelo GAM, en el cual se incluyen las variables de `p_glp` y `p_crudo`, adicionales a las variables macroeconómicas, de efecto calendario y covid, explicadas previamente para el escenario base. Adicionalmente, se incluyen tres rezagos para la variable `pib`, con el fin de capturar el efecto autoregresivo no estacional que presenta la variable de crecimiento económico.

Es de anotar que la inclusión de la variable *p_glp*, se debe a la lógica económica en la cual se tiene que la demanda de un bien y el precio del mismo poseen una relación negativa, en donde uno aumenta cuando el otro disminuye, o viceversa, siempre y cuando la elasticidad precio de la demanda del bien no sea inelástica. La variable *p_crudo*, entra en el modelo como una variable proxy del precio de otros combustibles fósiles que podrían ser considerados como sustitutos de GLP.

Similar a como se realizó en la Sección 7.1 para la demanda de ACPM/Diésel y la Sección 7.2 para la demanda de Fuel Oil, en este caso se presentan cuatro gráfico que buscan mostrar el ajuste, validación y proyecciones obtenidas por el modelo GAM, a saber, en la Figura 7.21 se ilustra el ajuste general del modelo GAM en todas las fases del estudio, en la Figura 7.22 se presenta específicamente el ajuste del modelo GAM para datos dentro de muestra, en la Figura 7.23 se muestra el nivel de ajuste del modelo respecto a los datos de validación, y en la Figura 7.24 se ilustran las proyecciones obtenidas por el modelo GAM junto con sus

correspondientes intervalos de confianza bootstrap.

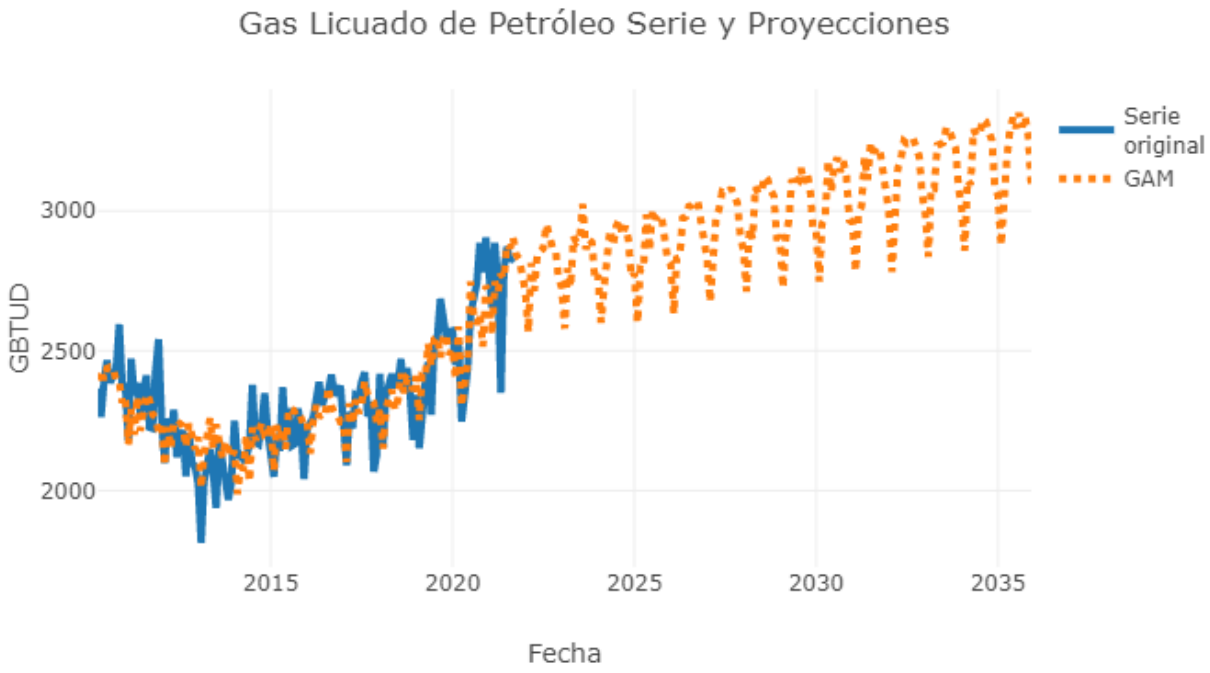

Figura 7.21: Ajuste del modelo GAM a la demanda de GLP para el periodo 2010/01 - 2035/12

De la Figura 7.21 se presenta en términos generales el ajuste del modelo GAM a la demanda de GLP, en donde se observa que a pesar de que el ajuste observado por el modelo GAM para el periodo 2010 - 2014 para datos de entrenamiento, no pareciera ser tan preciso como se esperaría, se observa que luego del 2014, el ajuste del modelo GAM mejora significativamente, en donde logra capturar los picos y los valles que se registran para la demanda GLP.

En el caso de las proyecciones, se evidencia que la variabilidad que expone el modelo GAM, es similar a la que se experimenta luego de 2014, en donde se exhiben valles marcados en los meses de Febrero y consumos relativamente constantes entre Mayo y Octubre.

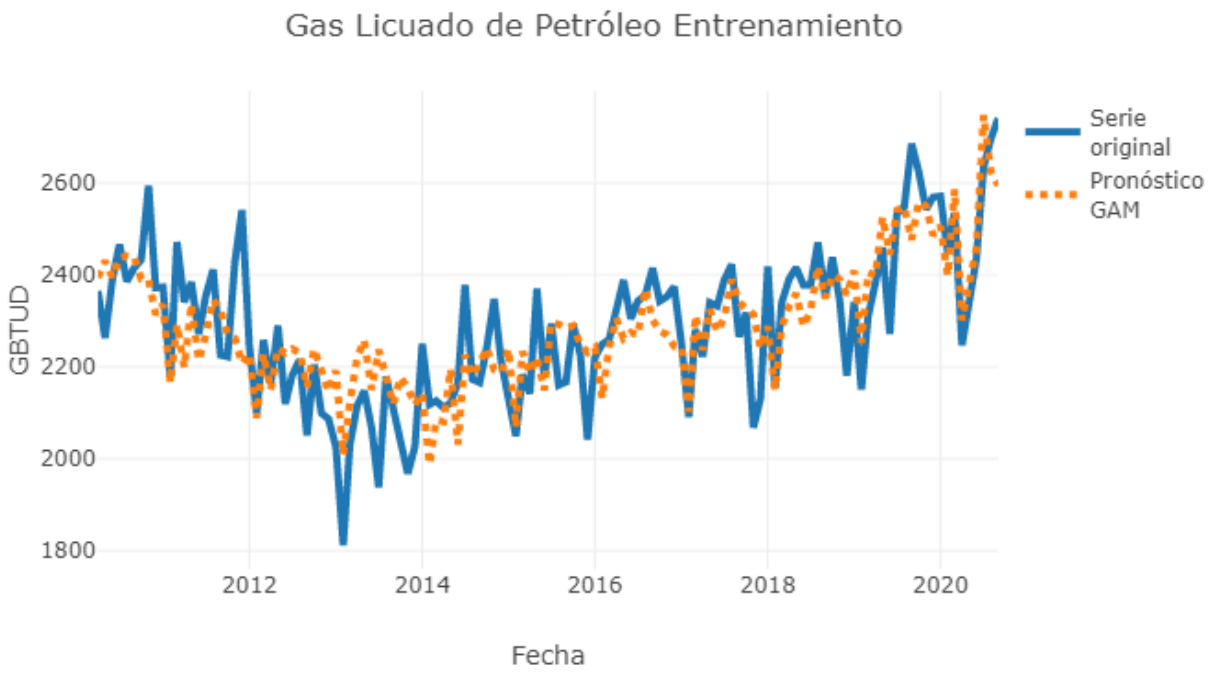

Figura 7.22: Ajuste del modelo GAM para datos de entrenamiento de la demanda GLP para el periodo 2010/01 - 2020/09

Con el fin de hacer un análisis más detallado para el ajuste del modelo GAM a los datos de entrenamiento, se presenta la Cuadro 7.17, en la cual se evidencia que la demanda de GLP, exhibe un cambio de pendiente en el año 2013, pasando de un comportamiento decreciente, a un comportamiento creciente. El comportamiento expuesto que sufre la demanda de GLP durante el 2013, puede ser explicada posiblemente por el cambio en la pendiente que sufren las reservas de crudo y de gas natural en el país, en donde como se observa en el documento de reservas históricas presentado por la Agencia Nacional de Hidrocarburos (ANH) (ANH, 2020), tanto las reservas de crudo, como de gas natural alcanzan su pico en entre 2012-2013, y a partir de ese momento, las reservas de ambos combustibles comienzan a decrecer.

Adicionalmente, en la Cuadro 7.17 se observa el nivel de ajuste del modelo GAM, una vez ocurre el cambio de pendiente de la demanda de GLP en 2013, en donde, el modelo a jsutado, logra capturar de forma satisfactoria la mayoría de los picos que se dan entre los meses de Mayo - Octubre de cada año, y los valles que se registran en los meses de Febrero.

Ahora, con el fin de cuantificar el nivel de ajuste que presenta el modelo para los datos de entrenamiento, y comparar dicho ajuste con el escenario que se presenta posteriormente, se presentan en el Cuadro 7.17 las medidas de bondad de ajuste, MAPE, AIC y BIC obtenidas por el modelo, tal como se hizo en los combustibles presentados anterioremente.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
3.33445	1196.96188	1253.68752

Cuadro 7.17: Medidas de bondad de ajuste del modelo GAM para datos de entrenamiento de la demanda de GLP

En el Cuadro 7.17 se encuentra que, a pesar de que el ajuste del modelo GAM no parece ser tan preciso para las observaciones registradas antes de 2013, se logra evidenciar que el MAPE de entrenamiento es muy bajo, siendo el valor registrado por este modelo de 3.33445 %, valor que según la escala de precisión de Cuadro 7.2 es muy precisa debido a que posee un valor inferior al 10 %. Por su parte los valores obtenidos por el AIC y BIC, iguales a 1196.96188 y 1253.68752 serán empleados en el análisis de la Subsección 7.3.2, debido a que estos criterios de decisión se emplean para la comparación de modelos, tal como se expone en la Sección 6.12.

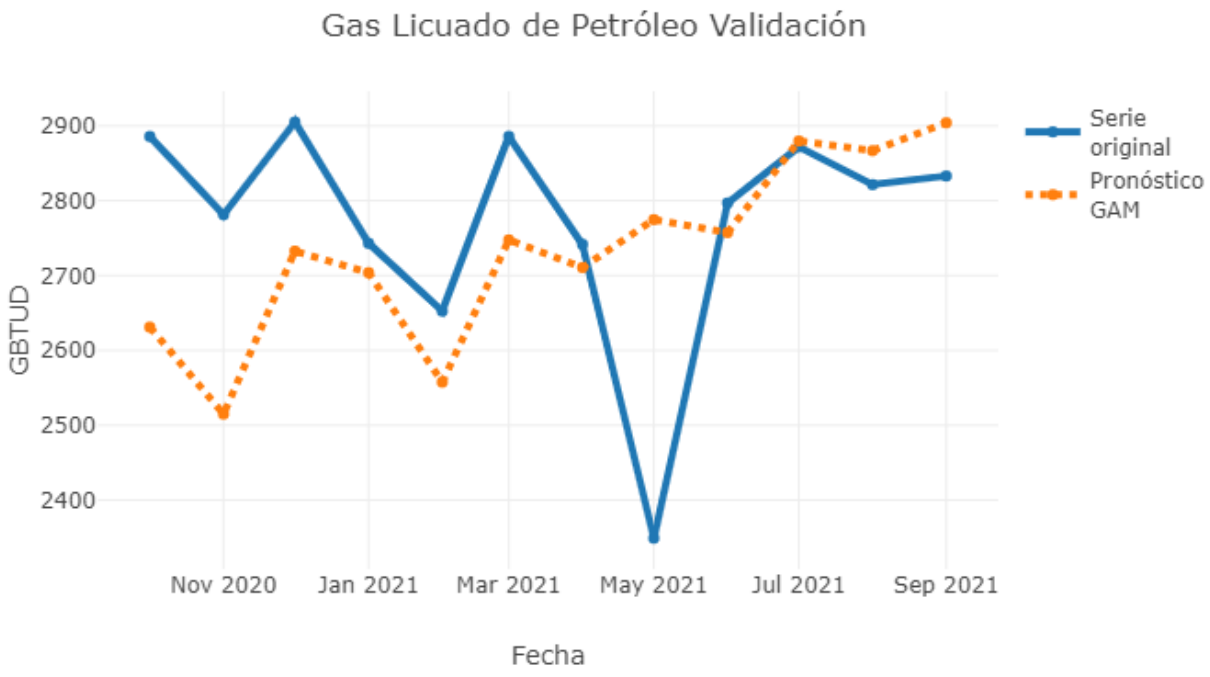

Figura 7.23: Ajuste del modelo GLP para datos de validación de la demanda GLP para el periodo 2020/10 - 2021/09

En el caso del nivel de ajuste del modelo GAM respecto a la información de validación que se dejó por fuera, se presenta la Figura 7.23, con el fin de presentar de forma gráfica el nivel de ajuste del modelo respecto a los 12 meses que se dejaron para medir el desempeño predictivo del modelo.

En dicha gráfica se observa, que el ajuste del modelo GAM pareciera no ser del todo adecuado para la demanda observada de GLP, puesto que, aunque los picos y valles que se registra por el modelo son similares en términos de incrementos y decrementos, los valores

puntuales son muy precisos respecto a los valores efectivamente observados, a excepción de los meses de Junio-Septiembre de 2021, en donde la serie logra estabilizarse y arrojar resultados similares a los efectivamente observados.

Dada la evidencia gráfica presentada en la Figura 7.23, se decide presentar algunas medidas de bondad de ajuste que permitirán cuantificar mediante medidas estadísticas el nivel de ajuste que se registra para estas proyecciones. Estas medidas son presentadas en el Cuadro 7.18.

Validación		
MAPE (%)	**AIC**	**BIC**
4.93333	164.55145	174.24959

Cuadro 7.18: Medidas de bondad de ajuste del modelo GAM para datos de validación de la demanda de GLP

Contrario a lo esperado, el Cuadro 7.18 muestra que el desempeño predictivo obtenido por el modelo GAM respecto a la demanda de GLP que se dejó por fuera de muestra es buena, puesto que, el MAPE obtenido por este ajuste fue del 4.93333 %. La razón por la cual se registra un MAPE tan bajo, aún cuando en la Figura 7.23 no se observa un ajsute tan preciso de parte del modelo GAM, puede ser explicado por la escala de medición de la variable GLP, en donde se observa que la escala de medición de la variable observada oscila entre 2400 y 2900 GBTU, y mientras que las observacionesestimado oscilan entre 2500 y 2900 GBTU, Además de los valores estimados por el modelo, aciertan de forma muy aproximada las observaciones registradas para Febrero, Abril, Junio, Julio, Agosto y Septiembre de 2021.

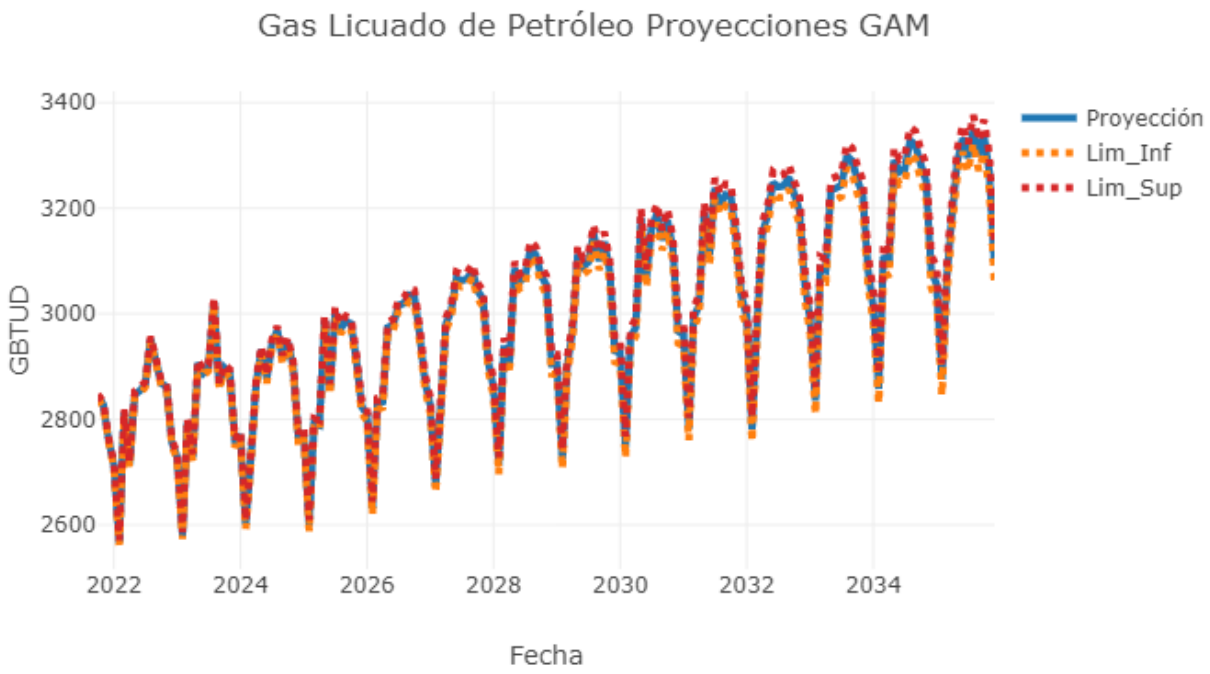

Figura 7.24: Proyecciones del modelo GAM para la demanda de GLP para el periodo 2021/10 - 2035/12

A pesar de que en la Figura 7.24 pareciera evidenciarse una alta variabilidad para los valores proyectados por el modelo GAM, al comparar la variabilidad de estas proyecciones con la variabilidad de la serie original reportada en la Figura 7.22, se evidencia que ambas variabilidades son similares luego del año 2013. Adicionalmente, en la Figura 7.24 se observa que durante los meses de Febrero de cada año, hay un caída en la demanda de GLP, caídas que son consecuentes con el comportamiento de la demanda original.

Finalmente en términos de la tendencia presentada en la Figura 7.24, se observa que la demanda proyectada de GLP incrementa presenta una pendiente un poco más leve que la registrada por la serie original entre 2013 y 2019, en donde la serie proyectada pasa de un valor mínimo de 2566 GBTU en Febrero de 2022, a un valor máximo de 3349 GBTU en Agosto de 2035, mientras que, la serie original pasa de un valor mínimo de 1812 GBTU en Febrero de 2013, a un valor máximo de 2685 en Septiembre de 2019.

Ahora bien, con el objetivo de presentar un vistazo rápido sobre los valores estimados por fuera de muestra y proyectados para la demanda de GLP, se presenta en el Cuadro 7.19 el encabezado de los datos de validación y las proyecciones obtenidas por el modelo GAM estimado en este escenario, con el objetivo de dar un vistazo rápido entre la diferencia entre los valores proyectados por el modelo y el intervalo de confianza bootstrap construido al realizar 1000 réplicas.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	2626.69851	2631.04995	2635.40138
2020-11-01	2510.47736	2514.98397	2519.49059
2020-12-01	2728.58824	2732.64125	2736.69427
2021-01-01	2699.64088	2703.88492	2708.12897
2021-02-01	2553.50562	2557.55145	2561.59728
...	...	...	...
2035-08-01	3322.45853	3349.42554	3376.39255
2035-09-01	3271.06974	3299.64965	3328.22955
2035-10-01	3306.15882	3336.90424	3367.64966
2035-11-01	3236.65478	3269.56379	3302.47280
2035-12-01	3058.59889	3092.56412	3126.52935

Cuadro 7.19: Encabezado proyecciones modelo GAM para la demanda de GLP

7.3.2 GLP: Escenario 2

El segundo escenario seleccionado para proyectar la demanda de GLP, se calcula al igual que el escenario 1 empleando un modelo GAM. A diferencia del escenario 1, en este escenario se emplean como variables adicionales al escenario base, el `p_glp`, el `p_crudo` y la `d_gnresi`, siendo la variable diferenciadora entre los dos escenarios la `d_gnresi`. Adicional a estas variables, en este escenario se incluyen dos rezados de la variable `pib` y dos rezagos de la variable `d_gnresi`.

Es de anotar que la inclusión del `p_glp` y el `p_crudo`, se explicó previamente en la Subsección 7.3.1, mientras que, la inclusión del `d_gnresi`, se basa en los hallazgos encontrados en la revisión de la literatura, en donde algunos autores hacen énfasis en que el GLP se suele emplear principalmente para consumo residencial con fines de calefacción y cocción, es decir, como un bien sustituto del GLP, y por tanto, se espera que su inclusión contribuya en el nivel de ajuste obtenido por el modelo.

Para probar la hipótesis de que la inclusión de la `d_gnresi` contribuye al ajuste del modelo, se plantean al igual que en los otros combustibles, un total de cuatro gráficas, a saber, en la Figura 7.25 se presenta el ajuste general y proyecciones que se obtienen con el modelo GAM, en la Figura 7.26 se hace énfasis en el ajuste del modelo GAM en los datos usados para entrenamiento, en la Figura 7.27 se presenta el ajuste del modelo GAM en los datos usados para validación, y finalmente, en la Figura 7.28, se presenta las proyecciones finales obtenidas por el modelo junto a sus correspondientes intervalos bootstrap del 95 % de confianza.

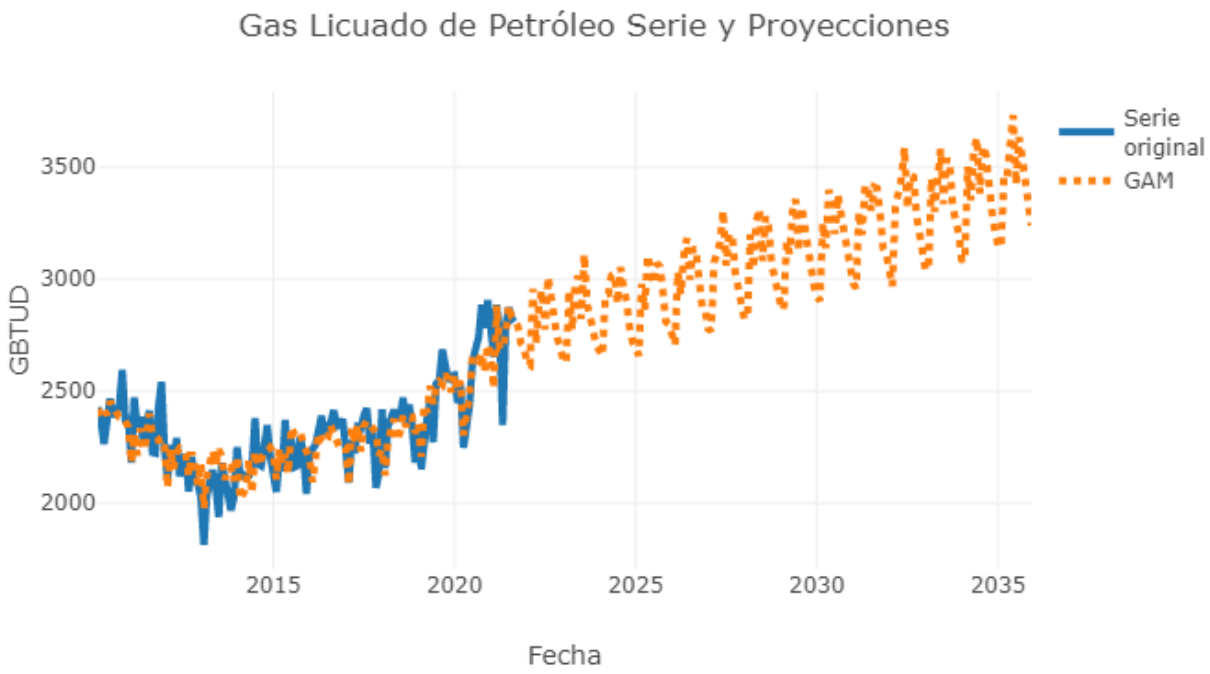

Figura 7.25: Ajuste del modelo GAM a la demanda de GLP para el periodo 2010/01 - 2035/12

En la Figura 7.25, se los valores estimados por el modelo GAM, en entrenamiento y en validación, junto con las proyecciones realizadas por el mismo al horizonte de 2035. En dicha figura, se evidencia que exceptuando los picos registrados antes de 2013, el modelo GAM logra capturar en general el comportamiento estacional que presentan la variable de demanda de GLP, además de capturar de forma adecuada el cambio de pendiente que se observa en el año 2013, impulsado posiblemente por la reducción en las reservas que registraron a partir de ese año (ANH, 2020).

Adicionalmente, en términos de proyección, se evidencia que la variabilidad de las proyecciones tiende a aumentar ligeramente con el tiempo, tratando de imitar la variabilidad que se presenta en la demanda de GLP entre 2013 a 2020, en donde se evidencia también un aumento en su variabilidad a medida que transcurre el tiempo. También puede notarse, que el comportamiento estacional de la serie original se mantiene en las proyecciones, puesto que, se registran valles en los meses de Febrero y picos entre Mayo y Octubre.

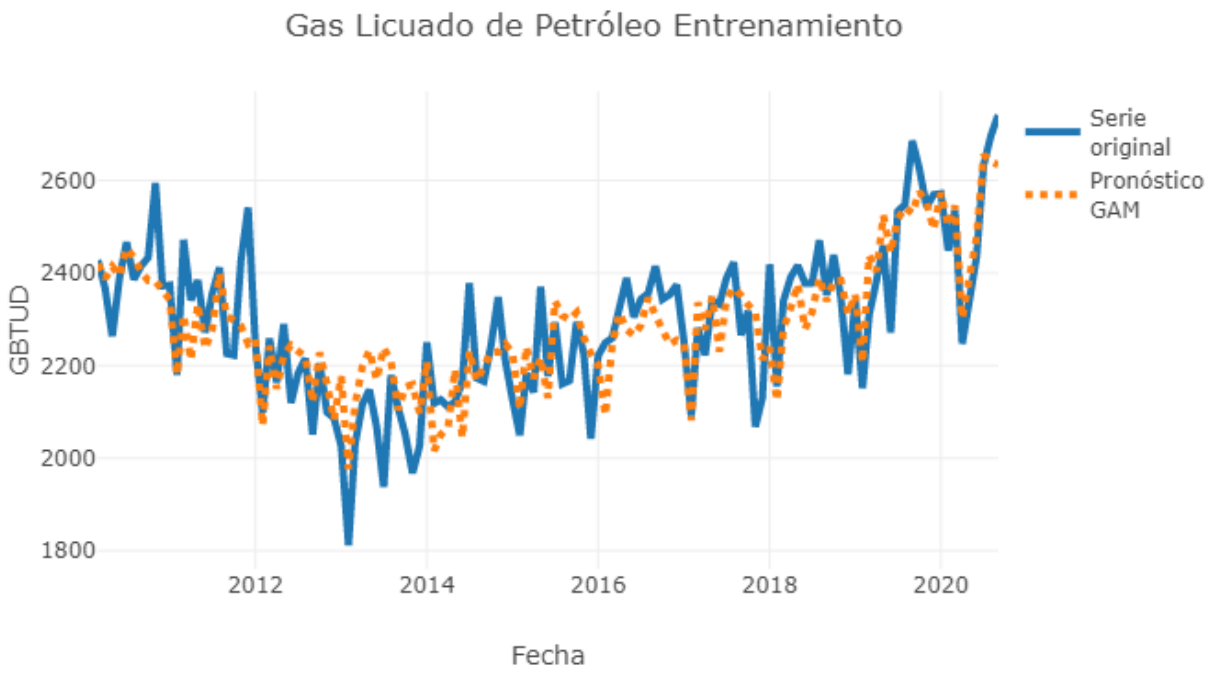

Figura 7.26: Ajuste del modelo GAM para datos de entrenamiento de la demanda GLP para el periodo 2010/01 - 2020/09

Al hacer un zoom, al ajuste que ofrece el modelo GAM respecto a los datos de entrenamiento que fueron utilizados para realizar el ajuste del modelo, se observa que efectivamente, el modelo GAM, logra capturar el comportamiento de la tendencia tanto decreciente como creciente que se registra en la serie original, además de capturar los valles que se registran antes de 2013 de forma casi perfecta, mientras que, en el ajuste posterior a 2013, se observa como logra adaptarle de forma adecuada a los valles que se registran en los meses de febrero, y a los periodos largos de consumo aproximadamente constante.

Para cuantificar el nivel de ajuste que ofrece el modelo GAM para este escenario respecto a los datos de entrenamiento, se presenta en el Cuadro 7.20, en el que se exhibe el MAPE del ajuste, junto a los criterios de información AIC y BIC.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
3.10334	1193.14223	1255.71435

Cuadro 7.20: Medidas de bondad de ajuste del modelo GAM para datos de entrenamiento de la demanda de GLP

De los resultados obtenidos en el Cuadro 7.20, destaca el hecho de que el MAPE obtenido es menor al registrado en el Cuadro 7.17, asociado al ajuste del modelo presentado en el escenario 1, en donde en este caso el MAPE registrado es del 3.10334 %, mientras que en el escenario 1, en el cual no se consideró la `d_gntran` fue del 3.33445 %.

Caso similar ocurre con el AIC registrado por el escenario 2, más no con el BIC, puesto que, en este escenario el AIC fue de 1193.14223, lo cual es menor al registrado en el escenario 1, cuyo valor fue de 1196.96188, pero en donde, en el caso del BIC, se observa que en este caso fue de 1255.71435, mientras que en el escenario 1 fue de 1253.68752.

Es decir, el MAPE y el AIC indican que el modelo ajustado con el escenario en el que se incluye la `d_gntran` es mejor que en el escenario en el cual se omite dicha variable, mientras que en el caso del BIC, se concluye lo contrario.

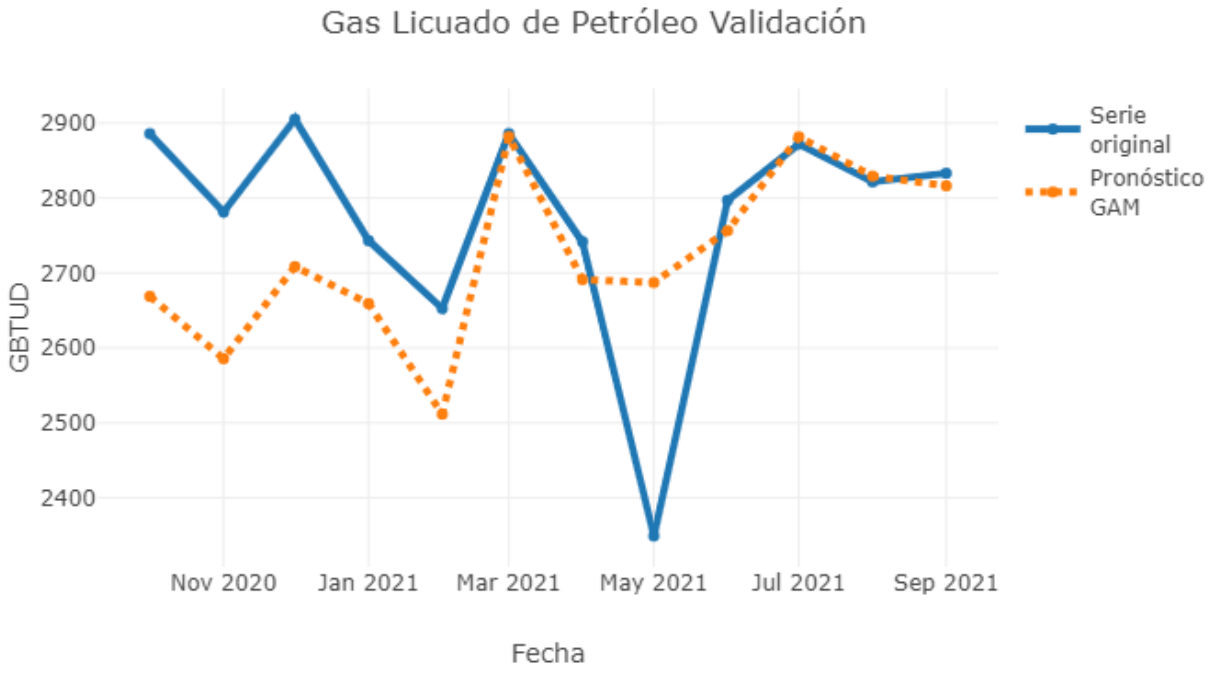

Figura 7.27: Ajuste del modelo GLP para datos de validación de la demanda GLP para el periodo 2020/10 - 2021/09

Por otro lado, al observar en la Figura 7.27 el ajuste que ofrece el modelo GAM planteado en este escenario, se evidencia una mejora significativa en términos de validación, en donde se evidencia inmediatamente, el ajuste casi perfecto de la observación registrada en Marzo, Abril, Junio, Julio, Agosto y Septiembre de 2021, siendo los valores más desfasados en el ajuste los meses de Octubre, Noviembre, Diciembre de 2020 y Enero de 2021.

Para cuantificar la mejora en el nivel de ajuste que se evidencia en la Figura 7.27, respecto a la exhibida en la Figura 7.23, se presentan como medidas de bondad de ajuste para datos fuera de muestra, el MAPE, el AIC y el BIC de los ajustes, con el fin de comparar el desempeño predictivo entre los dos escenarios presentados para GLP.

Validación		
MAPE (%)	**AIC**	**BIC**
4.06090	164.25737	174.92531

Cuadro 7.21: Medidas de bondad de ajuste del modelo GAM para datos de validación de la demanda de GLP

En el Cuadro 7.21 se ilustra el valor del MAPE, AIC y BIC que registra el modelo GAM al incluir la variable `d_gnresi`, en donde al comparar el MAPE obtenido por este modelo (4.06090 %) respecto al obtenido en el escenario 1 (4.93333 %), se aprecia una mejoría cercana al 1 %, lo cual corrobora lo que se esperaba luego del análisis gráfico, en donde la inclusión de la `d_gnresi` mejora el ajuste del modelo respecto a los datos de validación.

Similar a los resultados obtenidos para el ajuste dentro de entrenamiento, en la Cuadro 7.21 se evidencia que el AIC registrado por el ajuste del modelo GAM presentado en este escenario (164.25737) es menor a registrado en la Cuadro 7.18 asociado al escenario 1 (164.55145), sin embargo, caso contrario ocurre nuevamente con el BIC, en donde es mayor en este escenario (174.92531) que en el escenario 1 (174.24959).

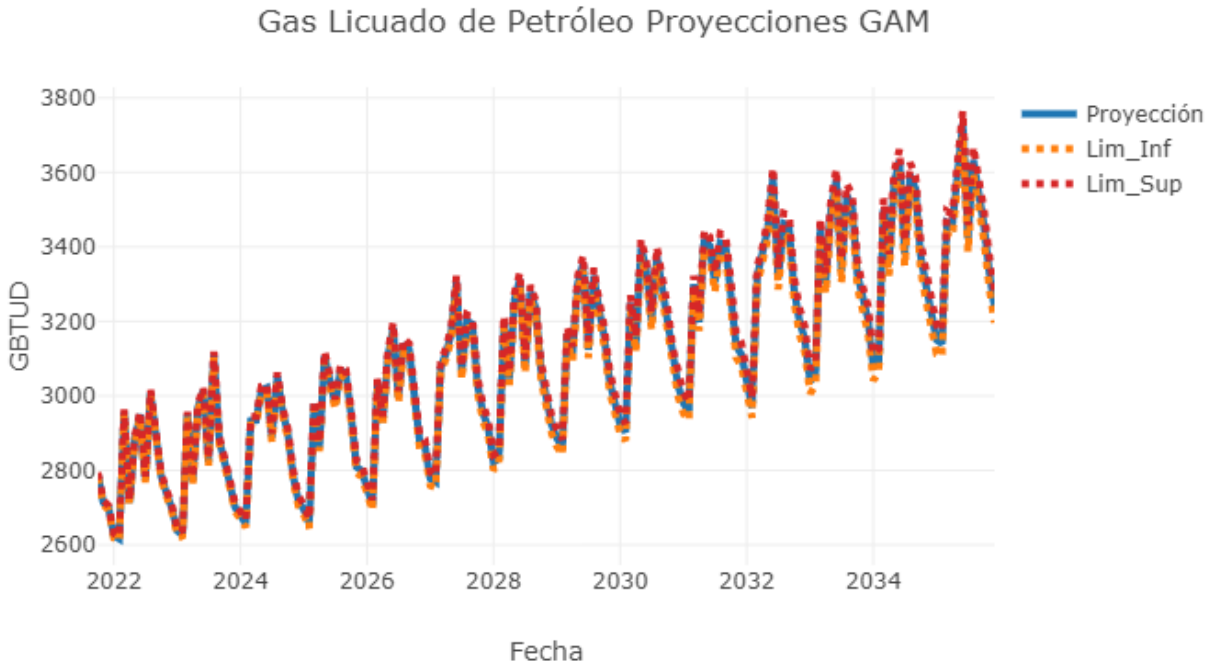

Figura 7.28: Proyecciones del modelo GAM para la demanda de GLP para el periodo 2021/10 - 2035/12

Finalmente, en la Figura 7.28 se ilustra las proyecciones obtenidas por el modelo GAM, junto con sus correspondientes intervalos de confianza. En esta figura se observa que la demanda de GLP presenta un comportamiento estacional creciente a lo largo del tiempo con valles en los meses de Febrero y picos relativamente constantes entre los meses de Mayo y Octubre.

Es de anotar que tanto la tendencia creciente que presenta esta variable, como la variabilidad de la misma puede ser explicada por el comportamiento creciente y la forma que presentaba la serie original entre el periodo 2013 - 2020, donde se observa que la demanda del GLP incrementa en aproximadamente en 900 GBTUD, durante este periodo, con una variabilidad similar a la presentada en la Figura 7.28.

Al revisar las proyecciones de otros casos y escenarios planteados, se evidencia que las proyecciones del GLP presentan en todos los casos comportamiento similar, cambiando solo un poco la pendiente en las cuales se registra pendientes más plantas como en la Figura 7.24 del escenario 1, o pendientes incluso más elevadas que la registra en la Figura 7.28 presentada en este escenario.

Similar a los escenarios planteados anteriormente para todos los combustibles, se ilustra en el Cuadro 7.22 los encabezados que se registran para los ajustes en validación y de las proyecciones realizadas, con el fin de observar el comportamiento que registra el modelo respecto a sus intervalos de confianza bootstrap que se generaron durante el proceso de proyección del modelo GAM.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	2704.39645	2708.86270	2713.32895
2020-11-01	2633.83026	2638.67991	2643.52956
2020-12-01	2738.28701	2742.53741	2746.78781
2021-01-01	2724.11361	2728.40445	2732.69530
2021-02-01	2636.08635	2641.17103	2646.25571
...	...	...	...
2035-08-01	4041.71277	4068.55607	4095.39938
2035-09-01	3955.96257	3985.08268	4014.20279
2035-10-01	4026.76479	4058.96146	4091.15812
2035-11-01	3923.12040	3957.63122	3992.14203
2035-12-01	3839.82830	3874.65557	3909.48283

Cuadro 7.22: Encabezado proyecciones modelo GAM para la demanda de GLP

Luego de analizar todos los resultados expuestos en el escenario 1 y el escenario 2, se tiene que, a pesar de que el criterio de información BIC favorece los ajustes obtenidos en el escenario 1, se concluye que la inclusión de la demanda del gas natural residencial dentro de las estimaciones de la demanda de GLP, permite mejorar los ajustes obtenidos por el modelo GAM ajustado, tanto dentro como fuera de muestra de entrenamiento.

Adicionalmente, la inclusión de dicha variable contribuye en el ajuste del comportamiento de la pendiente que se registra para las proyecciones, en donde se evidencia un leve incremento de la pendiente de las proyecciones, lo cual hace que la pendiente que registra la serie proyectada

tenga un comportamiento más constante a través del tiempo partiendo del periodo de Febrero de 2013, en donde se observa el cambio de pendiente que registra la demanda de GLP.

7.4 Gasolina Motor (GM)

Para el pronóstico de la demanda de Gasolima Motor (GM), además de plantear el escenario base compuesto por las variables de efecto calendario, variables macroeconómicas y la variable de cierres económicos, descritas en la Capítulo 7, se plantean 17 casos adicionales al caso base.

Dado lo anterior, se lograron analizar un total de 220 escenarios en los que se probaron diferentes combinaciones de variables, en las cuales se tiene, `p_gm`, `p_crudo`, `d_energ`, `d_gntran`, `d_gntot`, `rel_aut_elect/gm`, `rel_mot_elect/gm` `tot_aut_gm`, `tot_mot_gm`, `tot_aut_elect`, `tot_mot_elect` y `aut_liv_gm`; junto a un número diferente de rezagos para las variables `d_acpm`, `pib`, `d_gntran` y `d_gntot`, todas escritas en el Cuadro 6.1.

De los 220 escenarios estimados, se han seleccionado dos de ellos con el soporte de la UPME para ilustrar en este reporte. Estos escenarios satisfacen las expectativas de la UPME en términos de variabilidad en resultados, las variables explicativas y la trayectoria de proyección resultante al horizonte de diciembre 2035.

7.4.1 GM: Escenario 1

En este escenario, las variables explicativas adicionales a las del escenario base son el `p_gm`, `p_crudo` y `d_gntran`. El precio de la GM aporta a tener una representación de una función de demanda de la GM. Tal como en los casos anteriores, la variable `p_crudo` también ha mostrado ser útil para explicar el comportamiento de la demanda de GM.

La proyección del precio del crudo plantea que aumentaría de forma continua, aunque no aceleradamente, hasta 150 dólares/barril en 2035. La demanda de gas natural en el sector transporte puede considerarse como un bien sustituto de la GM. El modelo que mejor ha logrado explicar la demanda de GM en este escenario es MARS. Las especificaciones del mejor modelo en este escenario consideran incluir los dos primeros rezagos de la variable `d_gm` los cuales capturan la historia de dos meses del consumo de GM. El comportamiento estacional se sigue capturando a través de las variables de efecto calendario.

En la Figura 7.29 se presenta a nivel general el ajuste del modelo MARS. Como en todos los casos anteriores, el objetivo de esta gráfica es observar el ajuste que tiene el modelo dentro de las observaciones de entrenamiento, el ajuste para las observaciones de validación y las proyecciones obtenidas. En marzo y abril de 2020 se observa la caída en el consumo de GM debido al confinamiento causado por la pandemia del COVID-19. También se ha encontrado que la tasa de crecimiento en el consumo de GM decrece ligeramente en el largo plazo.

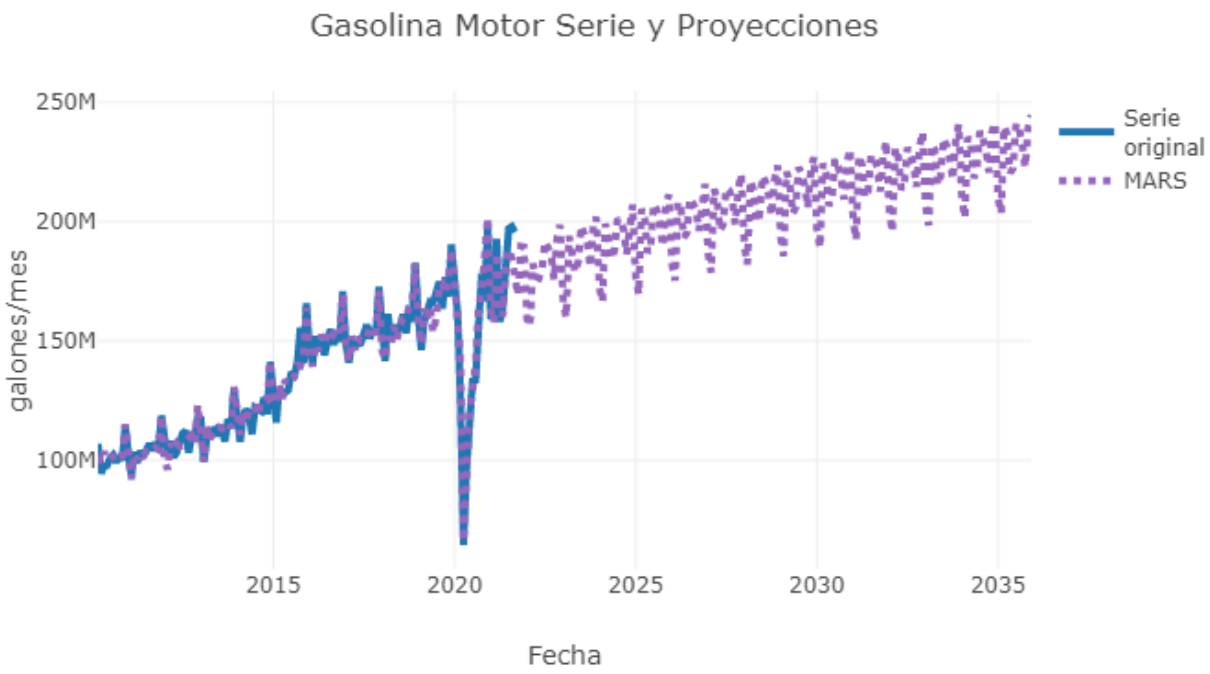

Figura 7.29: Ajuste del modelo MARS a la demanda de GM para el periodo 2010/01 - 2035/12

En la Figura 7.30 se ilustra la serie original y proyección obtenida a través del modelo MARS para los datos dentro de muestra. Allí también se observa un ajuste satisfactorio con este modelo MARS. La serie histórica de GM ha mostrado que en los meses de diciembre se presentan los máximos; mientras que los mínimos de consumo se han presentado en los meses de febrero. Como se observa en la Figura 7.30, el modelo MARS también captura tales fenómenos estacionales. Además, con apoyo de la variable de COVID-19, también ha sido posible capturar la fuerte disminución del consumo de GM en marzo y abril de 2020.

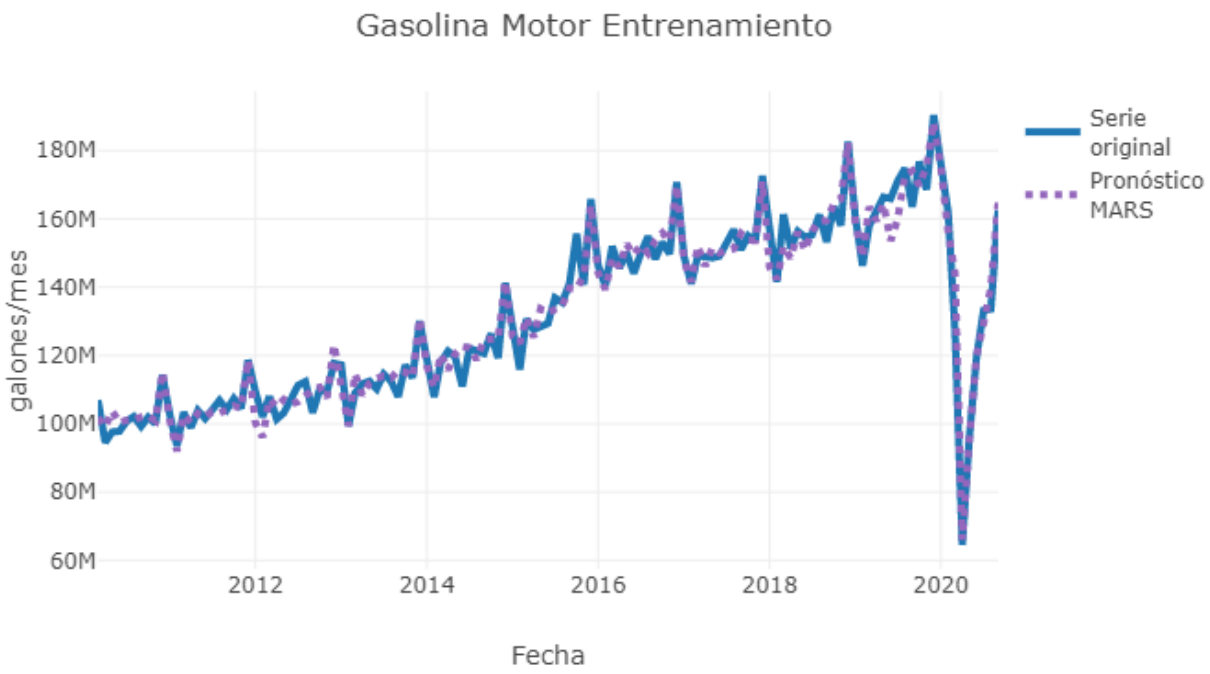

Figura 7.30: Ajuste del modelo MARS para datos de entrenamiento de la demanda GM para el periodo 2010/01 - 2020/09

En el Cuadro 7.23 se presenta el MAPE, el AIC y el BIC para este escenario de proyección de GM. El MAPE de entrenamiento obtenido es del 2.51897 %, lo que valida la precisión del ajuste de datos dentro de muestra.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
2.51897	3937.19340	3994.07714

Cuadro 7.23: Medidas de bondad de ajuste del modelo MARS para datos de entrenamiento de la demanda de GM

La Figura 7.31 presenta el desempeño del modelo MARS fuera de muestra. El MAPE de validación en este escenario es del 5.02701 % como se observa en el Cuadro 7.24. La demanda de GM es pronosticada con altos niveles de precisión para nueve de los doce meses del año de prueba.

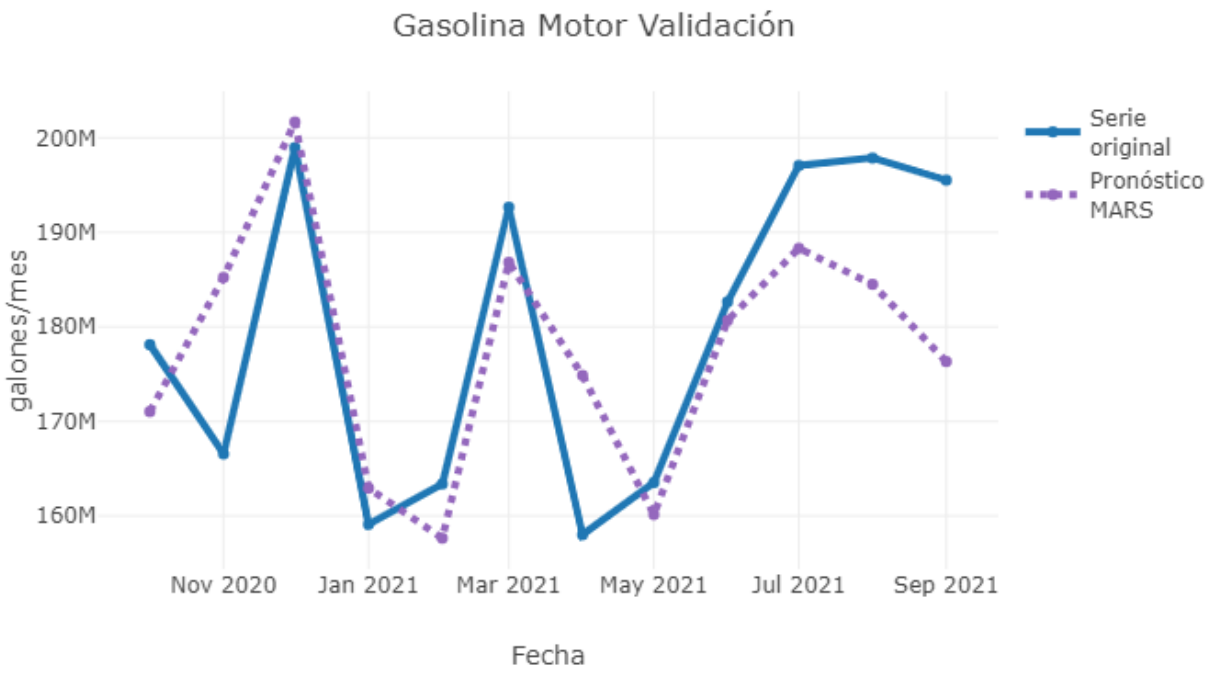

Figura 7.31: Ajuste del modelo MARS para datos de validación de la demanda GM para el periodo 2020/10 - 2021/09

Validación		
MAPE (%)	**AIC**	**BIC**
5.02701	428.77491	438.47304

Cuadro 7.24: Medidas de bondad de ajuste del modelo MARS para datos de validación de la demanda de GM

En la Figura 7.32 se presentan las proyecciones obtenidas por el modelo MARS junto con sus intervalos de confianza bootstrap con 1000 replicas. En las proyecciones obtenidas se observa que la demanda de GM registra un comportamiento estacional con demandas máximas anuales dadas en diciembre y valores mínimos en los meses de febrero como se ha presentado en la historia de la variable. Adicionalmente, también se sigue evidenciando que la variabilidad de las proyecciones es estable; de hecho es ligeramente mayor que la de la serie original como se observó en la figura Figura 7.29.

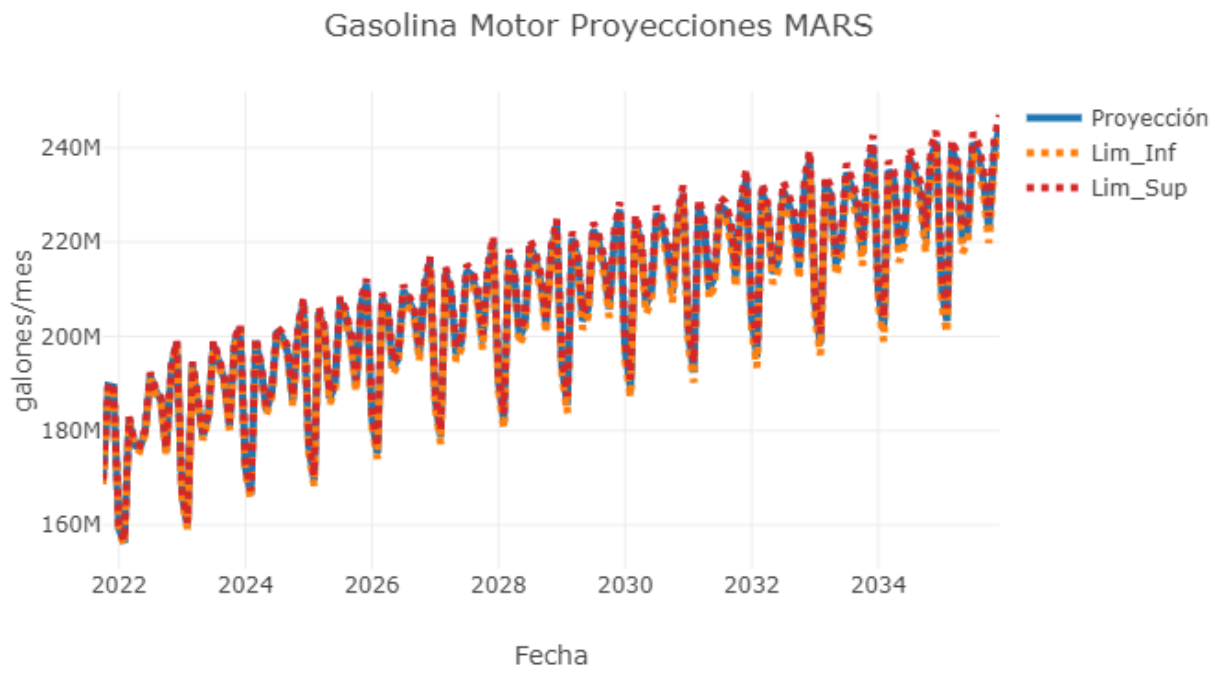

Figura 7.32: Proyecciones del modelo MARS para la demanda de GM para el periodo 2021/10 - 2035/12

Finalmente, la diferencia entre los intervalos de confianza bootstrap y el valor proyectado es pequeña como se presenta en el Cuadro 7.25. Los anchos de los intervalos de confianza de las proyecciones de consumo de GM oscilan entre 581387 gal/mes en 2020 y 4822268 gal/mes en 2035.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	170744890.33690	171035584.01428	171326277.69167
2020-11-01	184811791.76653	185229568.30497	185647344.84341
2020-12-01	201314007.55395	201670019.75202	202026031.95008
2021-01-01	162479960.80379	162937049.69276	163394138.58173
2021-02-01	156996806.87267	157645182.98164	158293559.09062
...	...	...	...
2035-08-01	236586299.41458	238722497.94638	240858696.47818
2035-09-01	229752736.60545	232551209.07714	235349681.54883
2035-10-01	219820206.15400	222656293.84287	225492381.53173
2035-11-01	236757875.25032	239259596.67736	241761318.10439
2035-12-01	242127116.10437	244538250.18821	246949384.27206

Cuadro 7.25: Encabezado proyecciones modelo GAM para la demanda de GM

7.4.2 GM: Escenario 2

En este escenario, las variables explicativas adicionales a las del escenario base son el `p_gm`, `d_gntran` y `tot_aut_gm`. Este nuevo escenario ha eliminado el precio del crudo pero sí considera la flota total de automóviles que consumen GM. Por tal razón, esta variable tiene un impacto directo sobre el consumo de GM mes a mes. El modelo que mejor ha logrado explicar la demanda de GM en este escenario es la combinación entre MARS y las redes LSTM. Las especificaciones del mejor modelo en este escenario consideran incluir los dos primeros rezagos de la variable `d_gm` los cuales capturan la historia de dos meses del consumo de GM. La red LSTM contiene tres capas ocultas. La primera, segunda y tercera capa tienen 30, 25 y 10 neuronas respectivamente. La función de activación en cada neurona es la tangente hiperbólica. El comportamiento estacional se sigue capturando a través de las variables de efecto calendario.

En la Figura 7.33 se presenta a nivel general el ajuste del modelo MARS. Como en todos los casos anteriores, el objetivo de esta gráfica es observar el ajuste que tiene el modelo dentro de las observaciones de entrenamiento, el ajuste para las observaciones de validación y las proyecciones obtenidas. En este escenario, también se ha obtenido que la tasa de crecimiento en el consumo de GM decrece ligeramente en el largo plazo.

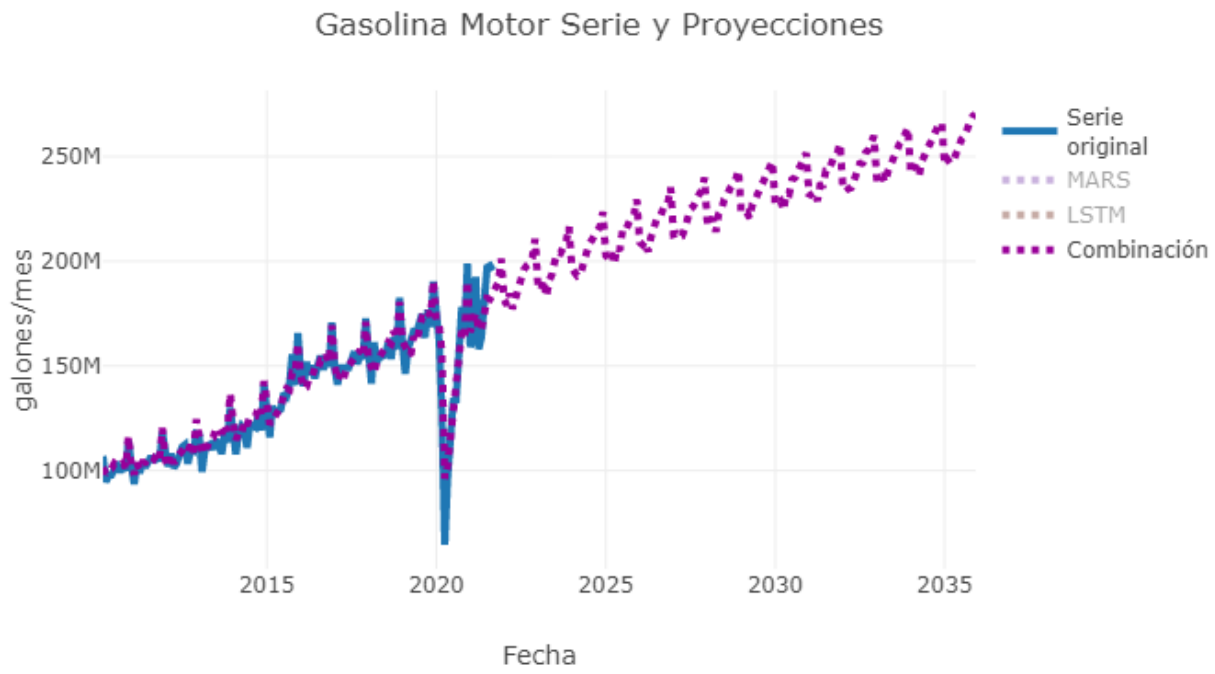

Figura 7.33: Ajuste del modelo de combinación de pronósticos a la demanda de GM para el periodo 2010/01 - 2035/12

En la Figura 7.34 se ilustra la serie original y proyección obtenida a través del modelo MARS para los datos dentro de muestra. El ajuste es satisfactorio con este modelo de combinación entre MARS y LSTM. Este modelo también captura tales fenómenos estacionales. Además, con apoyo de la variable de COVID-19, también ha sido posible capturar la fuerte

disminución del consumo de GM en marzo y abril de 2020.

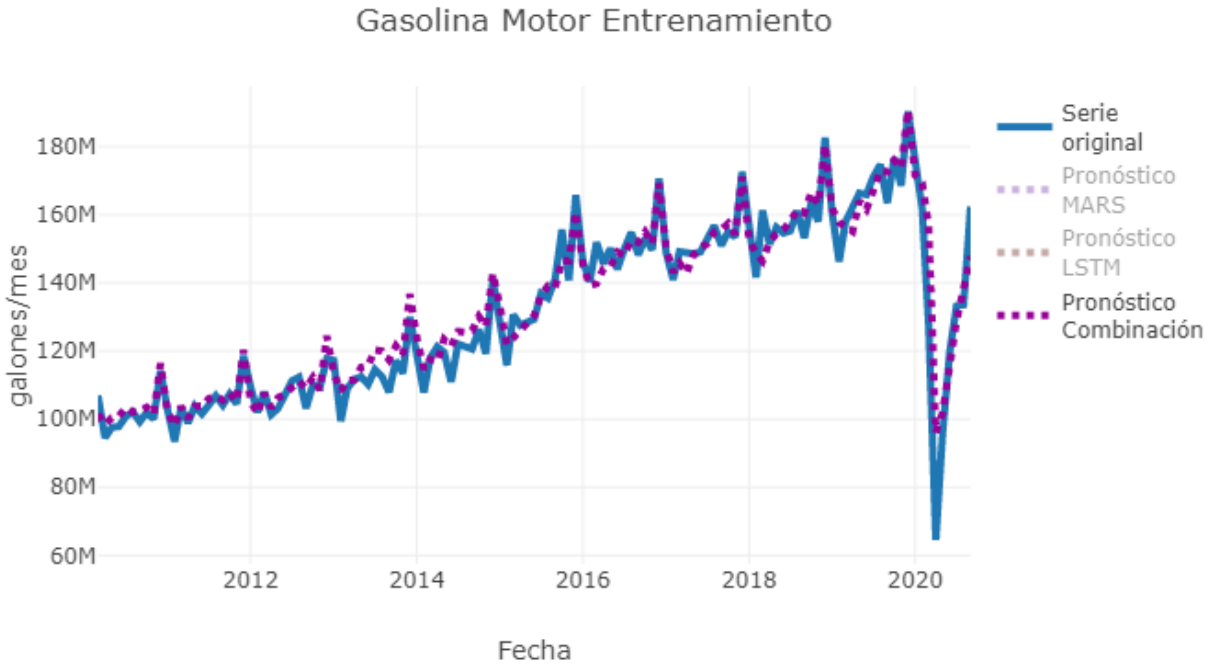

Figura 7.34: Ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda GM para el periodo 2010/01 - 2020/09

En el Cuadro 7.26 se presenta el MAPE, el AIC y el BIC para este escenario de proyección de GM. El MAPE de entrenamiento obtenido es del 3.29923 %, lo que valida la precisión del ajuste de datos dentro de muestra.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
3.29923	4005.46437	4062.34811

Cuadro 7.26: Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de entrenamiento de la demanda de GM

La Figura 7.35 presenta el desempeño del modelo de combinación MARS y LSTM fuera de muestra. El MAPE de validación en este escenario es del 5.30205 % como se observa en el Cuadro 7.27. La demanda de GM es pronosticada con altos niveles de precisión para aproximadamente seis de los doce meses del año de prueba. El comportamiento de la demanda entre noviembre 2020 y febrero 2021 proyectado es muy similar a la serie original. Sin embargo, el modelo de combinación propone una demanda menor en marzo 2021 comparada contra la serie original.

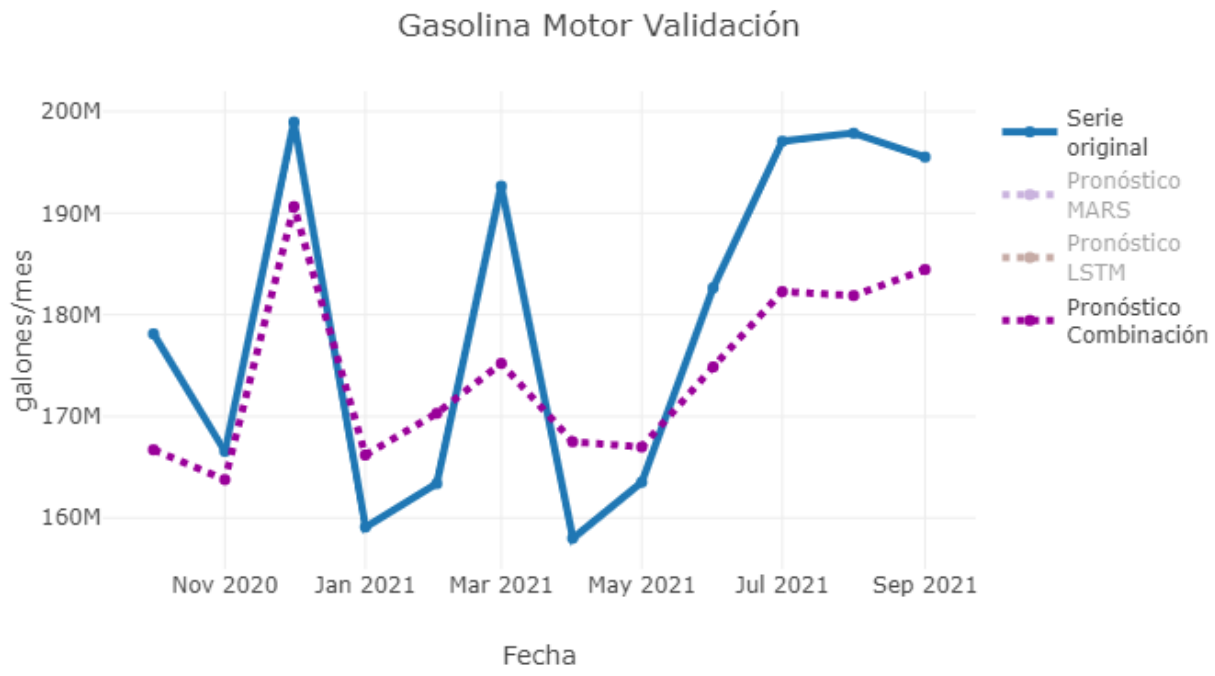

Figura 7.35: Ajuste del modelo de combinación de pronósticos para datos de validación de la demanda GM para el periodo 2020/10 - 2021/09

En la Figura 7.36 se presentan las proyecciones obtenidas por el modelo resultante de combinar MARS y redes LSTM junto con sus intervalos de confianza bootstrap con 1000 replicas. En las proyecciones obtenidas se observa que la demanda de GM registra un comportamiento estacional con demandas máximas anuales dadas en diciembre y valores mínimos en los meses de febrero como se ha presentado en la historia de la variable. En este escenario se proyecta una demanda ligeramente mayor que la del escenario 1. Una posible explicación es que la proyección de la flota de automóviles plantea un crecimiento continuo, aunque ligeramente desacelerado, hasta casi 7 millones de vehículos a 2035. El precio del galón de gasolina, según las proyecciones, se mantiene casi constante superando apenas los 10 mil pesos en 2035. Adicionalmente, también se sigue evidenciando que la variabilidad de las proyecciones es estable como se observó en la figura Figura 7.33.

Validación		
MAPE (%)	**AIC**	**BIC**
5.30205	428.41220	438.11033

Cuadro 7.27: Medidas de bondad de ajuste del modelo de combinación de pronósticos para datos de validación de la demanda de GM

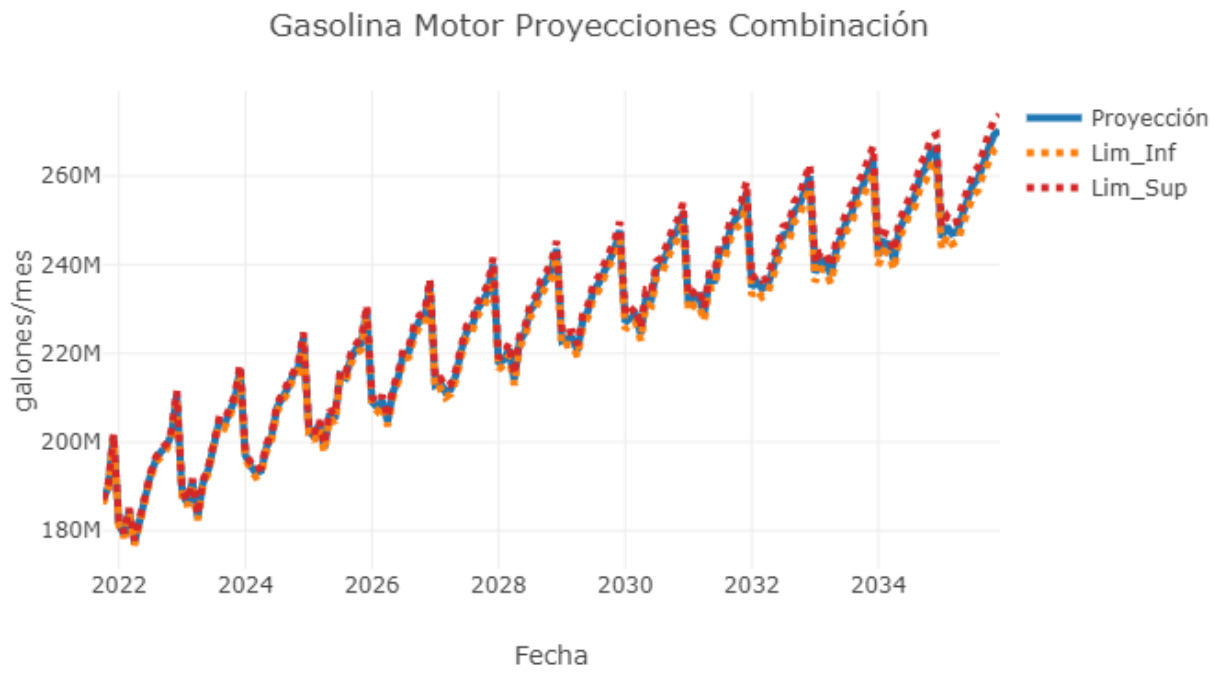

Figura 7.36: Proyecciones del modelo de combinación de pronósticos para la demanda de GM para el periodo 2021/10 - 2035/12

Finalmente, la diferencia entre los intervalos de confianza bootstrap y el valor proyectado es pequeña como se presenta en el Cuadro 7.28. Los anchos de los intervalos de confianza de las proyecciones de consumo de GM oscilan entre 714322 gal/mes en 2020 y 7001824 gal/mes en 2035.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	166355120.64393	166712281.76817	167069442.89241
2020-11-01	163380140.95432	163776016.16316	164171891.37200
2020-12-01	190020217.17666	190643824.44335	191267431.71005
2021-01-01	165813800.72003	166199676.57424	166585552.42845
2021-02-01	169828234.23930	170303699.75495	170779165.27059
...	...	...	...
2035-08-01	256810657.96921	259439669.19932	262068680.42944
2035-09-01	260889432.64418	263736881.84631	266584331.04845
2035-10-01	263207269.95294	266450801.50476	269694333.05658
2035-11-01	266046366.19773	269357211.23299	272668056.26825
2035-12-01	266725679.21700	270226591.37302	273727503.52903

Cuadro 7.28: Encabezado de validación y proyecciones del modelo de combinación de pronósticos para la demanda de ACPM/Diésell

7.5 Jet Fuel

Para el pronóstico de la demanda de Jet Fuel, se plantea el escenario base compuesto por las variables de efecto calendario y las variables macroeconómicas descritas en el Capítulo 7 tal como se hizo en los demás combustibles.

Adicionalmente como se mencionó en la introducción de la Capítulo 7, para el caso del Jet Fuel, se decide reemplazar la variable de cierres económicos (covid), por la variable (covidjet) en cada uno de los escenarios dentro de la varaible base de estimación, debido a que el efecto que tuvo la pandemia sobre el sector aeronáutico tuvo un impacto más prolongado que en otros sectores guiado en gran parte por el cierre temporal de aeropuertos que se realizó a nivel nacional e internacional como medida de contingencia para evitar la propagación del virus del COVID-19.

Respecto al número de casos y escenarios que se plantearon para la demanda del Jet Fuel, se tiene que se construyeron solo 5 casos diferentes repartidos en un total de 65 escenarios, siendo este combustible el que menos casos y escenarios registró, debido a la escasez de información que se tenía para la inclusión de variables adicionales que pudieran explicar el comportamiento de dicho combustible.

Dado esto, se incluyeron solo 3 variables adicionales a las planteadas en el escenario base, a saber, el `p_jetfuel`, el `p_crudo` y el `p_acpm`, en donde, la decisión de incluir el precio del ACPM/Diésel como variable explicativa en el modelo de Jet Fuel radicó en la práctica que se realizó en algunos lugares de Europa y el mundo durante la época de pandemia, en donde se trataron de impulsar iniciativas que buscaban mezclar el excedente sobrante de queroseno del combustible de los aviones que se encontraban inactivos en el momento, con el gasóleo de los vehículos diésel.

Debido a esta práctica, se decide incluir entre los escenarios evaluados, el precio del ACPM como un proxy de la demanda de ACPM/Diésel, con el fin de observar si efectivamente dicha variable contribuía en el nivel de la demanda del Jet Fuel durante los meses de pandemia, sin conseguir resultados relevantes.

Es de anotar que existen variables que según la revisión de la literatura si podrían contribuir en la explicación de la demanda del Jet Fuel, en donde se destacan las variables de, el número de pasajeros, el factor de carga y el número de operaciones nacionales e internacionales. Siendo dichas variables no incluidas en los análisis aquí realizados debido a que no se poseían proyecciones fidedignas de estas variables, y en consecuencia, se decide mejor omitirlas, y tal vez tratar de tenerlas en cuenta en futuras proyecciones del combustible.

Ahora bien, entre los 65 escenarios evaluados, se seleccionan 2 de ellos, debido a que éstos fueron los que mejor comportamiento presentaron, en términos de trayectoria y variabilidad en sus proyección, por lo cual fueron sugeridos por la UPME para su presentación en este reporte.

7.5.1 Jet Fuel: Escenario 1

En el primer escenario presentado para el pronóstico de la demanda de Jet Fuel, se plantea el escenario base, en cual solo se agregan los dos primeros rezagos de la variable `d_jetfuel` y el primer rezago de la variable `pib`. En dicho escenario se encuentra que el modelo que lográ ofrecer un mejor ajuste al comportamiento de la variable por fuera de los datos de entrenamiento, es decir en validación, es el modelo de redes neuronales LSTM con una sola capa oculta con un total de 10 neuronas y función de activación tangencial hiperbólica.

Con el objetivo de evidenciar cuál fue el nivel de ajuste del modelo LSTM, que fue estimado solo con las variables del escenario base, y algunos rezagos propios y del PIB, se presenta inicialmente, la Figura 7.37 en la que se resume el ajuste general del modelo LSTM dentro y fuera de muestra. junto a las proyecciones resultantes del proceso de estimación. En segundo lugar, se presenta la Figura 7.38, en donde se pretende expandir la explicación realizada sobre el ajuste que presenta el modelo respecto a los datos de entrenamiento.

En tercer lugar, se ilustra la Figura 7.39 con el fin de evaluar de forma más concreta el nivel de ajuste o desempeño predictivo del modelo LSTM. En último lugar, se presentan la Figura 7.40 las proyecciones resultantes del proceso de estimación junto a la debido interpretación de los resultados observados.

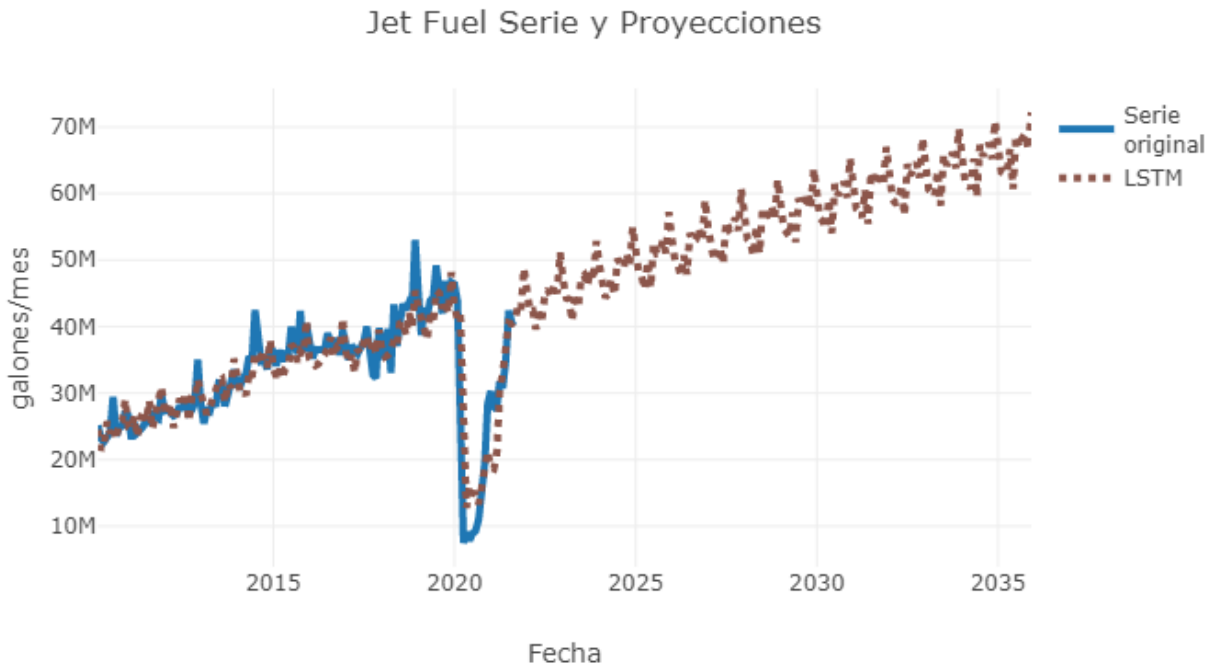

Figura 7.37: Ajuste del modelo LSTM a la demanda de Jet Fuel para el periodo 2010/01 - 2035/1

En la Figura 7.37 se ilustra el nivel de ajuste que tiene el modelo LSTM respecto a la demanda observada de Jet Fuel, evidenciándose que en términos de ajuste, el modelo de redes neuronales pareciera ofrecen un ajuste adecuado a la serie original, puesto que los valores

estimados por el modelo logran capturar el comportamiento de la tendencia de la serie, y adicionalmente ajustar de forma relativamente adecuada la caída prolongada que se observó en el año 2020 a causa de la pandemia del COVID-19.

Es de anotar, que el buen ajuste que registra el modelo ajustado en el periodo prolongado de pandemia que se registra para este combustible, se logra gracias al reemplazo de la variable explicativa `covid` que se usó en el escenario base de los demás combustibles, por la variable dummy `covidjet`, la cual se construye específicamente para capturar tiempo prologando que tardó el sector aeronáutico en recuperarse de la pandemia y retornar al nivel natural registrado antes en épocas pre-pandemia.

Respecto a las proyecciones que se presentan en la Figura 7.37, se observa que luego de la época de recuperación post-pandemia, los valores proyectados exhiben la trayectoria creciente que se venía registrando para la demanda de este combustible. Adicionalmente se logra evidenciar que las proyecciones presentan un comportamiento estacional marcado, en donde dicho análisis se ampliará con más detalle en la Figura 7.40.

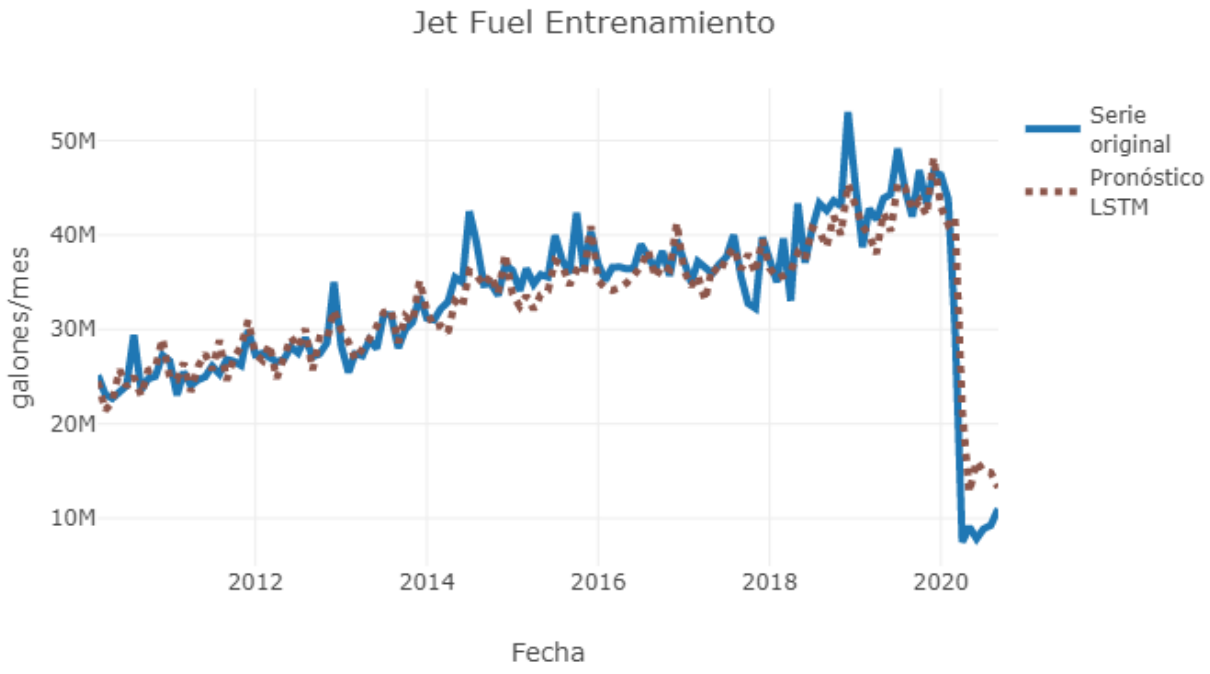

Figura 7.38: Ajuste del modelo LSTM para datos de entrenamiento de la demanda Jet Fuel para el periodo 2010/01 - 2020/09

Por su parte, en la Figura 7.38 se muestra un poco más de detalle el ajuste ofrecido por el modelo LSTM a los datos de entrenamiento que se usaron para el ajuste del modelo, en donde, se observa un detalle importante sobre la serie original, y es que, aunque no se registra un comportamiento estacional marcado en ésta, los picos en la demanda del combustible ocurren regularmente en los meses de Julio y Diciembre de cada año, siendo dichas fechas en las cuales se debería registrar mayor cantidad de operaciones nacionales e internacionales debido a la época de vacaciones que tienen los estudiantes de educación primaria, secundaria, media

y universitaria. De este punto se deriva la razón por la que la inclusión de la variable de operaciones toma importancia dentro de los análisis, y por lo cual se sugiere su inclusión en futuros análisis.

Dejando este aspecto de lado, y centrándose nuevamente en el análisis del ajuste del modelo LSTM, se evidencia en la Figura 7.38 que el modelo ajustado logra detectar la existencia del comportamiento estacional aquí expuesto, y captura de forma satisfactoria los picos que se registran en los meses de Julio y Diciembre, aunque en muchos casos la magnitud del incremento que se registra en la serie original, no sea la misma que en el modelo ajustado.

Adicional a lo anterior, se destaca el hecho de que el modelo LSTM logra capturar en su proceso de ajuste la caída generada por el cierre de aeropuertos y restricción que tuvieron los vuelos a causa del COVID-19, en donde se observa que a pesar de que la magnitud de la estimación no es igual al valor realmente registrado, se resalta el hecho que el valor estimado es bastante cercano.

Para cuantificar el nivel de ajuste que presenta el modelo LSTM, se construye al igual que en los demás combustibles, un cuadro donde se registra el MAPE, el AIC y el BIC reportados por el modelo ajustado, con el fin de cuantificar el nivel de ajuste del modelo dentro de muestra, y poder compararlo con el que se registrará en el escenario secundario evaluado. Dado lo anterior, se registra en el Cuadro 7.29 las medidas de bondad de ajuste mencionadas.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
8.81343	3818.71399	3869.90935

Cuadro 7.29: Medidas de bondad de ajuste del modelo LSTM para datos de entrenamiento de la demanda de Jet Fuel

Del Cuadro 7.29, se observa que el nivel de ajuste que registra el modelo LSTM en este escenario para datos de entrenamiento es del 8.81343 %, en donde al ser dicho valor inferior al 10 % significa que el ajuste alcanzado por el modelo respecto a los datos de entrenamiento es muy preciso. Es de anotar que esta afirmación se realiza basado en la escala de precisión del MAPE, que propone Lewis (1982) y que se presenta en la Cuadro 7.2, la cual dice que valores MAPE por debajo de 10 % están asociados a pronósticos precisos, o en este caso estimaciones precisas.

Respecto al valor del AIC y el BIC registrado en la Cuadro 7.29, no se realizará ningún tipo de análisis en su momento, ya que su propósito está más guiado respecto a la comparación de modelos que al nivel de ajuste que ofrece un modelo, y por tanto, se analizarán dicho valores previamente cuando se realice el análisis del escenario 2 de la demanda de Jet Fuel.

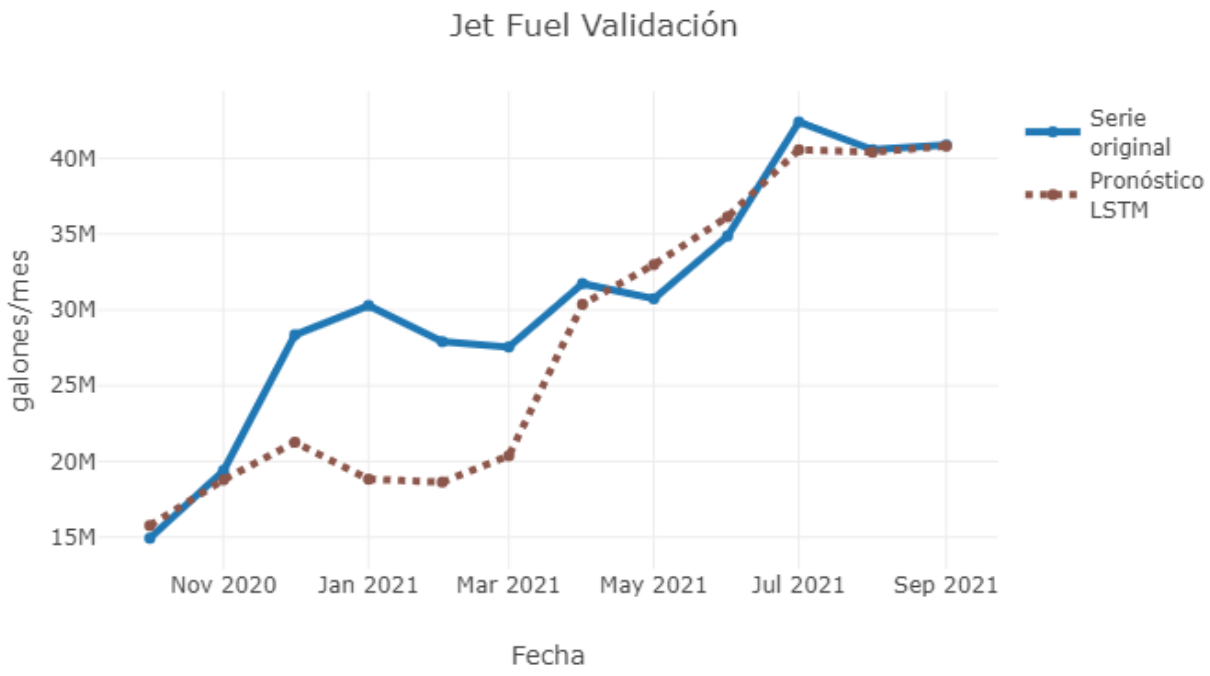

Figura 7.39: Ajuste del modelo LSTM para datos de validación de la demanda Jet Fuel para el periodo 2020/10 - 2021/09

Al analizar la Figura 7.39 que presenta el ajuste obtenido por el modelo LSTM para los datos dejados para validación del modelo y correspondientes al periodo 2020/10 - 2021/09, se encuentra hay un desfase significativo entre los valores efectivamente registrados en los primeros meses de 2021, y los valores ajustados por el modelo LSTM, en donde se registran diferencias de hasta 10 millones de galones de combustible durante el mes de enero de 2021.

Es de anotar que para el caso de Jet Fuel, durante los dos últimos meses del 2020, y los tres primeros meses del 2021, se seguía registrando en la demanda de este combustible efecto negativos generados por la pandemia del COVID-19, siendo estos meses parte de la fase de recuperación que registró el combustible. Por dicha razón es que se observa una pendiente positiva pronunciada durante el periodo de validación del modelo.

De la Figura 7.39, también es posible notar que el ajuste ofrecidpo por el modelo LSTM para meses posteriores a Abril de 2021, es muy preciso, por que se esperaría no obtener un valor MAPE muy alto, debido a que el mal ajuste que se observa durante los meses de Enero-Marzo de 2020, es compensado de cierta manera por el buen ajuste que se evidencia durante los otros meses.

Con el fin de corroborar si efectivamente el MAPE obtenido por el ajuste del modelo no es significativamente alto, se presenta el Cuadro 7.30 en el cual, además de registrar el MAPE del ajuste, se presenta también el valor de los criterios de información AIC, y BIC.

Validación		
MAPE (%)	**AIC**	**BIC**
12.58586	407.42142	416.14974

Cuadro 7.30: Medidas de bondad de ajuste del modelo LSTM para datos de validación de la demanda de Jet Fuel

En el Cuadro 7.30, se muestra como el MAPE obtenido por el modelo LSTM dentro por fuera de es igual al 12.58586 %, siendo el mayor registrado para los ajustes de los modelo para datos de validación. Sin embargo, aunque parezca un valor MAPE muy alto, Lewis (1982) establece en su escala de precisión del criterio MAPE, que valores MAPE que se encuentren entre el 10 % y el 20 % son sinónimo de ajustes buenos, no los los más precisos pero pero si lo suficientemente buenos como para explicar el comportamiento de la variable real.

Similar a lo señalado en el Cuadro 7.29, los valores AIC y BIC registrado en el Cuadro 7.30, no sirven para realizar ningún tipo de análisis en el momento, ya que su principal característica es ser empleados en el proceso de selección de modelo, tal que como se especificó en la Sección 6.12, y por tanto, serán utilizados en posteriores análisis una vez se plantee el escenario secundario para la proyección de la demanda del Jet Fuel.

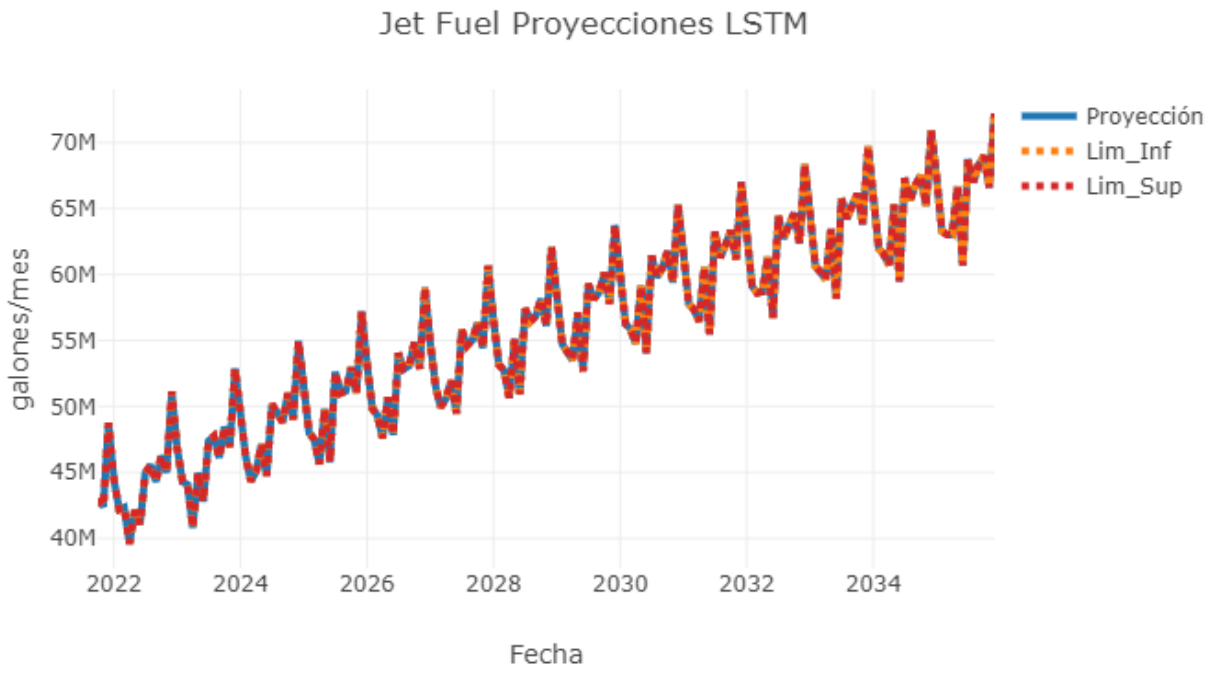

Figura 7.40: Proyecciones del modelo LSTM para la demanda de Jet Fuel para el periodo 2021/10 - 2035/12

Como punto final en el análisis de este escenario, se presenta la Figura 7.40, en la cual se exhiben los valores proyectados por el modelo LSTM ajustado en este escenario. En dichas proyecciones se logra evidenciar un comportamiento estacional marcado que es consecuente con

el comportamiento de la serie original, en la cual se presentaban picos más pronunciados en el mes de Diciembre, y un poco menos pronunciados durante el mes de Julio de cada año.

En adición, como se explicó antes, los intervalos de confianza bootstrap generados por el modelo LSTM suelen ser muy precisos debido a la robustez que exhibe el modelo LSTM, en donde el efecto de errores aleatorios en las observaciones originales no genera una desviación significativa en las réplicas, por lo cual se encuentra que la amplitud del intervalo suele ser bastante precisa, más no igual a los valores efectivamente proyectados.

Con el fin de soportar la afirmación aquí realizada, se presenta el Cuadro 7.31, en el cual se presenta un encabezado de las estimaciones realizadas por el modelo para los primeros meses ajustados respecto a los datos de validación, y los últimos cinco meses asociados a las proyecciones, para de esta manera ver la diferencia que hay entre los valores estimados y los intervalos de confianza bootstrap generados.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	15759672.46021	15763758.27516	15767844.09010
2020-11-01	18775987.17859	18784536.47165	18793085.76471
2020-12-01	21260403.29998	21262650.36343	21264897.42688
2021-01-01	18819720.86428	18825694.38105	18831667.89782
2021-02-01	18623911.57786	18633937.12680	18643962.67574
...	...	...	...
2035-08-01	66971846.13075	66975926.53896	66980006.94718
2035-09-01	68302394.65026	68303438.55895	68304482.46764
2035-10-01	68977808.86848	68980234.29562	68982659.72275
2035-11-01	66583890.08326	66588322.35969	66592754.63613
2035-12-01	72208804.08563	72209809.58897	72210815.09231

Cuadro 7.31: Encabezado proyecciones modelo LSTM para la demanda de Jet Fuel

7.5.2 Jet Fuel: Escenario 2

Para el segundo escenario de Jet Fuel, se decide incluir el `p_jetfuel` como variable explicativa adicional a las usadas en el escenario base, para proyectar la demanda del combustible. La razón de incluir esta variable dentro de las estimaciones, se debe a la relación económica que existe entre el precio y la demanda de un bien, en donde se tiene que al ser la demanda de Jet Fuel un bien elástico respecto al precio, a causa de la existencia de combustibles alternativos, entonces se tendrá que esta demanda se verá afectada por cambios en el precio del combustible.

Al realizar la estimación de los diferentes escenarios en los que se incluye el `p_jetfuel` como variable adicional a las variables del escenario base, con el fin de encontrar el modelo que ofrece el mejor ajuste para este caso, se encuentra que el modelo LSTM es el que ofrece el

mejor ajuste a la demanda de Jet fuel, cuando se incluye adicionalmente el primer rezago de la variable `d_jetfuel`, se emplea la función de activación ReLu, y se usa una capa oculta con un total de 5 neuronas.

Ahora bien, con el propósito de ser consistente con la estructura planteada en los demás combustibles, el análisis de este escenario se divide en cuatro partes, en la primera parte, se presenta la Figura 7.41, en la cual se ilustra el comportamiento general que presenta la serie original junto a los ajustes y proyecciones realizadas por el modelo LSTM.

En la segunda parte se plantea la Figura 7.42 con el fin de presentar el nivel ajuste que tiene el modelo LSTM dentro de muestra. Por su parte, en la tercera parte se plantea la Figura 7.43 en la cual se realiza un análisis similar a la presentara en la segunda parte, pero con la diferencia de que en lugar de analizar el ajuste dentro de muestra, se realiza el análisis del ajuste del modelo LSTM por fuera de muestra. Finalmente, en la cuarta parte se plantea la Figura 7.44, en el propósito de presentar las proyecciones finales generadas por el modelo LSTM.

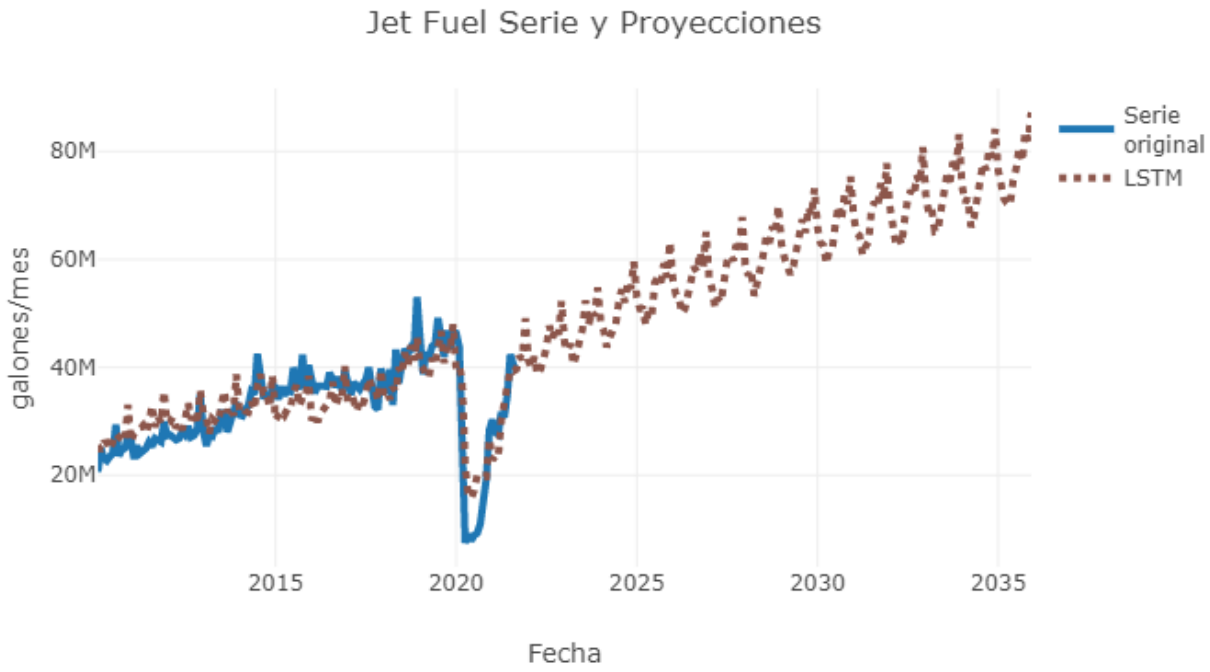

Figura 7.41: Ajuste del modelo LSTM a la demanda de Jet Fuel para el periodo 2010/01 - 2035/1

De la Figura 7.41 se observa inicialmente que la inclusión del `p_jetfuel` deteriora el ajuste del modelo dentro de muestra, ya que se evidencia que las estimaciones generadas por el modelo en los primeros años de 2010 se encuentran muy por encima de la serie original.

En segundo lugar, se observa que a pesar de que el ajuste obtenido por el modelo LSTM dentro del conjunto de datos de entrenamiento para este escenario, es más malo que el presentado en el escenario 1, no se evidencia que haya diferencias significativas entre escenarios en términos de ajuste para datos de validación.

En tercer lugar, de la Figura 7.41 se observa que las proyecciones realizadas por el modelo LSTM al incluir la variable `p_jetfuel`, provoca que la pendiente de las proyecciones sea mayor, que en aquellos ajustes en los cuales se omite dicha variable.

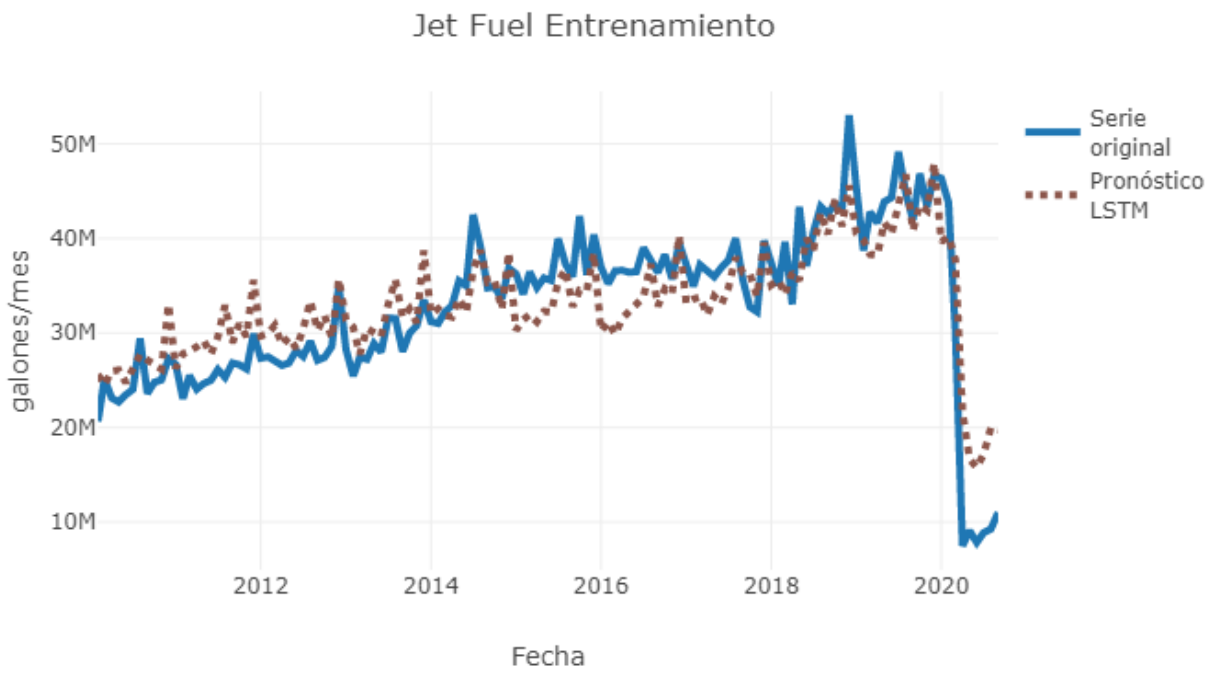

Figura 7.42: Ajuste del modelo LSTM para datos de entrenamiento de la demanda Jet Fuel para el periodo 2010/01 - 2020/09

Por su parte, al analizar el ajuste del modelo LSTM presentado en la Figura 7.42, respecto al conjunto de observaciones de entrenamiento, se corrobora lo ya mencionado sobre que las estimaciones del modelo se encuentran por encima de las observaciones registradas entre el año 2010 y 2011.

Adicionalmente al comparar el ajuste aquí planteado respecto al realizado en el escenario 1, se evidencia que incluir la variable `p_jetfuel` genera el ajuste del modelo respecto a la caída que se registra en la demanda a causa de la pandemia del COVID-19, sea menos precisa que la registrada en el escenario 1.

Para cuantificar si efectivamente la inclusión de la variable `p_jetfuel` dentro del modelo afectó significativamente el nivel de ajuste para datos dentro de muestra, se presenta el Cuadro 7.32, en el cual se expone el valor del MAPE, el AIC y el BIC obtenido por el modelo LSTM.

Entrenamiento		
MAPE (%)	**AIC**	**BIC**
13.81820	3929.22287	3977.70739

Cuadro 7.32: Medidas de bondad de ajuste del modelo LSTM para datos de entrenamiento de la demanda de Jet Fuel

Al comparar las tres medidas de bondad de ajuste presentadas en el Cuadro 7.32, respecto a las presentadas en el Cuadro 7.29 asociado al escenario 1, se observa que efectivamente, la inclusión del `p_fueloil` afecta negativamente el nivel de ajuste del modelo para datos dentro de muestra, puesto que el MAPE se incrementa desde un 8.81343 % en el escenario 1, hasta un 13.81820 % en el escenario 2. En el caso del AIC, este pasa de 3818.71399 en el escenario 1, a 3929.22287 en el escenario 2, mientras que, el BIC pasa de 3869.90935 en el escenario 1 a 3977.70739 en el escenario 2 en el cual se incluye la variable `p_fueloil`.

Es de anotar que al evidenciarse mayores valores para los tres estadísticos dentro del escenario base, se concluirá que la inclusión de la variable `p_fueloil` dentro del proceso de ajuste del modelo LSTM, afecta negativamente el nivel de ajuste resultante para datos entro de muestra.

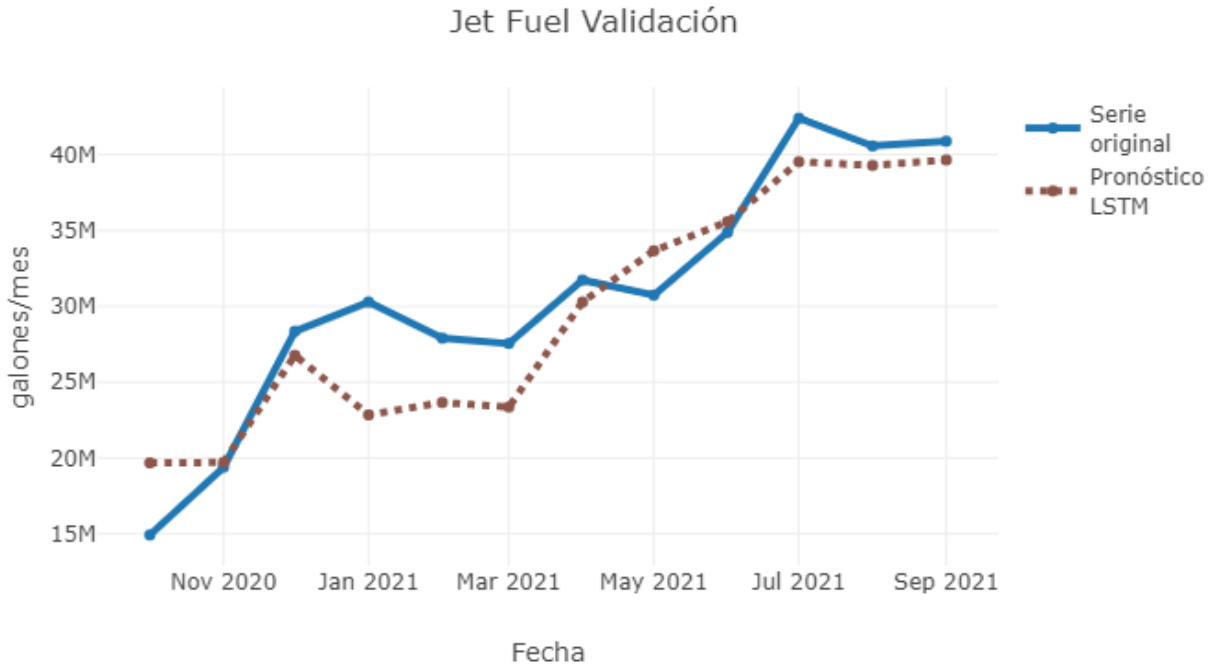

Figura 7.43: Ajuste del modelo LSTM para datos de validación de la demanda Jet Fuel para el periodo 2020/10 - 2021/09

Ahora bien, al analizar el ajuste del modelo LSTM presentado en la Figura 7.43, respecto al conjunto de observaciones que se dejaron para validación, se observa que la inclusión del `p_fueloil` dentro del modelo, contribuye en la mejora del desempeño predictivo que tiene el

modelo LSTM, puesto que se observa que las estimaciones realizadas por el modelo logran capturar la tendencia creciente que presenta la serie original asociada a la fase de recuperación pos-pandemia, además de que logra capturar el comportamiento de algunos picos y valles que exhibe la serie original.

Al comparar los resultados aquí expuestos respecto a los presentados en el escenario 1, se observa que incluir la variable `p_fueloil` dentro de la estimación del modelo LSTM genera que los ajustes dados por el modelo mejores, respecto al escenario 1 en donde no se considera dicha variable.

Esta afirmación puede ser corroborada en el Cuadro 7.33, en el cual se presenta el MAPE, el AIC y el BIC registrado por el modelo LSTM ajustado en el escenario aquí estudiado.

Validación		
MAPE (%)	**AIC**	**BIC**
10.25512	394.94022	403.18363

Cuadro 7.33: Medidas de bondad de ajuste del modelo LSTM para datos de validación de la demanda de Jet Fuel

Del cuadro Cuadro 7.33, se observa que efectivamente el desempeño predictivo del modelo LSTM mejora cuando se incluye la variable `p_fueloil` dentro de sus estimaciones, puesto que, tanto el MAPE como los criterios de información AIC y BIC calculados en este escenario, son inferiores a los presentados en el escenario 1, en el cual no se considera dicha variable.

Por tanto, se puede concluir que el desempeño predictivo del modelo LSTM para datos por fuera de muestra es mejor cuando se incluye la variable `p_fueloil` que cuando ésta es omitida dentro del proceso de ajuste del modelo.

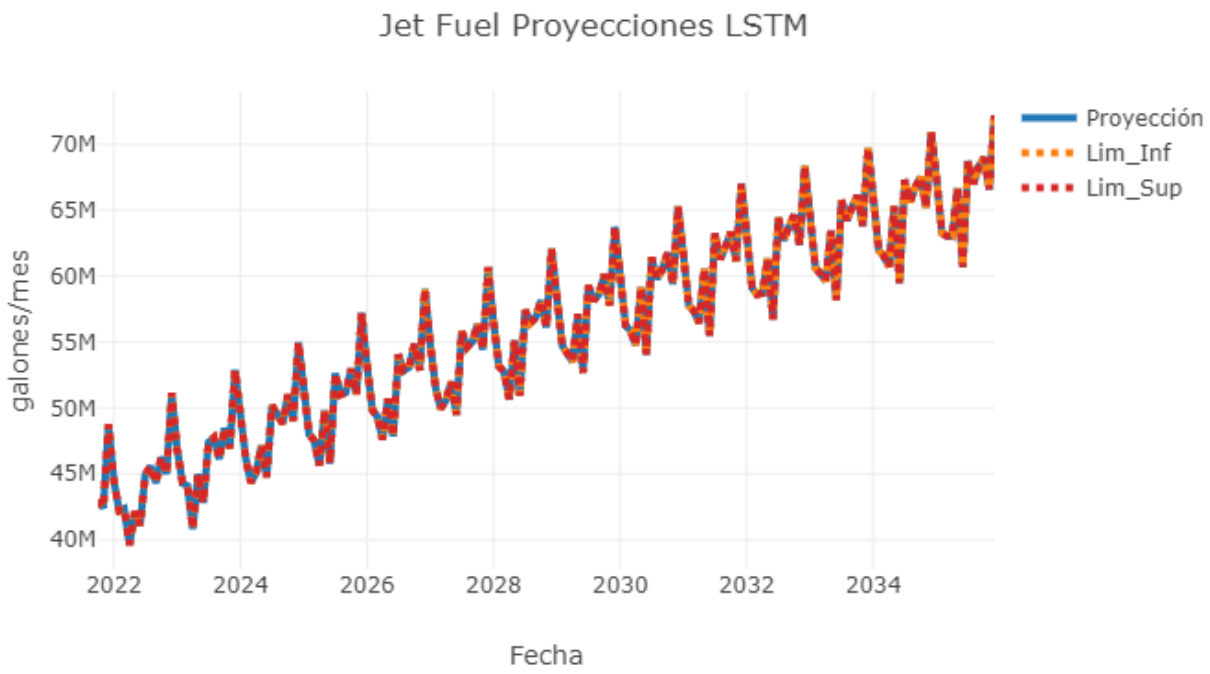

Figura 7.44: Proyecciones del modelo LSTM para la demanda de Jet Fuel para el periodo 2021/10 - 2035/12

Finalmente se presenta la Figura 7.44, en la cual se exhiben las proyecciones obtenidas por el modelo ajustado en este escenario, en donde se observa que dichas proyecciones poseen el mismo comportamiento creciente que el presentado por las observaciones de entrenamiento y validación.

Adicionalmente, se observa que las proyecciones de este escenario, también exhiben un comportamiento estacional con pico pronunciado en el mes de Diciembre, valle en el mes de Febrero y un comportamiento relativamente estable en el resto del año.

Es de anotar que al igual que en el escenario 1, en este escenario se tiene que los intervalos de confianza bootstrap generados por el modelo LSTM, son demasiado precisos, debido a la robustez que caracteriza a este modelo. Por tanto, con el fin de cuantificar la cuantía de la diferencia que hay entre los valores proyectados y los intervalos de confianza, se presenta el Cuadro 7.34, en el cual se registran las cinco primeras estimaciones del modelo LSTM para los datos de validación, y las últimas cinco proyecciones que genera dicho modelo, junto a sus correspondientes intervalos de confianza.

Encabezado Proyección			
Fechas	**Lim_Inf**	**Proyección**	**Lim_Sup**
2020-10-01	19677910.57259	19682084.87560	19686259.17860
2020-11-01	19706483.57825	19713089.84485	19719696.11144
2020-12-01	26763654.97173	26770324.25307	26776993.53441
2021-01-01	22842667.50725	22845895.71185	22849123.91645
2021-02-01	23660494.08912	23661212.42427	23661930.75942
...	...	...	...
2035-08-01	80681924.54311	80688475.97151	80695027.39991
2035-09-01	78118465.47743	78123012.19371	78127558.90998
2035-10-01	83745379.39146	83749373.37701	83753367.36257
2035-11-01	81974661.43353	81974711.35410	81974761.27466
2035-12-01	87255705.30508	87257610.75024	87259516.19540

Cuadro 7.34: Encabezado proyecciones modelo LSTM para la demanda de Jet Fuel

Luego de realizar el análisis de los resultados presentados en la Subsección 7.5.1 y Subsección 7.5.2, se concluye que la inclusión de la variable `p_jetfuel` contribuye positivamente en las proyecciones generadas por el modelo ajustado, puesto que, a pesar de que se reduce el nivel de ajuste dentro de muestra del modelo, se mejora en contra parte el nivel de ajuste del modelo por fuera de muestra, lo cual tiene un mayor peso en los modelos de pronósticos, en donde se desea que las proyecciones realizadas sean lo más certeras posibles.

7.6 Conclusiones y recomendaciones

Entre las principales conclusiones que se desprenden del desarrollo del capítulo de resultados de combustibles líquidos se encuentra que:

- En este trabajo se ha considerado una diversidad de modelos para la proyección de demandas de combustibles líquidos y GLP. Se han ajustado modelos estadísticos paramétricos (MLR - LASSO), semiparamétricos (GAM - MARS) y no paramétricos (LSTM). Los modelos paramétricos han ofrecido ajustes más limitados que los mostrados por modelos semiparamétricos y no paramétricos, o la combinación de éstos (a través de la metodología descrita en la Sección 6.7). Luego de ajustar y re-ajustar múltiples modelos, la proporción de escenarios en los cuales los modelos paramétricos mostraron resultados superiores a los modelos semiparamétricos y no paramétricos fue significativamente baja.
- Para las proyecciones de demandas de combustibles líquidos y GLP se ha considerado que los últimos doce meses de las series suministradas sean usados para la evaluación del desempeño predictivo de los modelos, con el fin de incluir en parte el efecto que tuvo la pandemia del COVID-19, sobre la demanda de los combustibles líquidos y GLP. De

esta manera, se validó que incluir la variable explicativa `covid` o `covidjet` dentro del escenario base de estimación de los modelos, fue fundamental para lograr capturar de forma adecuada el impacto que tuvo la pandemia del COVID-19 dentro de la demanda de los diferentes combustibles líquidos y GLP. Así también, fue posible capturar adecuadamente la fase de recuperación, que se registra posteriormente fuera de muestra durante los últimos meses de 2020 y los primeros meses de 2021, donde las demandas retornan a sus niveles históricos de crecimiento.

- En nuestros ejercicios de proyección a 15 años, no se encontró una mejora en los resultados cuando se consideran las flotas de vehículos para modelar la demanda de combustibles, por lo menos en cuanto al proceso de validación se refiere. Sin embargo, los precios propios de los combustibles, en varios casos, resultaron ser importantes para especificar el mejor modelo de proyección en cuanto a las métricas de validación. En el caso de GM, uno de los modelos ilustrados con buen desempeño fuera de muestra si mostró una relación importante con la flota de vehículos.
- El fenómeno del Niño en Colombia, dada la alta dependencia de la hidrología en el sector eléctrico, trae repercusiones sobre la demanda de combustibles que también son usados por las centrales termoeléctricas. En particular, considerar el fenómeno del Niño, a través de la variable ONI, fue útil para modelar la demanda del fuel oil, puesto que, el cambio en la demanda de este combustible fue notorio durante el fenómeno meteorológico ocurrido entre el 2015 y 2016.
- La metodología de optimización de hiperparámetros implementada para realizar el ajuste de los modelos estadísticos, junto con la combinación de diferentes variables explicativas fue crucial para encontrar aquellos modelos de proyección de combustibles que ofrecieran los mejores ajustes tanto dentro de muestra (datos de entrenamiento) como por fuera de ésta (datos de validación). Sin embargo, es importante anotar que, no siempre los modelos de proyección que ofrecen las mejores medidas de bondad de ajuste son los más adecuados para explicar el comportamiento de la demanda de un combustible. Además del nivel de ajuste que ofrezca un modelo a una variable particular, existen otros criterios a tener en cuenta al momento de tomar una decisión sobre las proyecciones más adecuadas, criterios tales como la tendencia, la variabilidad, las variables explicativas seleccionadas, y el criterio experto, juegan un papel importante en dicha decisión.
- Buscar el mejor modelo en términos de MAPE fuera de muestra ha sido la estrategia más eficiente para clasificar el desempeño predictivo de los modelos. Sin embargo, es esencial analizar las características de las proyecciones de los modelos para determinar si éstas son coherentes o no con el comportamiento esperado. Las proyecciones presentadas en este documento fueron analizadas en conjunto con la UPME; de manera que el MAPE, el conocimiento experto y las variables explicativas fueran los criterios relevantes para la selección de los mejores modelos y proyecciones.

Bibliografía

Abdullahi, A. B. (2014). «Modeling Petroleum Product Demand in Nigeria Using Structural Time Series Model (STSM) Approach». En: *International Journal of Energy Economics and Policy* 4.3, págs. 427-441. URL: https://ideas.repec.org/a/eco/journ2/2014-03-12.html.

Abu-Rayash, A. y Dincer, I. (2020). «Analysis of the electricity demand trends amidst the COVID-19 coronavirus pandemic». En: *Energy Research & Social Science* 68, pág. 101682.

Ackah, I. y Frank, A. (2013). «Modelling gasoline demand in ghana: a structural time series analysis». En: *International Journal of Energy Economics and Policy* 4.1, págs. 76-82.

Adom, P. K., Amakye, K., Barnor, C., Quartey, G. y Bekoe, W. (2016). «Shift in demand elasticities, road energy forecast and the persistence profile of shocks». En: *Economic modelling* 55, págs. 189-206.

Afkhami, M., Ghoddusi, H. y Rafizadeh, N. (2021). «Google Search Explains Your Gasoline Consumption!» En: *Energy Economics* 99.C. DOI: 10.1016/j.eneco.2021.1053. URL: https://ideas.repec.org/a/eee/eneeco/v99y2021ics0140988321002103.html.

Algunaibet, I. M. y Matar, W. (2018). «The responsiveness of fuel demand to gasoline price change in passenger transport: a case study of Saudi Arabia». En: *Energy Efficiency* 11.6, págs. 1341-1358.

Alonso Cifuentes, J. C., Díaz, J. G., Estrada, D., Figueroa, C. A., Tamura, G. et al. (2019). «Empleando modelos jerárquicos para encontrar el mejor modelo para pronosticar los galones de gasolina corriente demandados en Bogotá (Colombia)». En.

Anagnostis, A., Papageorgiou, E. y Bochtis, D. (2020). «Application of Artificial Neural Networks for Natural Gas Consumption Forecasting». En: *Sustainability* 12.16. ISSN: 2071-1050. DOI: 10.3390/su12166409. URL: https://www.mdpi.com/2071-1050/12/16/6409.

Anđelković, A. y Bajatović, D. (2020). «Integration of weather forecast and artificial intelligence for a short-term city-scale natural gas consumption prediction». En: *Journal of Cleaner Production* 266, pág. 122096.

Angelopoulos, D., Siskos, Y. y Psarras, J. (2019). «Disaggregating time series on multiple criteria for robust forecasting: The case of long-term electricity demand in Greece». En: *European Journal of Operational Research* 275.1, págs. 252-265.

ANH (2020). «Histórico de Reservas de Petróleo 2007 - 2020». En: Agencia Nacional Hidrocarburos, págs. 1-2.

Atalla, T. N., Gasim, A. A. y Hunt, L. C. (2018). «Gasoline demand, pricing policy, and social welfare in Saudi Arabia: A quantitative analysis». En: *Energy policy* 114, págs. 123-133.

Atems, B. (2021). «The response of the U.S. aviation industry to demand and supply shocks in the oil and jet fuel markets». En: *Transportation Research Interdisciplinary Perspectives* 11, pág. 100452. ISSN: 2590-1982. DOI: https://doi.org/10.1016/j.trip.2021.100452. URL: https://www.sciencedirect.com/science/article/pii/S2590198221001573.

Azadeh, A., Arab, R. y Behfard, S. (2010). «An adaptive intelligent algorithm for forecasting long term gasoline demand estimation: The cases of USA, Canada, Japan, Kuwait and Iran». En: *Expert Systems with Applications* 37.12, págs. 7427-7437.

Azadeh, A., Boskabadi, A. y Pashapour, S. (2015). «A unique support vector regression for improved modelling and forecasting of short-term gasoline consumption in railway systems». En: *International Journal of Services and Operations Management* 21.2, págs. 217-237.

Bates, J. M. y Granger, C. W. J. (1969). «The Combination of Forecasts». En: *OR* 20.4, págs. 451-468. ISSN: 14732858.

Baumann, S. y Klingauf, U. (2020). «Modeling of aircraft fuel consumption using machine learning algorithms». En: *CEAS Aeronautical Journal* 11.1, págs. 277-287.

Bhattacharyya, S. C. y Blake, A. (abr. de 2009). «Domestic demand for petroleum products in MENA countries». En: *Energy Policy* 37.4, págs. 1552-1560. DOI: 10.1016/j.energy.2007.02.008. URL: https://ideas.repec.org/a/eee/enepol/v37y2009i4p1552-1560.html.

Bichpuriya, Y. K., Soman, S. y Subramanyam, A. (2016). «Combining forecasts in short term load forecasting: empirical analysis and identification of robust forecaster». En: *Sādhanā* 41.10, págs. 1123-1133.

Brockwell, P., Brockwell, P., Davis, R. y Davis, R. (2016). *Introduction to time series and forecasting*. Springer.

Burnham, K. P. y Anderson, D. R. (2002). «A practical information-theoretic approach». En: *Model selection and multimodel inference* 2.

Carbonell, J. F. y Semerena, R. E. (2014). «Demanda de gasolina en la zona metropolitana del Valle de México: análisis empírico de la reducción del subsidio». En: *Revista de Economía del Rosario* 17.1, págs. 89-117.

Cervantes, B. (2018). *Pronostico del crecimiento de demanda de energía eléctrica en el área caribe colombiana para proyectar la generación por seguridad de 2018 a 2032*. Universidad de la Costa, Barranquilla, Colombia.

Cervero, R. (1985). «Short-run forecasting of highway gasoline consumption in the United States». En: *Transportation Research Part A: General* 19.4, págs. 305-313.

Chai, J., Wang, S., Wang, S. y Guo, J. (2012). «Demand forecast of petroleum product consumption in the Chinese transportation industry». En: *Energies* 5.3, págs. 577-598.

Chai, J., Zhang, Z., Wang, S., Lai, K. y Liu, J. (2014). «Aviation fuel demand development in China». En: *Energy economics* 46, págs. 224-235.

Chernick, M. R. y LaBudde, R. A. (2014). *An introduction to bootstrap methods with applications to R*. John Wiley & Sons.

Chèze, B., Gastineau, P. y Chevallier, J. (2011). «Forecasting world and regional aviation jet fuel demands to the mid-term (2025)». En: *Energy Policy* 39.9, págs. 5147-5158.

Chow, G. y Lin, A. (1971). «Best linear unbiased distribution and extrapolation of economics times series by related series». En: *The Review of Economics and Statistics* 53, págs. 471-476.

Concentra (2021). *Pronosticos de demanda de gas natural*.

Considine, T. J. y Clemente, F. A. (2007). «Gas-Market Forecast: Betting on Bad Numbers». En: *PUBLIC UTILITIES FORTNIGHTLY*.

Correia, A., Lopes, C., Silva, E. C. e, Monteiro, M. y Lopes, R. B. (2020). «A multi-model methodology for forecasting sales and returns of liquefied petroleum gas cylinders». En: *Neural Computing and Applications* 32.16, págs. 12643-12669.

Dahl, C. A. (2012). «Measuring global gasoline and diesel price and income elasticities». En: *Energy Policy* 41, págs. 2-13.

DHL (2020). *Delivering pandemic resilience. How to secure stable supply chains for vaccines and medical goods during the COVID-19 crisis and future health emergencies*. Accessed: 2021-09-10.

Dincer, I. (2018). *Comprehensive energy systems*. Elsevier.

Ecopetrol (2021). *Generación de escenarios de evolución de consumo energético en Colombia*.

Ertuğrul, H. M., Güngör, B. O. y Soytaş, U. (2020). «The effect of the COVID-19 outbreak on the Turkish diesel consumption volatility dynamics». En: *Energy Research Letters* 1.3, pág. 17496.

Al-Fattah, S. M. (2020). «A new artificial intelligence GANNATS model predicts gasoline demand of Saudi Arabia». En: *Journal of Petroleum Science and Engineering* 194, pág. 107528. ISSN: 0920-4105. DOI: https://doi.org/10.1016/j.petrol.2020.107528. URL: https://www.sciencedirect.com/science/article/pii/S0920410520305994.

Forouzanfar, M., Doustmohammadi, A., Menhaj, M. y Hasanzadeh, S. (2010). «Modeling and estimation of the natural gas consumption for residential and commercial sectors in Iran». En: *Applied Energy* 87.1, págs. 268-274.

Franco, C., Velásquez, J. y Olaya, Y. (2008). «Caracterización de la demanda mensual de electricidad en Colombia usando un modelo de componentes no observables». En: *Cuadernos de Administración* 21.36.

Franses, P. (1991). «Seasonality, non-stationarity and the forecasting of monthly time series». En: *International Journal of forecasting* 7.2, págs. 199-208.

Galindo, L. M., Samaniego, J., Alatorre, J. E., Ferrer, J. y Reyes, O. (2015). «Meta-análisis de las elasticidades ingreso y precio de la demanda de gasolina: implicaciones de política pública para América Latina». En: *Revista CEPAL*.

García, J. J., Pérez, D., Orrego, M., Castaño, J. M. et al. (2016). «Un modelo Casi Ideal de Demanda de Combustibles para la Industria de Transporte». En.

Gareth, J., Daniela, W., Trevor, H. y Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Gareth, J., Daniela, W., Trevor, H. y Robert, T. (2021). *An introduction to statistical learning: with applications in R. Second Edition*. Spinger.

Gaweł, B. y Paliński, A. (2021). «Long-Term Natural Gas Consumption Forecasting Based on Analog Method and Fuzzy Decision Tree». En: *Energies* 14.16, pág. 4905.

Gesto, N. (2015). «Estimación de la demanda mensual del Gas Licuado de Petróleo para el Sector Residencial de Uruguay mediante modelos de Vectores Autorregresivos Cointegrados.»

Gesto, N. (2016). «Estimación de la demanda mensual del Gas Licuado de Petróleo para el Sector Residencial de Uruguay mediante modelos de Vectores Autorregresivos Cointegrados.» En.

Gómez-Villalva, E. y Ramos, A. (2004). «An algorithm for the mid-term forecast and scenario generation of natural gas and fuel oil prices». En: *IEEE Transactions on Power Systems*, págs. 00124-2004.

Goodfellow, I., Bengio, Y. y Courville, A. (2016). «Sequence modeling: recurrent and recursive nets». En: *Deep learning*, págs. 367-415.

Graves, A., Mohamed, A.-r. e Hinton, G. (2013). «Speech recognition with deep recurrent neural networks». En: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, págs. 6645-6649.

Greene, W. H. (2000). *Econometric analysis*. Fourth Edition. New Jersey: Prentice Hall.

Grimaldo, J., Mendoza, M. y Reyes, W. (2016). «Modelo para pronosticar la demanda de energía eléctrica utilizando los producto interno brutos sectoriales: Caso de Colombia». En: 38.22.

Güngör, B. O., Ertuğrul, H. M. y Soytaş, U. (2021). «Impact of Covid-19 outbreak on Turkish gasoline consumption». En: *Technological Forecasting and Social Change* 166.C. DOI: 10.1016/j.techfore.2021.1. URL: https://ideas.repec.org/a/eee/tefoso/v166y2021ics004016252100069x.html.

Gutiérrez, R., Nafidi, A. y Sánchez, R. (2005). «Forecasting total natural-gas consumption in Spain by using the stochastic Gompertz innovation diffusion model». En: *Applied Energy* 80.2, págs. 115-124.

Harris, R. y Sollis, R. (2003). *Applied time series modelling and forecasting*. Wiley.

Hastie, T. y Tibshirani, R. (2017). *Generalized additive models*. Routledge.

Hastie, T., Tibshirani, R. y Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

He, Y., Jiao, J., Chen, Q., Ge, S., Chang, Y. y Xu, Y. (2017). «Urban long term electricity demand forecast method based on system dynamics of the new economic normal: the case of Tianjin». En: *Energy* 133, págs. 9-22.

Hochreiter, S. y Schmidhuber, J. (1997). «Long short-term memory». En: *Neural computation* 9.8, págs. 1735-1780.

Hsing, Y. (1990). «On the variable elasticity of the demand for gasoline: The case of the USA». En: *Energy Economics* 12.2, págs. 132-136.

Huntington, H. (2007). «Industrial natural gas consumption in the United States: An empirical model for evaluating future trends». En: *Energy Economics* 29.4, págs. 743-759.

Huseyin, A. (2021). «Monthly natural gas demand forecasting by adjusted seasonal grey forecasting model». En: *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* 43.1, págs. 54-69.

Hylleberg, S., Engle, R., Granger, C. y Yoo, B. (1990). «Seasonal integration and cointegration». En: *Journal of econometrics* 44.1-2, págs. 215-238.

IEA (2020). *Global Energy Review 2020: The impacts of the Covid-19 crisis on global energy demand and CO2 emissions*. Accessed: 2021-09-10.

IEA (2021a). *Gas Market Report, Q3-2021*. Accessed: 2021-09-10.

IEA (2021b). *Net Zero by 2050: A Roadmap for the Global Energy Sector*. Accessed: 2021-09-10.

IEA, I. E. A. .-. (mar. de 2021c). *Oil 2021: Analysis and forecast to 2026*. IEA. URL: www.iea.org/t&c/.

IPCC (2021). *Climate Change 2021 The Physical Science Basis Summary for Policymakers*. Accessed: 2021-09-10.

Ivanin, O. y Direktor, L. (2018). «The Use of Artificial Neural Networks for Forecasting the Electric Demand of Stand-Alone Consumers». En: *Thermal Engineering* 65.5, págs. 258-265.

Iwayemi, A., Adenikinju, A. y Babatunde, M. A. (2010). «Estimating petroleum products demand elasticities in Nigeria: A multivariate cointegration approach». En: *Energy Economics* 32.1, págs. 73-85.

Jiang, P., Fan, Y. y Klemeš, J. (2021). «Impacts of COVID-19 on energy demand and consumption: Challenges, lessons and emerging opportunities». En: *Applied energy* 285, pág. 116441.

Jiang, P., Li, R., Lu, H. y Zhang, X. (2020). «Modeling of electricity demand forecast for power system». En: *Neural Computing and Applications* 32.11, págs. 6857-6875.

Jimenez, J., Navarro, L., Quintero M., C. G. y Pardo, M. (2021). «Multivariate Statistical Analysis for Training Process Optimization in Neural Networks-Based Forecasting Models». En: *Applied Sciences* 11.8. ISSN: 2076-3417. DOI: 10.3390/app11083552. URL: https://www.mdpi.com/2076-3417/11/8/3552.

Jiménez, J., Pertuz, A., Quintero, C. y Montaña, J. (2019). «Multivariate statistical analysis based methodology for long-term demand forecasting». En: *IEEE Latin America Transactions* 17.01, págs. 93-101.

Jiménez, J., Donado, K. y Quintero, C. G. (2017). «A Methodology for Short-Term Load Forecasting». En: *IEEE Latin America Transactions* 15.3, págs. 400-407. DOI: 10.1109/TLA.2017.7867168.

Kamrani, E. (2010). *Modeling and forecasting long-term Natural Gas (NG) consumption in Iran, using Particle Swarm Optimization (PSO)*.

Kayser, H. A. (2000). «Gasoline demand and car choice: estimating gasoline demand using household information». En: *Energy economics* 22.3, págs. 331-348.

Kazemi, A., Ganjavi, H. S., Menhaj, M., Mehregan, M., Taghizadeh, M. y Asl, A. F. (2009). «A multi-level artificial neural network for gasoline demand forecasting of Iran». En: *2009 Second International Conference on Computer and Electrical Engineering*. Vol. 1. IEEE, págs. 61-64.

Kim, J.-H., Lee, S. y Preston, J. (2006). «The impact of the fuel price policy on the demand for diesel passenger cars in Korean cities». En: *International Review of Public Administration* 10.2, págs. 61-73.

Koshala, R. K., Koshal, M., Boyd, R. G. y Rachmany, H. (1999a). «Demand for kerosene in developing countries: A case of Indonesia». En: *Journal of Asian Economics* 10.2, págs. 329-336.

Koshala, R. K., Koshal, M., Boyd, R. G. y Rachmany, H. (1999b). «Demand for kerosene in developing countries: A case of Indonesia». En: *Journal of Asian Economics* 10.2, págs. 329-336.

Kumar, S., Viral, R., Deep, V., Sharma, P., Kumar, M., Mahmud, M. y Stephan, T. (2021). «Forecasting major impacts of COVID-19 pandemic on country-driven sectors: challenges, lessons, and future roadmap». En: *Personal and Ubiquitous Computing*, págs. 1-24.

Lahiri, S. y Lahiri, S. (2003). *Resampling methods for dependent data*. Springer Science & Business Media.

Lee, H. S. (1992). «Maximum likelihood inference on cointegration and seasonal cointegration». En: *Journal of Econometrics* 54.1-3, págs. 1-47.

Lee, J. y Cho, Y. (2009). «Demand forecasting of diesel passenger car considering consumer preference and government regulation in South Korea». En: *Transportation Research Part A: Policy and Practice* 43.4, págs. 420-429.

Lewis, C. (1982). «International and Business Forecasting Methods Butterworths: London». En.

Li, R., Chen, X., Balezentis, T., Streimikiene, D. y Niu, Z. (2021). «Multi-step least squares support vector machine modeling approach for forecasting short-term electricity demand with application». En: *Neural computing and applications* 33, págs. 301-320.

Liu, B., Fu, C., Bielefield, A. y Liu, Y. (2017). «Forecasting of Chinese primary energy consumption in 2021 with GRU artificial neural network». En: *Energies* 10.10, pág. 1453.

Liu, J., Wang, S., Wei, N., Chen, X., Xie, H. y Wang, J. (2021). «Natural gas consumption forecasting: A discussion on forecasting history and future challenges». En: *Journal of Natural Gas Science and Engineering*, pág. 103930.

Liu, L. y Lin, M. (1991). «Forecasting residential consumption of natural gas using monthly and quarterly time series». En: *International Journal of Forecasting* 7.1, págs. 3-16.

López Valderrama, J. S. et al. (2015). «Relación entre el precio de los combustibles y la seguridad vial en Bogotá». Tesis de mtría. Uniandes.

Lütkepohl, H. (2013). *Introduction to multiple time series analysis*. Springer Science & Business Media.

Mardiana, S., Saragih, F. y Huseini, M. (2020). «Forecasting Gasoline Demand in Indonesia Using Time Series». En: *International Journal of Energy Economics and Policy* 10.6, pág. 132.

Mariño, M. D., Arango, A., Lotero, L. y Jiménez, M. (2021). «Modelos de series temporales para pronóstico de la demanda eléctrica del sector de explotación de minas y canteras en Colombia». En: *Revista EIA* 18.35, págs. 1-23.

Martin, V., Hurn, S. y Harris, D. (2013). *Econometric modelling with time series: specification, estimation and testing*. Cambridge University Press.

Marziali, A., Fabbiani, E. y Nicolao, G. (2021). «Ensembling methods for countrywide short-term forecasting of gas demand». En: *International Journal of Oil, Gas and Coal Technology* 26.2, págs. 184-201.

Medina, D. O. G. y Marulanda, G. A. (2017). «Estimación del consumo eléctrico colombiano en el corto y largo plazo empleando regresión multivariable y series temporales». En: *Avances: Investigacion en Ingeniería* 14.1, págs. 155-168.

Medina, S., Moreno, J. y Gallego, J. (2011). «Pronóstico de la demanda de energía eléctrica horaria en Colombia mediante redes neuronales artificiales». En: *Revista Facultad de Ingeniería Universidad de Antioquia* 59, págs. 98-107.

Melikoglu, M. (2014). «Demand forecast for road transportation fuels including gasoline, diesel, LPG, bioethanol and biodiesel for Turkey between 2013 and 2023». En: *Renewable Energy* 64, págs. 164-171. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2013.11.009. URL: https://www.sciencedirect.com/science/article/pii/S0960148113005879.

Melikoglu, M. (2017). «Modelling and forecasting the demand for jet fuel and bio-based jet fuel in Turkey till 2023». En: *Sustainable Energy Technologies and Assessments* 19, págs. 17-23.

Mikayilov, J. I., Joutz, F. L. y Hasanov, F. J. (2020). «Gasoline demand in Saudi Arabia: are the price and income elasticities constant?» En: *Energy Sources, Part B: Economics, Planning, and Policy* 15.4, págs. 211-229.

Mirjat, N., Uqaili, M., Harijan, K., Walasai, G., Mondal, M. y Sahin, H. (2018). «Long-term electricity demand forecast and supply side scenarios for Pakistan (2015–2050): A LEAP model application for policy analysis». En: *Energy* 165, págs. 512-526.

Moran Rugel, F., Zuñiga Bastidas, J. y Marriott Garcia, F. (2009). «Estimación de las elasticidades de la demanda de gasolina en el Ecuador: un análisis empírico». En.

Moreno Guerrero, H. (2000). *Estimación de la demanda de combustibles para el sector automotor: gasolina corriente, gasolina extra y ACPM*.

Nasr, G. E., Badr, E. y Joun, C. (2002). «Cross entropy error function in neural networks: Forecasting gasoline demand.» En: *FLAIRS conference*, págs. 381-384.

Nicholson, W. (2005). *Teoría microeconómica. Principios básicos y ampliaciones: principios básicos y ampliaciones*. Editorial Paraninfo.

NOAA (2021). *Climate Prediction Center*. Accessed: 2021-09-25. National Oceanic y Atmospheric Administration. URL: https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php.

Norouzi, N., Rubens, G., Choupanpiesheh, S. y Enevoldsen, P. (2020). «When pandemics impact economies and climate change: exploring the impacts of COVID-19 on oil and electricity demand in China». En: *Energy Research & Social Science* 68, pág. 101654.

Özbay, H. y Dalcali, A. (2021). «Effects of COVID-19 on electric energy consumption in Turkey and ANN-based short-term forecasting». En: *Turkish Journal of Electrical Engineering & Computer Sciences* 29.1, págs. 78-97.

Özmen, A. (2021). «Sparse regression modeling for short-and long-term natural gas demand prediction». En: *Annals of Operations Research*, págs. 1-26.

Özmen, A., Yılmaz, Y. y Weber, G. (2018). «Natural gas consumption forecast with MARS and CMARS models for residential users». En: *Energy Economics* 70, págs. 357-381.

Páez Martínez, E. (2009). «Determinantes de los precios de la gasolina y su impacto económico en México, 2010-2017.» En: págs. 1-49.

Pathak, N., Ba, A., Ploennigs, J. y Roy, N. (2018). «Forecasting gas usage for big buildings using generalized additive models and deep learning». En: *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, págs. 203-210.

Pérez, J. (2020). *Pronóstico de demanda de energía eléctrica de Colombia utilizando un modelo estadístico a partir de la metodología de Box-Jenkins*. Fundación Universitaria Los Fundadores, Bogotá, Colombia.

Perwez, U., Sohail, A., Hassan, S. y Zia, U. (2015). «The long-term forecast of Pakistan's electricity supply and demand: An application of long range energy alternatives planning». En: *Energy* 93, págs. 2423-2435.

Pessanha, J. y Leon, N. (2015). «Forecasting long-term electricity demand in the residential sector». En: *Procedia computer science* 55, págs. 529-538.

Rao, B. B. y Rao, G. (2009). «Cointegration and the demand for gasoline». En: *Energy Policy* 37.10, págs. 3978-3983.

Rao, R. D. y Parikh, J. K. (1996). «Forecast and analysis of demand for petroleum products in India». En: *Energy policy* 24.6, págs. 583-592.

Rasouli, N. (jun. de 2018). «Forecasting the fuel consumption based on the fuzzy linear regression models». En: *International Journal of Engineering Science and Generic Research (IJESAR)* 4, págs. 41-59. ISSN: 2456-043X.

Rehman, S., Cai, Y., Fazal, R., Das Walasai, G. y Mirjat, N. (2017). «An integrated modeling approach for forecasting long-term energy demand in Pakistan». En: *Energies* 10.11, pág. 1868.

Rey, J. (2018). *Pronóstico del consumo de energía eléctrica residencial para la ciudad de Bogotá*. Universidad Santo Tomás, Bogotá, Colombia.

Reyes Müller, J. F. (2015). «Pronóstico de la demanda de gasolina en Colombia empleando modelos estocásticos». En.

Rivera-González, L., Bolonio, D., Mazadiego, L. F., Naranjo-Silva, S. y Escobar-Segovia, K. (2020). «Long-Term Forecast of Energy and Fuels Demand Towards a Sustainable Road Transport Sector in Ecuador (2016–2035): A LEAP Model Application». En: *Sustainability* 12.2, pág. 472.

Rodrıguez, C. A. (2012). «Empirical analysis of the demand function for gasoline in Puerto Rico:(1999-2006)». En: *Munich Personal RePEc Archive*, págs. 1-28.

Rodríguez, S. y Da Silva, N. (2010). «Modelos estocásticos para predecir la demanda de gas licuado de petróleo en Uruguay». En: *Serie DT (10/01);*

Sakunthala, K., Iniyan, S. y Mahalingam, S. (2018). «Forecasting energy consumption in Tamil Nadu using hybrid heuristic based regression model». En: *Thermal Science* 23.5 Part B, págs. 2885-2894.

Sánchez, L. y Reyes, O. (2016). «La demanda de gasolinas, gas licuado de petróleo y electricidad en el Ecuador: elementos para una reforma fiscal ambiental». En.

Sapnken, E. F. (2018). «Modeling and forecasting gasoline consumption in Cameroon using linear regression models». En.

Shekarchian, M., Moghavvemi, M., Motasemi, F., Zarifi, F. y Mahlia, T. (2012). «Energy and fuel consumption forecast by retrofitting absorption cooling in Malaysia from 2012 to 2025». En: *Renewable and Sustainable Energy Reviews* 16.8, págs. 6128-6141.

Sigauke, C. (2017). «Forecasting medium-term electricity demand in a South African electric power supply system». En: *Journal of Energy in Southern Africa* 28.4, págs. 54-67.

Suhono, S. (2015). «Long-term electricity demand forecasting of Sumatera system based on electricity consumption intensity and Indonesia population projection 2010-2035». En: *Energy Procedia* 68, págs. 455-462.

Suykens, J., Lemmerling, P., Favoreel, W., De Moor, B., Crepel, M. y Briol, P. (1996). «Modelling the Belgian gas consumption using neural networks». En: *Neural Processing Letters* 4.3, págs. 157-166.

Trotter, M., Bolkesjø, T., Féres, J. y Hollanda, L. (2016). «Climate change and electricity demand in Brazil: A stochastic approach». En: *Energy* 102, págs. 596-604.

UPME (2014a). *Nota técnica # 004 proyecciones de demanda de gas natural en colombia. una revisión necesaria de metodología y paradigmas de análisis.*

UPME (2014b). *Nota técnica # 005, Proyecciones de demanda de energía eléctrica en Colombia. Revisión de la Metodología.* "Unidad de Planeación Minero Energética". Colombia.

UPME (2021). *Proyecciones de demanda de energía eléctrica y gas natural 2021-2035.* "Unidad de Planeación Minero Energética". Colombia.

Uri, N. D. y Herbert, J. H. (1992). «A note on estimating the demand for diesel fuel by farmers in the United States». En: *Applied Economic Perspectives and Policy* 14.2, págs. 153-167.

Varian, H. R. (1992). *Análisis microeconómico.* Antoni Bosch Editor.

Velásquez, J. D., Franco, C. J. y García, H. A. (sep. de 2009). «Un modelo no lineal para la predicción de la demanda mensual de electricidad en Colombia». es. En: *Estudios Gerenciales* 25, págs. 37-54. ISSN: 0123-5923.

Vera, V. D. G. (2016). «Pronóstico De La Demanda Mensual De Electricidad Con Series De Tiempo». En: *Revista EIA* 13.26, págs. 111-120.

Villamarin Lafaurie, E. J. et al. (2007). «Elasticidad precio de la demanda de gasolina corriente y acpm». B.S. thesis. Bogotá-Uniandes.

Waheed Bhutto, A., Ahmed Bazmi, A., Qureshi, K., Harijan, K., Karim, S. y Shakil Ahmad, M. (2017). «Forecasting the consumption of gasoline in transport sector in pakistan based on ARIMA model». En: *Environmental Progress & Sustainable Energy* 36.5, págs. 1490-1497.

Wood, S. N. (2017). *Generalized additive models: an introduction with R.* CRC press.

Wu, C.-F. J. (1986). «Jackknife, bootstrap and other resampling methods in regression analysis». En: *the Annals of Statistics* 14.4, págs. 1261-1295.

XM (2021). *Pronóstico oficial de demanda (definitivo).* Accessed: 2021-10-03.

Yucesan, M., Pekel, E., Celik, E., Gul, M. y Serin, F. (2021). «Forecasting daily natural gas consumption with regression, time series and machine learning based methods». En: *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, págs. 1-16.

Anexo A Manual Usuario

El archivo se ejecuta desde google colaboratory, un entorno de google que permite ejecutar código en python sin necesidad de realizar configuraciones especiales y fácil de compartir. Inicialmente se debe tener el archivo de base de datos (**BD Combustibles.xlsx**) guardado en el drive del usuario que va a ejecutar el código. Luego debe abrirse el archivo tipo colaboratory que está dividido varias secciones que a su vez se dividen en celdas. Para ejecutar el código en la celda, se hace click sobre ella para seleccionarla y luego presionar el botón de play a la izquierda de la celda o usar la combinación de teclas Çommand/Ctrl + Enter". Mientras se ejecuta la celda el botón play cambia por un botón de stop el cual puede presionarse en cualquier momento si se desea detener la ejecución. Si el botón cambia a color rojo es que la ejecución falló. Si no se detecta actividad por parte de google colab puede cerrar el entorno perdiendo parte de la información de la ejecución. El código en cada una de las secciones puede desplegarse o ocultarse al hacer click en la flecha al lado del titulo de cada sección. Las secciones del archivo se muestran a continuación.

A.1 Conexión Google Colab—Drive

Consta de una celda la cual al ejecutarse mostrará el mensaje "Go to this URL in a browser" con una dirección electrónica. Se debe hacer click en el link el cual abrirá una ventana nueva solicitando permiso para acceder al google drive del usuario. La cuenta que se seleccione debe ser la misma que contenga el archivo "BD Combustibles.xlsx". Una vez seleccionada la cuenta se debe abre un nuevo recuadro en el cual se debe dar click en "Iniciar sesión", Aparecerá un recuadro con un código que debe copiarse. El usuario debe volver al google colab y pegar el código en el recuadro con el mensaje "Enter your authorization code" y presionar la tecla enter. Con esto se finaliza la conexión con el drive.

A.2 Funciones

Esta sección contiene funciones que no deben modificarse o pueden llevar a un mal funcionamiento. El usuario deberá ejecutar la sección o ejecutar celda por celda para cargar cada una de las funciones.

A.3 Cargar Módulos y Base de Datos

Antes de ejecutar esta sección se debe verificar que en la celda ***Cargar base de datos*** en la variable ***dir_colab*** se tenga la siguiente dirección:

content/drive/MyDrive/DIRECCIÓN A BASE DE DATOS/BD Combustibles.xlsx
De ser así se ejecuta la sección.
NOTA: **DIRECCIÓN A BASE DE DATOS** son todas las carpetas personales del usuario donde se aloja el archivo **BD Combustibles.xlsx**:

A.4 Nombre y Descripción de Variables

En la celda de esta sección se encuentran las variables que pueden ser pronosticadas junto con las variables que pueden ser usadas para realizar el pronostico cada una con un a breve descripción. Ninguna de las variables que es usada para pronostico puede usarse como variable de entrenamiento de los modelos.

A.5 Información de entrada para modelar

En esta sección se ajustan los parámetros con los que se va a hacer la modelación. Cada parámetro se describe a continuación.

A.5.1 Ajustes generales

`ajuste_auto:` hace referencia a si se desea hacer un ajuste de los modelos de forma automática mediante una función (True) o un ajuste manual donde el usuario selecciona los parámetros de los modelos a su gusto (False). El ajuste automático varía los parámetros de los modelos en varios rangos y el conjunto de parámetros con menor error se selecciona para los pronósticos sin embargo este proceso puede tardar varios minutos.

`crit_ajusteMAPE:`. hace referencia al criterio de ajuste que se desea emplear para el ajuste de los modelos. Las opciones para esta variable se muestran a continuación:

“ME”: Error Medio
“MPE”: Error Porcentual Medio
“MAE”: Error Absoluto Medio
“MAPE”: Error Porcentual Absoluto Medio
“SSE”: Suma de Cuadrados del Error
“MSE”: Error Cuadrático Medio
“RMSE”: Raíz Cuadrada del Error Cuadrático Medio
“FPE”: Error de Predicción Final
“AIC”: Criterio de Información de Akaike
“AICc”: Criterio de Información de Akaike Corregido
“BIC”: Criterio de Información Bayesiano

"HQC": Criterio de Información de Hannan-Quinn

MAPE_val: Se desea ajustar el mejor modelo a partir del MAPE de validación ("True") o de entrenamiento ("False").

boots_rep: Número de replicas que se desea realizar en el bootstrap.

A.5.2 Modelos a estimar

GAM_mod: Si se desea estimar el modelo GAM ("True"), si no ("False").

MLR_mod: Si se desea estimar el modelo MLR ("True"), si no ("False").

LASSO_mod: Si se desea estimar el modelo LASSO ("True"), si no ("False").

MARS_mod: Si se desea estimar el modelo MARS ("True"), si no ("False").

LSTM_mod: Si se desea estimar el modelo LSTM ("True"), si no ("False").

Las siguientes variables a ajustar solo tendrán efecto en los modelos si se tiene la función de ajuste automático apagada (**ajuste_auto = False**).

GAM_n_spline: Número de splines (particiones) que hace el modelo GAM al momento de realizar las estimaciones, este número debe ser un entero mayor a 3, en donde no se recomienda usar valores superiores a 6.

GAM_spline_order: Orden del polinomio de los spline del modelo GAM. Este valor debe ser un entero mayor a 2, en donde el entero seleccionado debe ser estrictamente menor al valor seleccionado para **GAM_n_spline**.

MARS_max_degree: Número máximo de términos generados por el proceso de avance.

MARS_penalty: Parámetro de suavizado que se utiliza en Validación Cruzada Generalizada usada para determinar si se debe agregar una función de base lineal o de bisagra durante el proceso de avance.

LSTM_capas: Número de capas de la red neuronal, Este valor debe ser un entero preferiblemente no mayor a 3.

LSTM_activation: El método de activación.

LSTM_neur1: Número de neuronas en la primera capa del modelo de red neuronal.

LSTM_neur2: Número de neuronas en la segunda capa del modelo de red neuronal.

LSTM_neur3: Número de neuronas en la tercera capa del modelo de red neuronal.

A.5.3 Ajuste base de datos

y_dep: Variable respuesta (dependiente). Se selecciona de la lista en la sección A.4.

x_ind: Variables explicativas (independientes).Se puede seleccionar más de una variable de la lista que se presentó en la sección A.4.

var_rez: Nombre variables que se quiere rezagar. Se puede seleccionar más de una variable de la lista que se presentó en la sección A.4.

n_rez: Número de rezagos variables. En caso de tener más de una variable en **var_rez:** se debe asignar en forma de tupla donde cada número son los rezagos en cada variable.

rez_est: Periodicidad de los rezagos. Si se asigna un valor de 1 entonces se generarán rezagos simples, al asignar una valor de 2 se generarán rezagos bimensuales, si se asigna un valor de 3 se generarán rezagos trimestrales, etc.

val: Número de observaciones usadas para VALIDAR el modelo

A.6 Cálculo de Información de Entrada

En esta sección se ajustan todos los datos de acuerdo a la sección anterior por lo que debe ejecutarse toda la sección.

A.7 Entrenamiento y Validación de los modelos

A.7.1 Ajuste modelo e Hiperparámetros

A.7.1.1 Ajuste

Al ejecutarse la primera celda en caso de tener la función automática activada se inicia el proceso de optimización de los modelos de lo contrario se realiza el proceso con los parámetros previamente seleccionados, el avance se muestra en la consola.

A.7.1.2 Calculo de error

En la segunda celda se calcula el MAPE de cada uno de los modelos activados en la sección A.5

A.7.1.3 Datos de validación y proyección

En esta sección se muestra el resultado final del pronostico e intervalos de confianza de forma tabulada.

A.7.2 Gráficos Entrenamiento, Validación y Proyección

En esta sección se muestran las gráficas de los pronósticos de los modelos. Al hacer click en cada uno de los nombres de los modelos en las leyendas de las gráficas se puede activar o desactivar la vista de del modelo.

Anexo B Modelos de proyección de combustibles

B.1 Modelo de Regresión Lineal (LRM)

Un modelo de regresión lineal es una metodología que busca establecer la relación existente entre una variable dependiente o respuesta y_t respecto a una o más variables independientes $x_{j,t}$, con $j = 1, 2, 3, \ldots, k$.

Entonces de acuerdo con Gareth et al. (2021), si se tiene una muestra de tamaño T para las variables y_t y $x_{j,t}$, con $j = 1, 2, \ldots, p$, se tiene que la relación lineal entre estas variables puede plantearse de la forma:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \ldots + \beta_p x_{p,t} + \varepsilon_t$$

con $\varepsilon_t \sim iid\,(0, \sigma_\varepsilon^2)$.

En este caso y_t representa la t-ésima observación de la variable dependiente, $x_{j,t}$ representa la t-ésima observación para la j-ésima variable. Y los parámetros β_j representan los efectos parciales de las $x_{j,t}$ sobre y_t. El término de error aleatorio, ε_t, explica todo lo que afecta a y_t distinto a las $x_{j,t}$.

En forma matricial tenemos

$$y_t = x_t{}'\beta + u_t$$

con

$$x_t = \begin{bmatrix} 1 \\ x_{1,t} \\ x_{2,t} \\ \vdots \\ x_{p,t} \end{bmatrix}$$

y

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

Para los T datos tenemos:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p,T} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix} = X\beta + \varepsilon \tag{B.1}$$

De la ecuación B.1, el estimador de mínimos cuadrados se define como el vector b que minimiza

$$SSE = \sum_{t=1}^{T} \left(y_t - {x_t}^\top b\right)^2$$

la solución es

$$b = \left(X^\top X\right)^{-1} X^\top Y = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \tag{B.2}$$

La matriz de covarianzas estimada de b está dada por $\widehat{\text{var}}\,(b) = \hat{\sigma}^2 \left(X^\top X\right)^{-1}$.

De acuerdo con Gareth et al. (2021), la multicolinealidad hace referencia a las relaciones lineales entre las variables explicativas. Si dos variables explicativas o más están fuertemente relacionadas, tendremos multicolinealidad. En este caso, quizás tengamos más variables en el modelo de las que sean necesarias y por lo tanto estemos sobre ajustando el modelo, es decir, estamos haciendo que el modelo aprenda más comportamientos de los que son necesarios para que se tengan buenas predicciones (buena generalización). Desde este punto de vista,se analizará con mayor detalle el problema del sobre-ajuste más adelante.

B.2 Modelo Aditivo Generalizado

De acuerdo con la literatura revisada a lo largo de este proyecto, el modelo LRM es un modelo muy utilizado para hacer predicciones, no sólo de demanda de energía eléctrica, sino también de combustibles líquidos. Sin embargo, este modelo plantea que la relación entre y_t y $x_{j,t}$ se da a través del término $\beta \cdot x_{j,t}$. Es decir, si nos movemos en la curva de nivel de y_t y $x_{j,t}$, la relación es estrictamente lineal, lo cual no es necesariamente cierto. Para superar la restricción de linealidad entre la relación de y_t y $x_{j,t}$, Gareth et al. (2021) ha propuesto el Modelo Aditivo Generalizado (GAM).

Para explicar mejor este modelo, suponga que se tiene solo una relación de y_t con una sola variable explicativa como la que se muestra en la Figura B.1

Siguiendo a Gareth et al. (2021), se podría pensar en un modelo de regresión cúbica de y_t con Age_t. O mejor aún, se podría pensar en tres regresiones cúbicas en los tramos $(0, 50)$, $(50, 100)$ y $(100, 150)$, es decir, en tres regresiones por tramos. Los modelos en los dos casos

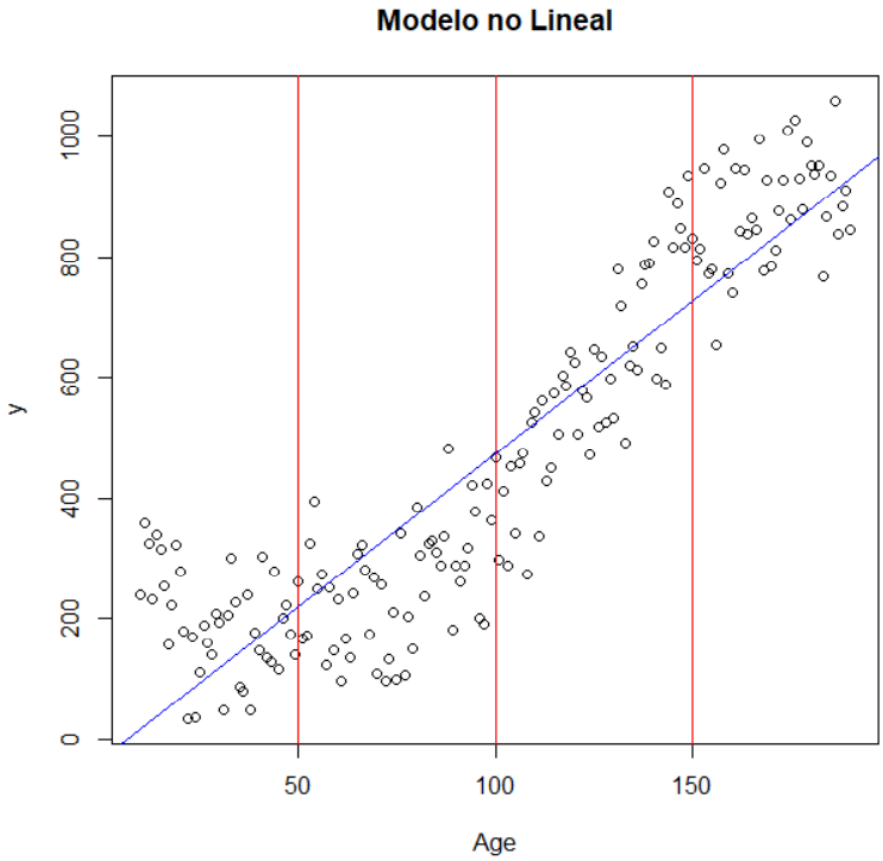

Figura B.1: Simulación Propia. Los puntos corresponden a un conjunto generado por el verdadero proceso simulado. El verdadero proceso se simuló como la concatenación de tres funciones distintas en los rangos de Age, $(0, 50)$, $(50, 100)$ y $(100, 150)$ La linea azul es la regresión lineal simple de y_t con Age_t

serían:
La regresión cúbica

$$y_t = \beta_0 + \beta_1 Age_t + \beta_2 Age_t^2 + \beta_3 Age_t^3 + \varepsilon_t$$

la regresión cúbica por tramos es

$$y_t = \begin{cases} \beta_{01} + \beta_{11} Age_t + \beta_{21} Age_t^2 + \beta_{31} Age_t^3 + \varepsilon_{1,t} & \text{si } x < 50 \\ \beta_{02} + \beta_{12} Age_t + \beta_{22} Age_t^2 + \beta_{32} Age_t^3 + \varepsilon_{2,t} & \text{si } 50 \leq Age < 100 \\ \beta_{03} + \beta_{13} Age_t + \beta_{23} Age_t^2 + \beta_{33} Age_t^3 + \varepsilon_{3,t} & \text{si } 100 \leq Age < 150 \end{cases}$$

De nuevo, basados en Gareth et al. (2021), en el segundo caso tendríamos la Fig. B.2 para los primeros dos tramos. En la parte superior izquierda se tendrían los polinomios en los rangos $(0, 50)$ y $(50, 100)$. Se puede imponer la restricción de continuidad en $Age_t = 50$ y se tendría la gráfica superior derecha. Se puede imponer, además, las restricciones de continuidad de la primera y segundas derivadas en $Age_t = 50$, y así se tendría la función spline en la parte baja izquierda de la Fig. B.2. Ahora el spline es un polinomio suave. En este caso, a $Age_t = 50$ se le conoce como el nudo.

Basados en Gareth et al. (2021), en general un spline de grado d es una función polinomial por tramos a los cuales llamamos nudos y con las primeras $d - 1$ derivadas continuas.

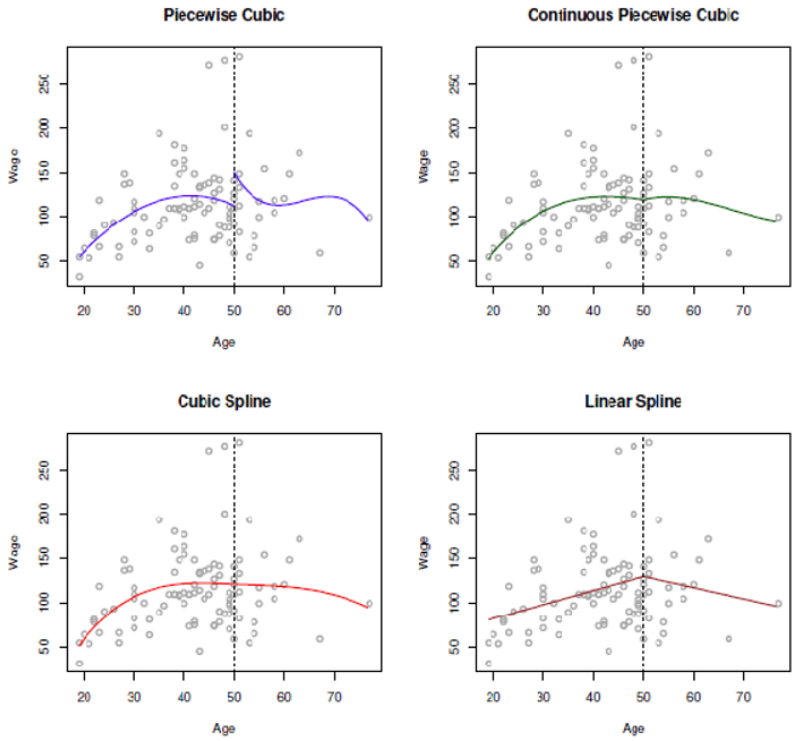

Figura B.2: Tomado de Gareth et al. (2021). La figura de la parte superior izquierda muestra el ajuste de dos polinomios cúbicos en los rangos de Age entre $(10, 50)$ y $(50, 80)$. Note como en el nudo $Age = 50$, se genera una discontinuidad. La figura superior derecha muestra los mismos polinomios de la parte superior izquierda, pero con la restricción de continuidad. Sin embargo, las pendientes en $Age = 50$ son distintas. La figura de la parte baja izquierda es un spline cúbico que corresponde a la figura superior derecha agregando las restricciones de que en $Age = 50$, las primeras y segundas derivadas son continuas. La figura inferior derecha es un spline lineal.

Si se definen los nudos $\xi_1, \xi_2, \xi_n, ..., \xi_K$, la función spline cúbica de la variable $x_{j,t}$ está dada por

$$s\left(x_{j,t}; \beta_j\right) = \left(1, x_{j,t}, x_{j,t}^2, x_{j,t}^3, h\left(x_{j,t}, \xi_1\right), h\left(x_{j,t}, \xi_2\right), ..., h\left(x_{j,t}, \xi_K\right)\right) \beta_j$$

donde β_j es de orden $K + 4$ y $h\left(x_{j,t}, \xi_i\right)$ se conoce como la función función base de potencia truncada y está dada por

$$h\left(x_{j,t}, \xi_i\right) = \left(x_{j,t} - \xi_i\right)_+^3 = \begin{cases} \left(x_{j,t} - \xi_i\right)^3 & \text{if } x > \xi_i \\ 0 & \text{En otro caso} \end{cases}$$

Entonces, ahora el modelo no es entre y_t y x_t sino entre y_t y $s\left(x_{j,t}\right)$.

Del mismo modo, para la relación de y_t con el vector

$$x_t = \begin{bmatrix} 1 \\ x_{1,t} \\ x_{2,t} \\ \vdots \\ x_{p,t} \end{bmatrix}$$

El GAM está dado por

$$y_t = \beta_0 + s\left(x_{1,t};\beta_1\right) + s\left(x_{2,t};\beta_2\right) + \cdots + s\left(x_{p,t};\beta_p\right) + \varepsilon_t \tag{B.3}$$

Este es un modelo sobre-parametrizado, no todos los términos en $s\left(x_{j.t}\right)$ son relevantes. Por tal razón, es posible que se tenga el problema del sobre-ajuste.

En cada uno de las p funciones spline $s\left(x_{j,t};\beta_j\right)$ tenderíamos para los T datos

$$X_j\beta_j = \begin{bmatrix} x_{j,1} & x_{j,1}^2 & x_{j,1}^3 & h\left(x_{j,1},\xi_1\right) & h\left(x_{j,1},\xi_2\right) & \cdots & h\left(x_{j,1},\xi_K\right) \\ x_{j,2} & x_{j,2}^2 & x_{j,2}^3 & h\left(x_{j,1},\xi_1\right) & h\left(x_{j,2},\xi_2\right) & \cdots & h\left(x_{j,2},\xi_K\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{j,T} & x_{j,T}^2 & x_{j,T}^3 & h\left(x_{j,T},\xi_1\right) & h\left(x_{j,T},\xi_2\right) & \cdots & h\left(x_{j,T},\xi_K\right) \end{bmatrix} \beta_j$$

Para todas las funciones spline de las variables explicativas en los T datos tenemos

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p + \varepsilon = \beta_0 + XB + \varepsilon \tag{B.4}$$

con

$$X = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}$$

y

$$B = \begin{bmatrix} \beta_1 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Por lo que de nuevo se llega al problema de minimizar

$$SSE = \sum_{t=1}^{T} \left(y_t - {x_t}^\top b\right)^2$$

B.3 Un Ejemplo con Datos Simulados

Para el modelo de la Figura B.1, se muestran un modelos lineal, GAM y un modelo polinomial de orden 10, un modelo más sobre ajustado que el GAM. Los modelos, ajustados en el software R, se muestran en las Figs. B.3, B.4, y B.5.

```
> summary(modelo1)

Call:
lm(formula = y/100 ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-267.72  -76.14   -6.48   83.07  334.77

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.9434    19.6585  -1.676   0.0955 .
Age           5.0713     0.1742  29.106   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 122.5 on 179 degrees of freedom
Multiple R-squared:  0.8256,    Adjusted R-squared:  0.8246
F-statistic: 847.2 on 1 and 179 DF,  p-value: < 2.2e-16
```

Figura B.3: Modelo Lineal

```
> summary(modelo2)

Call:
lm(formula = y/100 ~ bs(Age, knots = c(50, 100, 150), intercept = FALSE,
    degree = 3))

Residuals:
    Min      1Q  Median      3Q     Max
-193.05  -60.80    7.66   63.22  197.04

Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                         317.84      39.89   7.968 2.03e-13 ***
bs(Age, knots = c(50, 100, 150), intercept = FALSE, degree = 3)1   -209.41      71.07  -2.946  0.00366 **
bs(Age, knots = c(50, 100, 150), intercept = FALSE, degree = 3)2   -101.54      47.79  -2.125  0.03502 *
bs(Age, knots = c(50, 100, 150), intercept = FALSE, degree = 3)3    -59.61      59.70  -0.999  0.31942
bs(Age, knots = c(50, 100, 150), intercept = FALSE, degree = 3)4    559.29      53.82  10.391  < 2e-16 ***
bs(Age, knots = c(50, 100, 150), intercept = FALSE, degree = 3)5    601.26      59.58  10.092  < 2e-16 ***
bs(Age, knots = c(50, 100, 150), intercept = FALSE, degree = 3)6    584.89      55.58  10.523  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.5 on 174 degrees of freedom
Multiple R-squared:  0.9193,    Adjusted R-squared:  0.9165
F-statistic: 330.3 on 6 and 174 DF,  p-value: < 2.2e-16
```

Figura B.4: Modelo GAM

```
> summary(modelo3)

Call:
lm(formula = y/100 ~ poly(Age, 10))

Residuals:
    Min      1Q  Median      3Q     Max
-194.55  -66.40    2.18   63.41  197.51

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       474.19       6.26  75.747  < 2e-16 ***
poly(Age, 10)1   3564.83      84.22  42.327  < 2e-16 ***
poly(Age, 10)2    895.34      84.22  10.631  < 2e-16 ***
poly(Age, 10)3   -735.49      84.22  -8.733 2.26e-15 ***
poly(Age, 10)4   -229.15      84.22  -2.721  0.00719 **
poly(Age, 10)5    -89.01      84.22  -1.057  0.29206
poly(Age, 10)6    186.78      84.22   2.218  0.02790 *
poly(Age, 10)7   -122.65      84.22  -1.456  0.14717
poly(Age, 10)8    -21.78      84.22  -0.259  0.79621
poly(Age, 10)9     18.82      84.22   0.223  0.82349
poly(Age, 10)10  -159.75      84.22  -1.897  0.05956 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.22 on 170 degrees of freedom
Multiple R-squared:  0.9217,    Adjusted R-squared:  0.9171
F-statistic:   200 on 10 and 170 DF,  p-value: < 2.2e-16
```

Figura B.5: Modelo Polinomial

Los ajustes de los modelos se muestran en la Figura B.6. El modelo lineal sub-ajusta los datos, el modelo polinomial lo sobre-ajusta (tiene la mayor variabilidad, se presenta alta varianza) y el mejor modelo es GAM. Este tiene un R^2 ligeramente más bajo que el del modelo polinomial, lo que implica que el error de entrenamiento de este último es más alto. De hecho, de forma numérica el problema del sobre ajuste se observa en la Cuadro B.1

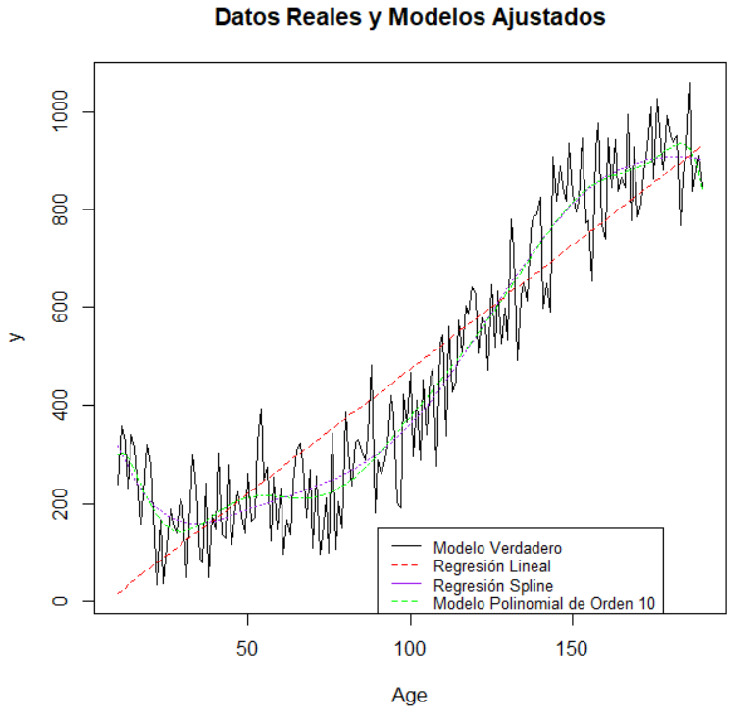

Figura B.6: Modelos Ajustados y Datos Reales. Note como el modelo lineal sub ajusta los datos, el modelo polinomial de grado 10 tiene más variabilidad que el GAM, lo que indica un sobre ajuste de los datos. El Modelo GAM es una función suave. Cálculos propios

El Problema del Sobre Ajuste			
Modelo	Flexibilidad	SSE de Entrenamiento	MAPE de Prueba
Modelo Lineal	2	2035.2461	0.004561438
Modelo GAM	6	964.2458	0.003240183
Modelo Polinomial de Orden 10	10	950.4631	0.003550421

Cuadro B.1: El MSE de entrenamiento disminuye con la Flexibilidad. El MAPE de prueba disminuye del Modelo lineal (Flexibilidad 2) al GAM (Flexibilidad=6) y aumenta con el Modelo Polinomial de grado 10 (Flexibilidad=10).

B.4 El problema del sobre-ajuste

En esta sección trataremos brevemente el problema del sobre ajuste. Si hacemos referencia a un modelo de regresión lineal, agregar una variable $x_{j,t}$ implica agregar el término $\beta_j x_{j,t}$ y si agregamos un nudo a una función spline cúbica, equivale a agregar una variable $h\left(x_{j,t}, \xi_i\right)$ con su respectivo coeficiente. En cualquier caso, agregamos un término que consiste de la variable y su coeficiente. Por esta razón haremos referencia únicamente al hecho de agregar términos o variables.

Como lo muestra Greene (2000), el R^2 de un modelo de regresión aumenta con el número de términos en el modelo. Esto implica que el error de entrenamiento $SSE = (1 - R^2)\,TSS$ disminuye con el aumento del número de términos. Por esta razón, el error de entrenamiento no es una medida adecuada para seleccionar entre modelos que tienen una misma variable dependiente y distinto número de variables explicativas, ya que siempre tendrá menor error el modelo con más términos. Cuando se agregan más variables a un modelo, la disminución del error de entrenamiento implica que el modelo ajustado pasa cada vez más cerca de todos los puntos, a lo que se le conoce como aumento de la flexibilidad, véase Gareth et al. (2021).

Sin embargo, lo que realmente importa es el poder de generalización de un modelo, que determina el poder predictivo fuera de muestra. Por está razón, hay tendencia a escoger el modelo con más bajo error de prueba. Entonces, lo más adecuado es encontrar el modelo que tenga menor error de prueba. Sin embargo, como lo muestra Gareth et al. (2021), este error, en un principio, disminuye con el número de términos (con el aumento de la flexibilidad), llega a un punto mínimo y luego comienza a aumentar como se muestra en la Fig. B.7

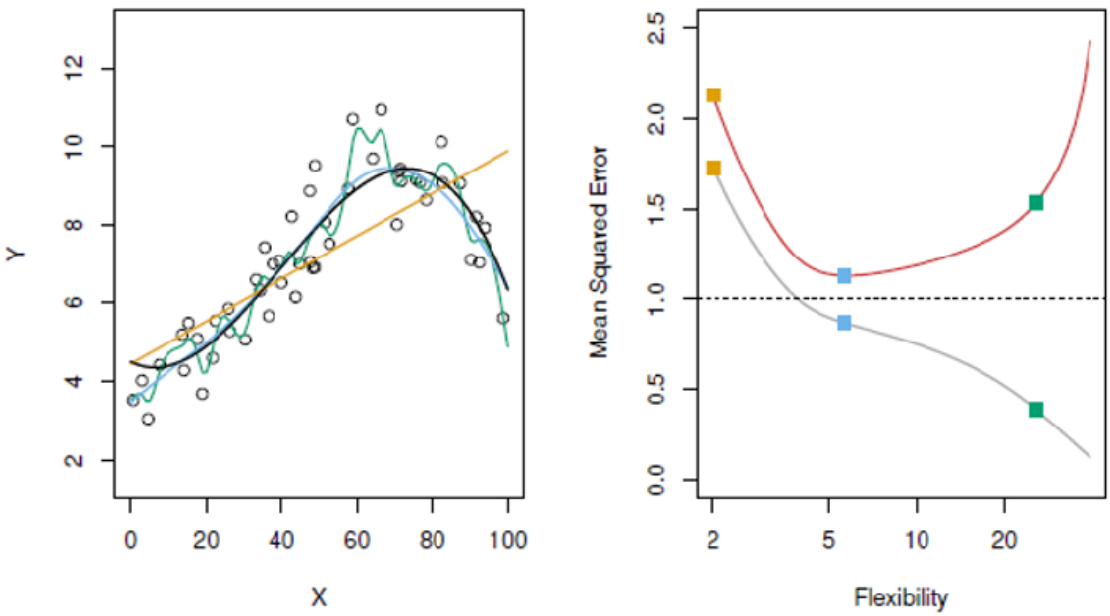

Figura B.7: Tomado de Gareth et al. (2021). El modelo verdadero es la linea negra. La linea naranja es la regresión lineal de y_t sobre x_t. La linea azul es la regresión spline de y_t sobre x_t. La linea verde que sobre ajusta el modelo es una función polinomial de grado 10 que sobre ajusta los datos. El mejor modelo es la función spline ya que tiene el menor error de prueba en la linea roja en la parte derecha. Note como el error de entrenamiento en la linea gris de la parte derecha, siempre disminuye con el número de términos en el modelo, mientras que el error de prueba disminuye al principio, alcanza un mínimo en la regresión spline y aumenta con la regresión polinomial que tiene más términos de los necesarios para dar buenas predicciones.

En la parte izquierda de la Fig. B.7, el verdadero modelo es la linea negra. Y se muestran tres modelos: regresión lineal (línea naranja), polinomio cúbico (línea azul), polinomio de grado cinco (línea verde). En el lado derecho se muestra el error de entrenamiento (línea gris) que

siempre disminuye con el número de términos. El error de prueba disminuye al pasar del modelo de regresión lineal al modelo cúbico y aumenta al pasar del modelo cúbico al modelo polinomial de grado cinco. Como en aplicaciones prácticas se debe encontrar el modelo cúbico (óptimo), se puede emplear la estimación penalizada.

B.5 Estimación penalizada

Aunque SSE no es el mismo en LRM y GAM, básicamente tenemos el mismo esquema. Gareth et al. (2021) define tres esquemas de penalización, a saber:

- Penalización RIDGE:
$$SSE^* = \sum_{t=1}^{T} \left(y_t - {x_t}^\top b\right)^2 + \lambda \sum_{j=2}^{p*} b_j^2$$
- Penalización LASSO:
$$SSE^* = \sum_{t=1}^{T} \left(y_t - {x_t}^\top b\right)^2 + \lambda \sum_{j=2}^{p*} |b_j|$$
- *Elastic-net*:
$$SSE^* = \sum_{t=1}^{T} \left(y_t - {x_t}^\top b\right)^2 + \lambda \sum_{j=1}^{p*} \left[\alpha b_j^2 + (1-\alpha)\,|b_j|\right]$$

Para cualquiera de los tres esquemas λ es el parámetro de penalización. Si $\lambda \to 0$, los coeficientes tienden a la estimación de mínimos cuadrados; y si $\lambda \to \infty$, los coeficientes tienden a cero. El objetivo es hallar el λ que deja los parámetros b necesarios para minimizar el error de prueba. Esto se consigue vía validación cruzada para una rejilla de valores de λ. Es importante mencionar que la validación por pliegues no es apropiada cuando se tiene una estructura temporal, ya que estos parten de selecciones aleatorias de los datos y esto altera la estructura temporal. En nuestro caso, se emplea una rejilla de valores para hallar el λ que minimiza el error de prueba para un conjunto de datos de validación.

B.6 Redes Neuronales Feed-Forward

Sea
$$x_t = \begin{bmatrix} x_{1,t} \\ x_{2,t} \\ \vdots \\ x_{p,t} \end{bmatrix}$$

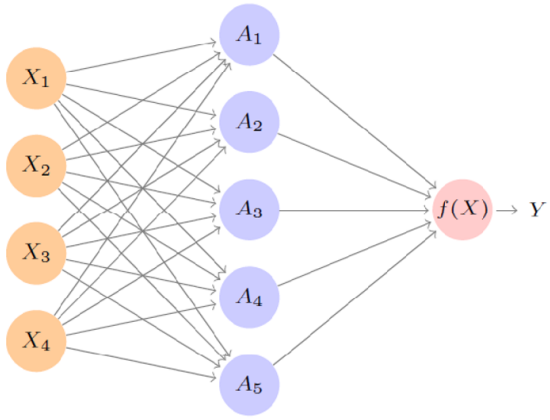

Figura B.8: Tomado de Gareth et al. (2021). La capa de entrada en el vector x_t. La capa oculta tiene cinco unidades de activación A_i. Las unidades de activación de la capa oculta alimenta la capa de salida $f(x)$.

la capa de entrada, que en realidad corresponde al vector de variables explicativas. Una red neuronal feed-forward de una sola capa oculta y k unidades de activación se define como

$$f(x_t) = \beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{j,t}\right) + \varepsilon_t$$

Para una red feed-forward de una capa oculta y una capa de entrada de cuatro variables explicativas se tiene

Las funciones de activación son funciones no lineales de las variables explicativas. Entre estas funciones de activación tenemos la sigmoidal dada por

$$g(z) = \frac{e^z}{1 + e^z}$$

Y la ReLU dada por

$$g(z) = (z)_+ = \begin{cases} 0 & \text{Si } z < 0 \\ z & \text{En otro caso} \end{cases}$$

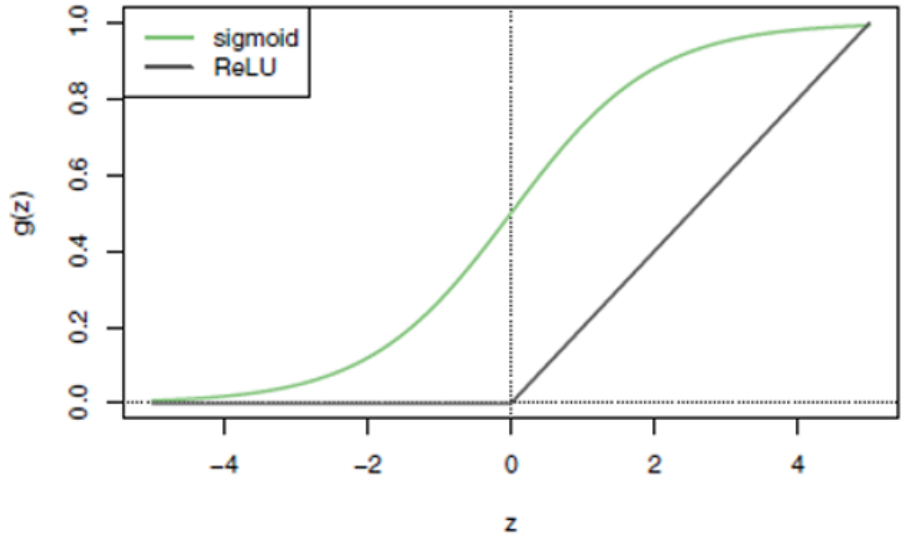

Figura B.9: Tomado de Gareth et al. (2021). La función de activación sigmoidal es la curva verde, La función de activación ReLU es la linea negra

Si se tienen dos capas ocultas como se observa en la Fig. B.10, la información de cada una de las variables de entrada va a cada una de las unidades de activación de la primera capa oculta como se observa en la siguiente ecuación:

$$g_k^1 = h_k^1(x_t) = g\left(w_{k0}^1 + \sum_{j=1}^{p} w_{kj}^1 x_{j,t}\right)$$

Cada una de estas funciones de activación, a su vez, se convierten en nuevas variables que alimentan cada una de las unidades de activación de la segunda capa oculta como se observa a continuación:

$$g_l^2 = h_l^1(x_t) = g\left(w_{l0}^2 + \sum_{k=1}^{K_1} w_{lk}^1 h_{k,t}^1\right).$$

El modelo final entonces es

$$y_t = f(x_t) + \varepsilon_t = \beta_0 + \sum_{k=1}^{K_2} \beta_k g\left(w_{k0} + \sum_{j=1}^{K_1} w_{kj} h_{k,t}^1\right) + \varepsilon_t$$

donde

$$f(x_t) = \beta_0 + \sum_{k=1}^{K_2} \beta_k g\left(w_{k0} + \sum_{j=1}^{K_1} w_{kj} h_{k,t}^1\right)$$

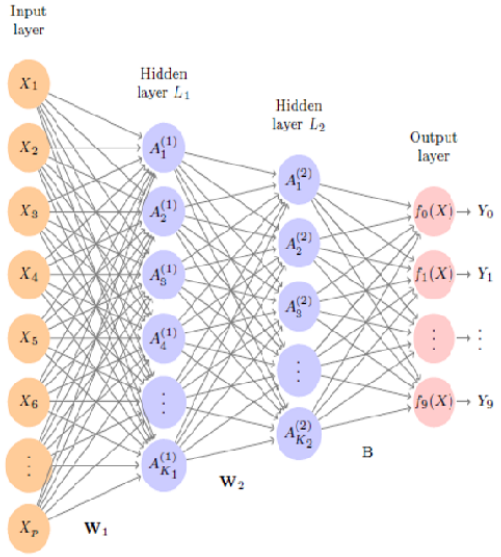

Figura B.10: Tomado de Gareth et al. (2021). La capa de entrada es x_t, la primera capa oculta recibe la información de la capa de entrada y la procesa con sus L_1 unidades de activación. La información de las L_1 unidades de activación de la primera capa alimentan las unidades de activación de la segunda capa oculta. Las L_2 unidades de activación de la segunda capa oculta alimentan cada una de las funciones de la capa de salida $f_i(x_t)$.

Como lo establece Gareth et al. (2021), esta red es capaz de aproximar *cualquier* función desconocida con muy buena precisión. La estimación de esta red se hace minimizando la función de pérdida:

$$SME = \sum_{t=1}^{T} (y_t - f(x_t))^2$$

B.7 Redes Neuronales Recurrentes (RNN)

Una red red neuronal recurrente, toma la información de forma secuencial. Desde este punto de vista, los puntos procesados por la RNN en cada periodo t no son (x_t, y_t) si no que también la red se alimenta de la información desde $t = 1, 2, \ldots, t-1$. Para ver la estructura de una RNN, de nuevo recurrimos a la Fig. B.11 del texto de Gareth et al. (2021).

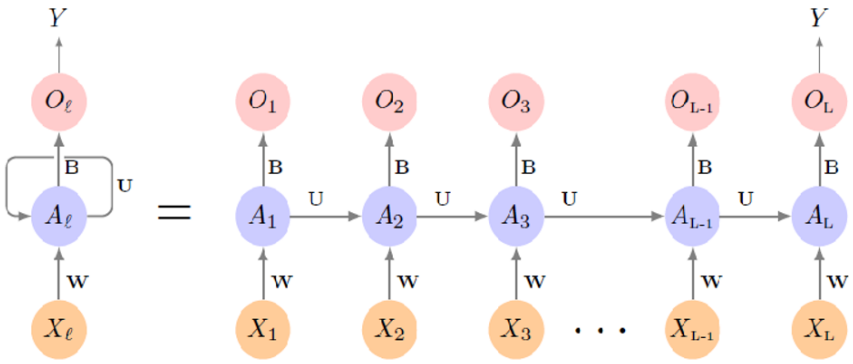

Figura B.11: Tomado de Gareth et al. (2021). La capa de entrada ahora es secuencial X_1 alimenta a A_1; X_2 y A_1 alimentan a A_2; X_3, A_1 y A_2 alimentan a A_3. AsíA_3 secuencialmente se alimentan las unidades de activación en la RNN.

Como se puede ver en la Fig. B.11, en el primer punto la red toma la información (x_1, y_1), en el segundo punto la red procesa $((x_2, A_1), y_2)$, en el tecer punto toma $((x_3, A_1, A_2), y_3)$. En este caso A_l es la función de activación de la capa oculta. Hay L funciones de activación y una sola capa oculta. Sin embargo, la función de activación L se alimenta de las $L-1$ capas anteriores. Los parámetros que conectan al vector x_t con las funciones de activación están en la matriz W. Los parámetros que conectan la función de activación l con las funciones de activación anteriores $1, 2, \ldots, l-1$ están en la matriz U y los parámetros que conectan a las funciones de activación con la capa de salida O_l están en la matriz B. Matemáticamente se tiene:

$$A_{lk} = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{lj} + \sum_{s=1}^{K} u_{ks} A_{l-1,s}\right)$$

y la capa de salida consiste de

$$O_l = \beta_0 + \sum_{k=1}^{K} \beta_k A_{lj}.$$

Aquí, $g(\cdot)$ es una función de activación como la ReLU. Teniendo en cuenta la información hasta el penúltimo rezago $L-1$, el modelo estadístico es ahora

$$y_t = \beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{tLj} + \sum_{s=1}^{K} u_{ks} a_{t,L-1,s}\right) + \varepsilon_t.$$

Aquí también es posible considerar varias capas, lo que da origen al nombre de *deep learning*.

La función de pérdida a minimizar es ahora

$$SSE = \sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} \left(y_t - \beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{tLj} + \sum_{s=1}^{K} u_{ks} a_{t,L-1,s}\right)\right)^2.$$

B.8 Redes Neuronales Recurrentes de Corta y Larga Memoria. LSTM

La arquitectura anterior se conoce como red Elman véase Godfellow 2018. De acuerdo con este último autor, esta arquitectura tiene el problema del gradiente desvaneciente, que consiste en que la derivada de SSE tiende a cero rápidamente (óptimos locales); y por lo tanto, no tiende a encontrar óptimos globales en el proceso de estimación de parámetros. De acuerdo con Godfellow 2018, un tipo de red neuronal recurrente libre del problema del gradiente desvaneciente es la red neuronal recurrente de corta y larga memoria, $LSTM$, cuya arquitectura se representa en la Fig. B.12.

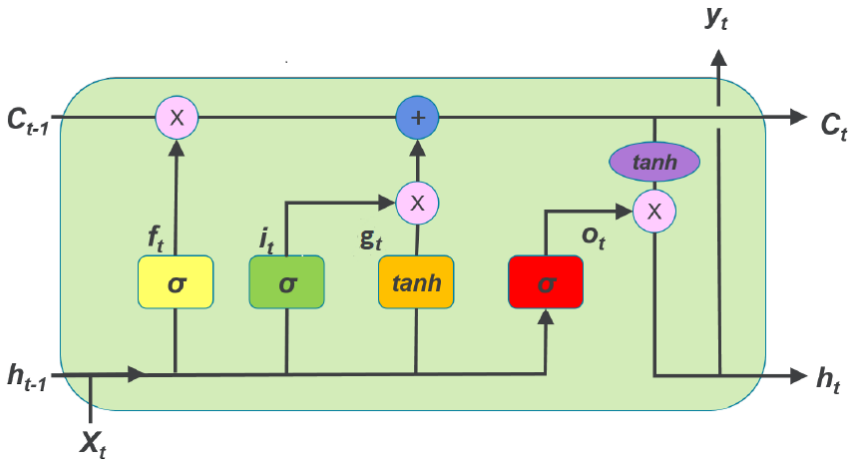

Figura B.12: Tomado de Gareth et al. (2021). Las funciones de activación σ son sigmoidales y deciden que información de x_t y h_{t-1} se mantiene para procesar la celda de estado c_t. Si las funciones sigmoidales están cerca de cero, la información en x_t y h_{t-1} se desecha; si están cerca de uno, se mantiene. Las funciones tangente hiperbólica al estar en el rango $(-1, 1)$ deciden si se toma la información de x_t y h_{t-1} de forma positiva o negativa en la celda de estado. La celda de estado alimenta la capa oculta h_t y esta última alimenta a y_t.

Dada la información de las variables de entrada, $X = [x_1, x_2, \ldots, x_T]$ y la información de las variables de salida $Y = [y_1, y_2, \ldots, y_T]$, la RNN calcula la secuencia de estados ocultos $H = [h_1, h_2, \ldots, h_T]$ y los valores ajustados $\hat{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T]$, iterativamente de forma que se minimice alguna función objetivo, e.g., la suma de errores al cuadrado en el caso de variables dependientes continuas.

La arquitectura de la RNN de corta y larga memoria (LSTM) fue propuesta por Hochreiter y Schmidhuber (1997). Como lo señalan estos autores, las redes LSTM tienen una arquitectura en la cual la información transmitida a una capa oculta es procesada iterativamente por una celda de estado c_t y dos componentes adicionales conocidos como puerta de entrada y puerta de olvido. Dadas las variables de entrada en x_t y el estado oculto en el periodo anterior h_{t-1}, la puerta de olvido se calcula como:

$$f_t = \sigma\left(w_{fx}x_t + w_{fh}h_{t-1} + b_f\right)$$

donde $\sigma(\cdot)$ es generalmente una función sigmoidal, aunque hay más posibilidades para esta función. f_t es una matriz de elementos entre cero y uno; si uno de los elementos de f_t está cerca de uno, este será un elemento que se mantendrá para los cálculos futuros de la celda de estado y si está cerca de cero, este será descartado. La puerta de entrada se calcula como:

$$i_t = \sigma\left(w_{it}x_t + w_{ih}h_{t-1} + b_i\right)$$

y

$$g_t = \tanh(w_{gx}x_t + w_{gh}h_{t-1} + b_f)$$

De nuevo, esta es una matriz con elementos entre cero y uno en el caso de i_t y entre $(-1, 1)$, en el caso de g_t. Esta decide que valores positivos o negativos de la entrada deberían ser utilizados para actualizar la celda de estado, si un elemento en i_t está cerca de uno se utilizará la mayoría de la información en g_t, y si está cerca de cero, no se utilizará la información en g_t para actualizar la celda de estado. Con c_{t-1}, f_t, i_t, y g_t la nueva celda de estado es calculada como:

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t.$$

El estado oculto se calcula recurriendo a la otra puerta, conocida como puerta de salida dada por:

$$o_t = \sigma\left(w_{ox}x_t + w_{oh}h_{t-1} + b_o\right)$$

la cual de nuevo entrega valores entre cero y uno. Si los elemntos en o_t están cerca de uno, indica que la información de la celda de estado debería ser utilizada para calcular la capa oculta y si están cerca de cero indica que tal información debería desecharse. La nueva capa oculta se calcula como

$$h_t = o_t \otimes tanh(c_t)$$

La información de esta capa oculta es enviada a la capa de salida mediante al siguiente transformación:

$$y_t = g(w_{yh}h_t + b_y)$$

donde $g(\cdot)$ aplicará una función que depende de la escala de la variable dependiente. Por ejemplo, si la variable es continua, será transformación lineal tipo:

$$y_t = b_y + w_{yh}h_t + \varepsilon_t.$$

Si estamos en un escenario de clasificación multinomial, $g(\cdot)$ será una transformación $\text{softmax}(b_y + w_{yh}h_t)$.

Como lo señala Goodfellow et al. (2016), una red LSTM puede ser implementada utilizando varias capas en cada una de las puertas de entrada, olvido y salida como también en la celda

de estado. Sin embargo, como lo señalan estos autores, esto último complica el proceso de aprendizaje (estimación) ya que la optimización de la función de pérdida (función objetivo a minimizar) se convierte en un proceso difícil. Una salida a este problema, lo sugiere Graves et al. (2013) quienes sugieren usar estructuras profundas en cada uno de los pasos iterativos de la red LSTM y sus estados de la forma:

$$h_t^n = g\left(w_{h^{n-1}h^n} h^{n-1} + w_{h^n h^n} h^n + b_n^n\right)$$

En realidad esta última ecuación corresponde al proceso de aplicar varias capas ocultas en la red LSTM, que es lo que en realidad hoy en día llamamos aprendizaje profundo (deep learning).

Para el caso de la capa de salida se tiene lo siguiente:

$$y_t = g\left(w_h^N h_t^N + b_y\right)$$

donde de nuevo $g(\cdot)$ es una función lineal para una variable dependiente con escala continua. La red LSTM, sobre todo cuando se tienen varias capas ocultas son modelos sobre parametrizados, por tal razón generalmente para evitar un poco el sobre ajuste se utilizan procedimientos de estimación penalizada tipo lasso, los detalles del procedimiento de estimación se encuentran en Goodfellow et al. (2016), uno de los textos clásicos hoy en día en el tema de redes neuronales.

B.9 Regresión Spline Adaptativa Multivariante (MARS)

El modelo MARS es muy similar al modelo GAM en el sentido de que parte de un modelo similar dado por:

$$y_t = \beta_0 + \sum_{j=1}^{L} \beta_j b_j(X) + \mu_t$$

donde

$$b_j(X) = \prod_{m=1}^{M_j} \Phi_{j,m}(X_{q(j,m)})$$

es el producto de M_j funciones multivariadas $\Phi_{j,m}(X)$, M_j es un número finito y $q(j,m)$ es un índice que depende de la m-ésima función base y la m-ésima función spline. De esta forma para cada j, $b_j(X)$ puede consistir en una sola función spline o un producto de dos o más funciones spline, y ninguna variable explicativa puede aparecer más de una vez en el producto. Estas funciones spline (para j impares) a menudo se toman como lineales de la forma $\Phi_{j,m}(X) = X - t_{l,m}$ y $\Phi_{l+1,m}(X) = (t_{l,m} - X)_+$ con:

$$(x-t)_+ = \begin{cases} x-t & si\, x > t \\ 0 & \text{en otro caso} \end{cases}$$

$$(t-x)_+ = \begin{cases} t-x & si\, x < t \\ 0 & \text{en otro caso} \end{cases}$$

donde $t_{l,m}$ es el nudo de $\phi_{l,m}(X)$ ocurriendo en uno de los valores de $X_{q(l,m)}$ con $m = 1, 2, 3, \ldots, M_t$ y $l = 1, 2, \ldots, L$.

La estrategia de construcción de MARS es como el de una regresión lineal hacia adelante. En lugar de usar las variables explicativas originales, se utilizan productos de sus funciones spline $\Phi_{l,m}(X) = (x - t_{lm})_+$, donde el spline de una variable puede aparecer sólo una vez pero se puede tener un producto de uno, dos o más splines. El modelo puede tener problemas de sobre-ajuste. Por esta razón, se minimizan criterios de validación cruzada generalizada similares al v_d en el caso del GAM. Para más detalles de esta metodología véase (Gareth et al., 2021). Similar al caso del modelo de regresión lineal, en MARS se minimiza la suma de residuales al cuadrado, pero penalizando esta última función. Similar al caso de GAM, la penalización busca combatir el sobre-ajuste.