PEDRO RODRIGUES FERRAREZI DA FONSECA

MODELOS DE PREVISÃO DE DEMANDA DE COMBUSTÍVEIS NO SETOR DE TRANSPORTES BRASILEIRO

São Paulo

2017

PEDRO RODRIGUES FERRAREZI DA FONSECA

MODELOS DE PREVISÃO DE DEMANDA DE COMBUSTÍVEIS NO SETOR DE TRANSPORTES BRASILEIRO

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Diploma de Engenheiro de Produção

Orientadora: Profa. Dra. Celma de Oliveira Ribeiro

São Paulo

2017

Catalogação-na-publicação

Fonseca, Pedro Rodrigues Ferrarezi da
Modelos de previsão de demanda de combustíveis no setor de transportes brasileiro / P. R. F. Fonseca -- São Paulo, 2017.
146 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1.Regressão Linear 2.ARIMA 3.Redes Neurais 4.Transporte 5.Combustíveis I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

À minha família.

AGRADECIMENTOS

À professora Celma, não só pela orientação em trabalhos acadêmicos, mas por todo apoio indispensável e por toda a sabedoria transmitida ao longo dos últimos anos.

Aos meus país, Pedro Fonseca e Elaine Ferrarezi, por toda a dedicação e afeto.

À minha irmã, Carolina Ferrarezi, pela parceria e por me mostrar caminhos.

Aos meus amigos da vida, por todos os momentos de felicidade e apoio incondicional.

Aos meus amigos com os quais convivi durante meu intercâmbio na França e cujos laços de amizade se mantém fortes desde então.

A todos os professores e funcionários do departamento de Engenharia de Produção.

Aos meus colegas de trabalho, por todos os ensinamentos de ordem profissional.

"*You judge a society by the decency of living of the weakest*"

Zygmunt Bauman (1925-2017)

RESUMO

A expansão vivida pelo setor de transportes deve se intensificar nos próximos anos, em função das políticas de incentivo, por parte do poder público, a grandes projetos de infraestrutura nessa área, da ainda carente estrutura logísticas do país e do crescente interesse do capital privado pelas atuais oportunidades de investimentos existentes no setor de transportes brasileiros. Nesse sentido, o objetivo desse trabalho foi avaliar o nível de atividade e as demandas energéticas do transporte nacional e compreender como funciona o mercado dos combustíveis mais demandados por esse setor no país. Por meio dessa análise, foi possível identificar variáveis quantitativas que mensuram as principais forças por trás do consumo a nível nacional dos combustíveis em estudo, o que daria fomento ao desenvolvimento de modelos de previsão da demanda desses combustíveis. A partir de um levantamento da literatura acadêmica, foram selecionadas algumas metodologias de modelagem matemática comumente utilizadas em problemas de previsão, as quais foram aplicadas nesse trabalho para a construção de modelos de previsão de demanda de gasolina C, um dos combustíveis mais utilizados pelo setor de transportes. Procurou-se, por fim, avaliar a capacidade preditiva dos modelos representantes de cada grupo de modelagem matemática por testes que simularam a aplicação dos modelos num contexto real.

Palavras-chave: Regressão Linear, ARIMA, Redes Neurais, Transporte, Combustíveis, Gasolina.

ABSTRACT

The ongoing expansion of the transportation sector is expected to ramp up over the course of the next years given the national policies to promote huge infrastructure projects, the relentless logistical hurdles in Brazil and the growing interests of private investors on the current capital allocation opportunities in the Brazilian transportation sector. In this regard, the main objective of this study was to assess the level of activity and the energy demands of the sector and to understand the dynamics behind the market of the most used fuels in transportation. From that perspective, it was possible to decide which quantitative variables are linked to the main drivers of transport fuels consumption in the country, which would be crucial for demand modeling and forecasting. Using the bibliographic survey, it was possible to select different mathematical modeling approaches commonly used for forecasting and apply them to developing demand forecasting models for gasoline, one of the most important fuels for the transportation sector. The aim was to assess the predictive quality of different mathematical modeling methods in order to propose a final forecasting model using the modeling technique whose performance was indicated as the highest by tests that simulate the actual usage of the model in a real world situation.

Key-words: Linear Regression, ARIMA, Neural Networks, Transportation, Fuels, Gasoline.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ABCR	Associação Brasileira de Concessionárias de Rodovias
ABEAR	Associação Brasileira das Empresas Aéreas
ABIFER	Associação Brasileira da Indústria Ferroviária
ANAC	Agência Nacional de Aviação Civil
ANFAVEA	Associação Nacional dos Fabricantes de Veículos Automotores
ANOVA	*Analysis of Variance* (Análise de Variância)
ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
ANTAQ	Agência Nacional do Transporte Aquaviário
ASK	*Available Seat Kilometers*
BEN	Balanço Energético Nacional
BEP	Barris Equivalentes de Petróleo
BLS	*Bureau of Labor Statistics*
CONAB	Companhia Nacional de Abastecimento
CPI	*Consumer Price Index*
CNT	Confederação Nacional de Transportes
FGV	Fundação Getúlio Vargas
IBC-Br	Índice de Atividade Econômica do Banco Central
IBGE	Instituto Brasileiro de Geografia e Estatística
IPCA	Índice de Preços ao Consumidor Amplo
Lasso	*Least Absolute Shrinkage and Selection Operator*
LF	*Load Factor*
MDIC	Ministério da Indústria, Comércio Exterior e Serviços
MLP	*Multilayer Perceptron*
MQO	Mínimos Quadrados Ordinários
PME	Pesquisa Mensal do Emprego
PNAD	Pesquisa Nacional por Amostra de Domicílio
PNLT	Plano Nacional de Logística dos Transportes
PPI	Programa de Parceria de Investimentos
RPK	*Revenue Passenger Kilometers*
tep	tonelada equivalente de petróleo
TKU	Toneladas Quilômetro Úteis

SUMÁRIO

1 INTRODUÇÃO

1.1 Descrição da área de atuação profissional do autor como estagiário

O estágio se deu em um banco de investimentos de porte global na área de *Equity Research*, cujo objetivo central é avaliar se o preço atual de ações de empresas cotadas em bolsa está de acordo com os fundamentos que a empresa apresenta e sua capacidade futura de gerar valor aos acionistas. A partir disso, as ações podem receber três classificações, que dizem respeito ao que o analista acredita serem as perspectivas futuras de valor para o papel: compra (*Buy*), neutro (*Neutral*) e venda (*Sell*). As teses de investimento são comunicadas aos investidores e clientes do banco por meio de relatórios de *research*, nos quais se explica, com dados e fatos, todo o racional por trás da tomada de decisão sobre as recomendações de compra ou venda das ações. Em geral, os analistas de ações nos bancos ficam alocados a em times que cobrem ações de setores específicos da economia, e o jargão da área utilizado para se referir a esse conjunto de ações é "cobertura". Nesse sentido, o autor fez parte do time que cobria empresas latino-americanas do setor de transportes, infraestrutura e *industrials*, entre elas CCR, Rumo S.A., Ecorodovias, LATAM Airlines e Nemak.

1.2 Motivações do estudo

O processo de análise do preço futuro de ações depende da projeção de vários fatores, como a demanda futura pelos serviços e produtos das empresas dentro da cobertura, as condições macroeconômicas que impactam o setor, a saúde financeira futura da empresa, a situação em que se encontram seus concorrentes diretos, entre outros. Diante disso, faz-se necessário o desenvolvimento de modelos de previsão que, a partir de uma base de dados histórica, consiga traduzir as dinâmicas econômicas e mercadológicas que movimentam a variável a ser estimada. A possiblidade de desenvolver modelos mais sofisticados de previsão, que pudessem fornecer estimativas mais robustas para fomentar as análises de investimentos do banco, foi uma das principais motivações para a escolha do tema do presente trabalho.

Não menos importante, porém, foi o objeto de estudo escolhido para servir de contexto de aplicação das metodologias de modelagem estudadas: a previsão da demanda nacional dos combustíveis mais utilizados pelo setor de transportes brasileiro.

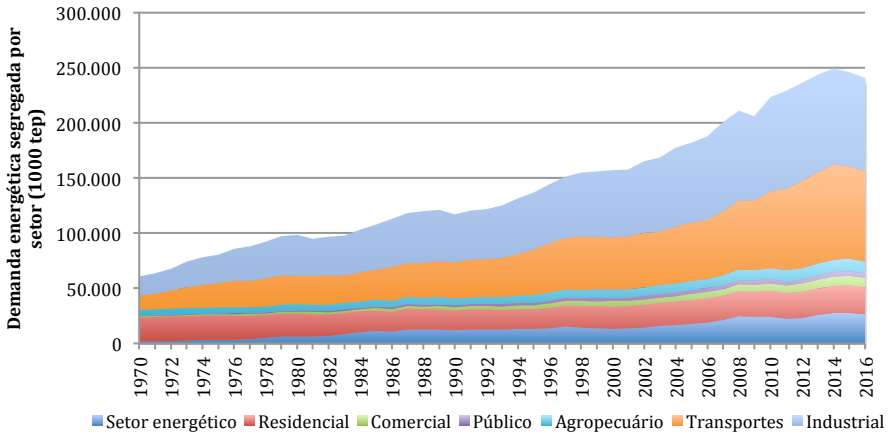

Figura 1.1 Gráfico da evolução da demanda energética nacional separada por setores da economia
Fonte: Balanço Energético Nacional de 2016

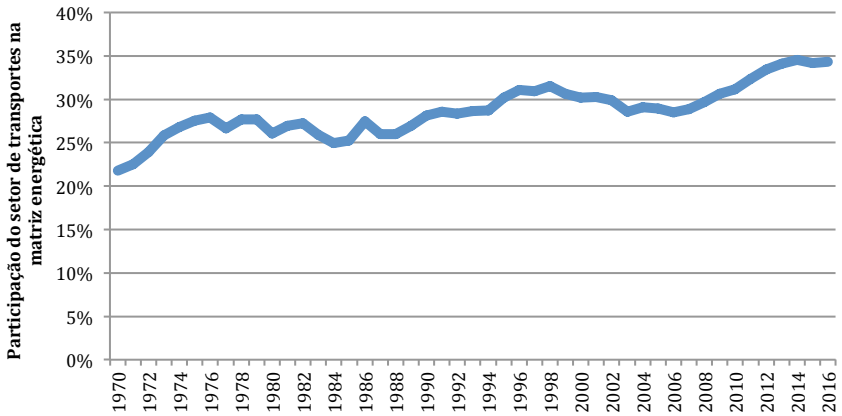

Figura 1.2 Gráfico da evolução da participação do setor de transportes na matriz energética
Fonte: Balanço Energético Nacional de 2016

A partir dos gráficos das Figuras 1.1 e 1.2, gerados com base nos dados históricos do Balanço Energético Nacional (BEN) desde 1970 até 2016, pode-se observar que o setor de transportes assume um papel cada vez mais importante dentro da matriz energética nacional, saltando de 21,7% em 1970 para 34,3% em 2016, atingindo praticamente o mesmo consumo de energia que o total da indústria. Isso equivale a uma demanda energética de 13,2 milhões de tep (ou tonelada equivalente de petróleo, que é uma medida de energia que corresponde à queima de uma tonelada de

petróleo cru e equivale a 42 bilhões de Joules) em 1970 evoluindo para 82,7 milhões de tep em 2016, um crescimento de aproximadamente 527% desde o início da série histórica.

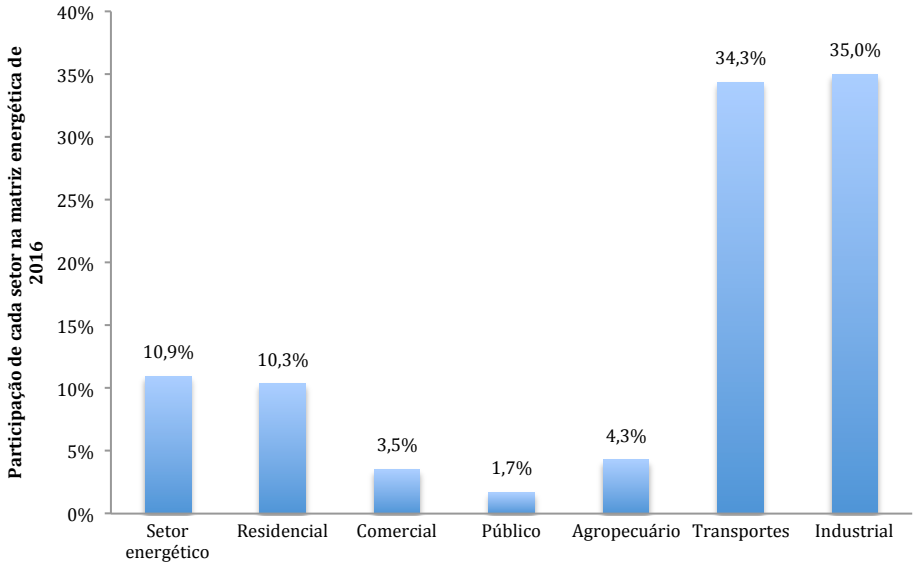

Figura 1.3 Gráfico da participação de cada setor no consumo energético total do país
Fonte: Balanço Energético Nacional de 2016

A demanda de combustíveis pelo setor de transportes deve continuar se mantendo em ritmo acelerado de crescimento tanto em termos absolutos como de participação relativa na matriz energética devido ao momento de expansão pelo qual o setor vive, com destaque para política governamental de incentivo a grandes projetos de infraestrutura. Isso atesta a atualidade do assunto e confirma a importância da compreensão, não só por parte dos e das empresas do setor de transporte, de energia e de óleo e gás como também por parte dos órgãos públicos competentes, sobre como a demanda pelos combustíveis em escala nacional vai evoluir, principalmente considerando o atual crescimento do setor de transportes. Essa informação forneceria embasamento quantitativo ao planejamento de ações públicas e privadas que garantam, em última estância, que a oferta de combustíveis consiga responder às novas necessidade energéticas nacionais.

1.3 Objetivos do trabalho

O objetivo do presente trabalho é compreender algumas abordagens de modelagem de previsão de demanda descritas na literatura e aplicar essas abordagens para construir modelos de previsão de um dos combustíveis mais relevantes para o setor de transportes: a gasolina C. A revisão da literatura fornece não somente o ferramental matemático para construção dos modelos em si, como também indica metodologias para avaliar o desempenho desses modelos frente ao problema proposto.

Também objetiva-se entender com maior profundidade o setor de transportes brasileiro e as suas demandas energéticas, o que inclui:

- Determinar os indicadores de nível de atividade mais relevantes para o setor;
- Identificar os fatores econômicos que mais o impulsionam (os chamados "drivers" do setor, segundo o jargão da área de *Equity Research*);
- Quantificar a relação da atividade do setor com a demanda nacional de combustíveis;
- Compreender as particularidades do setor de transportes no Brasil com relação ao de outros países e como isso impacta a demanda de combustíveis no mercado interno.

1.4 Estrutura do trabalho

O presente trabalho divide-se em seis capítulos. Nesse primeiro capítulo, apresentam-se os principais fatores determinantes para a escolha do objeto de estudo e quais são os objetivos que se pretende alcançar. Na sequência, o segundo capítulo consiste em uma síntese das publicações acadêmicas e dos autores que forneceram a base teórica da qual se desenvolve toda a metodologia aplicada ao longo do trabalho. Explora-se o ferramental estatístico e matemático por trás da modelagem de demanda de combustíveis e como os autores da área o aplicaram em seus respectivos contextos econômicos. Em seguida, o capítulo 3 introduz alguns conceitos sobre o setor de transportes brasileiro e como sua atividade pode ser traduzida em números, enquanto que o capítulo 4 procura compreender as fontes energéticas mais relevantes

para o setor e como se dá a dinâmica de consumo dessas fontes energéticas a nível nacional. Finalmente, a partir dos conhecimentos sobre modelagem adquiridos no capítulo 2 e com a compreensão do funcionamento do mercado dos combustíveis utilizados no setor de transportes a partir das questões levantadas nos capítulos 3 e 4, propõem-se modelos de previsão da demanda de um dos combustíveis mais importantes para o setor, a gasolina C. Esses modelos se originam de três grupos de técnicas de modelagem distintas, e, no final do capítulo 5, a performance de cada grupo de modelos é posta em comparação a partir de uma técnica que visa avaliar a capacidade desses modelos em prever no curto e longo prazo, simulando, portanto, um contexto de aplicação real desses modelos. Por fim, o capítulo 6 sumariza os principais pontos do trabalho e propõe futuras extensões sobre o tema.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo, inicia-se com a apresentação das bases teóricas por trás da modelagem das três grandes categorias de modelos que foram propostos no presente trabalho para prever a demanda volumétrica de gasolina C no país: regressão linear, modelos ARIMA e rede neural artificial. Posteriormente, identifica-se como a demanda por combustíveis, em especial gasolina C, vem sendo modelada na literatura especializada, num primeiro momento referindo-se a mercados de outros países e, em seguida, considerando o mercado brasileiro e suas particularidades.

2.1 Regressão Linear

2.1.1 Regressão Linear Simples

Um modelo de regressão genérico é uma função que descreve uma variável resposta (também chamada de dependente) Y a partir de um conjunto de variáveis explicativas (também chamadas de independentes) X. A regressão linear simples consiste em um modelo univariado na forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (2.1)$$

Sendo que:

β_0 é uma constante.

β_1 é o coeficiente linear associado à variável X.

ε é o erro associado ao modelo. O erro ε obedece as seguintes hipóteses:

1. Para qualquer par i e j, com $i \neq j$, ε_i e ε_j são independentes;
2. A variância de ε é constante no tempo;
3. A variável ε segue uma distribuição normal com média igual a zero e variância σ^2.

A determinação dos parâmetros do modelo, β_0 e β_1, é feita através do método dos mínimos quadrados ordinários (MQO), que obtém os estimadores não viesados de β_0 e β_1 que minimizam a soma quadrática dos erros do modelo.

2.1.2 Regressão Linear Multivariada

A regressão linear múltipla é um modelo linear multivariado da forma:

$$Y = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + \varepsilon \tag{2.2}$$

Em que K = o número de variáveis independentes do modelo.

Sejam Y, X, β e ε as matrizes abaixo:

$$Y = \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & (X_1)_1 & \cdots & (X_K)_1 \\ 1 & \vdots & \ddots & \vdots \\ 1 & (X_1)_n & \cdots & (X_K)_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Na forma matricial o modelo é dado por :

$$Y = X\beta + \varepsilon \tag{2.3}$$

É possível verificar que o estimador de mínimos quadrados ordinários (MQO) para os parâmetros β do modelo são obtidos através da resolução da equação matricial abaixo:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.4}$$

Em que:

X^T é a matriz transposta de X;

$(X^T X)^{-1}$ é a inversa da matriz resultante da multiplicação de $X^T X$.

Assim como na regressão linear simples, as hipóteses de normalidade, homocedasticidade e de independência dos erros também devem ser obedecidas. Porém, além dessas hipóteses, a regressão linear múltipla também pressupõe a ausência de colinearidade e multicolinearidade entre as variáveis independentes. Em outras palavras, as variáveis X do modelo não deveriam possuir correlação entre si. Os testes para diagnosticar colinearidade e multicolinearidade estão descritos no item 2.1.5.

2.1.3 Regressão *Ridge*

Segundo Gujarati e Porter (2011), é comum em aplicações de regressão linear utilizar como variáveis independentes os valores de séries temporais econométricas altamente correlacionadas entre si, uma vez que são geradas por uma a natureza econômica coincidente. Dessa forma, na prática, é difícil respeitar a hipótese de ausência de colinearidade e multicolinearidade em modelos de previsão. Ainda que quebrar esse pressuposto não significa inviabilizar o cálculo dos estimadores de MQO, segundo Reynaldo et al. (1997), a presença de multicolinearidade e colinearidade torna os valores estimados para esses coeficientes mais instáveis, de modo que pequenas mudanças nos valores das variáveis independentes geram altas diferenças nos valores estimados para esses coeficientes. Além disso, a presença de colinearidade e multicolinearidade estimula que as estimativas para a variância dos coeficientes de regressão se tornem altas, o que aumenta a possiblidade de que se aceite a hipótese nula para os coeficientes erroneamente (ou seja, indicar que o coeficiente da variável independente é estatisticamente igual a 0, ainda que não o seja).

Por essa razão, alguns autores sugerem maneiras para se trabalhar com variáveis de correlação significativa entre si em aplicações envolvendo regressão linear múltipla. Reynaldo et al. (1997) afirmam que é comum, por exemplo, a prática de eliminar algumas das variáveis do modelo que apresentem alta correlação entre si. Outra prática também recorrente, segundo os autores, é o uso de variáveis submetidas a funções de transformação, como, por exemplo, a transformação

logarítmica, que amenizam os efeitos da colinearidade e da multicolinearidade. Finalmente, uma terceira abordagem seria utilizar regularizadores que imponham restrições ao valor dos coeficientes estimados, de modo a reduzir a variabilidade de suas estimativas e evitar sobreajuste ao intervalo de calibração. Nesse trabalho, focou-se em dois métodos de regularização: a regressão *Ridge* e o *Lasso* (do inglês, *Least Absolute Shrinkage and Selection Operator*), conhecidas por regularização L2 e L1, respectivamente.

Os estimadores de MQO são obtidos através da minimização da soma quadrática dos erros do modelo de regressão linear, que assume a forma abaixo:

$$\mathrm{SQE}(\beta_0,\beta_1,\dots,\beta_N) = \sum_{i=1}^{n}\left(Y_i - \sum_{k=0}^{K}\beta_k(X_k)_i\right)^2 \tag{2.5}$$

A regressão *Ridge*, por sua vez, calcula os valores para os coeficientes lineares da mesma forma que os mínimos quadrados ordinários (minimizando a SQE), porém introduz uma restrição ao problema: a soma quadrática dos coeficientes estimados deve ser necessariamente menor que algum número real estritamente positivo:

$$\text{rs.:}\ \sum_{k=0}^{K}(\beta_k)^2 \le c^2, c \in \mathbb{R}^* \tag{2.6}$$

Resolver o problema de minimização da soma quadrática dos erros (2.5) com a restrição (2.6) é equivalente a resolver o problema abaixo, em que n é o número de observações para obter a regressão:

$$\min z(\beta_0,\beta_1,\dots,\beta_K) = \sum_{i=1}^{n}\left(Y_i - \sum_{k=0}^{K}\beta_k(X_k)_i\right)^2 + \lambda\sum_{k=0}^{K}(\beta_k)^2, \text{sendo } \lambda > 0$$

A solução do problema de minimização acima gera as estimativas dos coeficientes da regressão *Ridge*:

$$\hat{\beta}^{RIDGE} = (\mathrm{X^T X} + \lambda \mathrm{I})^{-1} \mathrm{X^T Y}, \text{ sendo } \lambda > 0 \qquad (2.7)$$

Vale destacar que se λ fosse igual a zero, o estimador de *Ridge* seria o mesmo dos que o estimador de mínimos quadrados ordinários (MQO).

A principal dificuldade ao aplicar a regressão *Ridge* é obter um valor eficiente para λ. Reynaldo et al. (1997) explicam que uma das maneiras de fazer isso é encontrar o mínimo da função do erro quadrático médio total (EQMT).

Sabe-se que a soma das variância dos coeficientes de *Ridge*, chamada de variância total, é uma função de λ e sempre diminui com o aumento de λ:

$$\mathrm{VART}(\lambda) = \hat{\sigma}^2 \sum_{\mathrm{i}=1}^{\mathrm{K}} \frac{\lambda_\mathrm{i}}{(\lambda_\mathrm{i} + \mathrm{k})^2}$$

$$\hat{\sigma}^2 = \frac{\mathrm{SQE}}{\mathrm{n} - \mathrm{K} - 1}$$

Em que:

λ_i^{-1} são os autovalores de $(\mathrm{X^T X})^{-1}$, para $\mathrm{i} = 1, \ldots, \mathrm{K}$

Já o chamado o vício-quadrado é uma função da distância dos estimadores *Ridge* aos estimadores de mínimos quadrados; portanto, sempre aumenta quanto maior o valor de λ:

$$\mathrm{VQ}(\lambda) = \lambda^2 \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\alpha^2}{(\lambda_\mathrm{i} + \lambda)^2}$$

A abordagem proposta por Reynaldo et al. (1997) é utilizar como valor para λ aquele que minimize a função EQMT, que é a soma de VQ com VART.

$$\mathrm{EQMT}(\lambda) = \mathrm{VQ}(\lambda) + \mathrm{VART}(\lambda)$$

$$\mathrm{EQMT}(\lambda) = \lambda^2 \sum_{i=1}^{N} \frac{\alpha^2}{(\lambda_i + \lambda)^2} + \hat{\sigma}^2 \sum_{i=1}^{N} \frac{\lambda_i}{(\lambda_i + \lambda)^2} \qquad (2.8)$$

Pode-se estimar o valor de λ que minimiza a função EQMT a partir uma heurística de busca de pontos de mínimo local, ou a partir de simulação.

2.1.4 Regressão *Lasso*

O *Lasso* (*Least Absolute Shrinkage and Selection Operator*) é bastante semelhante à regressão *Ridge*, já que também introduz uma penalização para o problema de minimização da soma quadrática dos erros do modelo. No entanto, como explicado por Tibshirani (1996), o termo penalizador do *Lasso* lhe confere a propriedade de zerar os coeficientes das variáveis independentes quando elas possuem baixa relevância na composição do modelo, algo que os regressores *Ridge* não fazem. Portanto, o *Lasso* pode também ser aplicado como método de seleção de variáveis independentes para compor um modelo de regressão.

A restrição do *Lasso* ao problema de minimização da soma quadrática dos erros da regressão é de que a soma dos valores absolutos estimados para os coeficientes lineares não ultrapasse algum número real estritamente positivo:

$$\text{sr.:} \sum_{k=0}^{K} |\beta_k| \leq c \,, c \in \mathbb{R}^{+*} \qquad (2.9)$$

A figura 2.1 ilustra como a restrição acima torna o *Lasso* capaz de zerar o coeficiente de alguma variável. As curvas vermelhas são as curvas de nível da função da soma quadrática dos erros da regressão e as coordenadas do ponto $\hat{\beta}$ são os valores que para β_1 e β_2 que minimizam a função quando ela não possui nenhuma restrição (ou seja, os valores dos coeficientes de MQO). É mais fácil que o ponto de mínimo da soma quadrática dos erros em função de β_1 e β_2 coincida com o eixo da maneira como a restrição é definida no *Lasso* (soma dos módulos dos coeficientes menor que algum número estritamente positivo) do que como ela é definida na

regressão *Ridge* (soma quadrática dos coeficientes menor que algum número estritamente positivo).

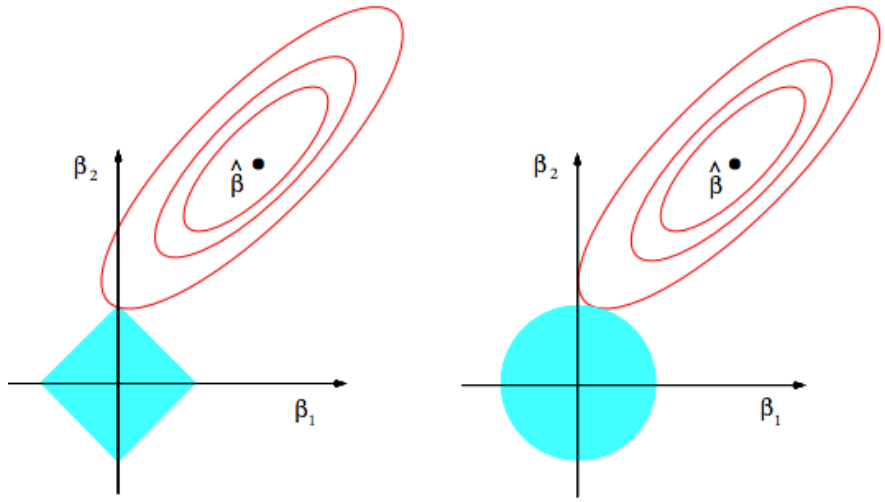

Figura 2.1 Estimativa para o coeficiente linear com a restrição do *Lasso* e da regressão *Ridge*
Fonte: Hastie, Tinshirani e Freidman (2002)

A solução do problema de minimização da soma quadrática dos erros com a restrição do *Lasso* é equivalente a resolver o problema de minimização abaixo:

$$\min z(\beta_0,\beta_1,\dots,\beta_N) = \sum_{i=1}^{n}\left(Y_i - \sum_{k=0}^{N}\beta_k(X_k)_i\right)^2 + \lambda\sum_{k=0}^{N}|\beta_k|$$

Diferentemente do que ocorre na regressão *Ridge*, o método *Lasso* não conta uma equação matricial única para calcular as estimativas dos coeficientes $\hat{\beta}_k^{Lasso}$. Deve-se aplicar, portanto, alguma heurística de minimização para obter os valores dos parâmetros.

2.1.5 Seleção e diagnóstico dos modelos de regressão linear

Análise de variância

A análise da adequação ou não de um modelo de regressão pode ser realizada através da Análise de Variância (ANOVA, sigla que vem do termo em inglês *Analysis of Variance*). Para definir as estatísticas que precisam ser calculadas para realizar essa análise, é preciso esclarecer alguns conceitos iniciais.

Na ausência de variáveis independentes para compor um modelo de regressão linear, uma maneira de modelar uma variável Y qualquer com base numa amostra de n elementos seria utilizando a média dos valores de Y da amostra. A esse modelo, dá-se o nome de modelo nulo:

$$\widehat{Y} = \sum_{i=1}^{n} \frac{Y_i}{n} = \overline{Y} \tag{2.10}$$

O erro desse modelo nulo é o que se denomina, na análise de variância, a soma quadrática total (SQT):

$$SQT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{2.11}$$

Ao criar um modelo de regressão linear com variáveis exógenas, pode-se calcular a diferença quadrática entre a estimativa desse novo modelo em cada instante t (ou seja, $\widehat{Y}_t$) com relação à estimativa do modelo nulo ($\overline{Y}$ para todos os instantes t). O valor obtido, que mede a variação devido à regressão, é chamado de soma quadrática da regressão (SQR):

$$SQR = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2 \tag{2.12}$$

Lembrando-se da equação da soma quadrática dos erros (2.11), pode-se provar a validade da equação (2.14):

$$SQE = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \tag{2.13}$$

$$SQT = SQR + SQE \tag{2.14}$$

Munido dessas definições, pode-se construir a tabela ANOVA:

Variação devido a	**Soma dos quadrados**	**Graus de liberdade**	**Quadrado Médio**	**Fator F_{obs}**
Regressão	SQR	K	$\mathrm{QMR} = \frac{\mathrm{SQR}}{\mathrm{K}}$	$\mathrm{F_{obs}} = \frac{\mathrm{QMR}}{\mathrm{QME}}$
Erro	SQE	K-n-1	$\mathrm{QME} = \frac{\mathrm{SQR}}{\mathrm{K}-\mathrm{n}-1}$	
Total	SQT	n-1		

Tabela 2.1 Tabela ANOVA
Fonte: Elaborado pelo autor

Sabe-se que o fator F segue uma distribuição de Fischer. O valor dessa estatística permite realizar o teste de hipótese abaixo, que é o chamado teste F. Caso a hipótese nula seja aceita, significa que o modelo de regressão linear não é significativamente superior do que a média amostral ($\overline{\mathrm{Y}}$) para modelar a variável dependente, indicando que o modelo não é adequado.

$$\begin{cases} \mathrm{H_0}: \beta_\mathrm{k} = 0, \mathrm{k} = 1{,}2 \ldots, \mathrm{K} \\ \mathrm{H_1}: \text{caso contrário} \end{cases}$$

Análogo ao teste F, tem-se o teste F-parcial. Esse teste avalia estatisticamente se a contribuição da inclusão de uma ou mais variáveis independentes ao modelo de regressão linear é significativa, a partir do cálculo da estatística $\mathrm{F_{parcial}}$ associada a essa inclusão.

$$\mathrm{F_{parcial}} = \frac{\frac{\mathrm{SQR_{restrito}} - \mathrm{SQR_{irrestrito}}}{\mathrm{g}}}{\frac{\mathrm{SQR_{irrestrito}}}{\mathrm{n}-\mathrm{K}}} = \frac{\mathrm{QMR_{adicional}}}{\mathrm{QMR_{irrrestrito}}} \qquad (2.15)$$

em que:

$\mathrm{SQR_{restrito}}$ e $\mathrm{SQR_{irrestrito}}$ são as somas quadráticas dos resíduos dos modelos restrito e irrestrito, respectivamente;

g é o número de restrições do modelo restrito em relação ao irrestrito e K o número de variáveis independentes do modelo restrito.

Finalmente, outro teste que vale ser destacado é o teste t, que é usado principalmente para verificar se um coeficiente β_k associado a uma variável k é significativamente diferente de zero (ainda que possa ser utilizado para provar se o coeficiente é significativamente diferente de qualquer número real b, num caso mais geral):

$$\begin{cases} H_0: \beta_k = b \\ H_1: \beta_k \neq b \end{cases}$$

O Stepwise

A partir da definição do teste F-parcial, pode-se definir os passos do método *Stepwise* de seleção de variáveis. O *Stepwise* baseia-se no cálculo da estatística F-parcial associada à saída ou à entrada de uma variável num modelo de regressão linear. A rotina que se seguirá nesse trabalho para aplicar o *Stepwise* pode ser resumida nos seguintes passos:

1. Calcular a correlação entre a variável resposta e todas as variáveis externas candidatas a serem incluídas no modelo;
2. Montar um modelo de regressão linear simples com a variável de maior correlação (que será chamada de X_1).
3. Calcular o valor da estatística F para essa variável. Se $F_1 > F_{crit}^{ent}$, incluir X_1 no modelo e ir para o passo 6. Caso contrário, aceita-se a hipótese nula (nenhuma variável entra e devemos modelar a variável resposta pela media das observações) e finaliza-se o algoritmo;
4. Obter a correlação parcial de todas as variáveis restantes para com a variável resposta. Essa correlação mede o quanto da variação da variável resposta que não é explicada pelas variáveis independentes já inclusas no modelo pode ser explicada pela inclusão de uma determinada variável independente nova. Calcular a correlação parcial é equivalente a calcular a correlação dos resíduos do modelo com os

resíduos de um modelo de regressão entre a variável candidata a ser incluída no modelo e as que já o compõem.

5. Passo "*forward*": selecionar a variável X_i com maior correlação parcial e calcular o valor da estatística F-parcial associada à sua inclusão no modelo. Se $F_i > F_{crit}^{ent}$, incluir X_i no modelo e reajustar os coeficientes lineares com a entrada da nova variável independente. Ir para o passo 6.
6. Passo "*backward*": calcular a estatística F para todas as variáveis do modelo. Selecionar a variável X_j com menor F. Se $F_j < F_{crit}^{sai}$, retirar X_j no modelo e ajustar os coeficientes das variáveis remanescentes. Ir para o passo 5.

O processo de análise de resíduos

Para que os resultados de ajuste do modelo de regressão linear sejam confiáveis, é necessário analisar os resíduos produzidos pelo modelo e verificar se valem as hipóteses:

- Independência dos erros (ausência de autocorrelação);
- Homocedasticidade (variância constante) e;
- Normalidade: $\varepsilon \sim N(0, \sigma^2)$.

O diagnóstico de independência dos erros pode ser realizado por um gráfico de dispersão que relacione o valor dos resíduos com a ordem em que eles se apresentam. Nesse trabalho, os resíduos dos modelos seguem uma ordem temporal. Caso eles sejam independentes, espera-se que, ao traçar o gráfico, os pontos estejam distribuídos aleatoriamente, não devendo haver nenhum tipo de tendência. Caso isso ocorra, uma das hipóteses sobre os resíduos da regressão linear não estaria sendo respeitada, indicando que o modelo não é adequado. Também há testes estatísticos mais objetivos para diagnosticar de independência dos erros. Um deles é chamado de teste de Durbin-Watson, que verifica a presença de autocorrelação de primeira ordem. Na prática, no entanto, a autocorrelação pode ocorrem em ordem mais avançadas. Nesse caso, é interessante sempre olhar para a função de autocorrelação da amostra (fac), definida no item 2.2.2., para verificar a presença de autocorrelação

em ordens mais elevadas.

Assim como o diagnóstico de independência dos resíduos, o diagnóstico de homocedasticidade pode-se ser realizado por meio de análise gráfica: ao traçar os erros do modelo em função dos valores ajustados ($\widehat{Y}$), caso não haja nenhuma tendência aparente no gráfico, sendo os pontos distribuídos aleatoriamente, tem-se um indício de variância constante. Também existem testes estatísticos mais objetivos para detectar a presença ou não de homocedasticidade nos resíduos, como teste de Breusch-Pagan, que consiste basicamente em fazer uma regressão linear simples entre a variável de valores ajustados e uma variável de erros padronizados e testar se a regressão é significativa.

Por fim, para testar se os resíduos estão normalmente distribuídos em torno do zero, pode-se recorrer a uma abordagem gráfica como o uso papel de probabilidade normal, por exemplo. Ao traçar os resíduos no papel de probabilidade normal, confirma-se a hipótese de normalidade se resíduos formarem uma reta. Como abordagem estatística, pode-se citar o teste de Kolmogorov-Smirnov, que consiste em comparar a função de distribuição acumulada empírica dos erros com os valores da função de distribuição acumulada da normal teórica.

Diagnóstico de colinearidade e multicolinearidade

Partindo-se da definição adotada por Reynaldo et al. (1997) de que colinearidade é a existência de correlação entre cada par de variável independente, o grau de colinearidade pode ser avaliado através de uma matriz de correlação:

$$\rho = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,K} \\ \rho_{2,1} & 1 & \ddots & \rho_{2,K} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{K,1} & \rho_{K,2} & \cdots & 1 \end{bmatrix}$$

$$\rho_{a,b} = \frac{\sum_{i=1}^{n}(X_{a,i} - \bar{X}_a)\,(X_{b,i} - \bar{X}_b)}{(n-1)s_a s_b} \tag{2.16}$$

Quanto mais o valor absoluto das correlações entre duas variáveis independentes ($|\rho_{a,b}|$, sendo $a \neq b$), se aproximarem de 1, maior é o grau de colinearidade entre elas. Testes de hipótese podem ser feitos para verificar quais correlações são significativamente diferentes de zero.

Com relação à multicolinearidade, Reynaldo et al. (1997) adotam que esse termo denota a existência de uma relação linear aproximada entre as variáveis independentes do modelo. Segundo os autores, essa característica pode ser atestada por meio dos fatores de inflação da variância, denominados de VIF (do inglês, *variance inflation factor*). Esse valor é associado a cada variável independente e mede o grau de correlação da variável com as demais variáveis independentes do modelo. Esse indicador é numericamente equivalente aos valores da diagonal principal da matriz $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$, que faz parte da equação matricial para estimação dos coeficientes lineares.

$$\mathrm{VIF} = \mathrm{diag}((\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}) \tag{2.17}$$

Segundo Reynaldo et al. (1997), assumem-se problemas de multicolinearidade significativos se algum elemento do vetor VIF for superior a 10.

2.2 Modelos ARIMA

Modelos ARIMA e seus derivados vem sendo utilizados na literatura sobre modelos de previsão de demanda, e seus resultados são frequentemente comparados com os obtidos por meio de regressão linear e outros tipos de modelagem. Nesse sentido, vale destacar os trabalhos de Suganthi et al. (2012) e Ediger et al. (2007), os quais trataram de modelos de previsão de demanda no mercado de combustíveis, o primeiro no mercado chinês e o segundo no mercado turco.

Nesse item, inicia-se o tema apresentando alguns conceitos importantes para a modelagem ARIMA (como os conceitos de processos estocásticos e estacionários, função de autocorrelação, entre outros). Na sequência, seguindo a abordagem de Morettin e Toloi (2006), é introduzida a noção de modelos auto-regressivos (AR) e

modelos de médias móveis (MA), para posteriormente definir os modelos deles derivados, como ARMA (modelos auto-regressivos e de médias móveis), ARIMA (modelos auto-regressivos integrados de médias móveis) e SARIMA (ARIMA sazonal).

2.2.1 Processos estocásticos e processos estacionários

Um processo estocástico é um conjunto de variáveis aleatórias que evoluem em função do tempo. É importante destacar a diferenciação desse tipo de processo com um processo determinístico, já que, no caso deste último, conhecendo-se as condições de contorno iniciais, pode-se obter com certeza toda a trajetória das variáveis no tempo. As variáveis de um processo estocástico, por sua vez, possuem várias (às vezes infinitas) evoluções possíveis ao longo do tempo.

Os processos estocásticos podem ser subdivididos em processos estacionários e não estacionários. Um processo estacionário é aquele no qual as propriedades estatísticas das variáveis aleatórias do processo não se alteram em função do tempo. Ou seja, sendo Z_1, Z_2, ..., Z_n um conjunto de séries que compõem juntas um processo estocástico, obter-se-iam as mesmas propriedades estatísticas nessas séries se elas fossem transladadas no tempo. Em outras palavras, sendo F a função de distribuição conjunta das variáveis do processo estocástico e h um número qualquer pertence ao conjunto dos números inteiros:

$$F(x_1, x_2, \dots, x_n) = F(x_{1+h}, x_{2+h}, \dots, x_{n+h})$$

Disso decorre que média e variância num processo estocástico estacionário é constante em relação ao tempo.

2.2.2 Função de autocovariância (facv), função de autocorrelação (fac) e função de autocorrelação parcial (facp)

Como será visto posteriormente, para a identificação dos modelos ARIMA mais adequados para modelar uma determinada série temporal, deve-se fazer uso de uma metodologia que depende da identificação dos comportamentos das funções de autocorrelação e de autocorrelação parcial da série e da comparação com os comportamentos dessas funções em modelos teóricos de ARIMA. Faz-se necessário, portanto, definir essas duas funções. Para isso, inicia-se com a apresentação da função de autocovariância. Baseando-se no trabalho e nas notações utilizadas por Morettin e Toloi (2006), define-se a função de autocovariância (facv) como sendo:

$$\gamma_\tau = E\{X_t X_{t+\tau}\} \tag{2.18}$$

em que $\{X_t, t \in Z\}$ trata-se de um processo estacionário real discreto de média zero.

Morettin e Toloi também elucidam que a facv γ_τ tem as seguintes propriedades

(i) $\gamma_0 > 0$,
(ii) $\gamma_{-\tau} = \gamma_\tau$
(iii) $|\gamma_\tau| \leq \gamma_0$
(iv) γ_τ é não negativa definida, no sentido que

$$\sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \gamma_{\tau_j - \tau_k} \geq 0$$

para quaisquer números reais $a_1, \ldots, a_n$ e $\tau_1, \ldots, \tau_n$ de Z.

Já a função de autocorrelação (fac) é definida como

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{2.19}$$

em que $\gamma_0 = E\{X_t^2\} = \text{Var}(X)$.

A função de autocorrelação tem as mesmas propriedades que a função de autocovariância.

Finalmente, a função de autocorrelação parcial (facp), proposta por Box, Jenkins e Reinsel (1994), é obtida a partir da resolução das equações de Yule-Walker. Como será visto posteriormente no item 2.2.3, essas equações assumem a forma de um modelo AR(k):

$$\rho_j = \emptyset_{k1}\rho_{j-1} + \emptyset_{k2}\rho_{j-2} + \cdots + \emptyset_{kk}\rho_{j-k}, j = 1, \ldots, k$$

As equações acima formam um sistema linear. O valor calculado para $\emptyset_{kk}$ é o valor da função de autocorrelação parcial. Morettin e Toloi (2006) adotam a seguinte notação simplificadora:

$$\phi_{kk} = \frac{|P_k^*|}{|P_k|} \tag{2.20}$$

em que P_k é a matriz de autocorrelações e P_k^* é a matriz P_k com a última coluna substituída pelo vetor de autocorrelações.

2.2.3 Modelos auto-regressivos – AR(p)

Um modelo auto-regressivo de ordem p, doravante denominado AR (p), trata-se de uma equação linear discreta com erro que assume a forma:

$$\tilde{X}_t = \alpha_1\tilde{X}_{t-1} + \cdots + \alpha_p\tilde{X}_{t-p} + \varepsilon_t$$

no qual $\tilde{X}_t = X_t - \mu$, sendo μ a média da processo X_t e ε_t é um erro, cujas hipóteses são similares às dos resíduos da regressão linear: ε_t é uma variável aleatória, com

média igual a zero e variância constante (homocedasticidade), que não apresenta autocorrelação.

Geralmente, a equação de um modelo auto-regressivo costuma ser apresentada utilizando-se de operadores de defasagem. Seja S o espaço de todas as sequências $Y^t, t \in Z$ possíveis com números pertencente ao conjunto dos números reais. Definimos o operador de defasagem L a seguinte transformação $L: S \rightarrow S$:

$$L^k Y^t = Y_{t-k}, \forall\, t \in Z \tag{2.21}$$

Logo, utilizando-se dos podemos operadores de defasagem L, podemos escrever a equação dos modelos auto-regressivos AR(p) da seguinte forma:

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) \tilde{X}_t = \varepsilon_t, \qquad \forall\, t \in Z \tag{2.22}$$

2.2.4 Modelos de médias móveis – MA(q)

Nos modelos de médias móveis MA(q), o elemento da série no instante t é descrito a partir dos erros ponderados de q instantes anteriores a t:

$$\tilde{X}_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}$$

A equação acima também pode ser reescrita a partir dos operadores de defasagem (2.21):

$$\tilde{X}_t = \left(1 + \sum_{k=1}^{1} \beta_k L^k\right) \varepsilon^t, \qquad \forall\, t \in Z \tag{2.23}$$

2.2.5 Modelos auto-regressivos e de médias móveis – ARMA(p,q)

Nos modelos auto-regressivos e de médias móveis ARMA(p,q), um item da série no instante t é calculado a partir dos erros ponderados de q instantes anteriores a t e da média ponderada dos valores em si da série em p instantes anteriores:

$$\widetilde{X}_t = \alpha_1 \widetilde{X}_{t-1} + \cdots + \alpha_p \widetilde{X}_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \cdots - \beta_q \varepsilon_{t-q}$$

Novamente, pode-se reescrever a equação acima utilizando-se dos operadores de defasagem (2.21):

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) \widetilde{X}_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon^t, \forall\, t \in Z \qquad (2.24)$$

Para simplificação, costuma-se utilizar o operador auto-regressivo de ordem e o operador de médias móveis de ordem q:

$$\phi(L)\widetilde{X}_t = \theta(L)\varepsilon^t \qquad (2.25)$$

2.2.6 Modelos auto-regressivos integrados de médias móveis – ARIMA(p,d,q)

Morettin e Toloi (2006) explicam que os modelos ARMA(p,q) são adequados apenas para modelar processos estacionários. As funções de autocovariância, por exemplo, da maneira como foram definidas no item 2.2.2., são supostas a não mudar com o decorrer do tempo, o que não ocorreria num processo não estacionário.

O problema é que a maioria das séries temporais que descrevem elementos econômicos e financeiros (como PIB, índices acionários, etc.) possuem tendência e sazonalidade, o que descarta a possibilidade de se tratarem de séries temporais estacionárias. Disso decorre a necessidade de tornar essas séries estacionárias para que se possa utilizar os modelos ARMA.

Segundo Morettin e Toloi (2006), uma das maneiras de se obter processos estacionários a partir de séries temporais não-estacionárias envolve diferenciar essas

séries. Uma vez estacionárias, pode-se modelar essas séries utilizando-se dos modelos ARMA. É preciso ressaltar, no entanto, que apenas algumas séries não-estacionárias se tornarão estacionárias por meio de diferenciação. Essas séries são chamadas de não-estacionárias homogêneas e poderão ser modeladas com uma categoria de modelos chamada ARIMA. Portanto, se a série temporal X_t for não-estacionária homogênea, há um valor d que pertence aos números naturais com o qual se obtém uma série estacionária W_t passível de ser modelada por uma ARMA.

$$W_t = \Delta^d X_t = (1 - L)^d X_t$$

Como $\Delta^d \widetilde{X}_t = \Delta^d X_t$ para todo e qualquer d > 1, pode-se substituir o valor $\widetilde{X}_t$ do modelo ARMA por $(1 - L)^d X_t$. Dessa forma, obtém-se o modelo genérico ARIMA(p,d,q):

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right)(1 - L)^d X_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right)\varepsilon^t, \forall\, t \in Z \qquad (2.26)$$

2.2.7 Modelos SARIMA (ARIMA Sazonal)

Quando se calcula a função de autocorrelação de uma determinada série temporal, é possível que ocorram algumas das situações abaixo, as quais inviabilizam a hipótese de série estacionária necessária para a aplicação da abordagem de modelos ARMA:

1. Autocorrelação significativa para "*lags*" de ordens mais baixas. Esse problema pode ser resultado da presença de tendência na série e geralmente pode ser resolvido por meio de diferenciação;
2. Autocorrelação significativa para "*lags*" espaçados em múltiplos de um determinado número natural s, que indica sazonalidade na série. Existem algumas maneiras de tentar eliminar essa sazonalidade, que variam para o caso de ela ser determinística (não varia com relação ao instante t da série) ou estocástica (varia em função do tempo).

Seja Z_t uma série temporal que possua um caráter sazonal com período igual a s. Supondo que o comportamento não-estacionário da série se dê por questões de sazonalidade, uma maneira que pode fazer com que a série se torne estacionária é subtraindo-lhe um componente μ_t, que conterá o caráter sazonal da série original.

$$N_t = Z_t - \mu_t \tag{2.27}$$

No caso da sazonalidade determinística, μ_t é uma função determinística periódica (ou seja, a sequência de valores de $\mu_1, \mu_2, \ldots, \mu_s$ vai se repetir para todos os ciclos sazonais da série):

$$\mu_t - \mu_{t-s} = 0 \tag{2.28}$$

A equação acima pode ser reescrita com operadores de defasagem:

$$(1 - L^s)\mu_t = 0$$

Com isso, se cada elemento da série Z_t for subtraído pelo valor μ_t que lhe for correspondente, obtém-se uma nova série N_t a qual, no caso de sazonalidade determinística, será um processo estacionário.

A solução de μ_t para o caso em que s = 12 segue a equação abaixo:

$$\mu_t = \mu + \sum_{j=1}^{6}\left[a_j \cos\frac{(2\pi jt)}{12} + b_j \operatorname{sen}\frac{(2\pi jt)}{12}\right] \tag{2.29}$$

Para se determinar os valores de μ_t que melhor se ajustam a uma determinada série temporal, determina-se preliminarmente os parâmetros $\hat{\mu}, \hat{a}_j$ e $\hat{b}_j$ por meio de uma regressão de Z_t com as variáveis $1, \cos\frac{(2\pi jt)}{12}$ e $\operatorname{sen}\frac{(2\pi jt)}{12}$, para j=1,...,6. Com isso, tem-se uma estimativa inicial para a série μ_t, que se chamará $\hat{\mu}_t$, com a qual obtém-se $\widehat{N}_t$. Na sequência, é preciso ajustar essa nova série dessazonalizada $\widehat{N}_t$ a

um modelo ARMA(p,q). Somando-se as duas séries modeladas $\widehat{N}_t$ e $\hat{\mu}_t$, obtém-se um primeiro modelo para Z_t, cujos parâmetros podem passar por um processo de refinamento conjunto posteriormente.

Em alguns casos, porém, uma série μ_t que se repete igualmente a cada ciclo s de períodos sem alterações no decorrer do tempo pode não conseguir transformar a série Z_t em uma série estacionária. Se for o caso, possivelmente a sazonalidade se altera em função do tempo e, por isso, não se pode modelar a sazonalidade como uma função determinística.

No caso da sazonalidade estocástica, a equação 2.27 não será mais igual a zero, mas sim a um processo estocástico Y_t.

$$(1 - L^s)\mu_t = Y_t \tag{2.30}$$

Multiplicando-se o operador de defasagem de um período $(1 - L^s)$ na equação (2.27) tem-se que:

$$(1 - L^s)\mu_t + (1 - L^s)N_t = (1 - L^s)Z_t \tag{2.31}$$

Substituindo (2.30) em (2.31), obtém-se:

$$Y_t + (1 - L^s)N_t = (1 - L^s)Z_t \tag{2.32}$$

com Y_t e N_t representados por modelos ARMA:

$$\phi_N(L)N_t = \Phi_N(L)\varepsilon^t$$
$$\phi_Y(L)Y_t = \Phi_Y(L)a^t$$

Dessa forma, pode-se provar que, de (2.32), obtém-se:

$$\left(1 - \Phi_1 L^s - \cdots - \Phi_p L^{sP}\right)(1 - L^s)^D(1 - L)^d Z_t = \left(1 - \Theta_1 L^s - \cdots - \Theta_p L^{sP}\right)\alpha^t$$

Essa expressão pode ser reescrita com o operador auto-regressivo sazonal de ordem P, o operador de médias móveis sazonal de ordem Q e com o operador diferença sazonal, sendo D o número de diferenças sazonais:

$$\Phi(L^s)\Delta_s^D\Delta^d Z_t = \Theta(L^s)\alpha^t \tag{2.33}$$

Partindo-se da suposição que o processo α^t possa ser ajustado a um ARIMA(p,d,q) tal que:

$$\emptyset(\mathrm{B})\Delta^d = \theta(\mathrm{B})a^t \tag{2.34}$$

Substituindo-se (2.33) em (2.34), obtém-se:

$$\emptyset(\mathrm{B})\Phi(\mathrm{L}^s)\Delta_s^{\mathrm{D}}\Delta^{\mathrm{d}}\mathrm{Z}_t = \theta(\mathrm{B})\Theta(\mathrm{L}^s)\mathrm{a}^t \tag{2.35}$$

A equação (2.35) é a que descreve o modelo chamado de ARIMA sazonal multiplicativo, que costuma ser representado pela notação SARIMA(p,d,q)x$(P,D,Q)_s$.

2.2.8 Construção dos modelos ARIMA – Metodologia de Box & Jenkins (1970)

Seguindo a sistemática introduzida por Box e Jenkins (1970), Morretin e Toloi (2006) detalham uma metodologia para obtenção dos modelos da classe ARIMA. Essa metodologia possui quatro etapas brevemente explicadas abaixo:

1. Especificação das classes de modelos a ser consideradas para modelar o problema (no caso, seriam os modelos ARIMA e seus derivados);
2. Identificação o modelo ARIMA que será utilizado, a partir da análise das funções de autocorrelação e autocorrelação parciais da série temporal que se pretende modelar;
3. Estimação dos parâmetros do modelo (ajuste aos dados de calibração por métodos como mínimos quadrados);

4. Diagnóstico do modelo ajustado (por meio da análise de resíduos, verificando se eles seguem a hipótese de ruído branco).

Desses quatro passos, é preciso destacar a etapa de identificação, que consiste, no caso dos modelos ARIMA especificados no presente trabalho, em determinar os valores para p, d, q, P, D, Q e s que melhor se ajustam à série temporal em estudo.

Como introduzir mais parâmetros do que o necessário para modelar uma série pode gerar problemas de *overfitting* (sobreajuste), dificultando a capacidade de generalização do modelo ARIMA para observações fora do período de calibração, a etapa de identificação é de fundamental importância.

O procedimento de identificação de modelos da classe ARIMA segue a lógica abaixo:

1. Verificar a função de autocorrelação da série temporal que se pretende modelar. Caso a função apresente um decaimento lento com o aumento do "*lag*", significa que a série não segue um processo estacionário.
2. Em caso de não estacionaridade, é preciso diferenciar a série quantas vezes forem necessárias até que ela se torne estacionária. O número de diferenças necessárias com relação a $\Delta = (1 - L)$ até que a série se torne estacionária é igual ao valor para d. Se o mesmo padrão de decaimento lento da fac se verificar em "*lags*" sazonais espaçados por s períodos, isso indica a presença a sazonalidade. Nesse caso, é preciso diferenciar a série com respeito a $\Delta_{12} = (1 - L^{12})$ quantas vezes forem necessárias até torná-la estacionária. O número de vezes equivale ao valor de D.
3. Com a série estacionária, calculam-se as funções de autocorrelação e autocorrelação parcial amostrais e verifica-se em que "*lags*" elas se mostram significativas e em que "*lags*" elas decaem. O comportamento da fac em 1,2,3… e em s,2s,3s,… determina os valores de p e P, respectivamente. Já o comportamento da facp em 1,2,3… e em s,2s,3s,… determina os valores de q e Q, respectivamente.

Uma vez identificado o modelo da classe ARIMA que se deve utilizar em função da amostra de dados de calibração, deve-se estimar os valores dos parâmetros do modelo, o que pode ser feito pelos estimadores de mínimos quadrados ordinários.

Finalmente, para diagnosticar o modelo ARIMA, deve-se que proceder com a mesma análise de resíduos da regressão linear, de modo a verificar se eles estão de acordo com as hipóteses de formulação do modelo: se o modelo proposto for adequado, os seus resíduos devem apresentar o comportamento de um ruído branco, o que significa que não podem estar correlacionados entre si, devem seguir uma distribuição normal em volta do zero e apresentaram caráter homocedástico. O não cumprimento dessas hipóteses ameaça a capacidade preditiva dos modelos ARIMA. O diagnóstico serve não somente para invalidar um modelo como também para ajudar a identificar pontos de mudança que possam ser feitos para corrigi-lo.

2.3 Redes Neurais Artificais

Segundo Hill, O'connor, Remus (1996) e Kovács (1997), as redes neurais artificias são algoritmos computacionais não-lineares de alta performance. Os autores afirmam que a aplicação de redes neurais para construção de modelos de previsão frequentemente supera outras modelagens mais tradicionais, como ARIMA e regressão linear, principalmente em problemas de elevada complexidade, com um conjunto de dados e variáveis extensos. Isso se deve ao fato de que o uma rede neural é composta por um conjunto de funções não-lineares que, em conjunto, são dotadas de alta capacidade de adaptação e generalização. A inspiração da denominação das redes neurais vem justamente da comparação com o cérebro humano, tanto no que tange à sua composição (já que a rede é formada por nós interconectados, de estrutura análoga às células nervosas) como também à sua capacidade primordial, que consiste no reconhecimento de padrões e classificação e na capacidade de generalização e flexibilidade (Wasserman, 1989).

Redes neurais têm sido utilizadas com sucesso como modelos para previsão. Zhang, Patuwo e Hu (1998) realizaram uma revisão bibliográfica bastante completa

dos variados tipos de redes neurais que podem ser aplicadas em modelos de previsão. Analisando o trabalho de alguns autores que haviam utilizado redes neurais nos mais variados tipos de problemas de previsão, Zhang, Patwo e Hu (1998) puderam delimitar alguns conjuntos de características das redes utilizadas para esse fim, as quais serão detalhadas a fundo nos próximos itens desse trabalho.

2.3.1 Os neurônios e a topologia de uma rede neural

Uma rede neural é constituída por um conjunto de nós interligados entre si, chamados de neurônios, os quais estão dispostos em camadas. Excluindo-se os nós da primeira camada, que recebem e emitem diretamente o valor das variáveis independentes, os nós são funções matemáticas do tipo $\varphi(x) = y$, na qual x é valor gerado a partir das entradas recebidas dos nós anteriores, e y é o valor que o nó gera como saída, como pode ser observado na figura 2.2.

Os elos entre neurônios conectados são chamados de sinapses ou elos de conexão. A cada sinapse entre dois neurônios, é atribuído um peso w. A soma de todos os sinais que saem dos neurônios precedentes, multiplicados pelo respectivo peso de cada elo sináptico, somados, ainda, um viés constante b, é o valor x da função $\varphi(x) = y$ do neurônio em questão.

$$x_j = \sum_{k=1}^{n} i_k w_{kj} + b_j \tag{2.36}$$

em que n é o número total de neurônios k que antecedem o neurônio j, e b_j é o viés do neurônio j e i_k são as saídas dos neurônios k. É importante frisar que os pesos w associados a cada sinapse, mais os vieses b de cada neurônio fora da primeira camada são os parâmetros do modelo de redes neurais que precisam ser estimados.

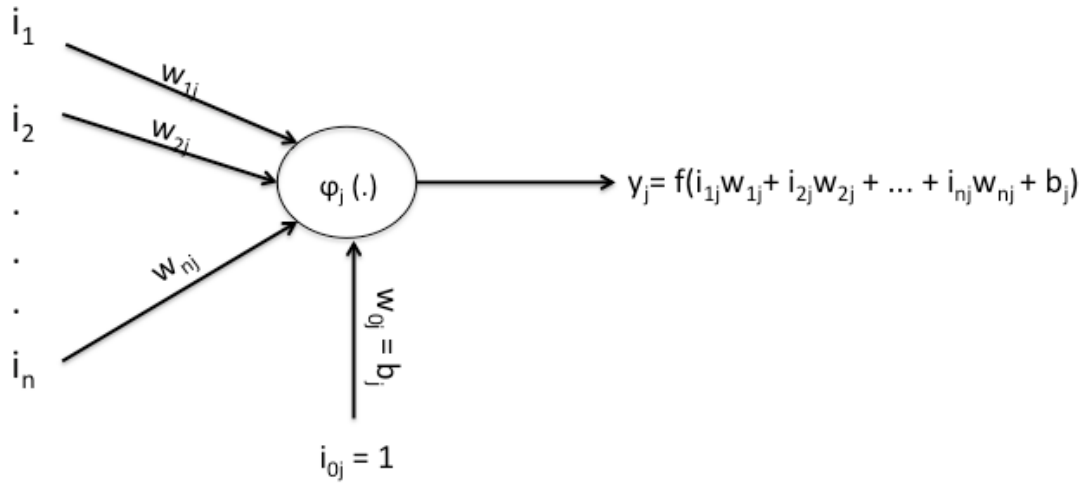

Figura 2.2 Esquema genérico de um neurônio transformador
Fonte: Elaborado pelo autor

As funções do tipo φ(x) = y presentes nos neurônios transformadores são chamadas de funções de ativação ou de transferência. Na tabela abaixo, têm-se as algumas funções de ativação presentes na literatura (Karlik e Olgac, 2011).

Função identidade	$\varphi(x) = x$
Função degrau	$\varphi(x) = \begin{cases} 0, \text{se } x < 0 \\ 1, \text{se } x \geq 0 \end{cases}$
Função ReLU	$\varphi(x) = \begin{cases} 0, \text{se } x < 0 \\ x, \text{se } x \geq 0 \end{cases}$
Função logística (ou função sigmoidal unipolar)	$\varphi(x) = \frac{1}{1 + e^{-x}}$
Função sigmoidal bipolar	$\varphi(x) = \frac{1 - e^{-x}}{1 + e^{-x}}$
Função logística	$\varphi(x) = \frac{1}{1 + e^{-x}}$
Função tangente hiperbólica	$\varphi(x) = \tanh(x)$
Função arco tangente	$\varphi(x) = \tan^{-1}(x)$

Tabela 2.2 Funções de ativação
Fonte: Elaborado pelo autor, modificado de Karlik e Olgac (2011)

Karlik e Olgac (2011) conduziram um estudo de comparação das performances entre modelos de rede neural que se diferenciaram apenas devido às funções de ativação não-lineares que possuíam. Resumidamente, os autores compararam o erro

obtido por essas várias redes neurais para modelar uma variável resposta. No cenário de menor número de iterações e com poucos neurônios na camada intermediária, as funções logísticas, bipolar sigmoidal e tangente hiperbólica não apresentaram diferenças significativas de performance. Porém, a tangente hiperbólica tanto no neurônio das camadas intermediárias quanto da última camada teve um resultado ligeiramente superior às demais funções de ativação no caso de uma rede com mais neurônios e maior número de iterações.

A maneira como os neurônios estão conectados entre si caracteriza a topologia (ou arquitetura) de uma rede neural. Haykin (1999) catalogou as arquiteturas de redes neurais e, para esse trabalho, dá-se destaque a uma categoria em específico, a chamada *Multilayer Perceptron* (MLP) com uma camada intermediária. Essa arquitetura de rede se caracteriza por ter três camadas no total:

1. Cada neurônio da primeira camada recebe o valor de uma das variáveis de entrada e não realiza nenhum tipo de transformação;
2. Cada neurônio da segunda camada (camada escondida ou intermediária) recebe de cada neurônio da primeira camada um sinal sináptico equivalente ao valor da variável de entrada e, a partir desses sinais e da função de ativação, gera um sinal sináptico para o(s) neurônio(s) da camada final;
3. A camada final possui o número de neurônios igual ao número de variáveis de saída. Esses neurônios também possuem funções de ativação, e a saída deles é o valor ajustado da rede para a variável de saída.

Outro fator que caracteriza a topologia de uma rede neural é sentido em que se propagam as informações entre os neurônios da rede. Nesse sentido, destacam-se nesse trabalho os dois tipos abaixo:

- Rede neural do tipo *feedfoward*: nas redes que seguem esse tipo de arquitetura, os neurônios estão dispostos em camadas e a informação se propaga exclusivamente de uma camada anterior até uma posterior (nunca

entre neurônios de uma mesma camada ou de uma camada posterior de volta a uma que lhe for anterior).

- Rede neural recorrente ou realimentada: nesse tipo de rede, o sentido de propagação das conexões sinápticas não ocorre somente de uma camada anterior para uma posterior, como também pode ocorrer dentro da mesma camada ou de uma camada posterior para uma anterior.

2.3.2 As redes neurais MLP como aproximadores universais

Hornik, Stinchcombe e White (1989) e Cybenko (1988) analisaram em trabalhos independentes as redes neurais do tipo *Multilayer Perceptron* e chegaram à conclusão, utilizando-se de abordagens matemáticas diferentes, que esse tipo de rede neural, utilizando apenas uma camada intermediária composta por um número finito de neurônios, é capaz de aproximar, com qualquer grau de exatidão desejado, qualquer função G(x) desde que, como também elucidado por Haykin (2001), G(x) seja uma função contínua limitada e cuja integral num conjunto compacto de R seja diferente de zero. Esse teorema é chamado de teorema da aproximação universal.

$$G: R^{m_o} \longrightarrow R \text{ uma função limitada integrável e contínua tal que}$$

$$\int_{R^{mo}} G(x)dx \neq 0$$

O teorema da aproximação universal respalda a ideia de que uma MLP seria capaz de constituir qualquer tipo de modelo de previsão com poucas premissas *a priori* sobre as variáveis a serem modeladas, o que lhes garante uma vantagem frente outras técnicas de modelagem, como a regressão linear, que se baseiam em pressupostos fortes.

2.3.3 Identificação e estimação de modelos baseados em redes neurais

O teorema da aproximação esclarece que uma MLP com número finito de neurônios na camada intermediária seria capaz de aproximar qualquer função G(x)

dadas certas premissas. No entanto, o teorema não fornece o número mínimo de neurônios que a rede deve possuir para aproximar satisfatoriamente a função. Assim como a identificação de outras características de uma rede neural (como sentido de propagação da informação, função de ativação, entre outras), a escolha para a quantidade de neurônios na camada intermediária de uma MLP pode ser obtida por meio de métodos de validação cruzada, como explicado por Haykin (2001). Uma dos métodos que se pode destacar é o *K-fold*, cuja aplicação no processo de identificação de redes neurais consiste no algoritmo abaixo:

1. Dividir toda a série histórica de dados disponíveis em K elementos do mesmo tamanho;
2. Selecionar K-1 dos elementos para calibrar uma determinada configuração de rede neural;
3. Calcular os valores de indicadores de acurácia de estimativa como RMSE e MAPE (que serão descritos em 2.4.1) no elemento não utilizado para calibrar a rede;
4. Repetir esse processo para vários valores de neurônios na camada intermediária, até que os valores dos indicadores comecem a piorar (indicando sobreajuste decorrente de alta parametrização no modelo);
5. Selecionar a configuração cuja performance tenha sido a melhor, considerando os valores obtidos para os indicadores de acurácia.

O processo descrito acima exige constante recalibração de um modelo de rede a uma dada amostra de valores. Ao processo de calibração de uma rede neural, costuma-se dar o nome de treinamento. Haykin (1999) separa os algoritmos de treinamento de uma rede neural em dois grandes grupos:

- Aprendizado supervisionado: Fornece-se à rede dados de entradas (*inputs*) e as respectivas saídas esperadas (*targets*) à rede, o aprendizado supervisionado calibra os valores dos parâmetros da rede (pesos e vieses dos neurônios transformadores) de modo a tentar aproximar a função real que relaciona as entradas com as saídas;

- Aprendizado não-supervisionado: os parâmetros são calculados apenas com dados de entradas. Nesse caso, a rede deveria ser capaz de reconhecer algum tipo de padrão nesses dados, de modo a classificá-los em grupos.

A performance de três algoritmos de treinamento é comparada por Kisi e Uncuoglu (2005) em dois casos de estudo, de acordo com algumas métricas distintas de performance. Os algoritmos eram o Levenberg-Maraquardt (LM), o Conjugate Gradient with Fletcher-Reeves (CGF) e o Resilient Backpropagation (RB). A diferença mais evidente entre os três algoritmos era o número de iterações que cada um precisava para chegar à solução final: enquanto que 50 iterações já eram suficientes para que o LM conseguisse convergir para uma boa solução, o CGF e o RB precisaram de 554 e 2000 iterações, respectivamente. Isso fazia com que o LM convergisse a uma solução final mais rapidamente que os demais algoritmos testados. Além disso, tanto LM quanto RB mostram boas performances de acurácia de previsão e de ajuste à amostra, superando CGF.

2.4 Seleção de modelos

Arlot et al. (2010) explicam que uma das dificuldades inerentes ao processo de seleção de modelos é a ausência de uma metodologia única para seleção do modelo com melhor performance, principalmente quando os modelos pertencem a famílias diferentes de modelagem. Além disso, como descrito por Zucchini (2000), a própria definição de "performance" de um modelo é bastante abrangente: existem atributos diferentes dos modelos que estão ligados à caracterização da sua performance, e não somente a acurácia das estimativas dos modelos com relação a uma determinada amostra de dados. Por exemplo, um modelo simples que apresenta um erro quadrático médio ligeiramente superior a um modelo mais complexo, com maior número de parâmetros, não necessariamente deve ser preterido frente ao modelo de menor erro quadrático, levando-se em conta o princípio da parcimônia, que supõe a preferência pela simplicidade.

Com base nisso, nessa seção do trabalho, introduzem-se alguns dos indicadores de ajuste e de acurácia de previsão para avaliar a performance relativa dos modelos a serem propostos ao longo do trabalho. Também descreve-se o método a ser aplicado no final para seleção dos modelos de cada uma das categorias de modelagem aplicadas (regressão linear, modelos ARIMA e redes neurais).

2.4.1 Indicadores de ajuste do modelo

Willmott (1982) mostra que os indicadores de performance mais comuns para avaliação de modelos envolvem a diferença ponto a ponto da estimativa obtida pelo modelo (P_i) e o valor real observado conhecido (O_i), chamada pelo autor de "quantidade fundamental", que origina todas as medidas de diferença que fazem parte dos indicadores de performance de um modelo. Cada indicador, no entanto, penaliza diferentemente as magnitudes das diferenças. Sendo n o número total de pontos que foram estimados pelo modelo, P_i cada ponto estimado e O_i o valor real que lhe é correspondente, Willmott (1982) indica alguns indicadores recorrentes na literatura:

$$\mathrm{MBE} = \frac{\sum_{i=1}^{n}(P_i - O_i)}{n} \tag{2.37}$$

$$s_d^2 = \frac{\sum_{i=1}^{n}(P_i - O_i - \mathrm{MBE})^2}{n-1} \tag{2.38}$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n} \tag{2.39}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|P_i - O_i|}{n} \tag{2.40}$$

$$\mathrm{MAPE} = \frac{\sum_{i=1}^{n}\frac{|P_i - O_i|}{O_i}}{n} * 100\% \tag{2.41}$$

O erro de viés médio (MBE, ou *Mean Bias Error*) é uma medida que visa quantificar o viés de um determinado modelo. Trata-se da média da variável "diferença", que é equivalente à subtração do valor previsto pelo modelo e o observado. Por sua vez, s_d^2 é a variância dessa variável. Enquanto o primeiro indicador de performance visa quantificar o viés do modelo, o segundo pode ser usado como uma medida da intensidade do desvio em si. Porém, Willmott (1982) afirma que para identificar essa intensidade de desvio, é preferível o uso dos indicadores MSE (*Mean Square Error*, ou erro quadrático médio), MAE (*Mean Absolute Error*, ou erro absoluto médio) ou MAPE (*Mean Absolute Percentage Error*, ou erro percentual absoluto médio).

Para facilitar a visualização do grau da diferença de performance entre dois modelos com relação a algum dos indicadores, introduziu-se a razão *Skill* (destreza):

$$\text{Skill}(\text{modelo A}, \text{modelo B}) = \frac{\text{Ind}_\text{A} - \text{Ind}_\text{B}}{\text{Ind}_{\text{modelo perfeito}} - \text{Ind}_\text{B}} \tag{2.42}$$

onde:

Ind_i é o indicador do modelo i

$\text{Ind}_{\text{modelo perfeito}}$ é o indicador do modelo teórico perfeito (sem erros). No caso de MAPE, MAE e MSE, o valor do modelo perfeito é zero.

Um outro indicador para comparar modelos de previsão é o coeficiente de determinação R^2. Kvalseth (1985), no entanto, identificou na literatura oito fórmulas diferentes para R^2, e detalhou os diferentes fatores aos quais cada definição para o indicador R^2 está exposta. O autor recomenda que R^2 seja definido pela fórumla abaixo, que pode ser aplicada para diferentes tipos de modelagem além da regressão linear:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} = 1 - \frac{SQ_{RES}}{SQ_{TOT}} \tag{2.43}$$

O coeficiente de determinação, definido dessa maneira, compara os resíduos quadráticos do modelo com os resíduos do modelo nulo (ou seja, modelar a variável resposta pela média das observações). A ideia é que modelos bons teriam valores de resíduos baixos, tendo coeficientes de correlação próximos a 1.

Willmott (1982) propõe também um outro indicador de performance chamado índice de exatidão de Willmott (d), argumentando que se trataria de uma medida mais sensível que R^2 às discrepâncias entre as médias e variâncias dos valores obtidos pelo modelo e dos valores observados correspondentes.

$$d = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(|O_i - \bar{O}| + |P_i - \bar{O}|)^2} \tag{2.44}$$

Todos esses indicadores procuram identificar, ainda que de maneiras diferentes, qual seria o melhor modelo a partir da "quantidade fundamental", não considerando o número de variáveis incluídas no modelo (ou seja, não se levando em conta o princípio da parcimônia). Outros indicadores, no entanto, procuram analisar a discrepância entre valores estimados e observados, porém sem deixar de penalizar, de alguma forma, modelos mais complexos frente aos mais simples. Para o processo de comparação de modelos, incluindo modelos de diferentes categorias como ARIMA e rede neural, Zou et Al. (2007) utilizaram, além dos indicadores RMSE, MAE e R^2, os indicadores AIC (Coeficiente de informação de Akaike), BIC (Critério de Informação Baynesiano) e o R^2 ajustado, que contêm alguma forma de penalização em função da complexidade do modelo. Considerando que p é o número total de parâmetros do modelo, esses indicadores podem ser definidos da seguinte forma:

$$\mathrm{R^2_{Adj}} = 1 - \left(\frac{\mathrm{n-1}}{\mathrm{n-(p+1)}}\right)(1 - \mathrm{R^2}) \tag{2.45}$$

$$\mathrm{AIC} = \mathrm{n ln}\left(\frac{\sum_{i=1}^{n}(\mathrm{O_i - P_i})^2}{\mathrm{n}}\right) + 2\mathrm{p} \tag{2.46}$$

$$\mathrm{BIC} = \mathrm{n}\ln\left(\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}\right) + \ln(n)p \tag{2.47}$$

Melhores modelos possuem R^2_{Adj} mais altos e AIC e BIC mais baixos. Com relação à diferença entre esses dois últimos indicadores, é fácil observar que BIC penaliza comparativamente mais a maior quantidade de parâmetros num modelo do que AIC, devido ao termo multiplicativo do logaritmo natural (2.47).

Finalmente, Zou et Al. (2007) também introduziram no seu trabalho a estatística de mudança direcional D_{stat} como uma ferramenta extra para comparar os modelos. Essa estatística pretende avaliar se o modelo consegue captar o sentido da mudança de tendência:

$$D_{stat} = \frac{1}{N}\sum_{i=1}^{N} a_i \tag{2.48}$$

$$\text{em que } a_i = \begin{cases} 1, \text{se } (P_{i+1} - P_i)(O_{i+1} - O_i) \geq 0 \\ 0, \text{se } (P_{i+1} - P_i)(O_{i+1} - O_i) < 0 \end{cases}$$

É importante saber, no entanto, que a correta interpretação dos valores de uma grande parte dos indicadores de ajuste é principalmente comparativa. Por exemplo, um determinado valor de AIC por si só não gera uma ideia se o modelo é adequado para descrever a variável resposta, mas a comparação desse valor entre dois modelos distintos gera uma indicação de qual deles possui melhor performance. Zou et Al. (2007) seguem essa abordagem para selecionar as melhores configurações de redes neurais, variando no número de nós na camada intermediária e as variáveis exógenas que incluem.

2.4.2 Comparação entre categorias distintas de modelos e a questão da instabilidade dos parâmetros de um modelo em função do tempo

Inoue, Jin e Rossi (2017) explicam que, em problemas de modelagem de séries temporais, calcular os parâmetros de um modelo utilizando-se de toda amostra de observações de que se dispõe limita a capacidade do modelo de refletir a relação atual entre as variáveis explicativas e a variável resposta. Isso se deve ao fato de que os modelos costumam estimar os parâmetros considerando a mesma relevância para todas as observações (não importando o quão distante elas estejam do instante de tempo para o qual se dá a aplicação do modelo), o que pode não ser adequado, já que a relação entre as variáveis pode ser dinâmica no tempo (ou seja, o valor dos parâmetros reais altera-se com o tempo). Por isso, no caso de se objetivar realizar estimativas futuras de séries temporais a partir de um determinado modelo, os autores sugerem que o modelo seja calibrado com uma janela de tempo de tamanho fixo, considerando apenas as w últimas observações da amostra.

Levando-se em conta essa questão, os autores propõem um método de validação cruzada para avaliação dos modelos que utiliza como intervalo de calibração apenas os w últimos valores anteriores ao intervalo de tempo em que o modelo começa estimar valores (ou seja, para o intervalo de teste). Conhecidos os valores reais da variável resposta nesses intervalos de testes, pode-se calcular os indicadores de acurácia de previsão do modelo. Aplicando-se esse procedimento para vários momentos da série histórica de dados, calcula-se a média da performance do modelo para cada um dos indicadores. Esses valores podem ser usados para comparar a performance preditiva de modelos, inclusive entre aqueles que são construídos com metodologias de modelagem distintas. Para esse trabalho, convencionou-se a denotar esse método de "validação cruzada com janela de tempo".

2.5 Estimação de modelos de previsão de demanda de combustíveis na literatura

A literatura na área de previsão de demanda combustíveis é extensa e alguns artigos relevantes foram publicados a partir da década de 1970. Dahl e Terne (1991 e

1992), Dix e Goodwin (1982), Goodwin (1992), Bohi e Zimmerman (1984) e Taylor (1977) procuraram identificar as variáveis que mais afetavam a demanda de combustíveis considerando modelos lineares em que as variáveis independentes representavam fatores econômicos conjunturais do mercado. Esses autores apontam para alguns grandes grupos de fatores, incluindo: renda e poder de compra da população, dinâmica de preços do petróleo no mercado global, preços reais de combustíveis locais para o consumidor final, frota (não só considerando o número de carros em si, mas também fatores como eficiência de consumo e distância média percorrida) e perturbações macroeconômicas que tenham alterado subitamente a dinâmica de consumo de combustível (como interferências governamentais, políticas de incentivo, choques de preços, etc.). Trabalhando com dados de mercados de países diferentes, esses autores puderam comprovar que a relevância de cada fator para a modelagem varia com relação ao mercado em questão, sugerindo a importância de se considerar as dinâmicas de consumo local próprias.

Modelagem do consumo interno de combustíveis no mercado brasileiro

Para estudar a demanda nacional pelos combustíveis utilizados no setor de transporte, é preciso compreender algumas particularidades do mercado brasileiro que interferem na dinâmica por trás do consumo de combustíveis. Uma delas é a participação indireta, ainda que considerável, do governo na definição de preços dos combustíveis no mercado doméstico por meio da Petrobrás, que define os preços dos combustíveis nas refinarias, de modo que eles podem ficar descolados dos preços do petróleo negociados no mercado de capitais global, diferentemente do que ocorreria em países com economia menos intervencionista. Ainda assim, é interessante mencionar que, como explica Marjotta e Barros (2002), a participação do Estado na definição do preço dos combustíveis para o consumidor final foi se reduzindo paulatinamente durante a década de 90 e início do novo milênio, o que estava em consoante com a agenda política de menor intervencionismo estatal dos governos da época, destacando momentos como o fim do Instituto do Açúcar e do Álcool em 1990 e a liberação em 1996 do preço da gasolina para o consumidor final. Além dessa particularidade, outras características do mercado local afetam especificamente a demanda de certos combustíveis.

Com relação à dinâmica de demanda por gasolina e etanol, a particularidade mais relevante do mercado brasileiro é a introdução de motores *flex-fuel* no início da década passada e a subsequente expansão da participação do etanol hidratado no setor de combustíveis líquidos. Com relação a isso, vale destacar o trabalho de Silva et al. (2009), que procurou aferir a elasticidade cruzada da demanda entre gasolina e etanol no período posterior à introdução dos motores *flex-fuel* no mercado, mostrando que a demanda desses combustíveis tornou-se sensível ao preço relativo do outro combustível. Por esse motivo, a relação de preço entre etanol e gasolina foi introduzida nos modelos de previsão de demanda estudados por Freitas e Kaneko (2011), Santiago (2009) e Nappo (2007).

Com relação ao consumo de óleo diesel, Silva (2014) e Santiago (2009) indicam que as particularidades do mercado brasileiro são geradas por dois fatores: a grande importância do transporte rodoviário para a logística nacional (mesmo para grandes distâncias) e a baixa penetração de motores a diesel em veículos de passeio. Dessa forma, o consumo de óleo diesel no país é determinado principalmente pelo fluxo de veículos pesados e, portanto, tem grande exposição aos fatores que afetam diretamente esse fluxo em específico, como quebra de safras e variações na produção industrial.

3 O SETOR DE TRANSPORTES NO BRASIL

Dado que o objetivo do trabalho é estabelecer modelos de previsão da demanda nacional pelos combustíveis mais relevantes para o setor de transportes, é preciso tentar quantificar, a partir de séries históricas de dados, como seria a evolução do nível de atividade do setor. No entanto, notou-se que uma das dificuldades de analisar o setor de transportes como um todo é o fato de que as entidades principais que o representam (e que são responsáveis pela divulgação de estatísticas mensais sobre o setor) em sua maioria estão voltadas a modais específicos, ao invés de representar o setor conjuntamente. Além disso, a própria atividade de cada modal é analisada por variáveis consideravelmente diferentes, que podem não fazer sentido se aplicadas para analisar outro modal. Por isso, para entender o setor de transporte, precisa-se analisar separadamente os subsetores que o compõem (o rodoviário, o aquaviário, o aeroviário e o ferroviário). Além disso, dados diferentes mesmo dentro de um subsetor específico são fornecidos por organizações distintas, de modo que é necessário consultar várias autarquias diferentes para que se possa traçar um panorama do setor como um todo. Ainda assim, vale destacar a recente iniciativa tomada pela Confederação Nacional de Transportes (CNT) de unificar as estatísticas do setor de transporte como um todo num anuário único, cuja primeira publicação se deu para o ano de 2016. A CNT é uma entidade com sede em Brasília que reúne em um mesmo grupo várias associações, entidades e sindicatos de todos os subsetores relacionados à logística e à infraestrutura de transporte do país, e almeja que o estudo conjunto de todos os modais por meio do anuário será importante para identificar tendências e lacunas no setor de transportes como um todo e fornecer uma base para um planejamento mais eficiente de alocação de investimentos para o setor por parte do poder público e da iniciativa privada.

Outra dificuldade em entender a evolução do nível de atividade no setor de transportes é entender quais seriam as variáveis mais importantes para avaliar essa evolução, em especial considerando o objetivo final de modelar a demanda dos combustíveis mais relevantes para o setor. Por isso, optou-se por focar principalmente na movimentação de passageiros e cargas dentro do território nacional e na quantificação dos fatores que mais determinam essa movimentação.

3.1 O setor rodoviário

O Plano Nacional de Logística dos Transportes (PNLT) de 2016 mostra que 65% de todo o transporte inter-regional de carga no país foi feito pela malha rodoviária em 2015, o que equivale a 1548 bilhões de TKU (Toneladas Quilômetro Úteis, resultado da multiplicação da massa total transportada de carga em toneladas e a extensão percorrida em quilômetros). 27% de todo esse carregamento rodoviário foi constituído de carga a granel (ou seja, transportado sem embalagem), sendo 16% de granel sólido não agrícola (que inclui minério de ferro, carvão, bauxita, etc.) e 6% de granel agrícola (soja, milho, açúcar) e 5% de granel líquido (suco de laranja, óleo de soja, etanol). Os restantes 73% eram de carga geral (CG), sendo o modal com maior percentual de transporte desse tipo de carga (excluindo-se o aeroviário), o que indica o predomínio do transporte rodoviário para movimentação de produtos e mercadorias de volume unitário no país. Também é o maior poluidor entre os modais, correspondendo a 86% do total de emissão de CO_2 pelo setor de transportes de carga como um todo no país (o que é proporcionalmente mais do que sua participação no transporte de cargas em si). Segundo os dados da CNT, a malha rodoviária do país era de 1.720.644 km em 2015, sendo que apenas 12% encontram-se totalmente pavimentadas.

O modal rodoviário possui várias vantagens frente aos demais modais, como agilidade e possibilidade de entrega porta a porta. É ideal para conectar curtas distâncias e fazer a ligação entre modais diferentes. No entanto, é bastante poluidor, tem baixa capacidade de transporte e frete mais caro, se comparado ao transporte rodoviário e aquaviário. A prevalência da estrutura de transporte rodoviário no país frente a outros que seriam mais adequados para longas distâncias, em especial o ferroviário, possui motivos históricos, com destaque para a política de Juscelino Kubitschek de incentivo à construção de rodovias e de caracterização da ferrovia como tecnologia ultrapassada, visando, em última instância, atrair as montadoras de carros para o país.

3.1.1 Índices ABCR – Fluxo de veículos em rodovias privadas

Dentro das variáveis representativas do setor rodoviário em escala nacional, duas possuem publicação mensal e fornecem uma boa *proxy* para movimentação de cargas e pessoas em rodovias privadas: o índice ABCR Brasil para veículos leves (que incluem os automóveis e os veículos comerciais leves) e o índice ABCR Brasil para veículos pesados (ônibus e caminhões). Esses índices, publicados pela Associação Brasileira de Concessionárias de Rodovias (ABCR), quantificam numa variável adimensional o fluxo de veículos leves e o de veículos pesados que passam por estradas com pedágio no Brasil, sendo que 100 seria o valor médio para os índices originais da série em 1999. É importante que se analise o fluxo segregado por tipo de veículo devido às características intrínsecas da frota brasileira: enquanto a frota de veículos pesados é composta basicamente por veículos com motores a diesel, esse tipo de motor só tem participação relevante nos veículos comerciais da frota de veículos leves, onde a predominância maior é de motores *flex-fuel* e a gasolina.

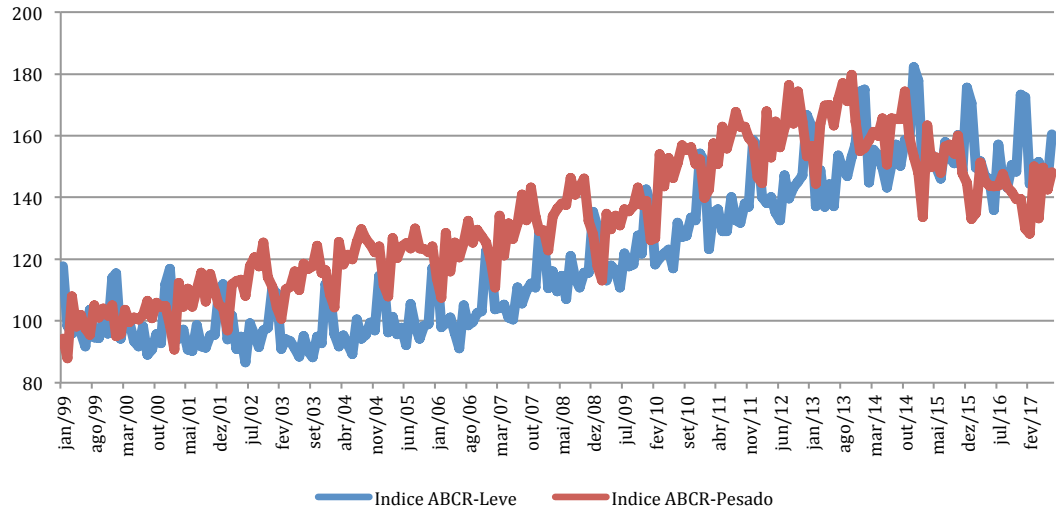

Figura 3.1 Gráfico da evolução dos índices ABCR-Leve e ABCR-Pesado
Fonte: ABCR

A ABCR também possui um índice que indica o fluxo agregado todos os tipos de veículos, o chamado índice ABCR-Total. Segundo a associação, em média, 70% do movimento desse índice é resultado do fluxo de veículos leves, enquanto 30% seria do fluxo de veículos pesados.

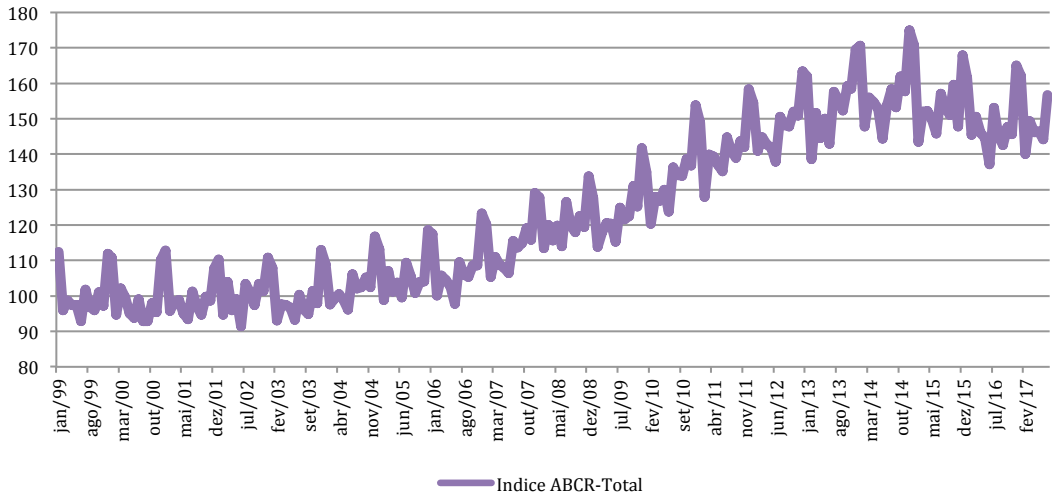

Figura 3.2 Gráfico da evolução do índice ABCR-Total
Fonte: ABCR

É fácil perceber pela a análise visual dos gráficos acima que o fluxo de veículos possui forte caráter sazonal. Para analisar o comportamento de cada índice (e, por extensão, do fluxo de veículos que ele representa) em cada mês excluindo-se os efeitos da tendência, calcularam-se os coeficientes de sazonalidade a partir da amostra de dados. Sendo $\Phi_{y,i}$ o valor do índice no mês i do ano y e Y o número total de anos da amostra, o valor do coeficiente de sazonalidade para o mês m será definido como:

$$s_m = \sum_{y=1}^{Y} \left(\frac{\Phi_{y,m}}{\sum_{i=1}^{12} \frac{\Phi_{y,i}}{12}} \right) \frac{1}{Y} \qquad (3.1)$$

Valores para o coeficiente de sazonalidade acima de 1 indicam que o mês teria fluxo superior à média. Valores entre 0 e 1 indicam fluxo inferior à média. Quanto mais próximo a 1 for o coeficiente de sazonalidade, mais o mês em questão está próximo à média.

Mês	Leve	Pesado	Total
Janeiro	1,12	0,92	1,08
Fevereiro	0,95	0,89	0,94
Março	0,98	1,03	0,99
Abril	0,97	0,98	0,97
Maio	0,95	1,02	0,97
Junho	0,92	0,98	0,93
Julho	1,02	1,02	1,02
Agosto	0,97	1,06	0,99
Setembro	0,96	1,03	0,98
Outubro	1,01	1,06	1,02
Novembro	0,99	1,01	1,00
Dezembro	1,16	0,99	1,12

Tabela 3.1 Coeficientes sazonais dos índices ABCR
Fonte: Elaborado pelo autor

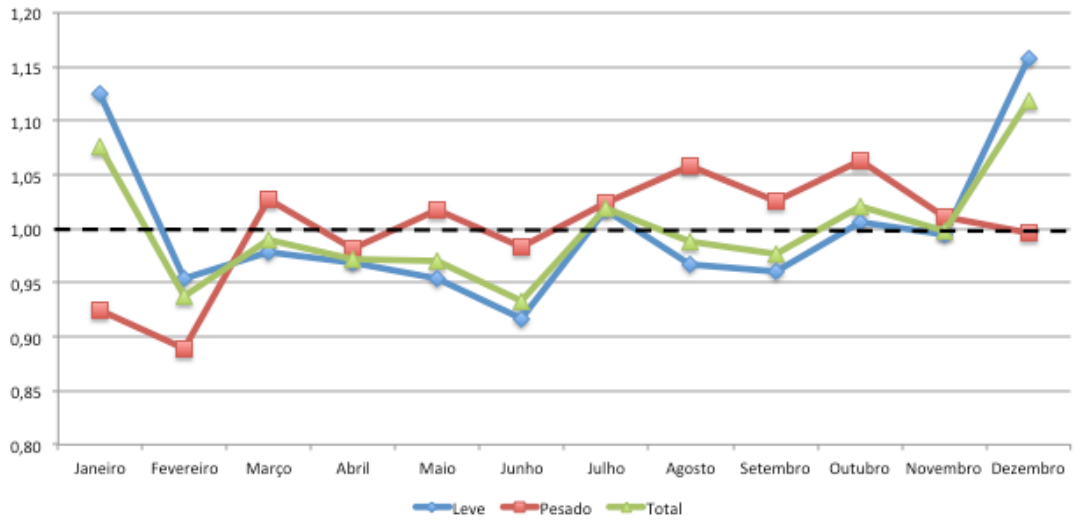

Figura 3.3 Gráfico dos coeficientes sazonais dos índices ABCR
Fonte: Elaborado pelo autor

O coeficiente de sazonalidade mostra que, nos meses de dezembro e janeiro, o fluxo de veículos leves nas estradas pedagiadas é mais intenso que a média anual, ao passo que, no caso de veículos pesados, o fluxo é abaixo da média para esses meses. Isso se deve ao fato de que esses são os meses coincidem com as férias, quando as

famílias tiram férias e costumam viajar para outras cidades. Em fevereiro o fluxo de todos os veículos cai consideravelmente, e isso é decorrência principalmente do carnaval.

No entanto, é preciso relembrar que o fluxo de veículos medido pela ABCR se refere ao movimento em rodovias pedagiadas. Isso significa que a movimentação dentro das cidades, por exemplo, não é levada em conta para o cálculo, o que prejudica o uso do índice ABCR-Leve original para prever demanda de gasolina e etanol, já que uma boa parte do consumo desses combustíveis se dá pelo uso dos carros nas cidades. Isso gera distorções bastante significativas no comportamento sazonal do índice ABCR-Leve com relação ao comportamento sazonal da demanda de etanol e gasolina, como é possível ver no gráfico abaixo:

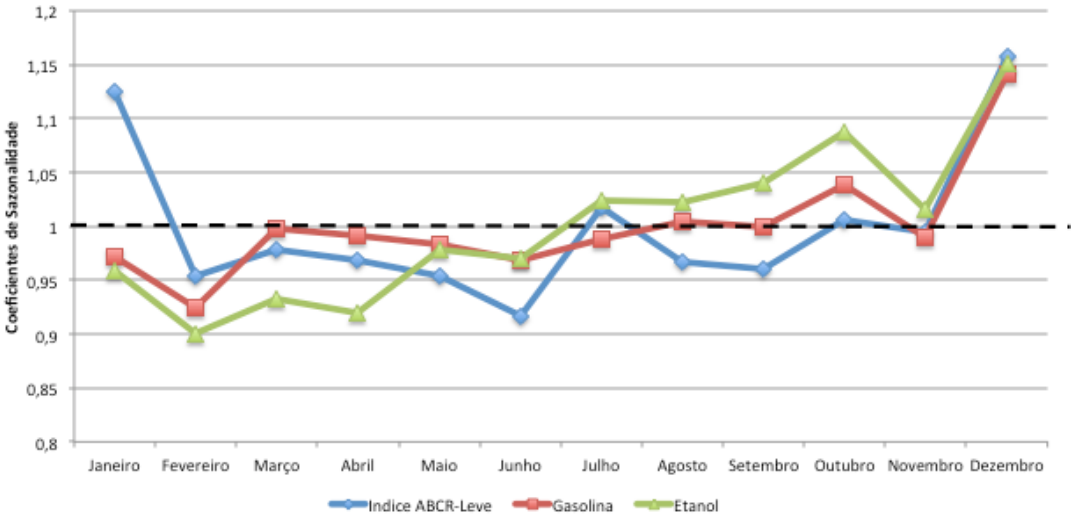

Figura 3.4 Gráfico dos coeficientes sazonais do índice ABCR-Leve original e da demanda volumétrica de etanol e gasolina
Fonte: Elaborado pelo autor

Mesmo com os descompassos gerados pelas diferenças sazonais, o cálculo do coeficiente de correlação de Pearson sugere que o índice ABCR-Leve teria correlação positiva significativa com a gasolina (0,86) e o etanol (0,65). Por isso, optou-se por utilizar o índice ABCR-Leve dessazonalizado (que também é fornecido pela ABCR) para modelar a demanda de gasolina ao invés do índice original. Com o

índice dessazonalizado, sua correlação com a demanda de gasolina e de etanol saltou para 0,95 e 0,76, respectivamente.

Já o índice ABCR-Pesado original possui comportamento sazonal e de tendência bastante em linha com a movimentação da demanda de diesel, o que indica que seria mais interessante utilizar o valor original desse índice para a modelagem da demanda de diesel. De fato, o coeficiente de correlação de Pearson se mostrou levemente superior com os valores originais do modelo (0,94) do que com os valores dessazonalizados (0,88).

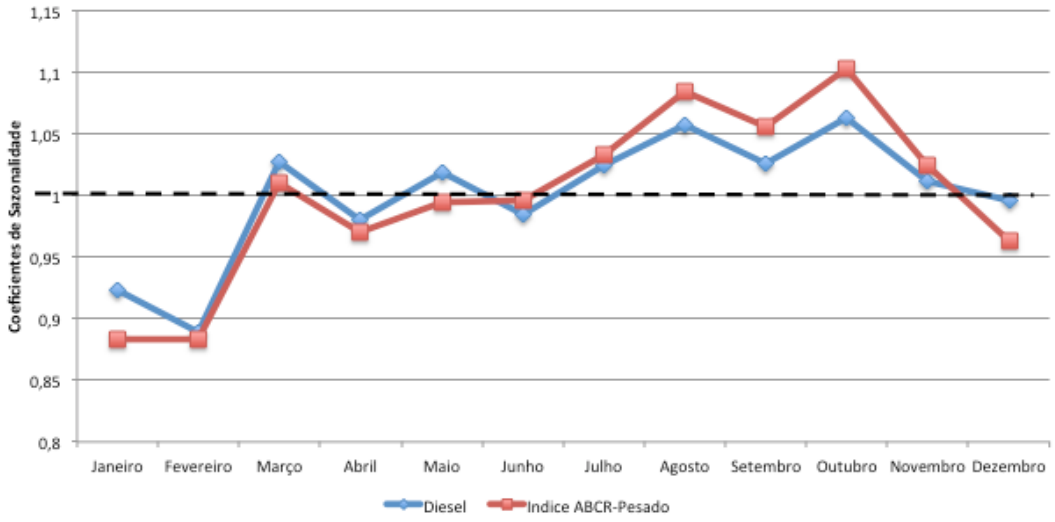

Figura 3.5 Gráfico dos coeficientes sazonais do índice ABCR-Pesado original e da demanda volumétrica de diesel
Fonte: Elaborado pelo autor

Vale destacar que, segundo a associação, o índice ABCR para veículos pesados possui uma alta correlação com a produção industrial do país. De fato, a análise da figura 3.6, que correlaciona Índice de Produção Industrial (de toda a indústria em geral) obtido da Pesquisa Industrial Mensal de Produção Física do IBGE (Instituto Brasileiro de Geografia e Estatística) com o índice ABCR-Pesado dessazonalizado, além do cálculo do índice de correlação de Pearson entre essas duas variáveis (0,85) confirmam essa afirmação, de modo que as estimativas futuras para variações na produção industrial dadas pelo mercado podem servir como base para o cálculo de estimativas para o fluxo de veículos pesados nas estradas e, por extensão, da demanda de diesel.

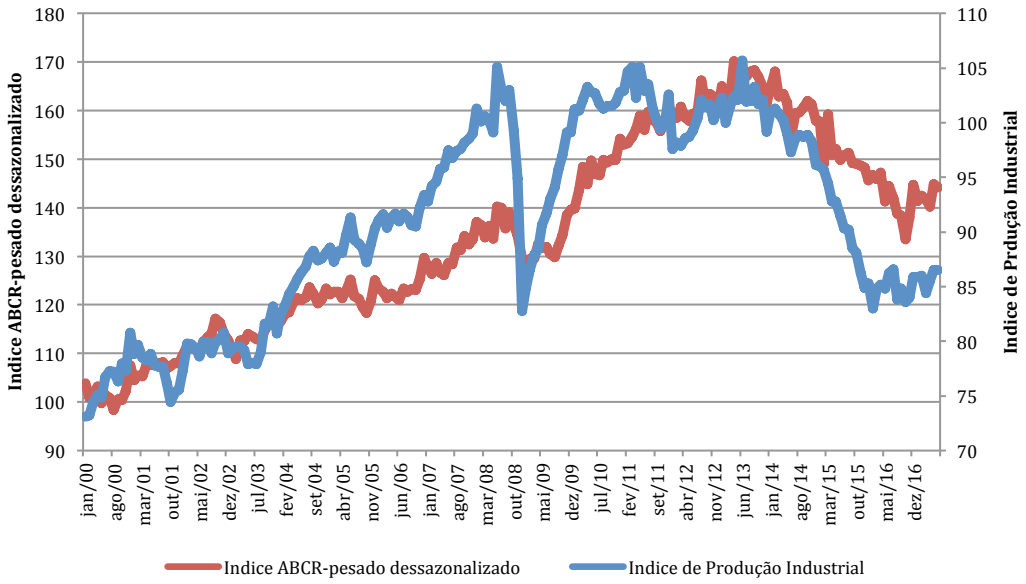

Figura 3.6 Gráfico dos coeficientes sazonais do índice ABCR-Pesado original e da demanda volumétrica de diesel
Fonte: Elaborado pelo autor

3.1.2 Frota de veículos no Brasil

Entender a frota de veículos é fundamental para compreender o potencial futuro da demanda pelo transporte rodoviário. Além disso, como explicado na revisão literária, a modelagem da demanda de combustíveis frequentemente inclui como variável explicativa a frota de veículos e suas características. Sabendo-se disso, e considerando o fato de que se trabalha com variáveis mensais nos modelos, é fundamental possuir uma série histórica mensal sobre frota segregada por tipo de combustível para que se possa utilizar essa variável nos modelos de previsão de demanda.

Ainda que haja estimativas mensais sobre frota divulgadas pelo DENATRAN, vários autores, como De Negri (1998), Castro (2012) e Losekann e Vilela (2012), concordam que essas estimativas não consideram devidamente o sucateamento da frota e, por isso, não seriam confiáveis. Para estimar valores mais realísticos para a frota nacional mensal, esses autores desenvolveram curvas de sucateamento a partir das informações sobre o perfil e a idade da frota contidas na PNAD (Pesquisa

Nacional por Amostra de Domicílio) contínua de 1988: dispondo dos dados mensais sobre licenciamento de veículos no mercado interno fornecidos pela Associação Nacional dos Fabricantes de Veículos Automotores (ANFAVEA), os autores calibravam os parâmetros de curvas de sucateamento de modo que a frota em 1988, que era resultante da entrada de veículos que ocorrem desde o início da série histórica sobre licenciamento, subtraído do respectivo sucateamento calculado pela curva, tivesse um perfil de idade em consoante aos valores fornecidos pela PNAD de 1988.

No entanto, é razoável se pensar que o perfil da frota trazido pela PNAD contínua de 1988 seja bastante distinto do atual. A expansão do poder aquisitivo da população vivida nos últimos anos, juntamente com as novas mudanças tecnológicas do setor automotivo e da modernização da frota com a abertura da economia vivida nos anos 90 são fortes indícios de que, possivelmente, o perfil de idade da frota de veículos possa ter se alterado significativamente nos últimos 30 anos.

Considerando todas as questões expostas, optou-se por desenvolver curvas de sucateamento da frota semelhantes àquelas propostas pelos autores, porém utilizando outros dados para calibrar os parâmetros dessas curvas: ao invés de utilizar os dados a PNAD de 1988, utilizaram-se as estimativas anuais para a frota de 2000 até 2016 fornecidas ANFAVEA, que consideram devidamente a questão de sucateamento e estarem próximas às estimativas de outras entidades privadas do setor automotivo como o Sindipeças.

A estimativa de frota para um determinado mês T é resultado da soma de todos os carros produzidos nos meses anteriores desde o início da série histórica de dados de licenciamento, menos a quantidade de carros de cada mês que já foi sucateada. Expressando-se matematicamente essa ideia, sendo F_T a estimativa da frota no mês T, E_t o licenciamento de veículos num mês t, $S_{t,T}$ a quantidade de veículos licenciados no mês t que já se encontravam sucateados no mês T e t = 1 o primeiro mês de que se tem estatísticas sobre o licenciamento de veículos na série histórica, a estimativa da frota é calculada pela fórmula abaixo:

$$F_T = \sum_{t=1}^{T} (E_t - S_{t,T}) \tag{3.2}$$

Sendo $s_{t,T}$ a porcentagem de veículos produzidos no mês t já sucateados no mês T, pode-se rescrever (3.2) da seguinte forma:

$$F_T = \sum_{t=1}^{T} (E_t - S_{t,T}) = \sum_{t=1}^{T} E_t(1 - s_{t,T}) \tag{3.3}$$

Losekann e Vilela (2012) partem da hipótese de que a porcentagem de sucateamento de um tipo de veículo depende apenas da sua idade em anos. Para modelar essa dependência, os autores indicam três funções que se aproximariam do comportamento não-linear do processo de sucateamento de um veículo em função da sua idade: a função logística, a curva de Gompertz e a função acumulada de probabilidade de Weibull. Os testes que foram conduzidos preliminarmente com os dados do presente trabalho não mostraram diferenças significativas de desempenho entre as funções, atingindo valores de erros quadráticos muito semelhantes. Porém, optou-se por utilizar a curva de Gompertz, uma vez que foi a função aplicada para estimar o sucateamento da frota no Inventário Nacional de Emissões Atmosféricas por Veículos Rodoviários de 2006, produzido pelo Ministério do Meio Ambiente.

Uma função Gompertz genérica assume a seguinte fórmula, com coeficiente L, a e b:

$$f(x) = Le^{e^{a-bx}} \tag{3.4}$$

Como queremos que a curva Gompertz equivalha à porcentagem de veículos com idade igual a x anos já foram sucateados, essa função, no limite, deve tender a 1. Portanto, L será necessariamente igual a um para o problema de sucateamento.

$$1 = \lim_{x\to\infty} Le^{e^{a-bx}} = Le^{0} = L \implies L = 1$$

Dessa forma, sendo $s(x)$ a porcentagem de veículos com idade x (em anos) que já foram sucateados, tem-se a equação abaixo:

$$s(x) = e^{-e^{a-bx}} \tag{3.5}$$

Sabendo que um veículo com x anos de idade foi licenciado há T – t meses, obtém-se a equivalência abaixo:

$$\frac{T-t}{12} = x \tag{3.6}$$

Finalmente, pela combinação de (3.3), (3.5) e (3.6), tem-se que a frota em um dado mês T é obtida pela seguinte equação:

$$F_T = \sum_{t=1}^{T} E_t(1 - s_{t,T}) = \sum_{t=1}^{T} E_t\left(1 - e^{-e^{a-bx}}\right) = \sum_{t=1}^{T} E_t\left(1 - e^{-e^{a-b\left(\frac{T-t}{12}\right)}}\right) \tag{3.7}$$

Os parâmetros da função de sucateamento são calibrados procurando minimizar a diferença quadrática entre as estimativas de frotas anuais fornecidas pela ANFAVEA de 2000 a 2016 e as médias das estimativas da frota dos 12 meses de cada ano estimadas pela equação (3.7). Em outras palavras, sendo $F_{y,m}$ a frota no mês m do ano y calculada pela equação (3.7) e F_y^{ANF} a estimativa da ANFAVEA para a frota no ano y, precisa-se encontrar os valores de a e b que minimizem o somatório do erro quadrático de estimativa da frota de cada ano y (e_y^2):

$$\min \sum_{y=2000}^{2016} e_y^2 = \sum_{y=2000}^{2016} \left(F_y^{ANF} - \frac{\sum_{m=1}^{12} F_{y,m}}{12}\right)^2$$

A ANFAVEA disponibiliza dados mensais de licenciamento e estimativas anuais sobre frota segregadas por quatro grandes tipos de veículo: automóveis, comerciais leves, ônibus e caminhões. Isso permite que cada tipo de veículo possa ter uma curva de sucateamento própria, o que aumenta a precisão das estimativas, já que é razoável

assumir que cada um desses tipos de veículos apresentam dinâmicas de sucateamento próprias. Resolvendo-se o problema de minimização utilizando uma heurística de otimização baseada num algoritmo genético do *software* MATLAB, obtém-se os valores abaixo para os parâmetros da curva Gompertz da equação (3.5):

	a	b
Automóveis	2,338	-0,146
Comerciais leves	2,015	-0,156
Ônibus	1,587	-0,123
Caminhões	3,768	-0,197

Tabela 3.2 Estimativas para os parâmetros da curva Gompertz para cada tipo de veículo
Fonte: Elaborado pelo autor

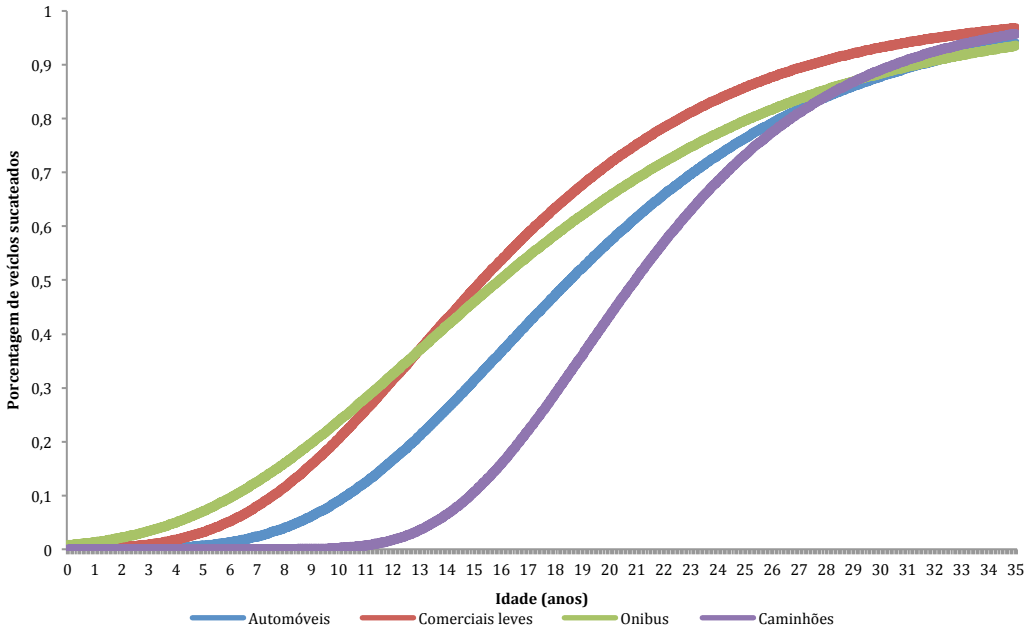

Figura 3.7 Gráfico das curvas de sucateamento calibradas para cada tipo de veículo
Fonte: Elaborado pelo autor

Partindo-se da hipótese simplificadora de que o tipo de combustível que cada segmento de veículo usa não altera a dinâmica de seu sucateamento, e utilizando-se da série histórica mensal de licenciamento de veículos no país, segregada por segmento e por tipo de combustível, pode-se gerar estimativas mensais de frota segregadas por tipo de combustível e segmento do veículo.

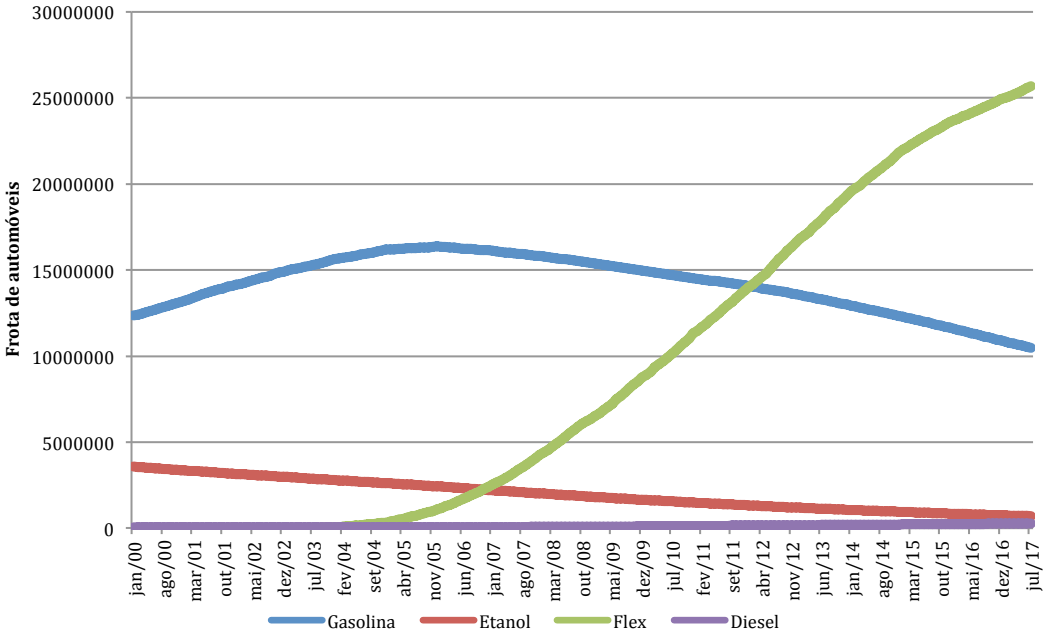

Figura 3.8 Gráfico da estimativa para a frota de automóveis segregada por tipo de combustível
Fonte: Elaborado pelo autor

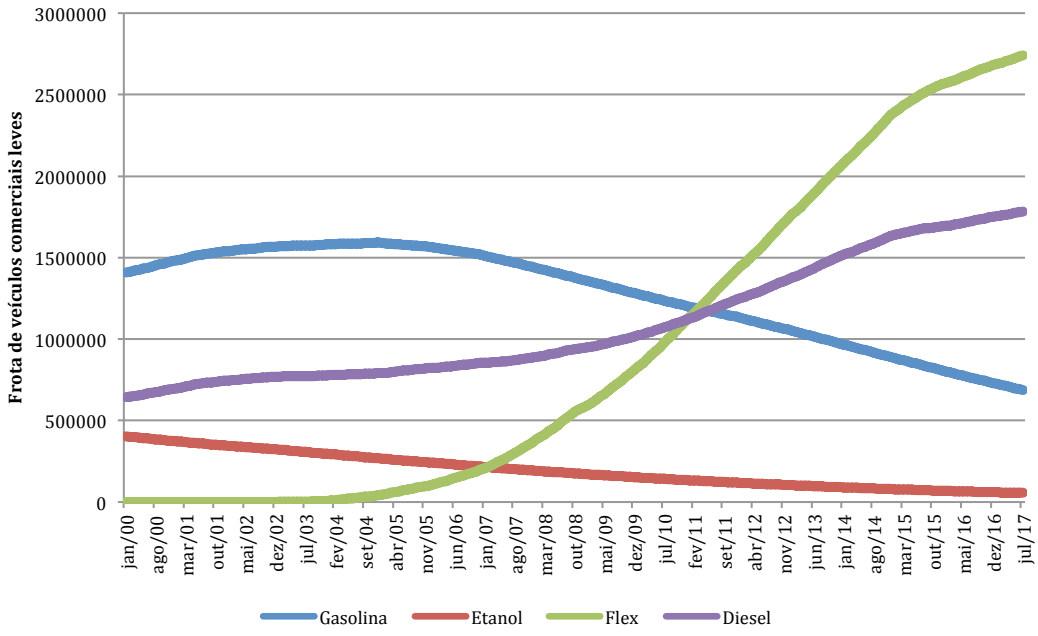

Figura 3.9 Gráfico da estimativa para frota de veículos comerciais leves segregada por tipo de combustível
Fonte: Elaborado pelo autor

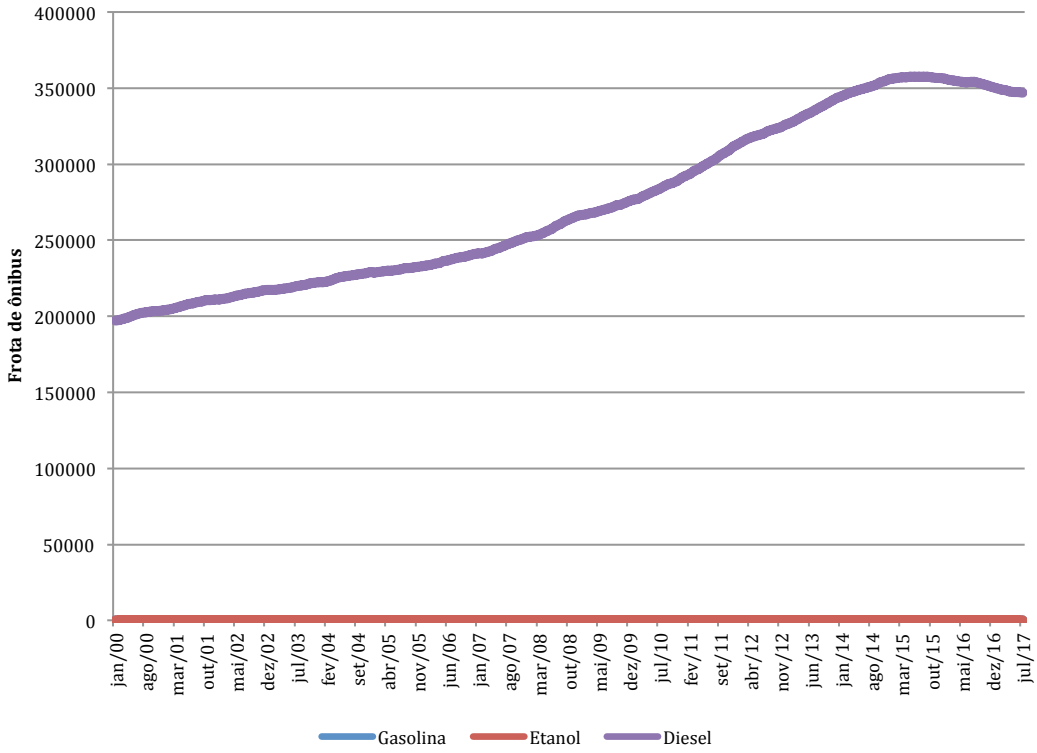

Figura 3.10 Gráfico da estimativa para a frota ônibus segregada por tipo de combustível
Fonte: Elaborado pelo autor

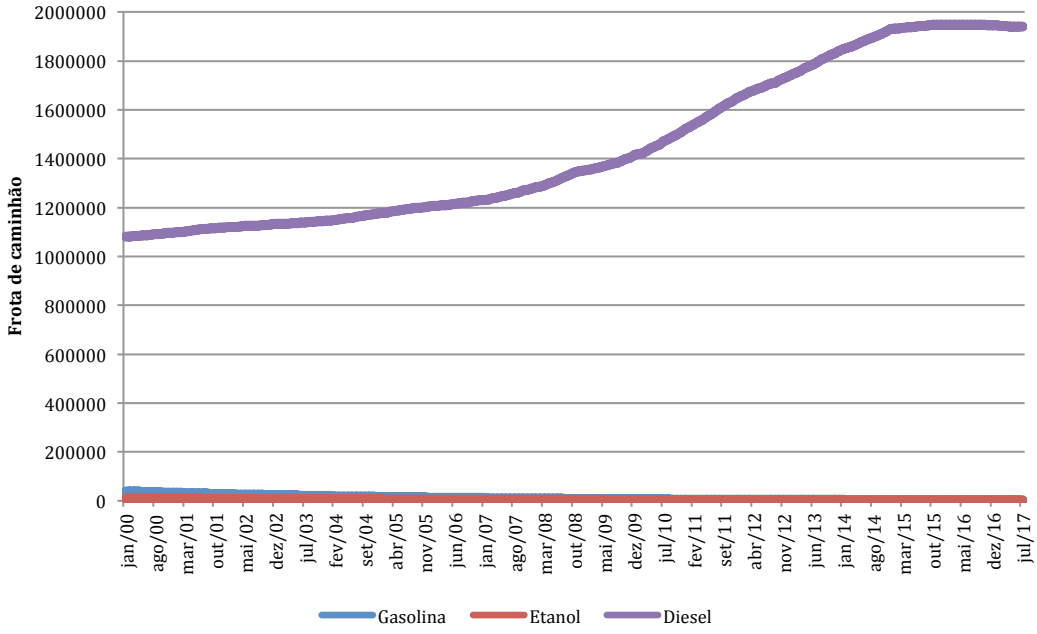

Figura 3.11 Gráfico da estimativa para a frota ônibus segregada por tipo de combustível
Fonte: Elaborado pelo autor

Com base nisso, pode-se eliminar a divisão por segmento de veículos e obter a frotas segregada apenas por tipo de combustível, já que essa informação é mais relevante considerando o objeto de estudo do trabalho.

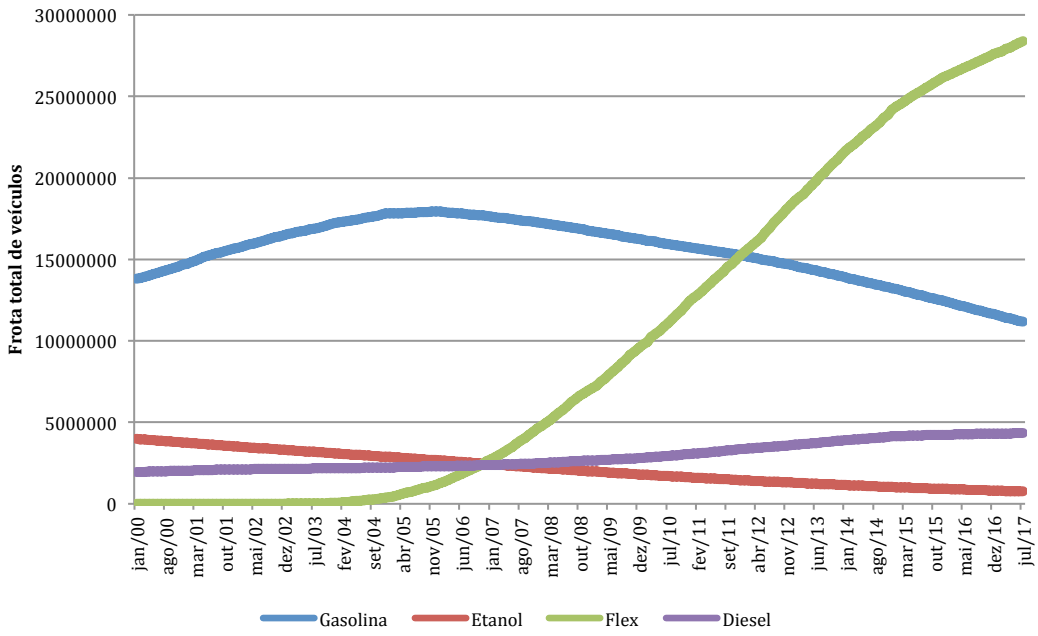

Figura 3.12 Gráfico da estimativa para a frota de veículos total segregada por tipo de combustível
Fonte: Elaborado pelo autor

3.2 O setor ferroviário

Ainda que fadado ao sucateamento devido à falta de incentivos públicos desde o governo de Juscelino Kubitschek, o setor ferroviário no país parece ganhar momento nos últimos anos com novos projetos de infraestrutura e crescimento das empresas do setor. Segundo os dados da ABIFER (Associação Brasileira da Indústria Ferroviária) e da CNT, de 2006 até 2016, o modal experimentou um crescimento de 43,1% do volume anual transportado por quilômetro útil (em TKU) e em 36,5% no número de locomotivas em operação. Os investimentos anuais no setor dispararam mais de 300% nos últimos dez anos, sendo mais de 750% em novas infraestruturas, indicando que a participação desse modal no setor de transportes deve crescer consideravelmente nos próximos anos.

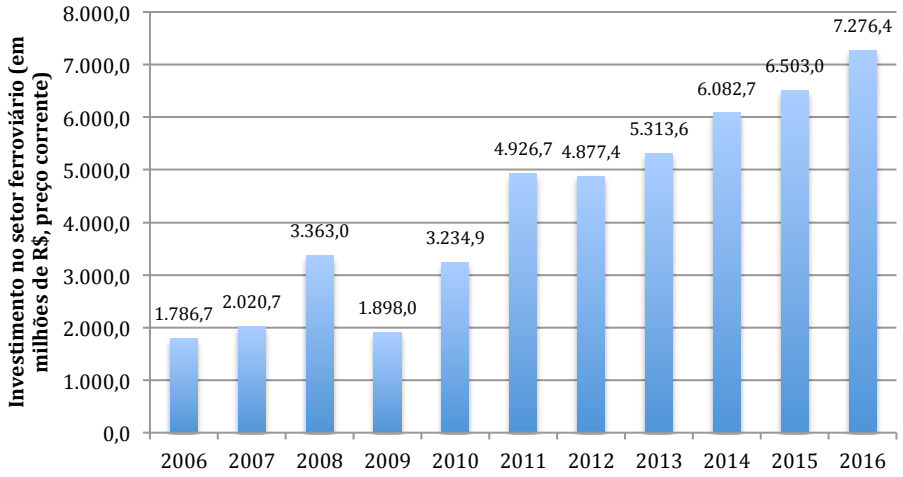

Figura 3.13 Gráfico da evolução anual do investimento no setor ferroviário brasileiro
Fonte: CNT

Os últimos dados do PNLT mostram que o setor ferroviário movimentou 356,8 bilhões de TKU de carga no Brasil em 2015, o que equivale a 15% do total do setor de transportes, mesmo respondendo por apenas 7% das emissões de CO_2 do setor para essa atividade. A maioria do carregamento ferroviário é composto por minérios, o que justifica que 81% da movimentação de carga pelo modal em 2015 correspondeu a granel sólido não agrícola. De fato, utilizando-se dos dados sobre exportação de minérios metalúrgicos (incluindo minério de ferro, de cobre, de manganês, de alumínio e de cromo) fornecidos pelo MDIC (Ministério da Indústria, Comércio Exterior e Serviços) e os dados da CNT e ABIFER sobre movimentação em ferrovias, obtém-se um coeficiente de correlação de Pearson de 0,83, o que indica forte correlação positiva.

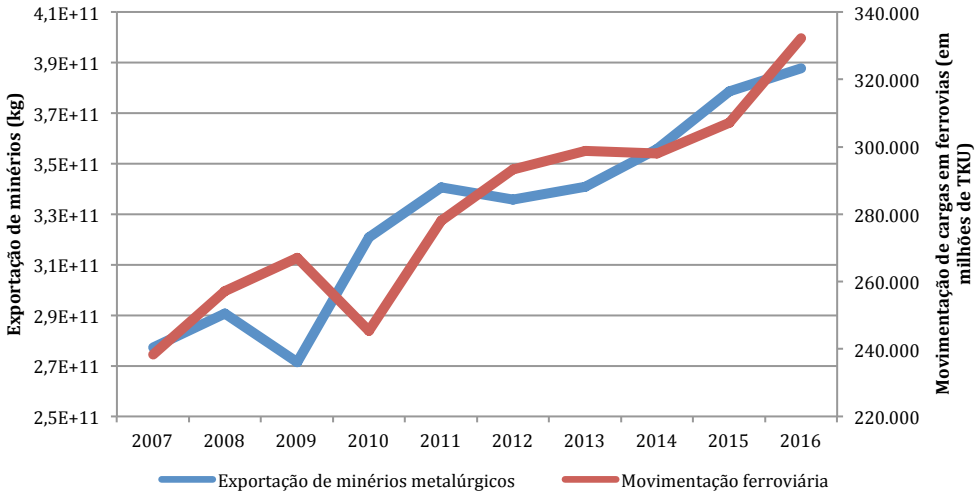

Figura 3.14 Gráfico da exportação de minérios (Kg) com relação à movimentação de cargas do setor ferroviário (em milhões de TKU)
Fonte: MDIC e ABIFER

Diferentemente do que se vê no transporte rodoviário, apenas 4% da movimentação de carga por ferrovias em 2015 foi voltada para carga geral. 14% foi voltada a granel sólido agrícola, enquanto que granel líquido não passa de 1%.

O frete ferroviário para grandes distâncias pode se tornar consideravelmente inferior ao rodoviário. Além disso, as ferrovias apresentam um risco de acidentes e congestionamento muito menor que as rodovias, além de ter maior eficiência em termos energéticos e ambientais. A pior desvantagem é a baixa flexibilidade desse transporte, tanto no que se refere aos lugares de transporte (sempre acaba dependendo do rodoviário para pequenas distâncias) quanto às disponibilidades de horários.

3.3 O setor aquaviário

A hidrografia nacional, a extensão territorial da costa brasileira e a forte concentração da população do país em cidades costeiras são características que favorecem o transporte aquaviário no Brasil. Considerando o transporte inter-regional, podemos dividir o setor aquaviário em dois subgrupos: o transporte por cabotagem (entre portos das áreas costeiras do Brasil) e o transporte por hidrovias

internas, correspondendo respectivamente a 10% (249,9 bilhões de TKU) e 5% (125,3 bilhões de TKU) do total de movimentação de carga dentro do país, ainda que só correspondam a 5% e 2% das emissões totais de gás carbônico para esse fim. O tipo de carga mais movimentado por cabotagem em 2015 foi granel líquido (61%) em função do transporte de combustível entre as regiões do país, seguido por carga geral (36%) e granel sólido não agrícola (3%). Já a movimentação por hidrovias internas foi dominada por transporte de carga geral (47%) e granel líquido (29%), sendo os demais 24% divididos igualmente entre granel sólido agrícola e não agrícola.

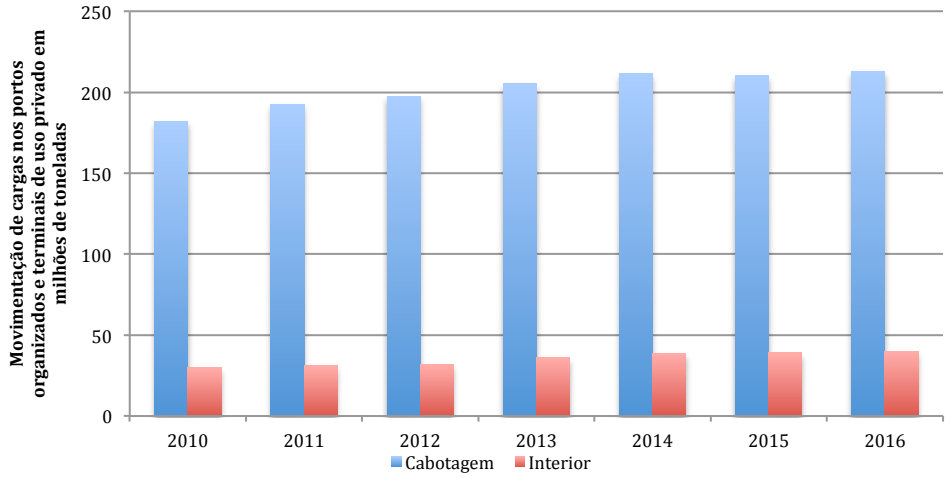

Figura 3.15 Gráfico da movimentação de cargas (por cabotagem e por hidrovias internas) em milhões de TKU
Fonte: MDIC e ANTAQ

Levando-se em consideração questões ligadas à exportação de carga, um dos temas mais relevantes atualmente para o setor aquaviário e portuário é o desenvolvimento da estrutura portuária do chamado "Arco Norte", que são portos localizados no norte e nordeste do Brasil, e da logística que liga os grandes centros produtores agrícolas das regiões interioranas do país a esses portos. As vantagens que o desenvolvimento do Arco Norte traria para a infraestrutura logística nacional é aliviar o excesso de demanda pelos portos das regiões sul e sudeste (destacando o porto de Santos), além do fato de que os portos do Arco Norte se localizam mais próximos aos principais mercados consumidores estrangeiros. Alguns dados obtidos pela ANTAQ (Agência Nacional do Transporte Aquaviário) mostram que os portos

do Arco Norte vem ganhando participação na exportação de grãos (milho e soja) nos últimos anos, e em 2016 responderam por 24% da movimentação total da produção nacional voltada ao mercado externo.

Entre todos os tipos de transporte, o aquaviário é o menos poluidor e é capaz de levar altas cargas por um preço bastante atrativo. Porém, assim como o ferroviário, é mais lento e menos flexível que o rodoviário, acabando dependendo desse último para as curtas distâncias.

3.4 O setor aeroviário

O setor aeroviário possui métricas bastante próprias para caracterizar o seu nível de atividade. Mensalmente, os dados de oferta e demanda no mercado aéreo são publicados pelas duas principais entidades do setor: a ABEAR (Associação Brasileira das Empresas Aéreas), que representa as companhias aéreas, e a agência reguladora ANAC (Agência Nacional de Aviação Civil). A oferta de voos é mensurada em ASK (*Available Seat Kilometers*), que é resultado do número de assentos disponíveis vezes a quantidade voada em quilômetros durante certo período, enquanto que a demanda é dada em RPK (*Revenue Passenger Kilometers*), que calcula o número de quilômetros que o total de passageiros voou em um determinado período de tempo. Esse número é divulgado mensalmente por cada uma das companhias abertas, e costuma ser assunto de publicações periódicas pelas corretoras de ações que cobrem o setor aéreo. Geralmente, a performance em cada mês de cada companhia aérea e do setor como um todo é avaliada pela comparação da demanda por voos no mês com aquela do mesmo mês no ano anterior, já que a comparação mês a mês é afetada pela sazonalidade. Costuma-se olhar também para a variação no acumulado dos últimos doze meses, em especial no fim do ano. Outro indicador que é alvo de análises é o chamado *Load Factor* (LF), que se trata de uma medida do aproveitamento total dos assentos disponibilizados, e é dada pela divisão da demanda pela oferta de voos. Quanto mais próximo a 1, maior a eficiência do setor.

Analisando o setor aéreo desde o começo dos anos 2000, pode-se perceber que houve uma expansão clara tanto da oferta quanto da demanda, em especial na de voos do mercado doméstico: a demanda de voos nacionais em 2000 foi de 25,5 bilhões de RPK, passando para 40,6 bilhões em 2006 e 89,0 bilhões em 2016, o que mostra uma expansão de 119% na demanda nos últimos 10 anos e de 156% desde o começo da década passada. A oferta também cresceu significativamente, mas em menor proporção, o que explica o ganho de eficiência do setor (aumento do LF, que variou de 58,6% em 2000, para 70,8% em 2006 e 80,0% em 2016). O mercado de voos domésticos, no entanto, não evoluiu uniformemente nesse período: o período de maior expansão ocorreu entre 2005 e 2011, quando a demanda e oferta de voos domésticos teve um crescimento percentual anual de quase dois dígitos. Posteriormente a esse período, o setor começou a apresentar estagnação, até cair em 6% em 2016 com relação a 2015 (tanto em demanda quanto em oferta), situação que vem se revertendo nos últimos meses de 2017.

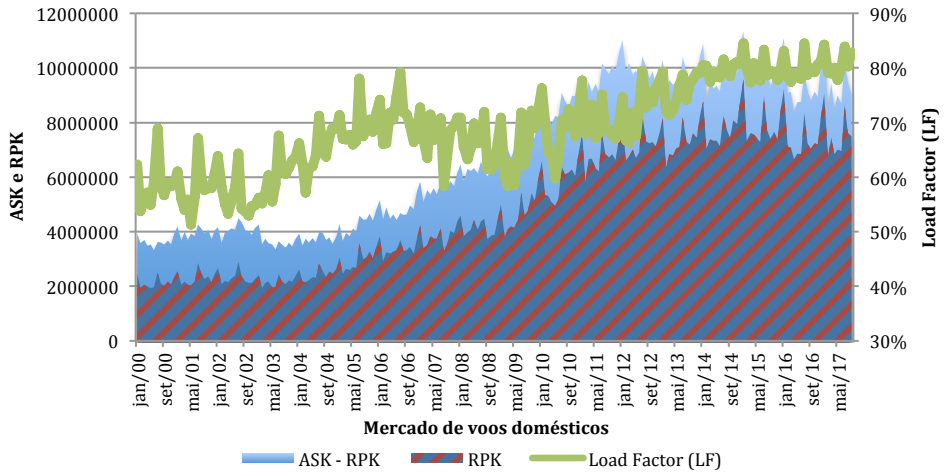

Figura 3.16 Gráfico da demanda (RPK), oferta (ASK) e *Load Factor* (LF) de voos domésticos
Fonte: ANAC

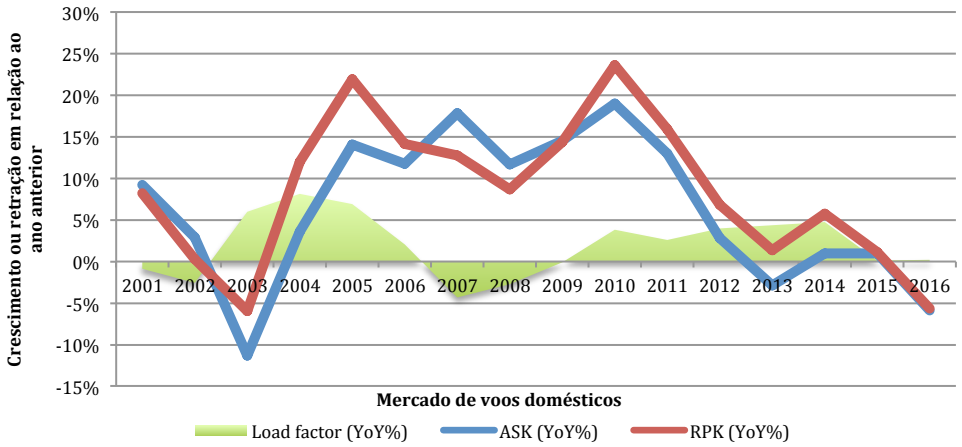

Figura 3.17 Gráfico da variação anual percentual da demanda (RPK), oferta (ASK) e *Load Factor* (LF) de voos domésticos
Fonte: ANAC

O mercado de voos internacionais também apresentou uma expansão no acumulado dos últimos anos, porém numa proporção bastante menor e de maneira mais instável que o mercado de voos domésticos: em 2000, a demanda por voos internacionais era de 31,4 bilhões de RPK, caindo em 29% em 2006 para 22,2 bilhões de RPK, porém que voltou a subir, chegando em 2016 em 39,5 bilhões de RPK, o que representa um crescimento de 26% com relação a 2000 e de 77% com relação a 2006.

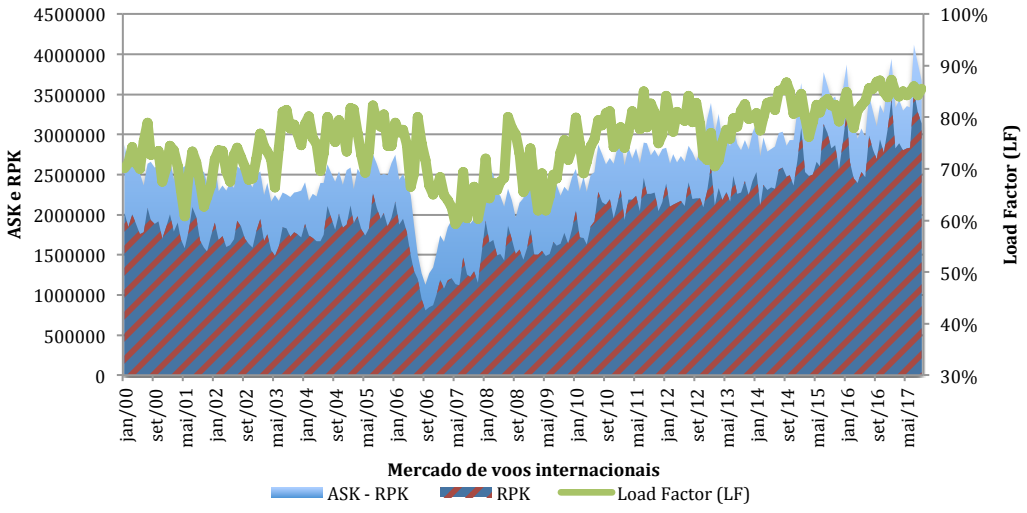

Figura 3.18 Gráfico da demanda (RPK), oferta (ASK) e *Load Factor* (LF) de voos internacionais
Fonte: ANAC

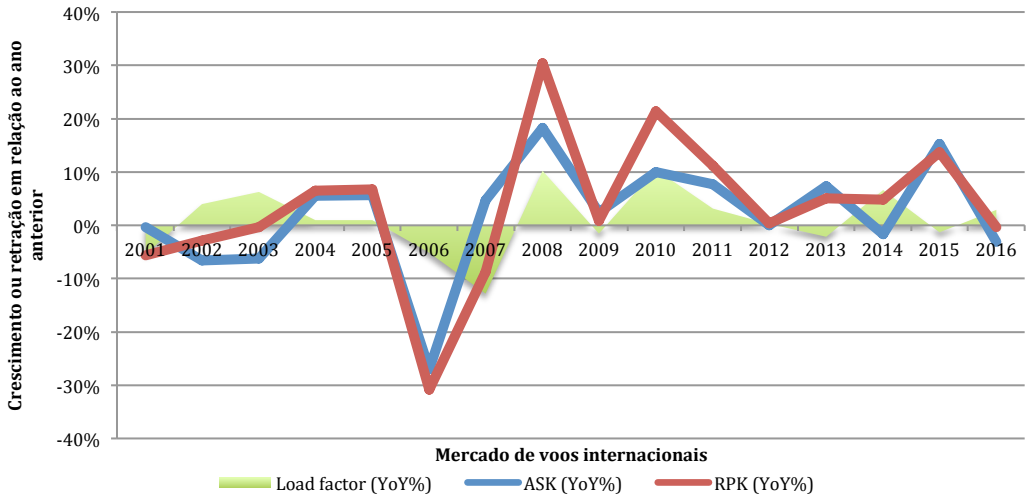

Figura 3.19 Gráfico da variação anual percentual da demanda (RPK), oferta (ASK) e *Load Factor* (LF) de voos internacionais
Fonte: ANAC

Vários fatores explicam a volatilidade desse mercado: semelhante ao que ocorre com a demanda por voos internos, o mercado depende muito do rendimento médio da população; porém, para o mercado internacional, a variação na cotação do dólar possui bastante relevância, o que sugere que essa variável possa servir como medida de previsão para o comportamento futuro da demanda por voos internacionais. Além disso, as empresas do setor aéreo como um todo também são bastante vulneráveis às oscilações de cotação pelo fato de que uma boa parte dos seus custos é dada em dólar, enquanto que a maior parte de suas receitas é dada em reais. Uma desvalorização do câmbio pode corroer a possiblidade das empresas aéreas de expansão ou manutenção de seu nível operacional, o que também sugere a importância de se considerar essa variável para compreender a dinâmica do setor.

É preciso ressaltar, no entanto, que o valor nominal do câmbio não deve ser levado em conta, já que ele não considera os diferenciais de inflação entre o mercado americano e o mercado brasileiro. Isso significa, por exemplo, que o dólar nominal comercial custar 3 reais em 2002 possui uma pressão cambial diferente do que se o dólar nominal comercial custasse os mesmos 3 reais em 2015. Evidentemente, a pressão cambial é muito mais significativa em 2002 que em 2015, já que a inflação acumulada no período é muito maior no mercado brasileiro do que no mercado americano. Por isso, é preciso utilizar o câmbio real para analisar a demanda de voos

internacionais de voos. Adotando-se a taxa de câmbio média de janeiro de 2000 como base (1,798 USD/BRL), e corrigindo as taxas nominais posteriores (obtidas pelo Ipeadata) pelas inflações acumuladas nos Estados Unidos e no Brasil a partir da fórmula abaixo (3.8), pode-se detectar certa correlação, que não é mais clara devido ao fato de que uma parte da demanda por voos internacionais é explicada também pelo aumento da renda local. De fato, o coeficiente de correlação de Pearson entre o câmbio real e a demanda de voos internacionais é de apenas -0,66. No entanto, a correlação parcial entre as demanda de voos internacionais e o câmbio real, excluindo-se os efeitos da *proxy* para renda, é de -0,79. Para calcular os valores para inflação mensal dos Estados Unidos, foi utilizada a variação mensal do *Consumer Price Index*, (CPI, ou índice de preços aos consumidor), que é fornecido pelo *Bureau of Labor Statistics* (BLS). Já para estimar a inflação no Brasil, utilizou-se a variação mês a mês do IPCA (Índice de Preços ao Consumidor Amplo), cujo cálculo é fornecido pelo IBGE.

$$\left(\frac{USD}{BRL}\right)_{REAL} = \frac{\left(\frac{USD}{BRL}\right)_{NOMINAL}(Inf_{US}+1)}{(Inf_{BR}+1)} \qquad (3.8)$$

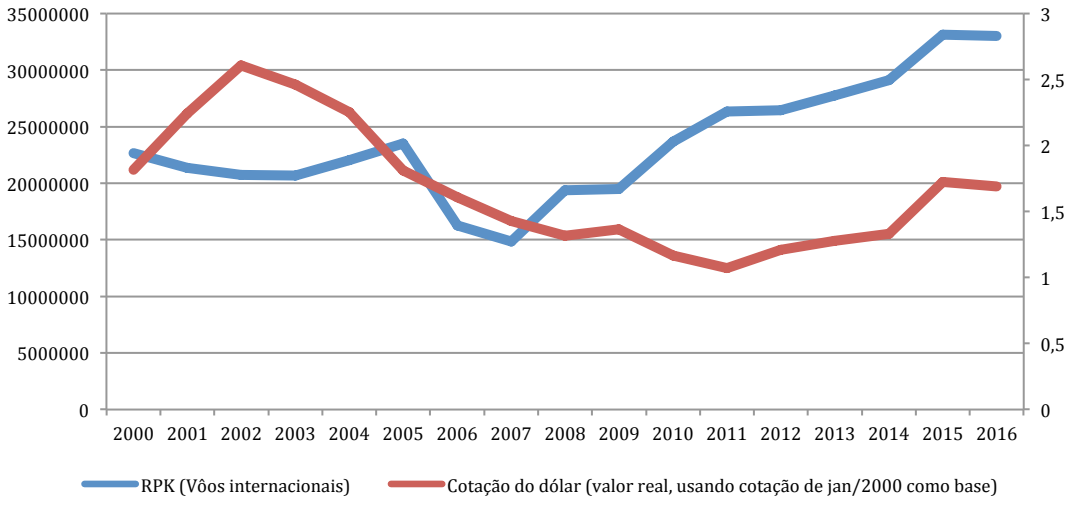

Figura 3.20 Gráfico da demanda (RPK) por voos internacionais anual e o valor da cotação do dólar a preços de janeiro de 2000
Fonte: ANAC, Ipeadata e IBGE

3.5 As perspectivas futuras para o setor de transporte

As dificuldades logísticas do Brasil constituem um dos principais fatores de perda da competitividade da produção nacional no mercado global. Além das questões tributárias e de segurança pública que afetam a movimentação de cargas no país, a própria precariedade da infraestrutura de transportes em si, principalmente levando-se em considerando a imensa extensão do território nacional e a participação do agronegócio de exportação na economia, são alguns dos problemas aos quais o poder público ainda precisa encontrar soluções de cunho definitivo. Os projetos de melhoria de infraestrutura geralmente envolvem obras complexas e que demandam vultuosos investimentos iniciais para ocorrerem, cujos retornos só aparecem no longo-prazo. Levando-se em consideração o desequilíbrio atual nas contas públicas e as dificuldades práticas em se implementarem soluções para aliviar a situação fiscal em vigência no país, é fácil verificar que os projetos de infraestrutura acabam perigando de não serem implementados nos prazos inicialmente estipulados, o que acaba postergando ainda mais a resolução dos problemas de eficiência que vem das limitações logísticas do país.

Nesse sentido, é interessante destacar que políticas públicas recentes andam tentando expandir a infraestrutura nacional por meio da redução da participação do Estado como financiador principal de projetos de infraestrutura, transferindo essa posição para a iniciativa privada, com especial destaque ao capital externo. Isso vem ao encontro de uma situação de alta liquidez vivida nos mercado internacionais, a qual impulsiona a procura por parte dos investidores externos por melhores rentabilidades nas alocações de seu capital e aumenta seu apetite por ativos de mercados emergentes, que são caracterizados por serem de maior risco e de maior rentabilidade. Com base nisso, destacam-se abaixo algumas medidas governamentais recentes que visam, em última instância, à melhoria da infraestrutura brasileira de transportes por meio do impulso da participação do capital privado no setor:

- Estímulo à desregulamentação do setor de transportes (algo bem visto pelos investidores e pelo mercado financeiro em geral, e, por conseguinte, estimularia o aporte de capital ao setor):
    - Desregulamentação do setor aéreo, com destaque para fim das regras sobre cobrança de bagagens extras e de serviços auxiliares como refeições a bordo e escolha de assentos (em linha com que já aconteceu em mercados desenvolvidos, onde gerou um aumento da oferta por voos nesses mercados e estimulou a competitividade no setor);
    - Permitir a aquisição ilimitada do capital de empresas aéreas brasileiras por investidores estrangeiros (antes limitados à participação de no máximo 20% no capital);
    - Venda de parte do capital dos aeroportos de Guarulhos (GCH), Confins (CNF), Galeão (GIG) e Brasília (BSB), atualmente nas mãos da Infraero (que detém atualmente 49% de participação acionária nessas empresas);
    - Desestatização da Companhia Docas do Espírito Santo (CODESA), que é a autoridade portuária encarregada de todos os portos do estado.
- Programa de Parceria de Investimentos (PPI), que é uma iniciativa do governo que se insere no contexto do "Projeto Crescer", que visa estimular a expansão e a melhora da infraestrutura nacional com a participação de investimentos privados por meio de concessões e leilões, privatização de organizações públicas e parcerias público-privadas (PPP). Os projetos que fazem parte do programa incluem não só o setor de transportes em geral, como também mineração, energia elétrica, óleo e gás, etc. Considerando o objetivo do trabalho, abaixo estão descritos alguns dos projetos de impacto no setor de transportes que fazem parte do PPI:
    - Ferrovia: expansão da malha ferroviária depois de anos de sucateamento e de primazia do transporte rodoviário. Alguns projetos incluem:
        - Ferrogrão (ferrovia EF-170), ligando as cidades da região agrícola de MT, começando por Sinop (MT), até a cidade

portuária de Mirituba (PA). O processo de concessão já se encontra em fase de consulta pública e o leilão deve ocorrem na segunda metade de 2018;

- Ferrovia Norte-Sul (mais especificamente o trecho EF-151), que terá uma posição bastante estratégica por conectar outras ferrovias que já foram ou que serão construídas, com destaque ao prolongamento norte (trecho Açailândia/MA – Bacarena/PA), que dará vazão da produção agrícola nacional a um porto mais próximo dos mercados internacionais, localizado no chamado "Arco Norte", e ao prolongamento sul, conectando a região agrícola do Centro-Oeste ao estado de São Paulo. Dentro do PPI, a concessão da ferrovia será composta por dois tramos:
  - O primeiro tramo conecta as cidades de Porto Nacional (TO) e Anápolis (GO), e já se encontra em operação, ainda que bastante abaixo do potencial da ferrovia;
  - O segundo tramo é compreendido entre as cidades de Estrela d'Oeste (SP) até Ouro Verde de Goiás (GO), e as obras já se encontram quase concluídas.
- Ferrovia de Integração Oeste-Leste (mais especificamente o trecho EF-334/BA), cujo objetivo principal é a interligação de áreas ao norte mais interioranas do país, desde Figueirópolis (TO) até a cidade litorânea de Ilhéus (BA), que possui um porto que também compõe o chamado "Arco Norte"). O PPI prevê a concessão do trecho da ferrovia entre Caetité (BA) e Ilhéus, visando principalmente de dar vazão à produção de minério de ferro da região e a futura conexão com a rodovia Norte-Sul.

- Rodovias: o PPI também prevê a renovação e expansão da malha rodoviária nacional, que inclui a renovação da concessão de várias rodovias:

- BR 364/RO/MT, que se estende desde Porto Velho (RO) até Comodoro (MT);
- BR 153/GO/TO, que vai do município de Aliança do Tocantins (TO) até Anápolis (GO);
- BR-364/365/MG/GO, de Uberlândia (MG) a Jataí (GO);
- BR-101/290/386/448/RS (Rodovia de Integração do Sul), que passa pelos municípios de Carazinho, Porto Alegre e Osório, todos em RS;
- BR-101/SC, de Paulo Lopes a São João do Sul, ambos em SC;
- BR-116/RJ/SP (Presidente Dutra), que liga Rio de Janeiro a São Paulo, atualmente sob concessão da CCR, com outorga para se encerrar em 2021.
- BR-040/MG/RJ, de Juiz de Fora (MG) a Rio de Janeiro;
- BR-116/RJ, que vai de Além Paraíba (RJ) e passa por cidades da Região dos Lagos.

- Aeroportos: o projeto também visa conceder à iniciativa privada a concessão de 13 aeroportos (que correspondem juntos a 10% da demanda de passageiros nacional) visando a sua ampliação e melhora operacional:
  - Aeroporto Eurico de Aguiar Salles, em Vitória (ES);
  - Aeroporto de Macaé, em Macaé (RJ);
  - Aeroporto Gilberto Freyre, em Recife (PE);
  - Aeroporto Orlando Bezerra de Menezes, em Juazeiro do Norte (CE);
  - Aeroporto Presidente Castro Pinto, em João Pessoa (PB);
  - Aeroporto Presidente João Suassuna, em Campina Grande (PB);
  - Aeroporto Santa Maria, em Aracaju (SE);
  - Aeroporto Zumbi dos Palmares, no em Maceió (AL);
  - Aeroporto Internacional Marechal Rondon, em Várzea Grande (MT);
  - Aeroporto de Rondonópolis, em Rondonópolis (MT);

- Aeroporto Presidente João Batista Figueiredo, em Sinop (MT);
- Aeroporto Piloto Oswaldo Marques Dias, em Alta Floresta (MT);
- Aeroporto de Barra do Garças, em Barra do Garças (MT).

- Portos: o PPI também inclui a licitação de novos ativos portuários e a renovação de alguns ativos já existentes.
  - Terminal de Cavaco no Porto de Santana (PA), ativo existente;
  - Terminal Portuário de Granéis Líquidos no Porto Vila do Conde (PA), ativo novo;
  - Terminais de GLP no Porto Miramar (PA), que inclui nova concessão de dois terminais que já existem e a construção de um novo terminal, ativo existente;
  - Terminais Portuários de Granéis Líquidos no Porto de Belém (PA), ativo existente;
  - Terminal de Carga Geral no Porto de Itaqui (MA), ativo novo;
  - Terminal Portuário de Granéis Líquidos no Porto de Vitória (ES), ativo novo;
  - Terminal de Celulose no Porto de Paranaguá (PR), ativo novo;
  - Terminais Portuários de Grãos no Porto Paranaguá (PR), ativo existente;
  - Terminal de Veículos no Porto de Paranaguá (PR), ativo novo.

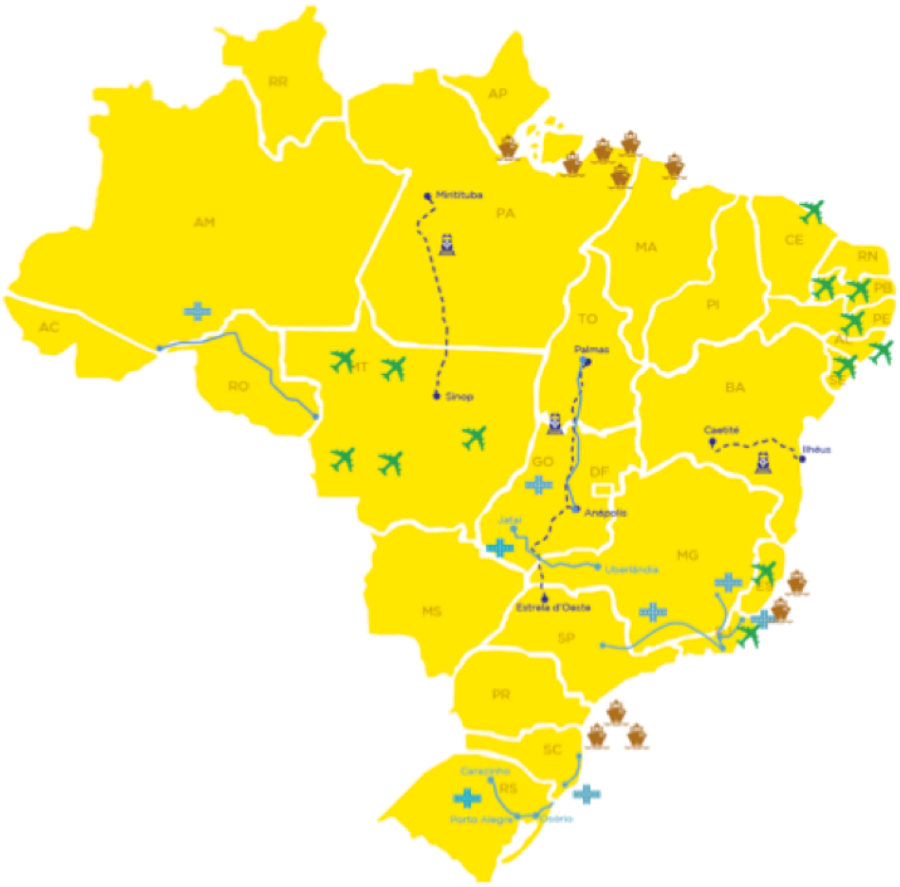

Figura 3.21 Mapa indicando os projetos de infraestrutura dentro do PPI
Fonte: http://www.projetocrescer.gov.br/

4 COMBUSTÍVEIS UTILIZADOS PELO SETOR DE TRANSPORTES

No presente capítulo, identificam-se quais são os combustíveis mais relevantes para o setor de transportes no Brasil e quais são os fatores que determinam a demanda por esses combustíveis numa escala nacional. Em seguida, sugerem-se variáveis representativas desses fatores como candidatas para serem incluídas num modelo de previsão. Finalmente, aplicam-se métodos de seleção de variáveis para determinar qual conjunto de variáveis deveria ser incluído num modelo de previsão que dependa de variáveis exógenas.

Utilizando-se das bases estatísticas do Balanço Energético Nacional (BEN), pode-se avaliar a evolução da demanda energética gerada pelo setor de transportes por cada tipo de combustível em cada ano desde 1970. No BEN, as fontes de energia utilizadas no setor de transportes (que se encontra subdivido em rodoviário, hidroviário, aeroviário e ferroviário) foram categorizados nos seguintes grupos:

- Gás Natural;
- Carvão vapor;
- Lenha;
- Óleo diesel, o que inclui as misturas com biodiesel puro (B100) e não apenas o chamado diesel de petróleo;
- Óleo combustível;
- Gasolina C (mistura de Gasolina A e álcool etílico anidro);
- Gasolina de aviação;
- Querosene;
- Eletricidade;
- Álcool etílico hidratado;
- Outros combustíveis gerados a partir de petróleo (como Nafta, Xisto, etc.).

Para avaliar relevância de cada fonte de energia na matriz energética do setor de transportes, obtiveram-se os gráficos das Figuras 4.1 e 4.2. Enquanto o gráfico da Figura 4.1 mostra a evolução do consumo energético (em 10^3 tep, ou toneladas

equivalentes de petróleo) de cada grupo de combustível gerado pelo setor de transportes de 1970 a 2016, o gráfico da Figura 4.2 analisa a participação relativa de cada combustível na matriz energética do setor de transportes para o ano de 2016. Para facilitar a visualização dos gráficos, optou-se por agregar as fontes de energia que não sejam diesel, gasolina, álcool etílico, querosene, gás natural e óleo combustível numa categoria denominada "outros", já que a demanda energética por esse conjunto não representa nem 1% do total de energia demandada pelo setor de transportes em 2016.

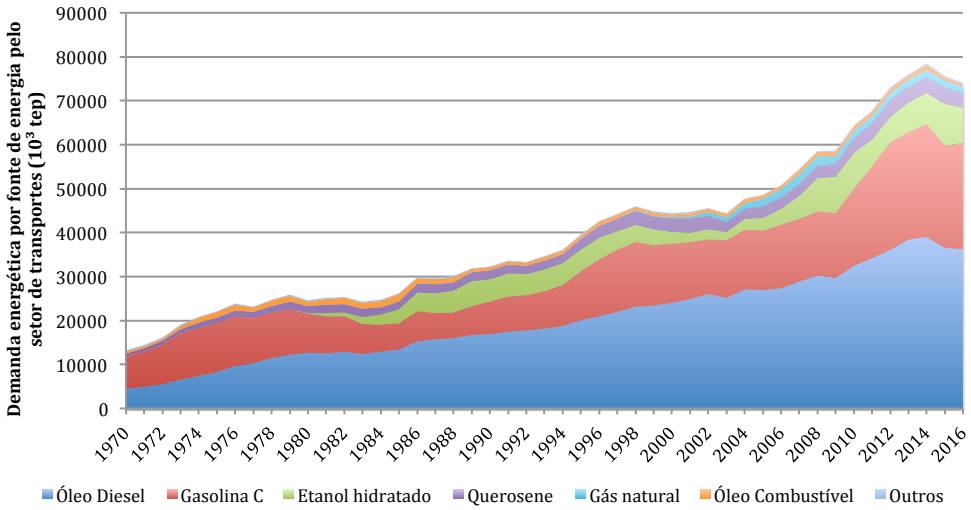

Figura 4.1 Gráfico da evolução da demanda energética do setor de transporte segregada por fonte de energia (em 10^3 tep)
Fonte: Balanço Energético Nacional de 2016

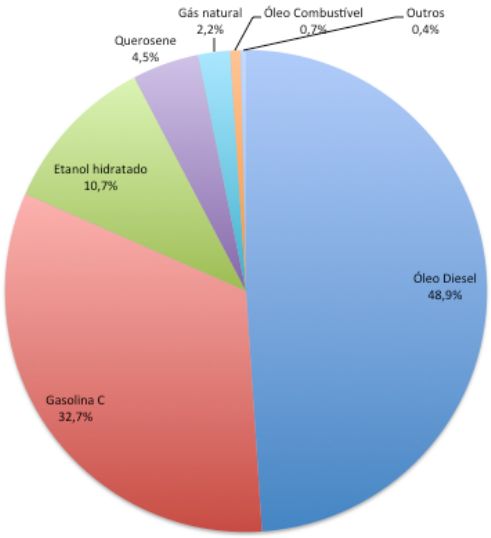

Figura 4.2 Gráfico da participação de cada combustível na matriz energética do setor de transportes para o ano de 2016
Fonte: Balanço Energético Nacional de 2016

Pela análise dos gráficos, é possível concluir que a maior parte da demanda de energia gerada pelo setor de transportes é pelos quatro tipos de combustível abaixo, que correspondem conjuntamente a aproximadamente 97% da matriz energética do setor de transportes em 2016:

- Óleo diesel (48,9% da matriz energética)
- Gasolina C (32,7% da matriz energética)
- Álcool etílico hidratado (10,7% da matriz energética)
- Querosene (4,5% da matriz energética)

Dessa forma, decidiu-se estudar mais a fundo apenas a demanda dos quatro combustíveis acima, desconsiderando os demais. Esse estudo é fundamental para entender o mercado desses combustíveis como um todo e estabelecer modelos de previsão da demanda por eles a nível nacional. Os valores de consumo volumétrico por cada combustível pode ser encontrado nas bases de dados da ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis).

Deve-se destacar, no entanto, que os modelos de previsão que serão propostos não vão apenas estimar a demanda gerada exclusivamente pelo setor de transportes: as aplicações futuras que se pode fazer do modelo, como, por exemplo, o planejamento

de ações de órgãos públicos e empresas privadas afim de responder à demanda futura de combustíveis e garantir o abastecimento, precisam das estimativas da demanda nacional por cada combustível como um todo, e não somente aquela gerada pelo setor de transportes. Nesse sentido, precisa-se explorar e quantificar todos os fatores significativos que geram demanda pelos quatro combustíveis a serem estudados.

Como exposto item 2.5, os autores referenciados na bibliografia costumam incluir dois tipos de variáveis em seus modelos de previsão de combustível: preço e renda. Com relação a preço, as variáveis serão geradas a partir das séries de média mensal do preço de revenda de cada combustível para o consumidor de varejo, as quais são fornecidas na base de dados da ANP. No entanto, é preciso determinar como esses preços devem entrar nos modelos de previsão. Por exemplo, no caso da demanda de gasolina e etanol, Silva et al. (2009) explicam que é preciso considerar nos modelos como os preços caminham um em relação ao outro, justamente por serem produtos substituíveis para os proprietários de carros *flex*. Outra questão que deve ser analisada é se o preço deve ser expresso em valores nominais ou reais (corrigido pela inflação). É possível que a influência do preço na demanda seja melhor considerada descontando-se os efeitos inflacionários, como sugere Bitencourt (2014). Levando-se em consideração essas questões, testaram-se a inclusão de variáveis de preços nominais e de preços corrigidos pela variação do IPCA acumulado (partindo como base o preço nominal do primeiro mês de observação) e avaliou-se, para cada combustível, qual delas gera uma performance melhor no modelo, considerando as demais variáveis exógenas candidatas a serem incluídas.

Com relação à renda, alguns autores aplicam o valor do PIB divulgado trimestralmente pelo IBGE como uma *proxy* para esse fator. No entanto, como essa variável possui divulgação trimestral e os modelos do presente trabalho possuem saídas e entradas mensais, sugerem-se outras variáveis de periodicidade mensal que podem ser usadas no lugar de PIB como *proxies* para a renda:

- Rendimento mensal real médio fornecido pela PNAD continua;
- Rendimento mensal real médio fornecido pela PME (Pesquisa Mensal do Emprego);

- A *proxy* mensal para o PIB com ajuste sazonal fornecida pelo monitor mensal de PIB da FGV (Fundação Getúlio Vargas). Esse índice é construído considerando-se os mesmos fatores usados para o cálculo do PIB e é reajustado sempre que o IBGE divulga os valores oficiais do PIB;
- Índice de atividade econômica do Banco Central (IBC-Br). Segundo o BC, esse índice capta as tendências de oscilação para o PIB, No entanto, é preciso ressaltar o fato de que seu cálculo envolve uma metodologia mais simplificada que a do PIB, levando-se em consideração um conjunto menor de fatores.

O único dos indicadores acima que possui dados mensais desde julho de 2001 (mês no qual se inicia a coleta de preços mensais dos combustíveis pela ANP e que, por essa razão, é o início da base de dados utilizada no presente trabalho) e que ainda é calculado é a *proxy* de PIB mensal fornecida pela FGV. A Pesquisa Mensal do Emprego foi encerrada em fevereiro de 2016, enquanto que a variável fornecida pela PNAD mensal contínua, por sua vez, só começou a ser calculada em 2012. De maneira semelhante, o índice IBC-Br também começou a ser calculado somente a partir de 2003. Logo, tendo em vista essas questões, optou-se por utilizar como variável de renda para o modelo a *proxy* mensal para o PIB com ajuste sazonal divulgada pela FGV.

Além desse conjunto de variáveis (que podem estar presente nos modelos de todos os combustíveis), analisou-se o mercado consumidor específico de cada um dos quatro combustíveis a serem estudados (gasolina C, do óleo diesel, do etanol hidratado e do querosene), visando identificar variáveis exógenas que quantificarão as forças relevantes por trás da geração de demanda que forem próprias de cada combustível. Uma vez identificadas variáveis de preço, de renda e outras variáveis específicas que podem ser relevantes para modelar a demanda de cada um dos quatro combustíveis em estudo, prossegue-se para a seleção de quais variáveis de fato devem compor o modelo final, a partir de um processo de seleção de variáveis que envolve a estimação dos coeficientes a elas atribuídos pelos métodos *Lasso* e

Stepwise (descritos no item 2.1) e o cálculo do coeficiente de correlação de Pearson entre a variável e a demanda do combustível em questão.

4.1 Óleo diesel

O BEN segrega os consumidores de óleo diesel em 8 grupos:

- Setor energético;
- Comércio;
- Setor público;
- Agropecuária;
- Setor rodoviário;
- Setor ferroviário;
- Setor hidroviário;
- Consumo final não energético (todos os demais setores utilizam óleo diesel para geração de energia).

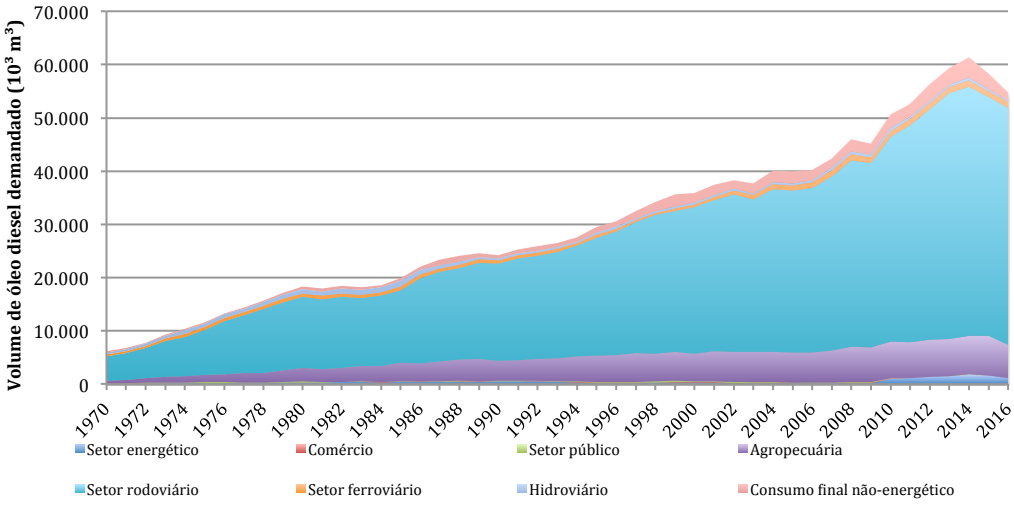

Figura 4.3 Gráfico da evolução da demanda de cada setor por óleo diesel (10^3 m^3)
Fonte: Balanço Energético Nacional de 2016

Consumidor final	Volume de óleo diesel demandado (10^3 m^3)	Participação no consumo total
Setor energético	1.149,39	2,10%
Comércio	9,52	0,02%
Setor público	2,98	0,01%
Agropecuária	6.179,35	11,30%
Setor rodoviário	44.552,97	81,45%
Setor ferroviário	1.122,60	2,05%
Hidroviário	264,73	0,48%
Consumo final não-energético	1.421,00	2,60%

Tabela 4.1 Volume demandado de óleo diesel por consumidor final e a participação relativa de cada consumidor em 2016
Fonte: Balanço Energético Nacional de 2016

Analisando os dados da figura 4.3 e da tabela 4.1, é fácil perceber que, historicamente, o setor rodoviário e o agropecuário são os que mais determinam o consumo de óleo diesel, representando conjuntamente 92,5% da demanda total do combustível em 2016.

Levando-se em conta as variáveis mencionadas no item 3.1 para descrever o setor rodoviário, e considerando a importância da inclusão das variáveis de preço e renda para descrever o consumo de combustíveis, sugerem-se como candidatas a compor os modelos de previsão de demanda de óleo diesel as variáveis exógenas abaixo:

- Índice ABCR-Pesado original (sem dessazonalização), já que esse índice reflete o fluxo da maioria dos veículos a diesel;
- Índice de Produção Industrial do IBGE, já que a produção industrial tem alta correlação com o fluxo de veículos pesados;
- Frota de veículos movidos a diesel (calculada a partir das estimativas de licenciamento de veículos e das curvas de sucateamento, como descrito no item 3.1.2);
- Volume de exportação agregada mensal do setor agropecuário. Ainda que possua participação bastante inferior ao setor rodoviário, a agropecuária corresponde a uma quantidade significativa de demanda de óleo diesel e, portanto, é possível que uma variável que mensure o nível de atividade desse setor em específico possa ser significativa para o modelo. A *proxy* sugerida para nível de atividade do setor agropecuário foi o volume de exportação

agregada do setor, principalmente levando-se em consideração que essa variável é frequentemente alvo de estudos de previsão por entidades do setor, como a Companhia Nacional de Abastecimento (CONAB), que divulga estimativas sobre a exportação do setor agrícola em publicações esporádicas. Os dados de exportação agrícola mensal podem ser encontrados nas bases de dados do MDIC (Ministério da Indústria, Comércio Exterior e Serviços);

- Preço do óleo diesel;
- Variável de renda (*proxy* mensal do PIB).

A produção e o licenciamento da frota de veículos futuros é constantemente estimada nos relatórios da ANFAVEA, enquanto que estimativas sobre o índice ABCR podem ser feitas a partir de previsões de tráfego divulgadas pelas concessionárias cotadas em bolsa (como a CCR e a Ecorodovias) ou mesmo pela própria ABCR. Esses valores podem ser usados pelo modelo para previsão da demanda de óleo diesel e para criação de cenários futuros de demanda a partir de certas hipóteses.

4.2 Etanol hidratado (álcool etílico hidratado)

O consumo de etanol hidratado atinge um número bem menor de setores que o óleo diesel. O BEN indica que, além do setor rodoviário, o etanol hidratado também é consumido diretamente pela agropecuária para geração de energia e também possui usos não-energéticos. No entanto, desde a introdução do etanol como combustível veicular no país, seu consumo acabou-se voltando principalmente para esse fim.

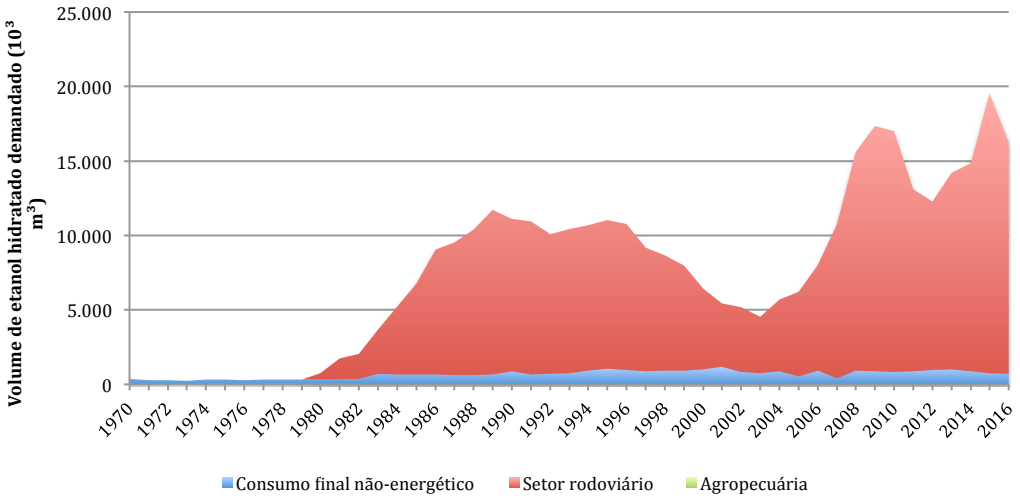

Figura 4.4 Gráfico da evolução da demanda de cada setor por etanol hidratado (10^3 m^3)
Fonte: Balanço Energético Nacional de 2016

Em 2016, a aplicação do etanol fora do setor rodoviário representou menos de 5% de todo o seu consumo:

Consumidor final	Participação no consumo total
Consumo final não-energético	4,27%
Setor rodoviário	95,62%
Agropecuária	0,10%

Tabela 4.2 Participação relativa de cada consumidor final na demanda volumétrica de etanol hidratado em 2016
Fonte: Balanço Energético Nacional de 2016

Novamente, considerando as questões expostas no item 3.1, sugerem-se algumas variáveis exógenas que refletem o nível das atividades do setor rodoviário que geram consumo de etanol. No entanto, testam-se duas formas de incluir as informações sobre frota no modelo, uma considerando a frota combinada de veículos *flex* e aqueles movidos exclusivamente a etanol e outra considerando essas frotas como duas variáveis exógenas distintas no modelo:

- Índice ABCR-Leve dessazonalizado;
- Frota de veículos *flex* e frota de veículos movidos exclusivamente a etanol, OU;
- Frota de veículos que podem ser abastecidos por etanol (frota agregada).

Além disso, seguindo a abordagem sugerida por Silva et al. (2009), procura-se introduzir uma variável que permita quantificar o efeito da elasticidade preço da demanda cruzada entre gasolina e etanol, já que esses produtos se tornaram substitutos para uma fatia crescente do mercado desde a introdução dos motores *flex-fuel*.

Teoricamente, a relação de preço entre gasolina e etanol que promoveria o equilíbrio entre as demandas desses combustíveis é aquela que permita que o consumidor disponha, com uma mesma quantia de dinheiro, da mesma quantidade de energia independentemente de qual dos dois combustíveis fosse adquirido. Para calcular qual seria essa relação de equilíbrio, recorre-se ao Anuário Estatístico Brasileiro de Petróleo, Gás Natural e Biocombustíveis de 2016 da ANP, que traz os fatores de conversão de cada combustível para a quantidade de barris equivalentes de petróleo (BEP), calculados a partir da densidade e do poder calorífico de cada tipo de combustível. Segundo o anuário, um metro cúbico de etanol hidratado equivale a 3,666 BEP, enquanto que um metro cúbico de gasolina C equivale a 5,101 BEP. Dessa forma, a relação de preço (R$/litro) que manteria a demanda entre os dois combustíveis em equilíbrio seria:

$$\left(\frac{\mathrm{P_{eta}}}{\mathrm{P_{gas}}}\right)_{Equilíbrio} = \frac{3{,}666}{5{,}101} \cong 0{,}7187 \qquad (4.1)$$

Isso significa que a variação no preço do etanol não impacta a demanda pelo combustível linearmente. Se o preço subir de tal forma que a relação $\frac{\mathrm{P_{eta}}}{\mathrm{P_{gas}}}$ se aproxime ou ultrapasse o patamar de 0,7187, o etanol deixa de ser competitivo com relação a gasolina e sua demanda seria penalizada por isso. Esse fenômeno se verifica na prática, como se pode observar no gráfico da Figura 4.5, que relaciona a demanda por etanol no tempo com o valor da razão $\frac{\mathrm{P_{eta}}}{\mathrm{P_{gas}}}$. Períodos de queda mais acentuada na demanda por etanol coincidem com os momentos em que $\frac{\mathrm{P_{eta}}}{\mathrm{P_{gas}}}$ atravessa o valor de equilíbrio de 0,719 (representada pela linha preta tracejada no gráfico).

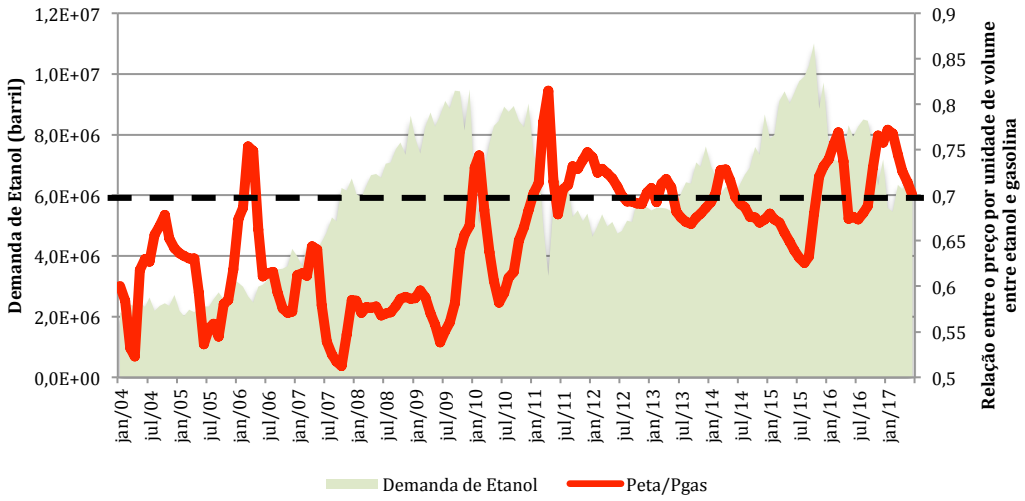

Figura 4.5 Gráfico da evolução da demanda nacional de etanol hidratado (barris) e da relação entre preço de etanol e preço de gasolina
Fonte: ANP

Dada a questão da elasticidade cruzada entre gasolina e etanol, propõem-se duas maneiras de introduzir o efeito da relação entre os preços no modelo de previsão de demanda de etanol:

- Introduzir a relação $\frac{P_{eta}}{P_{gas}}$ como uma variável do modelo, OU;
- Introduzir a variável de preço de gasolina e preço de etanol como duas variáveis separadas no modelo, que podem representar o valor nominal do preço do combustível ou o real (corrigido pela variação do IPCA).

Em suma, o modelo de previsão de consumo de etanol terá como candidatas as seguintes variáveis exógenas:

- Índice ABCR-Leve dessazonalizado;
- Frota de veículos *flex* e frota de veículos movidos exclusivamente a etanol;
- Frota de veículos que podem ser abastecidos por etanol;
- Relação de preços entre gasolina e etanol $\frac{P_{eta}}{P_{gas}}$;
- Preço de gasolina e preço de etanol (corrigidos ou não pela inflação);
- Variável de renda (*proxy* mensal do PIB).

4.3 Gasolina C

O BEN indica que toda a demanda de gasolina C é utilizada pelo setor rodoviário. Novamente, considerando as questões trabalhadas no item 3.1, sugerem-se variáveis para modelar o nível das atividades do setor rodoviário que geram consumo de gasolina. Analogamente ao etanol, testaram-se duas maneiras de incluir as informações sobre frota no modelo (uma única variável com todos os veículos que podem ser movidos a gasolina ou duas variáveis distintas de frota de veículos *flex* e frota de veículos movidos exclusivamente a gasolina). Também considerou-se a inclusão de variáveis que quantifiquem os efeitos de renda, de preço e da elasticidade cruzada entre gasolina e etanol. Dessa forma, as variáveis candidatas para compor os modelos de previsão de demanda de gasolina são as seguintes:

- Índice ABCR-Leve dessazonalizado;
- Frota de veículos *flex* e frota de veículos movidos exclusivamente a gasolina;
- Frota de veículos que podem ser abastecidos por gasolina;
- Relação de preço entre etanol e gasolina $\frac{P_{eta}}{P_{gas}}$;
- Preço de gasolina e preço de etanol (corrigidos ou não pela inflação);
- Variável de renda (*proxy* mensal do PIB).

4.4 Querosene

É possível verificar analisando o gráfico da Figura 4.6 que a perfil consumidor do querosene se alterou significativamente com o passar dos anos. De fato, devido à segurança e a facilidade de seu transporte, o querosene era muito utilizado nas antigas luminárias (querosene iluminante) e para aquecimento residencial. Em 2016, porém, 99,8% do consumo de querosene foi gerado pelo setor aeroviário (querosene de aviação, ou QAV).

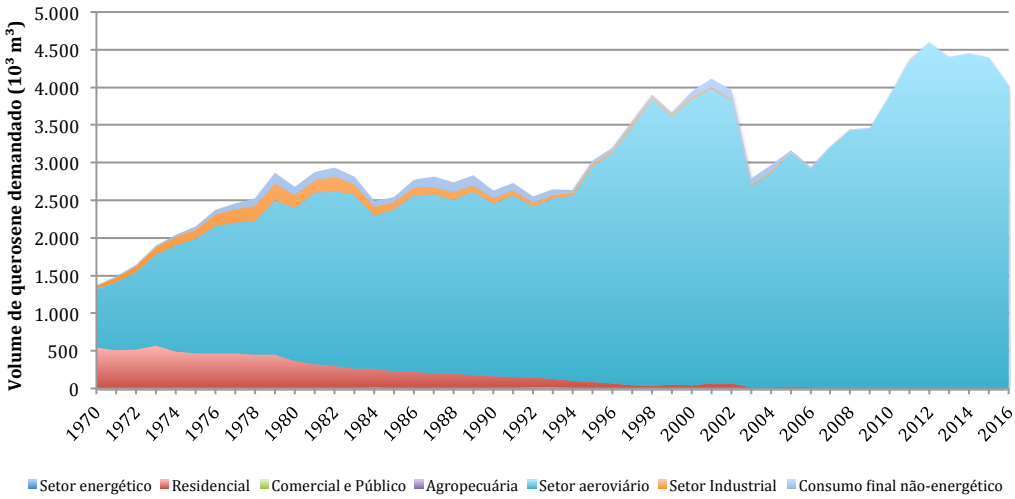

Figura 4.6 Gráfico da evolução da demanda de cada setor por querosene (10^3 m^3)
Fonte: Balanço Energético Nacional de 2016

As variáveis de demanda (RPK) e oferta (ASK) de voos nacionais e interacionais que foram discutidas no item 3.4 costumam ser estimadas para o curto e longo prazo pelas próprias empresas de capital aberto do setor (Azul, LATAM, Avianca e Gol), que divulgam e atualizam constantemente essas estimativas nos resultados trimestrais das empresas, nos relatórios mensais de tráfego e nas conferências com investidores. Assim, tem-se disponíveis estimativas futuras para demanda e oferta futuro de voos no país, que poderiam ser utilizadas num modelo para prever o consumo de querosene. Além disso, no item 3.4, constatou-se certa correlação negativa da atividade do setor aéreo com o valor do dólar real. Assim, propõe-se que o modelo de previsão de consumo de querosene possa incluir as seguintes variáveis:

- Valor real do dólar;
- Demanda (RPK) e oferta (ASK), segregadas por voos internacionais e nacionais;
- Demanda (RPK) e oferta (ASK) conjunta de voos internacionais e nacionais;
- Variável de renda (*proxy* mensal do PIB).

É preciso destacar que a ANP não realiza mais estimativas de preços mensais para o querosene de aviação, impossibilitando a inclusão da variável preço num eventual modelo de previsão de demanda de querosene.

4.5 Seleção final das variáveis independentes

Para facilitar a apresentação dos resultados do processo de seleção de variáveis entre as candidatas para cada um dos quatro combustíveis, introduz-se a notação abaixo. Cabe relembrar que os cálculos de todos os coeficientes foram obtidos usando os valores do algoritmo natural das variáveis abaixo, e não o valor original.

- $\mathbf{D_{gas}}, \mathbf{D_{eta}}, \mathbf{D_{od}}, \mathbf{D_{qav}}$ são as demandas volumétricas (barris) de gasolina, etanol, óleo diesel e querosene de aviação;
- $\mathbf{P_{n,gas}}, \mathbf{P_{n,eta}}, \mathbf{P_{n,od}}$ são as séries temporais de preços médios nominais de revenda (R$/l) de gasolina, etanol e óleo diesel;
- $\mathbf{P_{r,gas}}, \mathbf{P_{r,eta}}, \mathbf{P_{r,od}}$ são as séries temporais de preços médios de revenda (R$/l) corrigidos pela inflação, utilizando como ponto de partida o valor para o primeiro mês da série histórica de dados (julho de 2001);
- **Y** é a variável *proxy* mensal para PIB dessazonalidada fornecida pela FGV;
- **USD** é a cotação do dólar frente ao real corrigida pelo diferencial de inflação entre Estados Unidos e Brasil, utilizando-se como base o valor nominal da cotação no início da série de dados;
- $\mathbf{ABCR_L}, \mathbf{ABCR_P}$ são, respectivamente, o índice ABCR dessazonalizado para veículos leves e o índice ABCR original para veículos pesados;
- **IPI** é o Índice de Produção Industrial do IBGE;
- **Exp** é a exportação agrícola agregada em Kg do país;
- $\mathbf{ASK_{dom}}, \mathbf{ASK_{ext}}, \mathbf{RPK_{dom}}, \mathbf{RPK_{ext}}$ são, respectivamente, a oferta por voos domésticos e externos e a demanda por voos domésticos e externos;
- $\mathbf{ASK_+}, \mathbf{RPK_+}$ são, respectivamente, a oferta e demanda conjunta por voos internos e externos;
- $\mathbf{F_{gas}}, \mathbf{F_{eta}}, \mathbf{F_{od}}, \mathbf{F_{flex}}$ são as frotas de veículos com motores movidos exclusivamente a gasolina, etanol, diesel e com motores *flex*, respectivamente;
- $\mathbf{F_{gas+}}, \mathbf{F_{eta+}}$ são as frotas de veículos com motores que podem usar gasolina e dos veículos que podem usar etanol (não importando se são *flex* ou não).

No processo de seleção de variáveis independentes sugerido, calculou-se o coeficiente de correlação de Pearson entre demanda de combustível e as variáveis candidatas a integrarem o modelo e aplicou-se um teste de hipótese para verificar se a correlação é estatisticamente diferente de zero. Além disso, também foram obtidos os coeficientes que seriam atribuídos a essas variáveis caso fossem aplicados o método de regressão e seleção de variáveis *Lasso* e o *Stepwise* com a demanda de combustíveis sendo a variável resposta. Esses métodos foram detalhados no item 2.2. A decisão pela inclusão ou não de uma variável independente considerou, primeiramente, se o coeficiente de Pearson era diferente de zero e se os métodos de seleção de variáveis *Lasso* e *Stepwise* selecionaram a variável para compor o modelo. Em segundo lugar, comparam-se os sinais dos coeficientes com o que se esperava segundo a lógica do mercado. As variáveis que tinham valores opostos ao esperado foram descartadas. Finalmente, procurou-se não incluir no conjunto de variáveis para o modelo final duas ou mais variáveis que reflitam os mesmo fatores econômicos. Os resultados dos cálculos para cada um dos quatro combustíveis estão nas tabelas abaixo (Tabelas 4.3, 4.4, 4.5 e 4.6). Quando há zero para o coeficiente de correlação de Pearson, significa que se aceitou a hipótese nula para o coeficiente. Já quando há 0 para os coeficientes obtidos pelo método *Lasso* ou pelo *Stepwise*, significa que as variáveis não foram incluídas no modelo gerado por esses métodos.

Vale ressaltar que o coeficiente λ utilizado em cada regressão *Lasso* foi aquele que minimizou o erro quadrático médio total da regressão (EQMT). Para isso, simularam-se vários valores para o coeficiente λ e calcularam-se os EQMT correspondentes, utilizando-se do *software* MATLAB.

Óleo diesel				
Variáveis candidatas	Correlação de Pearson	Coeficientes *Lasso*	Coeficientes *Stepwise*	Descartar variável?
$P_{n,od}$	0.72	0	-0.058457	Sim
$P_{r,od}$	0	-0.0284	0	Não
Y	0.89	0.2074	0	Não
$ABCR_L$	0.95	0.1232	1.1459	Não
F_{od}	0.89	0.2903	0.44955	Não
Exp	0.43	0	0	Sim
IPI	0.56	0	0	Sim

Tabela 4.3 Seleção das variáveis independentes para compor modelos de previsão de óleo diesel
Fonte: Elaborado pelo autor

Etanol				
Variáveis candidatas	Correlação de Pearson	Coeficientes *Lasso*	Coeficientes *Stepwise*	Descartar variável?
$P_{n,eta}$	0.64	0	4.1459	Sim
$P_{n,gaso}$	0.76	-1.2442	0	Sim
$P_{r,eta}$	-0.47	0	-3.0813	Não
$P_{r,gaso}$	-0.72	2.5344	0	Sim
P_{eta}/P_{gas}	0	-1.8237	-3.3233	Não
Y	0.78	3.4233	0	Não
$ABCR_L$	0.76	0	-1.5105	Sim
F_{flex}	0.71	0.1955	0	Sim
F_{eta}	-0.81	0.8234	3.6372	Sim
F_{eta+}	0.73	1.6913	1.8995	Não

Tabela 4.4 Seleção das variáveis independentes para compor modelos de previsão de etanol
Fonte: Elaborado pelo autor

Gasolina C				
Variáveis candidatas	Correlação de Pearson	Coeficientes *Lasso*	Coeficientes *Stepwise*	Descartar variável?
$P_{n,eta}$	0.85	0	0	Sim
$P_{n,gaso}$	0.81	0	0	Sim
$P_{r,eta}$	0	0	-0.39526	Sim
$P_{r,gaso}$	-0.39	0	0	Sim
P_{eta}/P_{gas}	0.67	0.2931	0.80849	Não
Y	0.87	-0.0094	0	Sim
$ABCR_L$	0.95	0.5635	0.90366	Não
F_{flex}	0.91	-0.0184	0	Sim
F_{gas}	-0.85	-0.0167	0	Sim
F_{gas+}	0.94	1.0778	0.34044	Não

Tabela 4.5 Seleção das variáveis independentes para compor modelos de previsão de gasolina C
Fonte: Elaborado pelo autor

Querosene				
Variáveis candidatas	Correlação de Pearson	Coeficientes *Lasso*	Coeficientes *Stepwise*	Descartar variável?
Y	0.93	0	0	Sim
ASK_{dom}	0.97	0	0	Sim
ASK_{ext}	0.68	0	0	Sim
RPK_{dom}	0.96	0	0	Sim
RPK_{ext}	0.68	0	0	Sim
ASK_{+}	0.98	0.5841	0.71881	Não
RPK_{+}	0.97	0	0	Sim
USD	-0.66	0	0	Sim

Tabela 4.6 Seleção das variáveis candidatas para compor modelos de previsão de querosene
Fonte: Elaborado pelo autor

5 CONSTRUÇÃO E COMPARAÇÃO DE MODELOS

Nesse capítulo, as três grandes categorias de modelagem (redes neurais, modelos ARIMA e regressão linear) serão aplicadas para previsão da demanda volumétrica de gasolina C. Os mesmos procedimentos podem ser realizados para os demais combustíveis citados ao longo do trabalho.

Para facilitar a apresentação dos cálculos, vamos definir a seguinte notação para as variáveis que serão utilizadas na modelagem da demanda de gasolina C:

- Y ou ln (D_{gas}) é o logaritmo natural da demanda de gasolina em barris;
- X_1 ou ln($ABCR_L$) é o logaritmo natural do índice ABCR para fluxo de veículos leves original nas estradas pedagiadas, sem correção sazonal;
- X_2 ou ln (P_{eta}/P_{gas}) é o logaritmo natural da relação entre preço de etanol e preço de gasolina no tempo;
- X_3 ou ln(F_{gas+}) é o logaritmo natural do tamanho da frota agregada de todos os veículos que podem ser movidos a gasolina, o que inclui aqueles com motores *flex* e aqueles que o motor só aceita gasolina.

Além disso, também será adotada a seguinte denominação para as categorias de modelagem propostas:

- Grupo I: Regressão Linear
- Grupo II: Modelos ARIMA
- Grupo III: Rede Neural Artificial

Num primeiro momento, aplicam-se as metodologias de modelagem descritas no capítulo 2 para gerar um modelo representante de cada categoria de modelos. Posteriormente, a performance preditiva de curto e longo prazo desses três modelos é comparada por meio do método de validação cruzada com janela de tempo. Finalmente, o modelo final é selecionado e será aplicado para prever a demanda mensal dos próximos 12 meses.

5.1 Grupo I - Regressão Linear

A primeira metodologia de modelagem aplicada é a regressão linear. Considerando que todas as variáveis selecionadas no capítulo 4 devem ser incluídas no modelo, a única fonte de variação que pode ocorrer entre esses modelos é a maneira como os coeficientes para esse conjunto de variáveis foram estimados. No item 2.1, além do método de estimação de coeficientes a partir do método canônico de mínimos quadrados ordinários (MQO), foram propostos dois métodos de regressão com penalização (regressão *Ridge* e *Lasso*), que são indicados nos casos em que se verifica maior presença de colinearidade e multicolinearidade. Por isso, para saber qual método seria o mais adequado levando-se em conta as variáveis independentes a serem incluídas no modelo, deve-se realizar alguns testes para diagnosticar o grau de colinearidade e multicolinearidade entre elas.

Visando identificar a presença de colinearidade, obtiveram-se as correlações dois a dois entre todas as três variáveis independentes do modelo. Para isso, recorreu-se ao cálculo da matriz de correlação ρ:

$$\rho = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & 1 & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.594 & 0.862 \\ 0.594 & 1 & 0.603 \\ 0.862 & 0.603 & 1 \end{bmatrix} \qquad (5.1)$$

É fácil perceber que as variáveis possuem correlação entre si (o que indica a presença de colinearidade), principalmente entre as variáveis de fluxo de veículos leves dessazonalizado e de frota (X_1 e X_3, respectivamente), que se mostram fortemente correlacionadas entre si.

O próximo teste envolve o cálculo dos Fatores de Inflação da Variância (denotado por VIF, sigla de *Variance Inflation Factor*) de cada variável independente. Conforme apresentado por Reynaldo et al. (1997), esse vetor é um indicativo da existência de relação linear aproximada entre as variáveis independentes, o que configuraria multicolinearidade. Nesse caso, segundo as autores, a matriz (X^TX) tornar-se-ia próxima a uma matriz singular, o que não somente gera dificuldades

computacionais para invertê-la, como também torna os coeficientes de mínimos quadrados ordinários muito instáveis a pequenas variações da amostra X utilizada para calcular os coeficientes da regressão, o que poderia inviabilizar a capacidade preditiva desses modelos fora do período de calibração. Como já no item 2.1, Reynaldo et al. (1997) afirmam que se deva considerar a multicolinearidade entre as variáveis independentes significativa se algum os fatores de inflação de variância forem superiores a 10. O cálculo dos fatores de variância considerando as três variáveis propostas para modelar a demanda de gasolina encontram-se abaixo:

$$\mathrm{VIF} = \mathrm{diag}((\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}) = [0.727 \quad 0.07 \quad 0.837] \tag{5.2}$$

Portanto, pode-se afirmar que o cálculo dos VIF não acusou presença de multicolinearidade significativa. Isso provavelmente significa que os estimadores de MQO devam ser usados preferencialmente ao invés de estimadores viesados como os dos métodos *Ridge* e *Lasso*. No entanto, para efeitos de comparação e considerando que a correlação dois a dois das variáveis independentes é bastante considerável, optou-se por aplicar o método de regressão *Ridge* e comparar seus resultados com o método dos mínimos quadrados ordinários.

5.1.1 Regressão Linear com estimadores de mínimos quadrados ordinários (MQO)

Como visto no capítulo 2, o método dos mínimos quadrados ordinários fornece estimadores que minimizam a soma dos erros quadráticos do modelo linear sem que se imponha restrições ao valor dos coeficientes lineares obtidos. Utilizando-se de toda a série histórica de dados disponível para calibração, obtêm-se os seguintes resultados:

	Coeficientes	Desvio padrão	Estatística t	P-value
Intercepto (β_0)	8.5171	1,0919	7,8001	0,0000
$\ln(ABCR_L)$	0,4264	0,0685	6,2226	0,0000
$\ln(P_{eta}/P_{gas})$	0,8976	0,1936	4,6362	0,0000
$\ln(F_{gas+})$	0,2304	0,1164	1,9796	0,0492

Tabela 5.1 Resultados da regressão utilizando estimadores de MQO
Fonte: Elaborado pelo autor

A equação do modelo de previsão, portanto, assumiria a seguinte forma:

$$\ln(D_{gas}) = 8{,}52 + 0{,}43\ln(ABCR_L) + 0{,}90\ln(P_{eta}/P_{gas}) + 0{,}23\ln(F_{gas+}) \quad (5.3)$$

5.1.2 Regressão Linear com estimadores *Ridge*

Como visto no item 2.1.3, a regressão *Ridge* fornece estimadores que minimizam a soma quadrática dos erros do modelo, porém exige que a soma quadrática dos coeficientes lineares estimados seja inferior a uma constante positiva pré definida, o que é equivalente a resolver o seguinte problema de minimização:

$$\min z(\beta_0,\beta_1,\dots,\beta_N) = \sum_{i=1}^{n}\left(Y_i - \sum_{k=0}^{N}\beta_k(X_k)_i\right)^2 + \lambda\sum_{k=0}^{N}(\beta_k)^2$$

Cada valor de λ produzirá coeficientes de regressão diferentes. Para avaliar a influência desse parâmetro no valor dos coeficientes de cada uma das três variáveis independentes sugeridas para modelar a demanda de gasolina, simularam-se vários valores de λ e foram obtidos os coeficientes *Ridge* correspondentes.

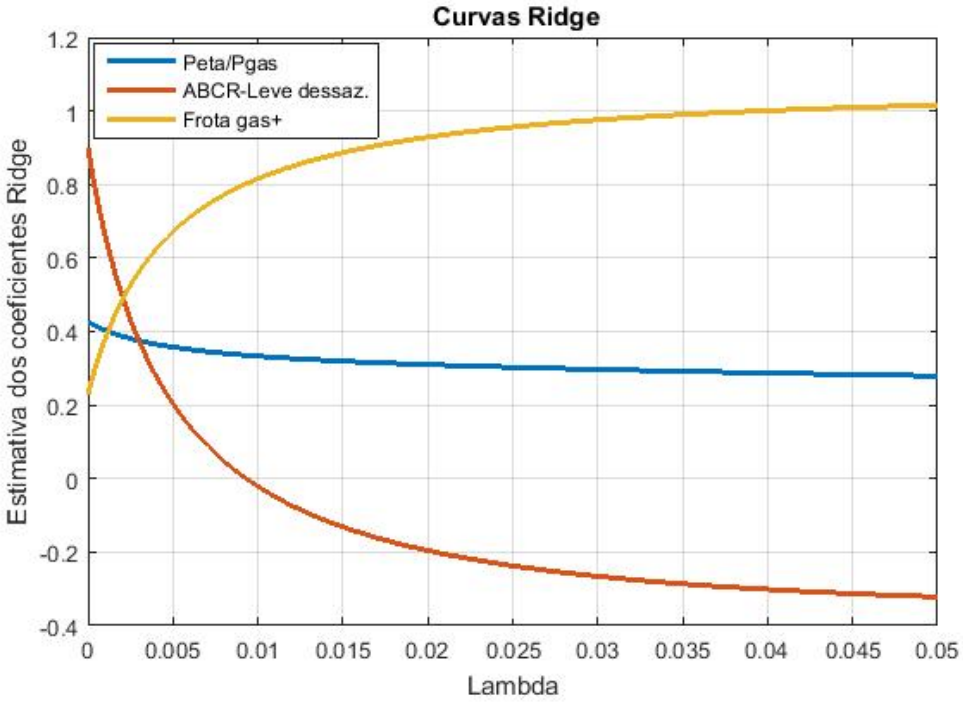

Figura 5.1 Gráfico dos valores estimados para os coeficientes *Ridge* em função do valor de λ
Fonte: Elaborado pelo autor

A Figura 5.1 apresenta o valor calculado para os coeficientes do modelo em função do valor de λ. Pode-se verificar que pequenos valores de λ (representado no eixo das abscissas, com valores entre 0 e 0,05) alteraram significativamente as estimativas dos MQO (que são iguais às estimativas para os coeficientes *Ridge* com $\lambda = 0$, ou seja, os valores do eixo das ordenadas do gráfico). Já a gráfico da Figura 5.2 mostra o valor do EQMT (Erro Quadrático Médio Total) em função de λ. Nota-se que o erro cresce rapidamente com o aumento do valor desse parâmetro. Isso é um indicativo de que as variáveis independentes não apresentam colinearidade e multicolinearidade e que, portanto, devem ser utilizados os regressores MQO ao invés dos regressores *Ridge* para esse problema.

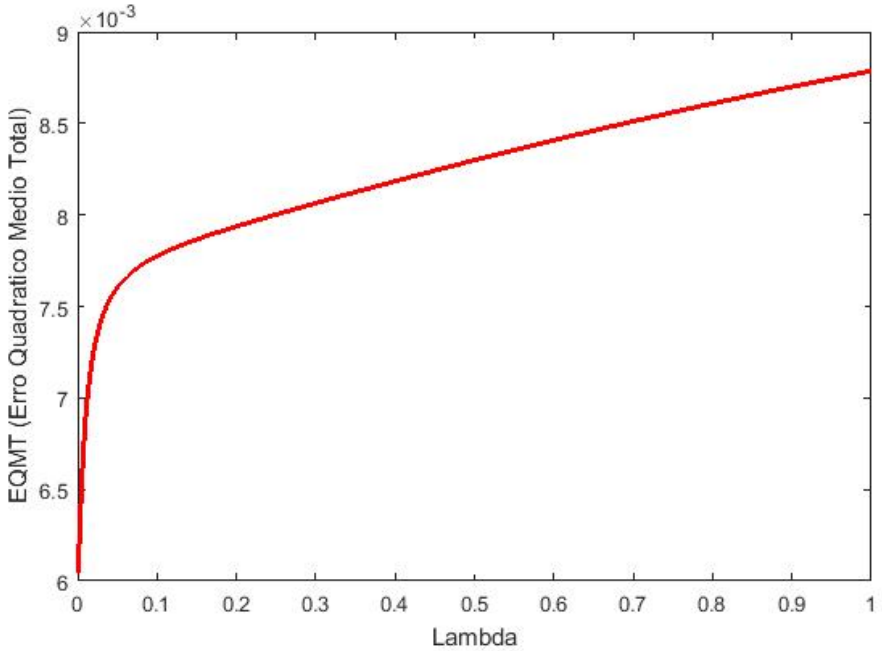

Figura 5.2 Gráfico do valor do EQMT do modelo em função do valor de λ
Fonte: Elaborado pelo autor

5.1.3 Verificação do modelo de regressão linear – Análise de resíduos

Uma vez estimados os parâmetros do modelo de regressão linear, deve-se prosseguir para a etapa de verificação do modelo. Em outras palavras, é preciso verificar se o modelo atende às hipóteses da regressão linear. A hipótese de ausência de multicolinearidade e colinearidade aparentemente é obedecida, considerando os resultados obtidos por meio do cálculo dos regressores *Ridge* e do cálculo dos fatores de inflação da variância. Para avaliar as demais hipóteses, deve-se analisar o comportamento dos resíduos do modelo calibrado.

5.1.3.1 Diagnóstico de normalidade

Iniciando pelo diagnóstico de normalidade dos resíduos, aplicou-se o teste de Kolmogorov-Smirnov. Não só se pode rejeitar a hipótese nula (ausência de normalidade) considerando um nível de significância $\alpha = 5\%$, como também obteve-se um *p-value* associado ao teste bastante reduzido $(3{,}52.10^{-31})$, o que indica que os erros seguem uma distribuição normal. Chega-se à mesma conclusão utilizando-se do papel de probabilidade normal, já que os resíduos no papel formam uma reta (Figura 5.3).

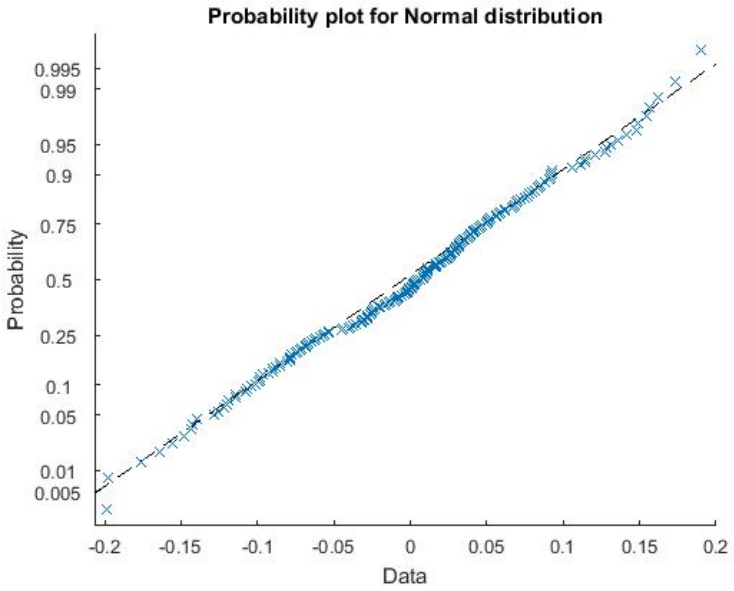

Figura 5.3 Papel de probabilidade normal dos resíduos do modelo de regressão linear
Fonte: Elaborado pelo autor

Também traçou-se a função de distribuição acumulada empírica dos resíduos com a função de distribuição acumulada teórica da normal (Figura 5.4) e fez-se o histograma dos resíduos (Figura 5.5), que também parecem confirmar a normalidade. No entanto, vale destacar que os valores dos resíduos parecem não apresentarem média nula: um viés levemente positivo pode ser observado.

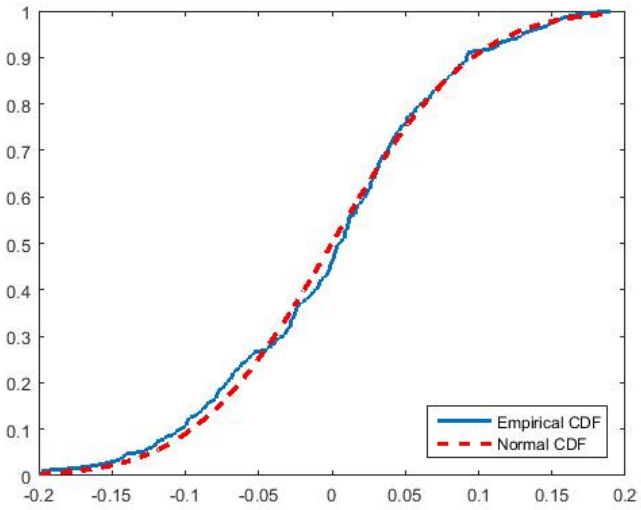

Figura 5.4 Gráfico da FDA empírica dos resíduos da regressão linear e a FDA teórica da normal
Fonte: Elaborado pelo autor

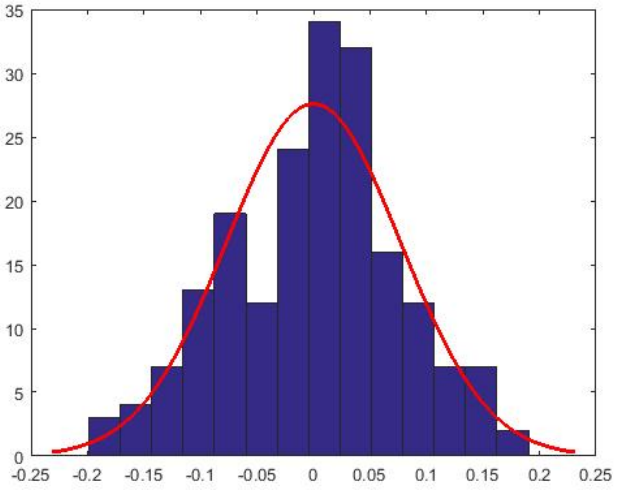

Figura 5.5 Histograma dos resíduos da regressão linear
Fonte: Elaborado pelo autor

5.1.3.2 Diagnóstico de autocorrelação

Por outro lado, não se pode afirmar que a hipótese de independência entre os resíduos está sendo respeitada. Pelo gráfico de dispersão (Figura 5.6), pode-se perceber que os resíduos não estão distribuídos aleatoriamente no tempo, o que é um forte indicativo de autocorrelação nos resíduos: parece haver uma concentração de erros positivos nos primeiros e nos últimos meses da série histórica, enquanto os erros negativos concentram-se nos meses intermediários.

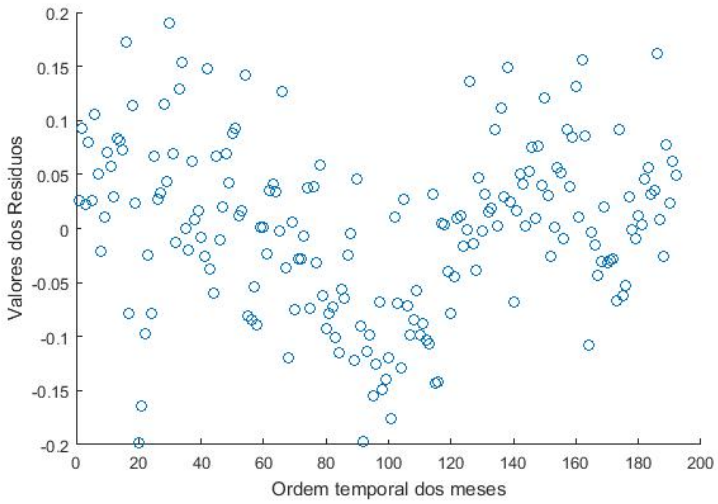

Figura 5.6 Gráfico de dispersão dos resíduos da regressão linear em função da ordem temporal dos meses
Fonte: Elaborado pelo autor

A presença de autocorrelação nos resíduos também é confirmada pelo teste de Durbin-Watson. Ao aplicar o teste, obteve-se um *p-value* de $2.038.10^{-18}$, que indica que a hipótese nula (ausência de autocorrelação de primeira ordem) deve ser rejeitada. O cálculo da função de autocorrelação (fac) dos resíduos reforça essa conclusão (Figura 5.7).

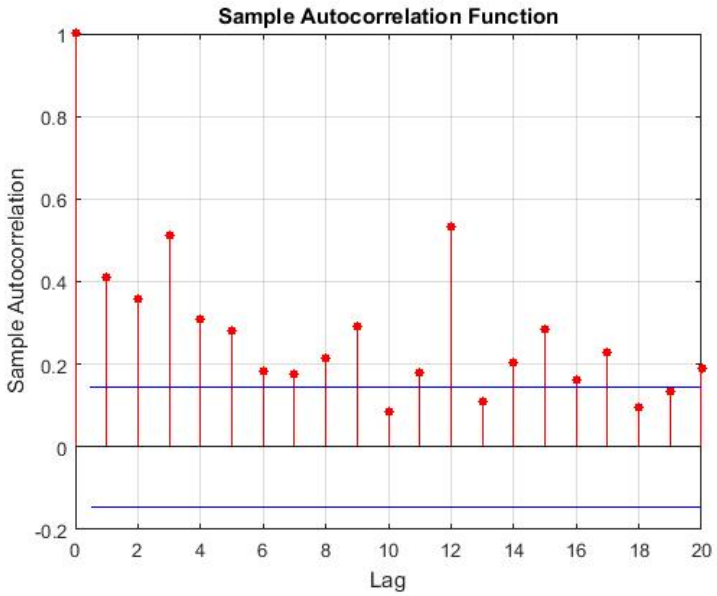

Figura 5.7 Gráfico da função de autocorrelação dos resíduos do modelo de regressão linear
Fonte: Elaborado pelo autor

É interessante destacar que a função de autocorrelação indicou correlação alta nos primeiros *lags* e no *lag* sazonal (de 12 períodos), o que pode sugerir que o modelo poderia ser melhorado com a introdução de variáveis com defasagem e variáveis sazonais.

Segundo Bollerslev (1986), a presença de autocorrelação nos resíduos põe em questão a confiabilidade dos testes t e F para a regressão linear, na medida que pode levar à subestimação da variância do erro e dos coeficientes. Portanto, há mais chances de o teste t e F indicar significância dos parâmetros do modelo, ainda que eles não sejam significativos.

5.1.3.3 Diagnóstico de homocedasticidade

Finalmente, parte-se para o diagnóstico da condição de homocedasticidade dos resíduos, que é a última hipótese da regressão linear a ser testada. Para isso, utilizou-se um gráfico que relaciona os resíduos com os respectivos valores previstos pelo modelo. É fácil de se verificar que os resíduos não estão dispostos aleatoriamente no gráfico, conforme se observa na Figura 5.8. Além de detectar heterocedasticidade, a dispersão parece sugerir a existência de relações não-lineares de segunda ordem entre as variáveis independentes e a demanda de gasolina, já que os resíduos parecem formar uma parábola.

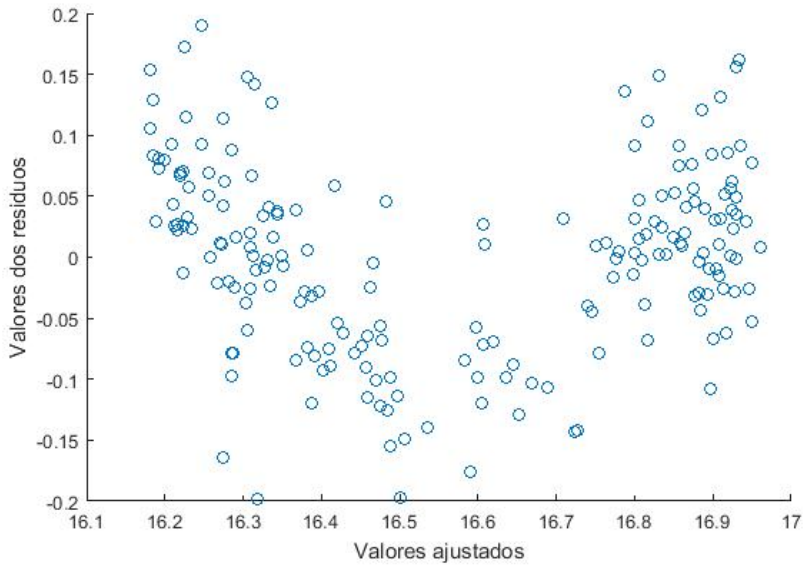

Figura 5.8 Gráfico dos resíduos em função dos valores ajustados do modelo de regressão linear
Fonte: Elaborado pelo autor

Além disso, a rejeição da hipótese de homocedasticidade também é reforçada pelo teste de Breusch-Pagan, em que se obtém um *p-value* associado de $1.044.10^{-4}$, indicando a rejeição da hipótese nula de que variância dos resíduos é constante.

Bollerslev (1986) afirma que a presença de homocedasticidade gera viés nos erros padrões estimados, o que torna as conclusões dos testes t e F duvidosas.

Conclui-se que, apesar de o modelo de regressão linear parecer cumprir os requisitos de normalidade dos resíduos e de ausência de multicolinearidade e colinearidade entre as variáveis independentes do modelo, as demais hipóteses da regressão linear não foram respeitadas (ausência de autocorrelação e variância constante dos resíduos). Isso pode ser um indicativo de que seja interessante que se opte por outras variáveis ou outros métodos de modelagem para gerar previsões futuras para a demanda de gasolina.

5.2 Grupo II: Modelos ARIMA

A segunda categoria de modelos analisada são os modelos ARIMA. Seguindo a abordagem sugerida por Box e Jenkins (1970), especificou-se que, para esse trabalho, os espectro de modelos que se incluem nessa categoria são os processos $\mathrm{AR(p)}$, $\mathrm{MA(q)}$, $\mathrm{ARMA(p,q)}$, $\mathrm{ARIMA(p,d,q)}$ e $\mathrm{SARIMA(p,d,q)_{\times}(P,D,Q)_s}$. Posteriormente, prosseguiu-se para a etapa de identificação do modelo, que consistiu em definir, qual processo e quais valores para p, d, q, P, D, Q e s que mais se adequaram ao comportamento da série temporal da demanda volumétrica de gasolina. Por fim, estimam-se os parâmetros do modelo com o auxílio do *software* MATLAB e verifica-se a adequação do modelo ao problema por meio da análise de resíduos.

5.2.1 Análise das funções de autocorrelação e autocorrelação parciais

Por meio dos gráficos das funções de autocorrelação (fac) e autocorrelação parcial (facp) amostral da variável de demanda de gasolina com transformação logarítmica, pode-se notar que a série não é estacionária, já que a função de autocorrelação não mostra decaimento exponencial rápido. Em outras palavras, a autocorrelação se mostra significativa mesmo para *lags* bastante altos, havendo uma queda bastante lenta na autocorrelação com o avanço de *lag*, como é possível ver pela Figura 5.9:

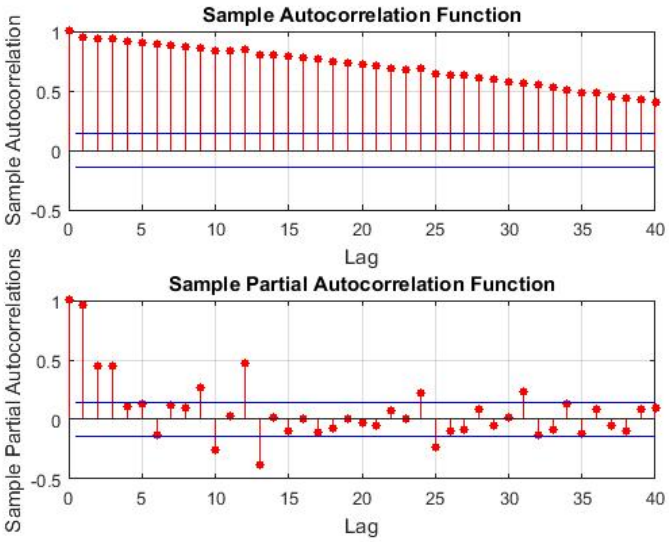

Figura 5.9 Gráfico da fac e facp da variável ln(D_{gas})
Fonte: Elaborado pelo autor

Visando tornar a série estacionária, aplicaram-se a primeira (Figura 5.10) e a segunda diferença (Figura 5.11) na variável de demanda de gasolina logarítmica.

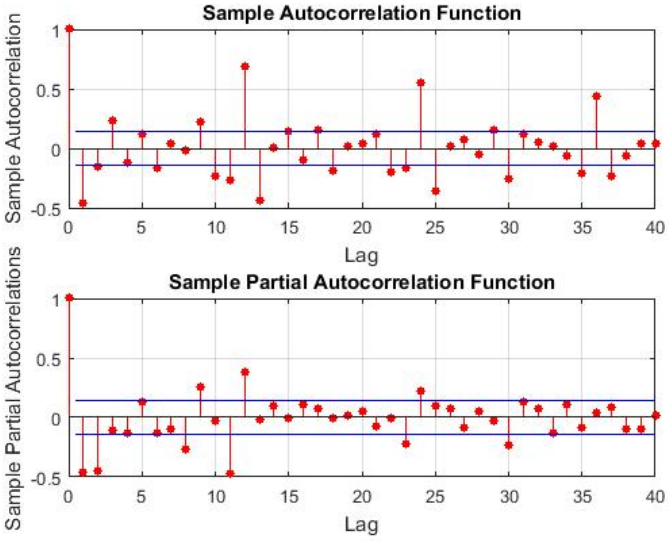

Figura 5.10 Gráfico da fac e facp da série diferenciada uma única vez (d = 1)
Fonte: Elaborado pelo autor

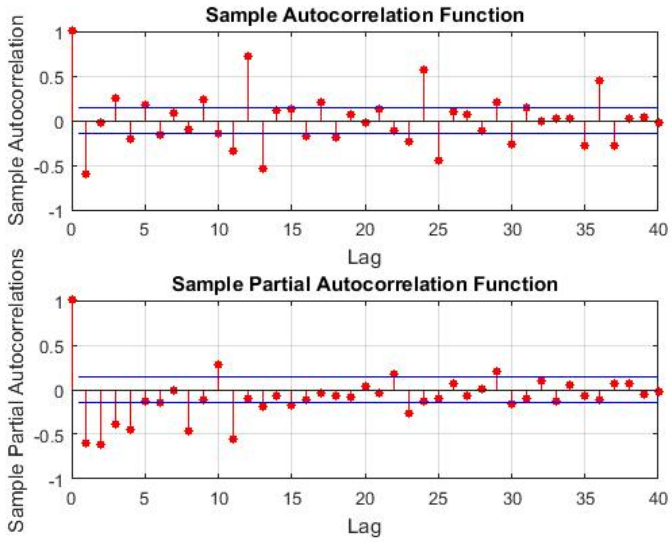

Figura 5.11 Gráfico da fac e facp da série diferenciada duas vezes (d = 2)
Fonte: Elaborado pelo autor

A transformação eliminou a significância da autocorrelação em altos *lags* que havia sido identificada na série original (com exceção nos *lags* sazonais, em 12, 24 e 36). Com relação ao número de diferenciações, o fato de a autocorrelação no *lag* 1 da série com duas diferenças, Figura 5.11, ser consideravelmente negativo (menor que -0.5) pode sugerir que a série foi diferenciada demais. Além disso, não há diferenças significativas na velocidade de decaimento entre as duas séries diferenciadas. Por isso, optou-se por apenas uma diferenciação (resultando em d = 1). No entanto, ainda se verifica alta correlação nos *lags* 12, 24 e 36 para ambas as séries diferenciadas, que decaem lentamente com o tempo, o que é coerente com o comportamento sazonal da demanda de gasolina e indica que a série ainda não é estacionária. Esse fato sugere que seja aplicado o modelo do ARIMA sazonal (SARIMA), com s = 12.

Em seguida, foram obtidas as funções de autocorrelação e autocorrelação parcial das séries originadas da aplicação em uma (Figura 5.12) ou duas vezes (Figura 5.13) do operador de diferença sazonal $\Delta_{12} = (1 - L^{12})$ na série já diferenciada uma única vez pelo operador de diferença $\Delta = (1 - L)$. Novamente se optou por aplicar apenas uma diferença sazonal (D = 1), já que o decaimento das séries com D = 1 e D = 2 não apresentou diferenças significativas. Além disso, a correlação no *lag* 12 tornou-se menor do que -0.5 com D = 2, o que pode ser um indicativo de diferenciação excessiva.

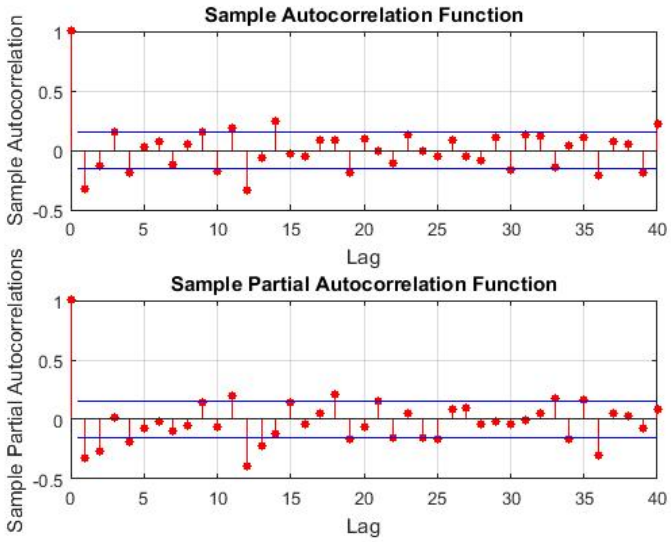

Figura 5.12 Gráfico da fac e facp da série com uma diferença sazonal (D = 1)
Fonte: Elaborado pelo autor

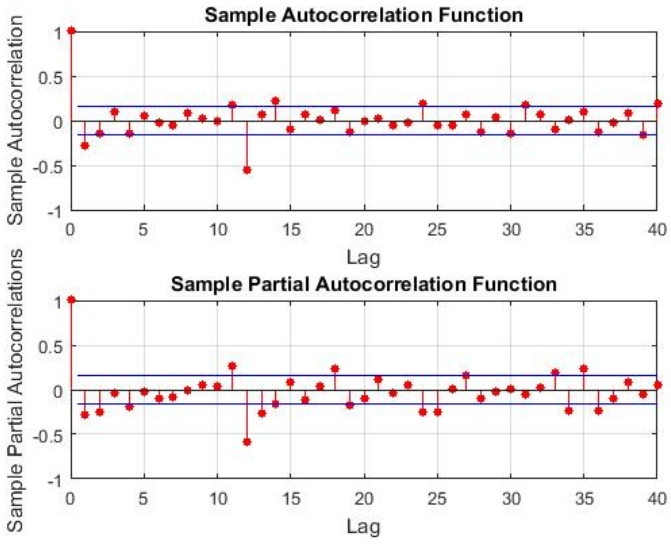

Figura 5.13 Gráfico da fac e facp da série com duas diferenças sazonais (D = 2)
Fonte: Elaborado pelo autor

Com d = 1 (nível de diferenciação) e D = 1 (nível de diferenciação sazonal), a série de demanda de gasolina logarítmica aparenta ter se tornado estacionária. Nesse momento, é preciso analisar o comportamento da fac e facp para que se possa identificar os valores dos parâmetros p (nível de auto-regressão), q (nível de médias móveis), P (nível de auto-regressão sazonal) e Q (nível de médias móveis sazonal).

Analisando novamente a Figura 5.12, pode-se perceber que tanto a autocorrelação quanto a correlação parcial não se mostram muito significativas nos *lag* 24 e 36: elas aparentam ser bastante significativas apenas no *lag* 12, o que é coerente com P = 1 e Q = 1, respectivamente.

Analisando os *lags* menores na fac e na facp da Figura 5.12, considerou-se que a autocorrelação parece perder relevância a partir do *lag* 4, o que é coerente com $p = 4$. Com relação ao valor de q, decidiu-se testar inicialmente dois valores, q = 2 e q = 4, já que o comportamento da facp nos lags 3 e 4 pode gerar dúvidas se os coeficientes MA(3) e MA(4) são de fato relevantes.

Em suma, o processo de identificação a partir da análise das funções de autocorrelação e autocorrelação parcial permitiu levantar dois modelos tentativos:

- $\text{SARIMA}(4,1,2)_{\times}(1,1,1)_{12}$ e;
- $\text{SARIMA}(4,1,4)_{\times}(1,1,1)_{12}$.

Para determinar qual dos dois será selecionado para representar o grupo dos modelos ARIMA, estimaram-se os valores dos coeficientes dos termos auto-regressivos (AR), de médias móveis (MA), auto-regressivos sazonais (SAR) e de médias móveis sazonais (SMA) de cada um dos modelos, além dos indicadores AIC e BIC. Os resultados para o modelo $\text{SARIMA}(4,1,4)_{\times}(1,1,1)_{12}$ e para o modelo $\text{SARIMA}(4,1,2)_{\times}(1,1,1)_{12}$ encontram-se, respectivamente, nas Tabelas 5.2 e 5.3.

$SARIMA(4,1,4)_{\times}(1,1,1)_{12}$			
	Coeficientes	Desvio padrão	Estatística t
AR{1}	-0,0038	1815,8	0,0000
AR{2}	0,7766	536434,0	0,0014
AR{3}	0,1289	1253,8	0,0001
AR{4}	-0,0963	599737,0	-0,0002
SAR{1}	0,2918	1816,0	0,0002
MA{1}	-0,1004	1,0152	-0,0989
MA{2}	-1,0256	0,2467	-4,1576
MA{3}	0,1004	1,0069	0,0997
MA{4}	0,0256	0,2611	0,0982
SMA{1}	-0,5602	1,2386	-0,4523
AIC	-576,6091	BIC	-544,0342

Tabela 5.2 Estimação dos coeficientes e dos indicadores AIC e BIC para o modelo com q = 4
Fonte: Elaborado pelo autor

$SARIMA(4,1,2)_{\times}(1,1,1)_{12}$			
	Coeficientes	Desvio padrão	Estatística t
AR{1}	-0,185	0,300	-0,616
AR{2}	0,714	0,148	4,829
AR{3}	0,257	0,206	1,245
AR{4}	-0,021	0,144	-0,149
SAR{1}	0,485	0,363	1,337
MA{1}	0,000	0,038	0,000
MA{2}	-1,000	0,041	-24,420
SMA{1}	-0,660	0,177	-3,730
AIC	-580,4356	BIC	-554,3756

Tabela 5.3 Estimação dos coeficientes e dos indicadores AIC e BIC para o modelo com q = 2
Fonte: Elaborado pelo autor

Os cálculos dos indicadores AIC e BIC, que levam em conta o princípio da parcimônia e penalizam os modelos mais complexos, sugerem que o modelo com apenas duas variáveis de média móvel é levemente superior ao modelo com quatro, já que eles indicam valores menores para modelos de maior performance. Além disso, os coeficientes do SARIMA com menos parâmetros possuem em geral estatística t maiores e variância menores do que seus pares no modelo com maior número de variáveis, em especial para as variáveis auto-regressivas. Por isso, optou-se pelo modelo $\mathrm{SARIMA}(4,1,2)_{\times}(1,1,1)_{12}$ para representar a categoria de modelos ARIMA na comparação entre as três categorias de modelagem propostas.

5.2.2 Verificação do modelo SARIMA – Análise de resíduos

Uma vez tendo ndentificado e estimado o modelo SARIMA proposto, procedeu-se para a verificação do modelo a partir da análise dos seus resíduos, utilizando testes semelhantes àqueles da regressão linear.

5.2.2.1 Diagnóstico de Normalidade

Novamente, traçando-se os resíduos no papel de probabilidade normal (Figura 5.14), obteve-se aproximadamente uma reta, o que é coerente com resíduos distribuídos normalmente. A semelhança da função de probabilidade acumulada empírica dos resíduos com a função de probabilidade teórica da distribuição normal aponta para a mesma conclusão (Figura 5.15). Também se pode identificar que os resíduos parecem estar normalmente distribuídos em torno do zero por meio do histograma (Figura 5.16). Finalmente, a hipótese de normalidade também foi confirmada pelo teste de Kolmogorov-Smirnov, com *p-value* de $2.0831.10^{-29}$;

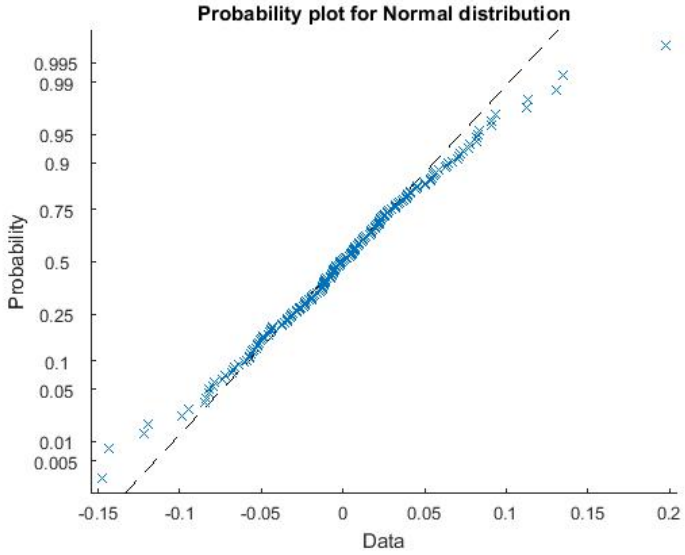

Figura 5.14 Papel de probabilidade normal dos resíduos do modelo SARIMA
Fonte: Elaborado pelo autor

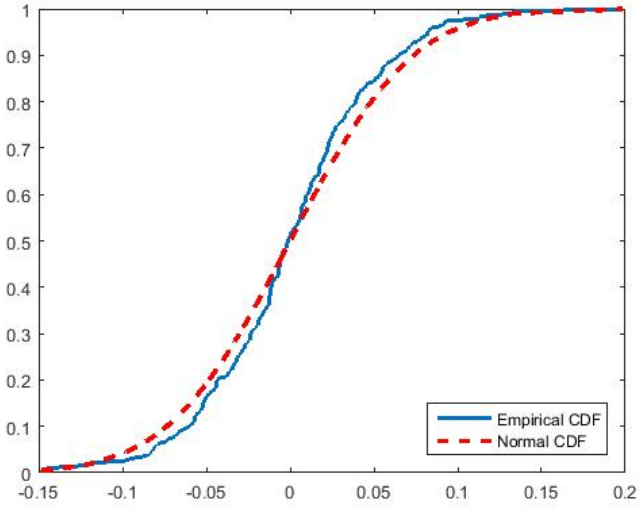

Figura 5.15 Gráfico da FDA empírica dos resíduos do modelo SARIMA e a FDA teórica da normal
Fonte: Elaborado pelo autor

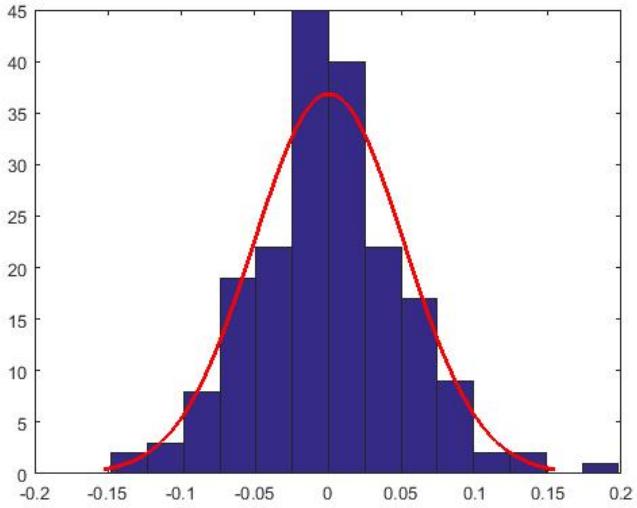

Figura 5.16 Histograma dos resíduos do modelo SARIMA
Fonte: Elaborado pelo autor

5.2.2.2 Diagnóstico de autocorrelação

Analisando o gráfico de dispersão dos resíduos no tempo (Figura 5.17), pode-se perceber que os pontos no gráfico aparentam estar aleatoriamente distribuídos, o que indica que os resíduos são independentes entre si. A função de autocorrelação (Figura 5.18) dos resíduos também parece confirmar essa hipótese.

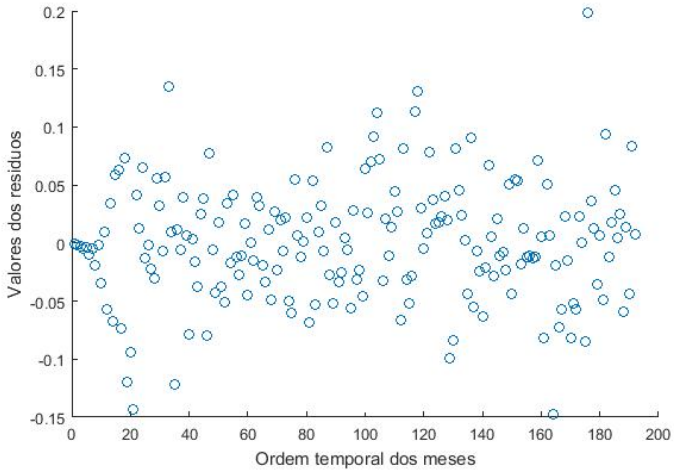

Figura 5.17 Gráfico de dispersão dos resíduos do modelo SARIMA em função da ordem temporal dos meses
Fonte: Elaborado pelo autor

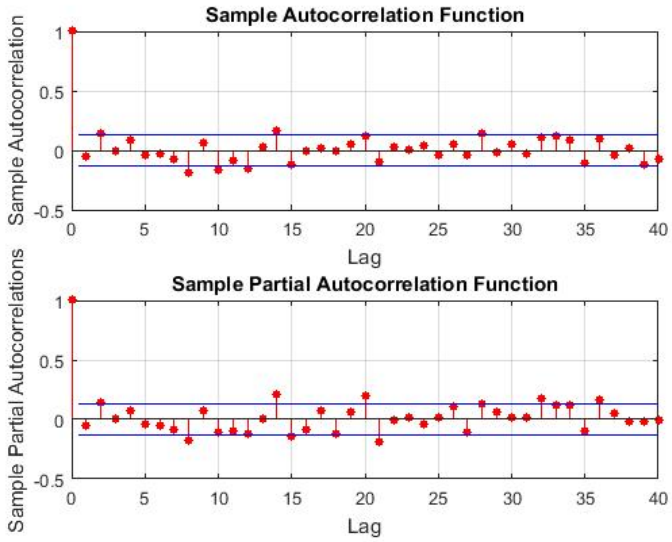

Figura 5.18 Gráfico da fac e facp da série dos resíduos do modelo SARIMA
Fonte: Elaborado pelo autor

5.2.2.3 Diagnóstico de homocedasticidade

Não aparentam existir tendências claras nos pontos do gráfico de dispersão dos valores dos resíduos do modelo SARIMA com os valores ajustados, o que indicaria homocedasticidade. Esse mesmo diagnóstico foi confirmado pelo teste de Breusch-Pagan que aceitou a hipótese nula (resíduos com variância constante) com um *p-value* associado de 0,2148.

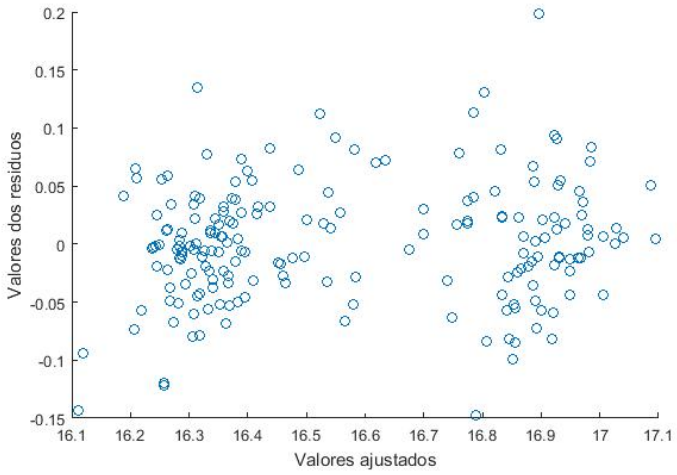

Figura 5.19 Gráfico dos resíduos em função dos valores ajustados do modelo de regressão linear
Fonte: Elaborado pelo autor

Por fim, pode-se concluir que os testes dos itens 5.2.2.1, 5.2.2.2 e 5.2.2.3 puderam confirmar que, diferentemente do que ocorreu com o modelo de regressão linear, o modelo SARIMA gerou resíduos brancos (normalmente distribuídos em torno do zero, homocedásticos e independentes entre si). Isso valida o modelo e respalda a capacidade da metodologia SARIMA de descrever a série temporal da demanda de gasolina.

5.3 Grupo III: Redes Neurais Artificiais

A última categoria de modelos que será aplicada é a rede neural. Em comparação com as anteriores, existe uma gama maior de fatores que fazem com que modelos baseados em rede neural se diferenciem entre si, o que gera a necessidade de um número consideravelmente mais extenso de testes para poder ter métricas de avaliação sobre qual modelo parece se adequar melhor às circunstâncias do problema. Como forma de facilitar esse processo, resolveu-se restringir o espectro de modelos a serem testados. Nesse sentido, com relação às arquiteturas possíveis, Zhang et al. (1998) afirmam que as redes do tipo *Multilayer Perceptrons* (MLP) mostram bons resultados para problemas de previsão, graças ao seu alto poder de mapeamento das relações entre *inputs* e *outputs*, o que lhe confere capacidade de generalização fora do período de calibração. Os autores indicam que essa capacidade deriva do fato de que as redes MLP, utilizando um número finito de neurônios e contendo pelo menos uma camada intermediária, são aproximadores universais, o que significa que podem aproximar qualquer função contínua com qualquer nível de exatidão que se deseje. Com esse respaldo teórico, optou-se por restringir os modelos de rede neural a uma arquitetura MLP com apenas uma camada intermediária

Zou et al. (2007) aplicaram esse mesmo tipo de rede para previsão do preço futuro de trigo na China. Os autores avaliaram a performance de um grupo de modelos de MLP com várias especificações em comum (incluindo função de ativação dos neurônios, algoritmo de treinamento, etc.), mas que se diferenciavam por uma

combinação de dois fatores: as variáveis *inputs* utilizadas e o número de neurônios da camada intermediária (n). Para selecionar o modelo dentro desse grupo delimitado, os autores testaram o desempenho de alguns dos modelos originados pelas diferentes combinações de variáveis e de número de nós na camada intermediária por um método de validação cruzada: as redes foram calibradas com uma parte da série histórica de dados e testou-se a performance do modelo dentro dessa parte de dados. Com o restante dos dados, pode-se testar a performance do modelo com observações fora do período de calibração, visando aferir sua capacidade preditiva. A combinação de número de nós e de variáveis de entrada que produziu o modelo com os melhores valores para os indicadores de performance foi selecionado. Os indicadores utilizados nesses testes pelos autores foram os seguintes:

- Para valores dentro do período de calibração: MAE, MAPE, MSE, AIC e BIC, cujas definições encontram-se no item 2.4.1.
- Para valores fora do período de teste: MAE, MAPE e MSE.

Para identificação do modelo de rede neural, optou-se por seguir a mesma abordagem de Zou et al. (2007), porém com os seguintes fatores de diferenciação entre os modelos:

- Sentido de propagação da informação dos impulsos sinápticos: as redes MLP testadas nesse trabalho podem ser do tipo *feedfoward*, em que o sentido de propagação é exclusivamente de uma camada anterior a uma posterior, ou podem ser recorrentes (*recurrent*), em que o *output* do nó da última camada retroalimenta os neurônios da camada intermediária, seguindo a mesma abordagem de Fernandez et al. (1990), com realimentação de apenas um período de defasagem. Zhang et al. (1998) destaca que redes neurais recorrentes possuem excelente performance para problemas de previsão, citando autores que aplicam esse tipo de rede para criação de modelos auto-regressivos não lineares;
- A quantidade de neurônios na camada intermediária. Zou et al. (2007) mostram que uma quantidade excessiva de neurônios poder gerar sobreajuste (*overffitting*), limitando a capacidade de generalização da rede para casos fora do período de validação, enquanto que poucos neurônios podem ser

insuficientes para conseguir aproximar a relação entre as variáveis de entrada e de saída.

Considerando a boa performance da função tangente hiperbólica atestada por Karlik e Olgac (2011) como função de ativação para problemas envolvendo redes MLP, optou-se por utilizá-la em todos os neurônios transformadores. Também fixou-se como método de aprendizagem o Levenberg-Marquardt com 50 iterações, considerando as características desse algoritmo indicadas por Kisi e Uncuoglu (2005): uma boa performance preditiva aliada a alta velocidade e necessidade de poucas iterações, o que se considerou importante em frente às necessidades computacionais envolvidas no processo de constante recalibração do modelo para os testes de validação cruzada com janela de tempo e de seleção dos fatores de diferenciação entre as redes.

5.3.1 Seleção entre os modelos baseados em redes neurais

Para identificar qual combinação de sentido de propagação de informações na MLP (recorrente ou *feedfoward*) e de quantidade de neurônios na camada intermediária (n) produziria o modelo com melhores resultados, aplicou-se um processo de validação cruzada semelhante àquele utilizado por Zou et al. (2007). A rede será calibrada com 4 anos (que equivale a 48 observações mensais) e testada para os 12 meses seguintes. O período de teste de um ano foi escolhido para simular o comportamento do modelo numa condição de aplicação real em que se precise estimar a demanda de combustível para o próximo ano corrente. Já a escolha da janela de 4 anos foi motivada para seguir um dos padrões de divisão que Zhang et al. (1998) afirmam ser comumente utilizada na literatura, em que 80% dos dados são para treinamento e os 20% restantes são para teste. Considerando que a série histórica de dados possui 16 anos, cada modelo foi testado com 12 pares de combinações de intervalos de calibração e de teste diferentes. Dessa forma, a capacidade de previsão da rede pode ser testada considerando intervalos de tempo distintos. Cada um dos testes foi repetido 5 vezes (para diminuir as diferenças de performance dentro de um mesmo modelo oriundas do algoritmo de treinamento),

resultando em 60 testes para cada modelo de rede neural. Os valores dos indicadores apresentados nas Tabela 5.4 e 5.5 são as médias desses 60 testes. Em negrito, estão os valores de melhor performance para cada indicador. Os resultados da Tabela 5.4 se referem à rede MLP *feedfoward*, enquanto os da Tabela 5.5 se referem à rede MLP recorrente:

MLP *feedfoward*								
n	Período de calibração					Período de teste		
	MAE (10^{-3})	MAPE (10^{-3})	MSE (10^{-3})	AIC	BIC	MAE (10^{-3})	MAPE (10^{-3})	MSE (10^{-3})
1	46.30	279.90	4.00	-250.17	-225.85	77.62	389.67	10.18
2	40.73	246	2.63	**-260.81**	**-233.89**	**69.85**	**372.45**	**7.79**
3	39.36	237.75	2.91	-246.50	-210.94	85.91	513.15	14.92
4	37.74	228.30	3.03	-230.88	-184.10	101.35	608.84	31.54
5	40.58	245.37	2.78	-222.81	-164.80	76.45	458.17	10.89
6	**37.39**	**226.09**	**2.45**	-216.19	-146.96	85.90	502.77	12.21
7	39.96	241.74	2.64	-203.23	-122.77	87.63	526.31	13.59
8	36.72	222.47	**2.45**	-193.30	-101.61	87.81	525.27	17.212
9	46.08	278.83	3.41	-166.51	-63.59	87.12	521.70	12.04
10	41.61	251.60	2.98	-160.79	-46.65	76.60	458.34	10.48

Tabela 5.4 Indicadores de performance de modelos de rede neural de arquitetura MLP *feedfoward* com diferentes números de neurônios na camada intermediária.
Fonte: Elaborado pelo autor

Considerando exclusivamente as redes *feedfoward*, a rede com dois neurônios na camada intermediária (n = 2) parece apresentar os melhores resultados frente às demais. Para o período de calibração, os valores de AIC e BIC dessa configuração foram os menores da tabela, ainda que MAE, MAPE e MSE tenham se mostrado piores que os da rede com n = 6. Isso se deve ao fato de que a rede com 6 neurônios na camada intermediária é menos parcimoniosa que a rede de apenas, sendo, dessa forma, penalizada por AIC e BIC. Além disso, para o período de teste, a rede n = 2 superou as demais em todos os indicadores.

MLP recorrente								
n	Período de calibração					Período de teste		
	MAE (10^{-3})	MAPE (10^{-3})	MSE (10^{-3})	AIC	BIC	MAE (10^{-3})	MAPE (10^{-3})	MSE (10^{-3})
1	40.28	243.55	2.70	**-269.35**	**-254.38**	**61.50**	**367.80**	**7.38**
2	37.61	227.37	2.42	-262.00	-233.93	74.89	389.30	7.99
3	42.98	259.85	3.07	-236.82	-195.66	80.18	479.36	11.39
4	38.59	233.37	2.72	-229.93	-175.66	73.36	439.21	8.84
5	38.29	321.59	2.91	-212.31	-144.95	89.57	536.50	16.36
6	42.00	253.93	2.30	-200.24	-119.78	95.39	570.72	15.20
7	**32.39**	**195.99**	**2.26**	-196.00	-102.33	89.51	536.53	16.39
8	37.77	228.58	2.66	-175.05	-68.39	111.70	667.97	24.24
9	48.13	290.27	3.79	-144.18	-40.17	85.13	509.19	14.67
10	34.40	208.26	2.44	-148.19	-15.33	86.87	519.34	16.20

Tabela 5.5 Indicadores de performance de modelos de rede neural de arquitetura MLP recorrente com diferentes números de neurônios na camada intermediária.
Fonte: Elaborado pelo autor

No grupo das redes MLP recorrentes, a que possui apenas um neurônio na camada intermediária superou as demais, pelos mesmos motivos que a de dois neurônios tinha superado no grupo das MLP *feedfoward*: apresentou os menores valores de AIC, BIC e dos três indicadores de ajuste de teste. Na comparação entre essas duas redes, a MLP recorrente de um neurônio teve um desempenho levemente superior, o que motivou a escolha por ela para representar o grupo III na comparação entre categorias de modelos.

Portanto, a rede neural que representa o grupo III possui uma topologia do tipo *multilayer perceptron* (MLP) recorrente e com um neurônio na camada intermediária.

5.4 Comparação entre os modelos representantes de cada grupo

Uma vez identificados os modelos que representam cada categoria de modelagem (regressão linear, modelos ARIMA e rede neural), a seleção do melhor modelo entre

os três será realizada por meio de um processo de validação cruzada com janela de tempo. Para se ter uma ideia mais abrangente da capacidade preditiva dos modelos, as previsões testadas foram feitas em dois horizontes de tempo diferentes: o de curto prazo (H = 1 mês) e o de longo prazo (H = 12 meses). Os parâmetros do modelo foram constantemente recalibrados com os 48 meses que antecedem o primeiro mês de teste, até que se tivesse utilizado toda a série histórica de dados. Isso gerou 12 intervalos de testes de longo prazo e 144 intervalos de testes de curto prazo. Para cada teste, computaram-se os valores de três indicadores de acurácia de previsão (MAE, MAPE e MSE). As médias dos valores desses indicadores em cada tipo de testes (H = 1 e H = 12) e para o representante de cada grupo de modelo estão indicadas na Tabela 5.6, com os melhores valores obtidos destacados em negrito. Além disso, também calculou-se a razão *Skill*, descrita no item 2.4.1., visando aferir com mais clareza a superioridade do melhor modelo com relação aos demais em cada um dos indicadores e cujos valores encontram-se na Tabela 5.7:

Grupo	H = 1 (Curto Prazo)			H = 12 (Longo Prazo)		
	MAPE (10^{-3})	MAE (10-3)	MSE (10-3)	MAPE (10-3)	MAE (10-3)	MSE (10-3)
Regressão linear (I)	311.3315	51.9491	4.3572	64.5513	386.7566	7.4678
SARIMA (II)	**294.6285**	**49.2148**	**3.9529**	**59.5513**	**353.3655**	**7.3240**
Rede Neural (III)	375.7593	62.6999	6.6725	62.5720	374.4591	7.9582

Tabela 5.6 Indicadores de performance de modelos representantes de cada um dos três grupos de modelagem
Fonte: Elaborado pelo autor

	H = 1			H = 12		
	MAPE	MAE	MSE	MAPE	MAE	MSE
$Skill_{II,III}$	5,37%	5,26%	9,28%	7,75%	8,63%	1,93%
$Skill_{II,I}$	27,54%	27,40%	68,80%	5,07%	5,97%	8,66%

Tabela 5.7 Valores de *Skill* calculados entre o modelo de melhor performance e os outros dois para cada um dos indicadores
Fonte: Elaborado pelo autor

Os gráficos com a previsão de curto prazo (H = 1) e longo prazo (H = 12) dos modelos em relação ao valor real da demanda de gasolina encontram-se na Figura 5.20. Como eram usados 48 meses para calibração dos modelos, a previsão inicia-se

no 49º mês. Os gráficos dispostos primeira coluna representam as previsões dos modelos no curto prazo, enquanto os da segunda representam as previsões de longo prazo. A primeira, segunda e terceira linha de gráficos referem-se, respectivamente, às previsões do modelo de regressão linear (I), do ARIMA (II) e da rede neural (III).

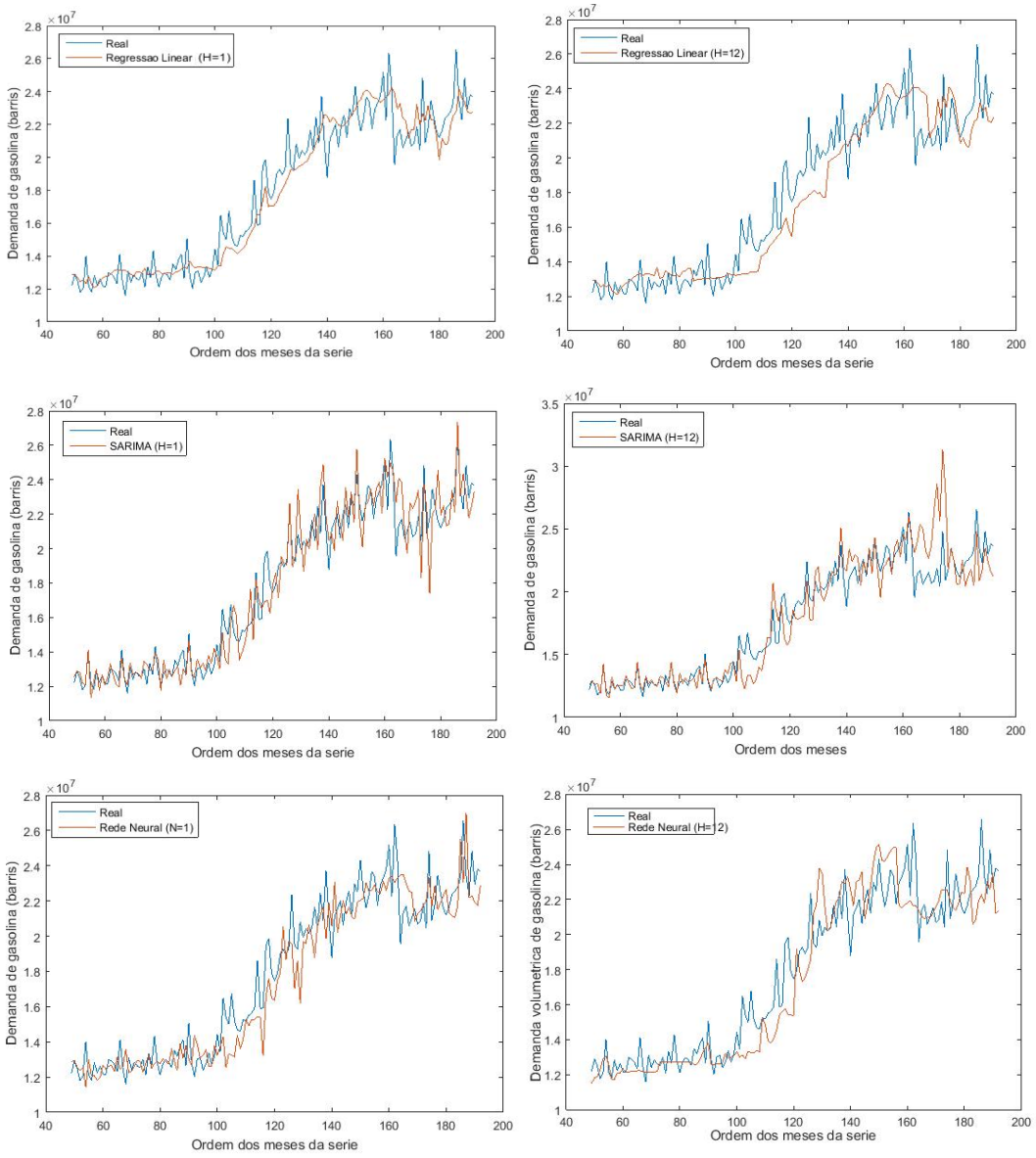

Figura 5.20 Gráfico dos valores previstos pelos modelos e dos valores reais para a demanda volumétrica de gasolina (em barris)
Fonte: Elaborado pelo autor

A análise da Figura 5.20 e das Tabelas 5.6 e 5.7 mostram que o modelo SARIMA superou os modelos representantes das demais categorias de modelagem sob todos os indicadores e aspectos analisados, em especial para as previsões de curto prazo. No longo prazo, ainda que o SARIMA tenha apresentado melhores resultados em média

que as demais modelagens, cabe ressaltar que, durante os doze meses posteriores ao 169º mês (que coincide com julho de 2015), a previsão do SARIMA se descolou consideravelmente do movimento real da demanda de gasolina, reaproximando-se novamente apenas em julho de 2016. Isso se deve ao fato de que o modelo previu os meses de julho de 2015 até junho de 2016 com os parâmetros calibrados com meses em que a tendência de crescimento se mostrava bastante em alta (julho de 2011 até junho de 2015). Porém, a queda na atividade econômica na época atingiu a demanda de gasolina, fazendo cair abruptamente a tendência de crescimento do consumo pelo combustível que vinha se observando até então. Os demais modelos não tiveram esse mesmo problema com a mudança de paradigma de consumo de gasolina devido à crise, provavelmente pelo fato de que as variáveis exógenas refletiram nos demais modelos que a tendência da demanda de gasolina se reverteria. Ainda assim, no geral, o SARIMA conseguiu ter uma performance preditiva superior, principalmente nos momentos em que as tendências eram estáveis ou no caso em que a série era recalibrada constantemente (ou seja, no caso da previsão de curto prazo).

Outro ponto que vale destacar é que, no curto prazo, a performance da rede neural foi bastante aquém do que se esperava, em especial com relação à média quadrática dos erros (MSE), que foi 68,8% inferior à obtida pelo SARIMA. Isso se deve principalmente aos desvios da rede no momento em que a tendência de demanda de gasolina começou a subir: observando o gráfico do canto inferior esquerdo da Figura 5.20, pode-se perceber que, a partir do 100º mês, quando a demanda de gasolina começa a acelerar sua tendência de crescimento, a rede previu em vários momentos justamente o oposto. Isso pode ser um indício de que a rede proposta não conseguiu generalizar suficientemente bem as relações entre as variáveis de entrada e a demanda de gasolina, ou que ela precisa de uma janela de tempo maior para ser calibrada (mais de 4 anos). Esse problema também se repetiu com o horizonte de previsão de longo prazo, mas a diferença para com os demais modelos, nesse caso, não foi tão considerável. Outra questão que chama a atenção é que a rede apresentou piores performances com o indicador MSE do que com MAPE e MAE: inclusive, a previsão de longo prazo da rede superou a regressão linear nos indicadores MAPE e MAE, porém foi superada por essa última no indicador MSE. Como MSE penaliza

comparativamente mais os desvios mais intensos devido ao fato de levar em conta o erro quadrático (e não o erro absoluto como MAPE e MAE), isso significa que a rede foi mais propensa a se desviar bastante do real do que as demais modelagens, o que também é um fator negativo.

Chama-se atenção o fato de que o modelo baseado em regressão linear não mostrou um resultado tão inferiores ao melhor modelo do trabalho (o que não foi o caso da rede neural), ainda que não tenha respeitado duas hipóteses da sua formulação sobre os seus resíduos (homocedasticidade e ausência de autocorrelação). As diferenças entre os valores do modelo SARIMA e da regressão linear não ultrapassaram 10% em nenhum dos indicadores de nenhuma das situações de teste dos modelos (previsão de curto ou longo prazo).

Em suma, a análise dos indicadores mostra que o modelo SARIMA superou os modelos representantes das demais categorias de modelagem tanto sob o aspecto de previsão de curto prazo quanto de longo. Além da boa performance em relação aos demais modelos, existem algumas outras vantagens que o SARIMA possui sobre as demais modelagens. Primeiramente, o modelo não precisa de variáveis exógenas como entrada: apenas a série temporal da demanda de gasolina é suficiente para realizar previsões. Também vale destacar que, em comparação com a rede neural, o SARIMA é mais simples computacionalmente de ser modelado e calibrado, além de ser mais fácil de construir intervalos de confiança estatísticos para os valores estimados. Finalmente, ao contrário do que se verificou com a regressão linear, a análise de resíduos mostrou que os erros do modelo SARIMA proposto possuem características que permitem confirmar a validade do modelo: erros homocedásticos, independentes entre si e distribuídos normalmente em volta do zero.

Para concluir esse capítulo, aplicou-se o modelo selecionado (SARIMA) para prever a demanda de gasolina para os próximos 12 meses. Até o fechamento do presente trabalho, haviam sido divulgadas estatísticas sobre venda mensal de gasolina C até setembro de 2017. Portanto, o modelo SARIMA proposto estimou a

demanda de outubro de 2017 até setembro de 2018. Os valores estimados (e seus respectivos intervalos de confiança com $\alpha = 5\%$), encontram-se no gráfico abaixo:

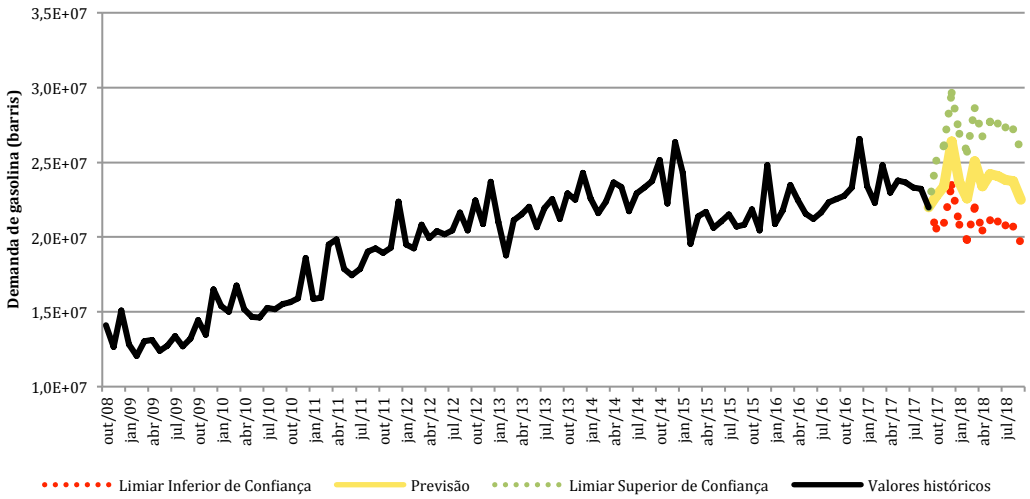

Figura 5.21 Gráfico da previsão de demanda de gaoslina para o próximo ano realizada pelo SARIMA
Fonte: Elaborado pelo autor

6 CONCLUSÕES

A partir da análise detalhada de relatórios e bases estatísticas de entidades ligadas ao setor de transporte, o trabalho conseguiu quantificar o nível de atividade e as demandas energéticas do setor. A partir disso, os mercados dos combustíveis mais relevantes para o transporte no país foram estudados, permitindo não só a identificação dos fatores que mais afetam a demanda por esses combustíveis a nível nacional, como também a seleção de variáveis exógenas que consigam traduzir esses fatores em números.

A partir de uma análise detalhada de algumas publicações anteriores sobre previsão de demanda na literatura nacional e internacional, aliadas ao conhecimento adquirido sobre o funcionamento do mercado doméstico de combustíveis utilizados no setor de transportes, puderam-se sugerir modelos para estimar a demanda de gasolina C, um dos combustíveis mais utilizados pelo setor. Esses modelos partiram de três abordagem de modelagem distintas e, no final, a capacidade preditiva de cada modelo foi comparada por meio de um método de validação cruzada que objetiva simular uma situação de aplicação real do modelo proposto.

Uma possível extensão futura do trabalho é aplicar as mesmas metodologias utilizadas para modelar a demanda de gasolina C para os outros três combustíveis mais relevantes para o setor de transportes brasileiro: o óleo diesel, o etanol hidratado e o querosene. Também poderia ser interessante expandir os métodos de modelagem de demanda para sanar algumas deficiências que os modelos apresentaram. Com relação ao modelo de melhor performance no trabalho (SARIMA), por exemplo, verificou-se que poderia ser interessante introduzir de alguma de sensibilizar o modelo a variáveis externas, à semelhança dos outros modelos propostos no trabalho (redes neurais e regressão linear). Na literatura, existem modelos derivados dos modelos ARIMA que possuem em sua formulação variáveis exógenas e sazonais, como o modelo SARIMAX (*Seasonal Autoregressive Integrated Moving Average Model with Exogenous Variables*). Outra sugestão

poderia ser combinar modelos que possuem variáveis exógenas com a modelagem SARIMA proposta, seguindo uma abordagem hierárquica.

Finalmente, espera-se que os modelos desenvolvidos nesse trabalho possam ser aplicados por entidades do setor de transporte, energia e óleo e gás, pelo poder público ou por investidores privados como uma ferramenta suplementar para ajudar a entender as perspectivas de evolução da demanda futura de combustíveis em um contexto de expansão do setor de transportes, o que proveria condições para um planejamento adequado das oportunidades de investimentos para garantir o abastecimento do setor.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDUL-WAHAB, Sabah A.; BAKHEIT, Charles S.; AL-ALAWI, Saleh M. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. **Environmental Modelling & Software**, v. 20, n. 10, p. 1263-1271, 2005.

ARLOT, Sylvain et al. A survey of cross-validation procedures for model selection. **Statistics surveys**, v. 4, p. 40-79, 2010.

AZEVEDO, B. S., 2007. Análise das elasticidades preço e renda da demanda por combustíveis no Brasil e desagregadas por regiões geográficas. **Dissertation in Economics**. **Ibmec**, Rio de Janeiro, March.

BENOIT, Kenneth. Linear regression models with logarithmic transformations. **London School of Economics, London**, v. 22, n. 1, p. 23-36, 2011.

BOHI, Douglas R.; ZIMMERMAN, Mary Beth. An update on econometric studies of energy demand behavior. **Annual Review of Energy**, v. 9, n. 1, p. 105-154, 1984.

BURNHAM, Kenneth P.; ANDERSON, David R. Multimodel inference: understanding AIC and BIC in model selection. **Sociological methods & research**, v. 33, n. 2, p. 261-304, 2004.

BURNQUIST, H. L., BACCHI, M.R.P., 2002. The demand for gasoline in Brazil: an analysis using techniques of co-integration. CEPEA, Discussion Paper. www.cepea.esalq.usp.br / pdf / DemandaGasolina.pdf.

CHEZE, Benoit; GASTINEAU, Pascal; CHEVALLIER, Julien. Forecasting Air Traffic and Corresponding Jet-Fuel Demand until 2025. 2010.

CHONG, Il-Gyo; JUN, Chi-Hyuck. Performance of some variable selection methods when multicollinearity is present. **Chemometrics and intelligent laboratory systems**, v. 78, n. 1, p. 103-112, 2005.

CYBENKO, George. Approximation by superpositions of a sigmoidal function. **Mathematics of Control, Signals, and Systems (MCSS)**, v. 2, n. 4, p. 303-314, 1989.

DAHL, Carol; STERNER, Thomas. Analysing gasoline demand elasticities: a survey. **Energy economics**, v. 13, n. 3, p. 203-210, 1991.

DIX, M. C.; GOODWIN, P. B. **Petrol prices and car use: a synthesis of conflicting evidence**. 1981.

DUARTE, Leonardo Tomazeli et al. Um estudo sobre separação cega de fontes e contribuições ao caso de misturas não-lineares. 2006.

DUDA, Richard O.; HART, Peter E.; STORK, David G. **Pattern classification**. Wiley, New York, 1973.

EDIGER, Volkan Ş.; AKAR, Sertac. ARIMA forecasting of primary energy demand by fuel in Turkey. **Energy Policy**, v. 35, n. 3, p. 1701-1708, 2007.

FLOM, Peter L.; CASSELL, David L. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In: **NorthEast SAS Users Group Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland**. 2007.

FREITAS, L. C., KANEKO, S., 2011. Ethanol demand under the flex-fuel technology regime in Brazil. **Energy Economics**. 33 (6), 1146-1154.

GRAHAM, Daniel J.; GLAISTER, Stephen. The demand for automobile fuel: a survey of elasticities. **Journal of Transport Economics and Policy (JTEP)**, v. 36, n. 1, p. 1-25, 2002.

GRÉGOIRE, G. Multiple linear regression. **European Astronomical Society Publications Series**, v. 66, p. 45-72, 2014.

HAYKIN, Simon. **Neural Networks: A Comprehensive Foundation**, v. 2, 1999.

HILL, Tim; O'CONNOR, Marcus; REMUS, William. Neural network models for time series forecasts. **Management science**, v. 42, n. 7, p. 1082-1092, 1996.

HORNIK, Kurt; STINCHCOMBE, Maxwell; WHITE, Halbert. Multilayer feedforward networks are universal approximators. **Neural networks**, v. 2, n. 5, p. 359-366, 1989.

INOUE, Atsushi; JIN, Lu; ROSSI, Barbara. Rolling window selection for out-of-sample forecasting with time-varying parameters. **Journal of Econometrics**, v. 196, n. 1, p. 55-67, 2017.

IOOTY, M., PINTO Jr., H., ROPPA, B., BIASI, G., 2004. An analysis of the competitive price of CNG compared to gasoline: estimation of elasticities of demand for CNG in Brazil in recent years. In: IE / UFRJ (Ed.), Rio Oil and Gas Expo and Conference. UFRJ, Rio de Janeiro.

KIŞI, Özgür; UNCUOĞLU, Erdal. Comparison of three back-propagation training algorithms for two case studies. 2005.

KVÅLSETH, Tarald O. Cautionary note about R 2. **The American Statistician**, v. 39, n. 4, p. 279-285, 1985.

MAHAJAN, Vijay; JAIN, Arun K.; BERGIER, Michel. Parameter estimation in marketing models in the presence of multicollinearity: An application of ridge regression. **Journal of Marketing Research**, p. 586-591, 1977.

MARJOTTA-MAISTRO, Marta Cristina; BARROS, GS de C. Relações comerciais e de preços no mercado nacional de combustíveis. In: **CONGRESSO BRASILEIRO DE ECONOMIA E SOCIOLOGIA RURAL**. 2002.

NAPPO, M., 2007. The demand for gasoline in Brazil: A Review of its elasticity after the introduction of flex fuel cars. Getulio Vargas Foundation School of Economics are Paulo-EESP/FGV, Sao Paulo. (March).

REYNALDO, Cristiane et al. Regressão" Ridge": um metodo alternativo para o mal condicionamento da matriz das regressoras. 1997

ROWLEY, Henry A.; BALUJA, Shumeet; KANADE, Takeo. Neural network-based face detection. **IEEE Transactions on pattern analysis and machine intelligence**, v. 20, n. 1, p. 23-38, 1998.

SANTIAGO, Flaviane Souza et al. Um modelo econométrico+ insumo-produto para a previsão de longo prazo da demanda de combustíveis no Brasil. 2009.
SILVA, José Almir da. Modelagem Multivariada para previsão de demanda de gasolina e óleo diesel no Brasil. 2014.

SHARDA, Ramesh. Neural networks for the MS/OR analyst: An application bibliography. **Interfaces**, v. 24, n. 2, p. 116-130, 1994.

STERNER, Thomas; DAHL, Carol A. Modelling transport fuel demand.
In: **International Energy Economics**. Springer Netherlands, 1992. p. 65-79.

SUGANTHI, L.; SAMUEL, Anand A. Energy models for demand forecasting—A review. **Renewable and sustainable energy reviews**, v. 16, n. 2, p. 1223-1240, 2012.

TAYLOR, Lester D. The demand for energy: a survey of price and income elasticities. **International Studies of the Demand for Energy**, p. 3-43, 1977.

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, p. 267-288, 1996.

TIBSHIRANI, Robert. The lasso method for variable selection in the Cox model. **Statistics in medicine**, v. 16, n. 4, p. 385-395, 1997.

WASSERMAN, Philip D. **Neural computing**. Van Nostrand Reinhold, New York, 1989.

WERNTGES, H.W., 1993. Partitions of unity improve neural function approximation, Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, vol 2, 914-918.

ZHANG, G. Peter. Time series forecasting using a hybrid ARIMA and neural network model. **Neurocomputing**, v. 50, p. 159-175, 2003.

WILLMOTT, Cort J. Some comments on the evaluation of model performance. **Bulletin of the American Meteorological Society**, v. 63, n. 11, p. 1309-1313, 1982.

ZOU, H. F. et al. An investigation and comparison of artificial neural network and time series models for Chinese food grain price forecasting. **Neurocomputing**, v. 70, n. 16, p. 2913-2923, 2007

ZUCCHINI, Walter. An introduction to model selection. **Journal of mathematical psychology**, v. 44, n. 1, p. 41-61, 2000.