FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

ADRIANA BEZERRA BESSA

PREVISÃO DE VENDAS NO VAREJO DE MODA COM MODELOS DE REDES NEURAIS

SÃO PAULO
2018

ADRIANA BEZERRA BESSA

PREVISÃO DE VENDAS NO VAREJO DE MODA COM MODELOS DE REDES NEURAIS

Dissertação apresentada à Escola de Economia da Fundação Getulio Vargas (FGV/EESP) como requisito para obtenção do título de Mestre em Finanças e Economia Empresarial

Campo de conhecimento:
Economia Aplicada

Orientador: Prof. Dr. Afonso de Campos Pinto

SÃO PAULO
2018

Bessa, Adriana Bezerra.
Previsão de vendas no varejo de moda com modelos de redes neurais / Adriana Bezerra Bessa. - 2018.
256 f.

Orientador: Afonso de Campos Pinto.
Dissertação (MPFE) - Escola de Economia de São Paulo.

1. Previsão de vendas – Modelos estatísticos. 2. Redes neurais (Computação). 3. Previsão comercial. 4. Planejamento estratégico. I. Pinto, Afonso de Campos. II. Dissertação (MPFE) - Escola de Economia de São Paulo. III. Título.

CDU 658.8.012.2

ADRIANA BEZERRA BESSA

PREVISÃO DE VENDAS NO VAREJO DE MODA COM MODELOS DE REDES NEURAIS

Dissertação apresentada à Escola de Economia da Fundação Getulio Vargas (FGV/EESP) como requisito para obtenção do título de Mestre em Finanças e Economia Empresarial

Campo de conhecimento:
Economia Aplicada

Data de aprovação: ___/___/_____

Banca Examinadora:

_____
Prof. Dr. Afonso de Campos Pinto
(Orientador)
EESP - FGV

_____
Prof. Dr. Ricardo Ratner Rochman
EESP - FGV

_____
Profa. Dra. Élia Yathie Matsumoto
POLI - USP

AGRADECIMENTOS

Agradeço, primeiramente, a Deus por ter me dado saúde e perseverança ao longo de todo o mestrado.

Agradeço aos meus Pais e aos meus Irmãos por todo amor, apoio e confiança.

Agradeço aos meus Colegas do curso por tornar esta jornada mais leve e o processo de aprendizado mais fácil.

Agradeço a cada um dos meus Professores pelo conhecimento compartilhado.

Agradeço ao meu Orientador por me guiar neste trabalho.

“Prediction is very difficult, especially if it’s about the future.”
Niels Bohr, laureado com Nobel em Física

RESUMO

A previsão de vendas é um aspecto crítico para maior parte das organizações, já que permite tornar o processo de planejamento mais eficiente, impactando assim nos resultados a serem obtidos pelas empresas. Entre as diversas técnicas de previsão, temos o grupo de métodos estatísticos clássicos e os métodos avançados, que trazem uma contribuição no tratamento das não linearidades. É neste contexto, que surge o problema desta dissertação: Quais são as técnicas que apresentam maior acurácia quando aplicadas para previsão de vendas no varejo de moda? Para responder a esta questão, esse trabalho avaliou dez métodos de previsão: Naive, SARIMA, SARIMA com exógenas, SARIMA GARCH, SARIMA GARCH com exógenas, método atual utilizado pela empresa estudada, rede neural MLP, rede neural MLP com exógenas, rede neural recorrente LSTM e rede neural recorrente LSTM com exógenas para quatro séries de quantidades vendidas de categorias de produtos distintas de uma empresa varejista do setor de moda.

É fundamental destacar, que de forma casual, a pesquisa identificou que as quatro séries semanais de vendas dos produtos analisados são estacionárias, considerando um período longo de dez anos, o que por si só já é um resultado relevante. A análise dos diversos métodos de previsão para cada série de produto mostrou que os métodos avançados superaram os métodos estatísticos clássicos e, mais especificamente, a rede neural recorrente LSTM foi a que apresentou a maior precisão.

Sendo assim, não há dúvidas que adoção dos métodos avançados para as empresas, que atuam no varejo de moda, pode trazer melhorias significativas em termos de gestão de estoque, de gestão da cadeia de abastecimento e de gestão de caixa, garantindo um aumento de eficiência e dos resultados das mesmas. De forma prática, para a empresa estudada foi obtido um incremento de acuracidade de 54,32%.

Palavras-chave: Previsão de Vendas; Métodos Quantitativos de Previsão; Redes Neurais; SARIMA; SARIMA GARCH; Redes Neurais Recorrentes; LSTM.

ABSTRACT

The sales forecasting is a critical aspect for most organizations, since it allows to make the planning process more efficient, thus impacting the results to be obtained by the companies. Among the various forecasting techniques, we have the group of classical statistical methods and the advanced methods, which make a contribution in the treatment of nonlinearities. It is in this context, that the problem of this dissertation arises: What are the techniques that present the greatest accuracy when applied to forecast sales in fashion retail? In order to answer this question, this study evaluated ten predictive methods: Naive, SARIMA, SARIMA with exogenous, SARIMA GARCH, SARIMA GARCH with exogenous, current method used by the studied company, MLP neural network, MLP neural network with exogenous, recurrent neural network LSTM and LSTM recurrent neural network with exogenous for four series of quantities sold from product categories distinct from a retailer in the fashion industry.

It is important to highlight that, on a casual basis, the research identified that the four weekly series of sales of the analyzed products are stationary, considering a long period of ten years, which in itself is already a relevant result. The analysis of the various prediction methods for each product series showed that the advanced methods overcame the classic statistical methods and, more specifically, the recurrent neural network LSTM was the one that presented the highest precision.

Therefore, there is no doubt that adoption of the advanced methods for companies that operate in fashion retail can bring significant improvements in terms of inventory management, supply chain management and cash management, ensuring an increase in efficiency and in its results. In practice, for the company studied, an accuracy increase of 54.32% was obtained.

Keywords: Sales Forecast; Neural Networks; SARIMA;SARIMA GARCH; Recurrents Neural Networks; LSTM.

FIGURAS

TABELAS

LISTA DE ABREVIATURAS

ADAM	Adaptive Moment Estimation
AD-ELM	Extreme Learning Machine com Métricas Adaptativas
ADF	Augmented Dickey-Fuller – Dickey Fuller Aumentado
AHFCCX	Automatic Hybrid Forecasting Model with Fuzzy Estimation of Corrective Coefficients of the Explanatory Variables Influence
AIC	Akaike Information Criteria – Critério de Informação de Akaike
ANFIS	Adjustment of Mean-term Forecast with Adaptative Neural Fuzzy Inference System
AR	Autoregressive - Autoregressivo
ARCH	Autoregressive Conditional Heterocedasticity – Autorregressivo com Heterocedasticidade Condicional
ARIMA	Autoregressive Integrated Moving-average – Autorregressivo Integrado de Média Móvel
ARMA	Autoregressive Moving-average – Autorregressivo de Média Móvel
BIC	Bayesian information criterion/ Schwarz Criterion
BPTT	Backpropagation Through Time – Retropropagação no Tempo
ELM	Extreme Learning Machine – Aprendizagem de Máquina Extrema
CPU	Central Processing Unit – Unidade de Processamento Central
DF	Dickey-Fuller
FAC	Função de autocorrelação
FACP	Função de autocorrelação parcial
FECOMERCIOSP	Federação do Comércio de Bens, Serviços e Turismo do Estado de São Paulo
FIS	Fuzzy Inference Systems – Sistemas de Inferência Fuzzy
FTDNN	Focused Time Delayed Neural Network – Rede Neural com Atraso de Tempo Focalizado

GARCH	Generalized Autoregressive Conditional Heterocedasticity – Autorregressivo com Heterocedasticidade Condicional Generalizada
GD	Gradient Descent – Gradiente Descendente
GPU	Graphics Processing Unit – Unidade de Processamento Gráfico
GRU	Gate Recurrent Unit – Unidade Recorrente de Portão
HC	Hanna-Quinn
HFCCX	Hybrid Forecasting Model with Fuzzy Estimation of Corrective Coefficients of the Explanatory Variables Influence
HIS	Hibrid Intelingence System - Sistema Inteligente Híbrido
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IPCA	Índice de Preços ao Consumidor
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
LSTM	Long Short Term Memory – Memória de Curto e Longo Prazo
MA	Moving Average – Média Móvel
MAE	Mean Absolute Error - Erro Médio Absoluto
MAPE	Mean Absolute Percentual Error - Média dos Erros Percentuais Absolutos
ME	Mean Error - Erro Médio
MLP	Multilayer Perceptron – Perceptron de Camadas Múltiplas
MPE	Mean Percentual Error - Erro Percentual Médio
MSE	Mean Squared Error – Erro Quadrático Médio
PAC	Pesquisa Anual do Comércio
PED	Pesquisa de Emprego e Desemprego
PMC	Pesquisa Mensal de Comércio
PNAD	Pesquisa Nacional por Amostra de Domicílios
PIB	Produto Interno Bruto
PP	Phillips-Perron
p-value	Menor nível de significância para o qual se rejeita a hipótese nula

ReLu	Função retificadora linear
RMSE	Root Mean Squared Error – Raiz Quadrada do Erro Quadrático médio
RN	Rede Neural
RNA	Rede Neural Artificial
RNN	Rede Neural Recorrente
SAMANFIS	Short-term forecasting model by adjustment of mean-term forecast with adaptive neural fuzzy inference system
SARIMA	Seasonal Autoregressive Integrated Moving Average - ARIMA Sazonal
SBVC	Sociedade Brasileira de Varejo e Consumo
SEAD	Fundação Sistema Estadual de Análise de Dados
SGD	Stochastic Gradient Descent – Gradiente Descendente Estocástico
SKU	Stock Keeping Unit – Unidade de Manutênção em Estoque
SVM	Support Vector Machine – Máquina de Vetores de Suporte
TLFN	Time-lagged Feedforward Neural Network – Rede com Atraso Alimentada Adiante
TPU	Tensor Processing Unit – Unidade de Processamento de Tensor

SUMÁRIO

1. Introdução

Diariamente, a maior parte das organizações confronta-se com questões críticas em relação ao seu crescimento e à sua competitividade. Sendo assim, as organizações estão inseridas em um ambiente extremamente dinâmico, onde as decisões gerenciais possuem um papel fundamental na definição das estratégias e representam orientações que serão seguidas com o intuito de garantir maior competitividade e resultados superiores. Para definição das decisões gerenciais, o processo dos planejamentos em diferentes horizontes no tempo – curto, médio e longo prazo – torna-se essencial. Mais especificamente, entender que fatores e/ou variáveis influenciam o planejamento nos diversos prazos é tão ou mais importante que o planejamento em si.

O planejamento aqui está colocado de uma forma ampla, faz parte dos diversos departamentos de uma empresa e está presente nos diversos setores de mercado. Conforme Montgomery et al. (1990) e Makridakis et al. (1998), para que seja bem-sucedida a formulação de planejamentos e o direcionamento estratégico das empresas, são fundamentais a identificação e a previsão correta das mudanças no ambiente de negócios, o que torna a previsão de demanda um elemento-chave na tomada de decisão gerencial. Ainda, segundo Armstrong (1983), as companhias podem melhorar sua eficiência se elas puderem antecipar problemas e desenvolver planos para responder a estes problemas. O mesmo autor coloca também que métodos de previsão de demanda têm sido desenvolvidos de forma a garantir a antecipação de estados futuros de alguns fatores e/ou variáveis que afetam o planejamento estratégico das organizações.

Deve-se ressaltar que, muitas vezes, conforme Armstrong (2001), a previsão de demanda é substituída pelo planejamento estratégico em si dentro das companhias. No entanto, esta substituição não é adequada, pois previsões são feitas sobre eventos externos, onde a empresa não tem total controle. Por sua vez a tomada de decisão se faz sobre eventos internos, os quais estão sobre total controle da empresa. Neste ambiente surge o planejamento, que faz a integração entre estes dois mundos. De forma mais detalhada, o planejamento gera estratégias e ações a serem realizadas, as quais levam a resultados que são registrados nos sistemas internos da empresa,

retroalimentando o sistema de planejamento e as previsões de demanda. Assim, o planejamento é muito mais amplo que a previsão de demanda em si.

Conforme Dias (1999), a importância da previsão de demanda está diretamente ligada à racionalização dos recursos envolvidos e à meta do nível de atendimento aos clientes. Qualquer atividade de planejamento pressupõe previsões sobre o futuro, as quais serão mais assertivas dependendo das hipóteses utilizadas. O processo de previsão pode ser definido como o conjunto de todas as atividades envolvidas para que seja possível a geração dos números finais da previsão, o que inclui desde a etapa de coleta de dados até a última revisão da previsão.

Neste trabalho, a previsão de demanda tem como foco a previsão de vendas para algumas categorias de produtos da empresa varejista Lojão do Brás. Apesar deste foco, como foi apresentada acima, a previsão de demanda é extremamente relevante para o planejamento estratégico de empresas dos mais diversos setores da economia e para os seus respectivos departamentos, garantindo competitividade e maior eficiência.

1.1. Motivação

O varejo é toda atividade econômica da venda de um bem ou um serviço para o consumidor final, ou seja, uma transação entre um CNPJ e um CPF. O varejo tem forte influência no PIB nacional, uma vez que o seu impacto é de 42,54%, considerando os dados de 2016, conforme a Sociedade Brasileira de Varejo e Consumo (SBVC), e é responsável por 22,3% dos empregos formais do país. Em relação ao varejo de moda, setor alvo deste trabalho, ele representa 7,1% do varejo total e é afetado por uma série de variáveis exógenas, sendo que as mais relevantes são: renda, emprego, crédito e confiança.

Em um cenário cada vez mais competitivo, onde temos cadeias globais de suprimentos estabelecidas para o atendimento das demandas das redes varejistas têxteis, a tarefa de previsão de demanda tem ganhado cada vez mais relevância dentro das organizações. Mais especificamente, uma previsão mais assertiva garantirá menores níveis de estoques, menores remarcações, menores rupturas, maiores

margens e um efeito chicote menor, o que significa uma cadeia de *supply chain* mais integrada e eficiente.

Cabe ressaltar, desde já, a complexidade e a especificidade do varejo têxtil, que torna o processo de previsão mais difícil do que em outros setores varejistas. As características que aumentam o grau de complexidade de previsão neste setor são: a definição do horizonte de previsão em si, se é de curto ou longo prazo; o ciclo de vida de produtos, pois temos produtos mais básicos e outros produtos com atributos mais *fashion,* que são renovados com maior frequência, diminuindo o volume de dados disponível para previsão; o impacto da sazonalidade e da temperatura; e a influência de uma série de outras variáveis exógenas, tais como promoções e preços da concorrência, entre outros.

Diante deste quadro pode-se dizer, então, que a principal motivação do trabalho é encontrar modelos de previsão mais assertivos com o intuito de garantir uma margem de erro menor do que os modelos tradicionais e, consequentemente, assegurarão melhores resultados para as companhias deste setor. Neste sentido, espera-se que técnicas mais contemporâneas: como as de redes neurais, que são técnicas mais flexíveis, genéricas e abrangentes, sejam também mais precisas.

1.2. Objetivos

O principal objetivo deste trabalho é aplicar técnicas de previsão e verificar suas respectivas acuracidades por meio de redes neurais, tanto diretas e não recorrentes, a exemplo das Multilayer Perceptron (MLP), como as redes neurais recorrentes, através das redes de Memória de Longo e Curto Prazo (LSTM, do inglês *Long Short Term Memory*). A aplicação foi feita utilizando um conjunto de dados reais e a comparação do nível de precisão foi feito contra os valores obtidos deste indicador nos métodos estatísticos clássicos, tais como SARIMA, SARIMA GARCH e *Naive*, com o intuito de verificar qual modelo apresenta maior precisão. Adicionalmente, também foi realizada uma comparação com o modelo atual utilizado pela empresa participante deste trabalho. Deve-se destacar que este tipo de trabalho, já foi desenvolvido para outros

setores varejistas no Brasil, mas não para o setor de moda e não exatamente da mesma forma.

Outro objetivo deste trabalho é descrever as restrições e as limitações existentes para o processo de previsão de vendas no varejo de moda. Embora a verificação da acuracidade dos modelos contemplados seja o principal objetivo deste trabalho, o processo de construção de cada um deles, juntamente com a especificação das variáveis e da arquitetura presente em cada uma das modelagens realizadas, é um aspecto fundamental para entendimento do melhor desempenho de um modelo em detrimento de outro. Sendo assim, para cada um dos modelos sugeridos é apresentada a justificativa para sua utilização e o seu detalhamento.

Considerando que o foco do trabalho é comprovar a superioridade dos modelos de redes neurais, evidencia-se uma revisão do funcionamento de cada um dos tipos de rede neural (RN) proposto para que seja possível aplicá-los com maior efetividade, já que suas vantagens e desvantagens são clarificadas.

Diante da exposição acima, de forma específica os objetivos deste trabalho são:

i. Verificar as restrições e limitações existentes para o processo de previsão de vendas no varejo de moda, considerando as características do setor;
ii. Identificar os métodos de previsão mais utilizados em pesquisas científicas para o varejo de moda;
iii. Construir os modelos de previsão para as séries temporais reais deste trabalho;
iv. Comparar a acuracidade dos métodos avançados, que inclui técnicas não lineares como as redes neurais versus modelos estatísticos clássicos;
v. Comparar os métodos avançados e estatísticos clássicos com o modelo utilizado, atualmente, pela empresa estudada;
vi. Verificar o impacto da inclusão de variáveis exógenas tanto nos modelos avançados como nos modelos estatísticos clássicos.

O primeiro e o segundo objetivo serão explicitados e detalhados ao longo do capítulo 2, onde é feita a revisão bibliográfica. Já os demais objetivos serão endereçados no capítulo 5, onde é feita a construção dos diversos métodos e a avaliação comparativa entre eles para cada categoria de produto presente nesta dissertação.

1.3. Estrutura do trabalho

Considerando os objetivos colocados anteriormente, a estrutura deste trabalho é composta por seis capítulos. No segundo capítulo, é apresentada a revisão bibliográfica que, mais especificamente, busca contextualizar o tema, enquadrá-lo dentro dos campos de estudo e evidenciar um resumo do que já produzido na literatura nacional e internacional sobre o assunto. No terceiro capítulo, é realizada uma breve revisão teórica dos modelos utilizados para previsão, tornando possível o entendimento para a aplicação dos mesmos. No quarto capítulo, é vista a metodologia de trabalho, onde são especificados detalhadamente os dados, os métodos de pesquisas, as ferramentas e os *softwares* que são empregados. No quinto capítulo, são mostrados os modelos propostos, os respectivos resultados obtidos com cada modelo para cada categoria de produto e a comparação entre eles, em termos de acuracidade. No último capítulo, são apresentadas as principais conclusões e algumas sugestões para trabalhos futuros.

2. Revisão Bibliográfica

Neste capítulo é apresentada a revisão da bibliografia do tema abordado nesta dissertação. Mais especificamente, o objetivo deste capítulo é mostrar um resumo do que já foi produzido no meio acadêmico, evidenciando os aspectos comuns e os aspectos distintos colocados pelos autores estudados sobre a temática de previsão de vendas no varejo de moda. Além disso, é definido o setor no qual este trabalho é desenvolvido.

A revisão bibliográfica está dividida em três partes: (1) o setor de varejo de moda no Brasil, contextualizando o trabalho, (2) previsão de vendas no varejo de moda e suas peculiaridades, e (3) apresentação de trabalhos realizados no Brasil e no setor sobre o tema.

2.1. O Setor de Varejo de Moda no Brasil

Neste item são conceituados e apresentados os principais aspectos do varejo brasileiro e do varejo de moda, que é o setor onde está situado o trabalho. Antes, porém, é interessante colocar de forma resumida a complexidade da indústria têxtil para entendermos como o produto é distribuído ao varejo.

Na Figura 1, é possível verificar como se dá a interligação entre as partes para a produção de uma peça de moda até ela chegar à ponta final, que é o varejo.

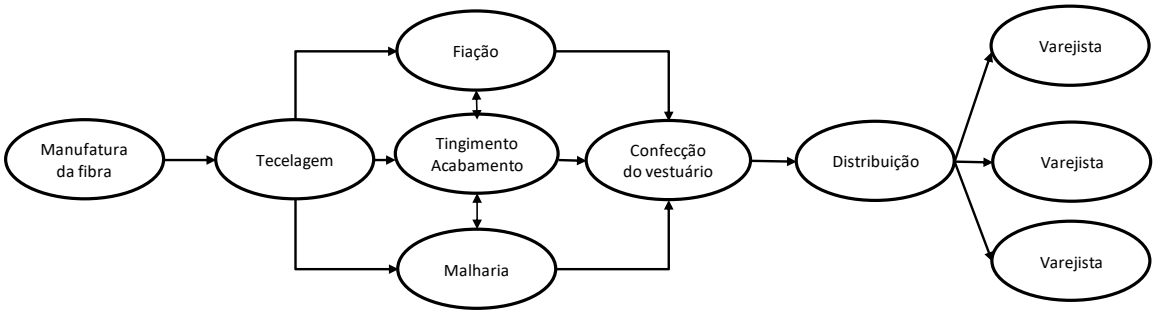


Figura 1 – Estrutura cadeia de abastecimento do setor têxtil
Fonte: Adaptado de CHOI et al. , 2014, p. 10.

No Brasil, a cadeia de abastecimento têxtil e de moda pode ser caracterizada por operar de forma tradicional. Isto significa dizer que ela apresenta os seguintes aspectos:

- pouca coordenação entre os vários elos;
- previsões independentes, geralmente não representando a realidade;
- prazos de entrega extensos, estoques volumosos com risco de obsolescência;
- elevado número de SKU's (*stock keeping unit*) em função de tamanhos, cores, detalhes;
- produtos impactados pela sazonalidade e pelo aspecto de moda de forma disitinta entre si (sujeitos a maior ou menor influência) e, geralmente, com ciclo de vida curto.

Tais características fazem com que a cadeia, de modo geral, não esteja operando de forma eficiente e esteja em um estágio de desenvolvimento inferior ao da cadeia de abastecimento dos hipermercados e supermercados. Apesar disto, as empresas que compõem o setor têxtil estão progredindo de forma a construir uma cadeia mais ágil e enxuta.

Segundo Cattini (2004), uma cadeia de abastecimento mais enxuta significa: ter foco no cliente e ele estar no comando da produção, quebra das barreiras entre elos da cadeia para que seja possível ter uma visão geral do processo, precisão e rapidez no atendimento dos pedidos. Com relação às características de agilidade que a cadeia deve possuir, isto significa: tempos de resposta mínimos, domínio sobre a capacidade (folga e qualidade), utilização da tecnologia da informação e a colaboração entre os elos da cadeia.

Esta nova visão da cadeia de abastecimento têxtil permitiria gerar uma operação mais eficiente, com maior integração entre as partes, redução de custos, sobretudo através da redução dos estoques e, fundamentalmente, permitiria atender melhor às necessidades do consumidor, já que este determinaria o funcionamento da cadeia. Na verdade, esta visão da cadeia de abastecimento busca a constituição de uma cadeia integrada e flexível, que seja capaz de colocar o produto certo e no local certo, respondendo assim às mudanças e necessidades do seu consumidor final.

Segundo alguns autores, o varejo é:

"Todas as atividades que englobam o processo de venda de produtos e serviços para atender a uma necessidade pessoal do consumidor final" (PARENTE, 2000, p. 22).

“Todas as atividades envolvidas diretamente na venda de bens e serviços diretamente aos consumidores finais para uso pessoal” (KOTLER, 2000, p. 540).

Berman e Evans (1998) estabelecem três características do varejo que representam, em pontos-chave, os conceitos de varejo vistos em Parente (2000) e Kotler (2000). Neste sentido, de acordo com Berman e Evans (1998), o varejo se distingue dos outros negócios pelos seguintes aspectos: (1) as vendas são de pequenas quantidades, (2) na maioria das vezes os consumidores compram por impulso, não existindo uma compra planejada, e (3) geralmente, as pessoas necessitam frequentar pontos de vendas para adquirirem produtos/serviços.

A partir disto, pode-se afirmar que o varejista é “um negociante que vende produtos e serviços de uso pessoal ou familiar aos consumidores” (LEVY; WEITZ, 2000, p. 26).

O varejo, apesar de intermediário, possui um papel fundamental na cadeia de valor. Ao manter o contato direto com o consumidor final, ele é responsável por identificar suas necessidades e desejos, definir o que será produzido e oferecer experiências prazerosas que fortaleçam a sua marca, conforme Parente (2010).

O consumo das famílias é o principal indicador de relevância do setor, representando cerca de 63,4% no PIB, em 2016. Verifica-se então, que o setor varejista é uma das principais atividades econômicas da economia brasileira. Mas especificamente, segundo estudo do SBVC de 2016, o impacto do varejo ampliado, que considera as categorias de material de construção, no PIB, é de 42,54% e do varejo restrito é de 20,84%.

Conforme o IBGE, o setor do comércio, que compreende tanto atacado como o varejo, é o maior gerador de empregos formais do país e emprega 22,3% dos trabalhadores formais brasileiros, sendo o maior empregador brasileiro e, assim, um pilar fundamental da economia nacional.

Para verificarmos as possibilidades de interação e os canais possíveis de distribuição, é primordial apresentar a classificação e os formatos varejistas. O varejo é muito mais amplo e diversificado do que se imagina, passando pelo varejo físico, on-line, de produto e de serviço e de franquias dos shoppings centers.

Segundo Parente (2000), as empresas varejistas podem ser classificadas a partir de vários critérios. Considerando um primeiro critério, pode-se dividir os estabelecimentos varejistas conforme seu tipo de propriedade: independentes, redes, franquias, departamentos alugados e sistemas verticais de marketing.

Outro tipo de classificação é o de varejo com lojas e sem lojas. O varejo com lojas pode ser dividido em três tipos: alimentício, não alimentício e serviços. E o varejo sem lojas pode ser identificado como: marketing direto, vendas diretas, máquinas de venda e varejo virtual. Na Figura 2, é apresentado um resumo desta classificação.

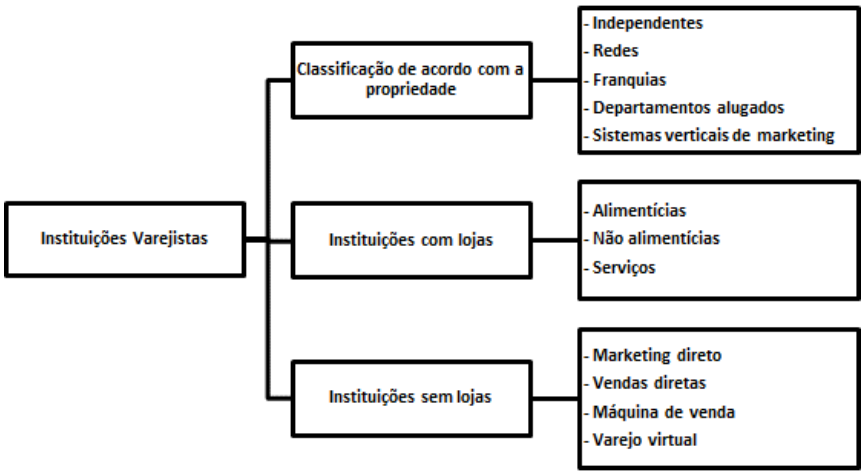


Figura 2 – Classificação das instituições varejistas
Fonte: PARENTE, 2000, p. 25.

Ainda segundo Parente (2000), os formatos existentes de varejo alimentício com lojas são: bares, mercearias, padarias, minimercados, lojas de conveniência, supermercados compactos, supermercados convencionais, superlojas, hipermercados e clubes atacadistas.

Já os tipos de varejo não alimentício com lojas são os seguintes: lojas especializadas, que vendem uma única linha de produtos e podem ter vários tamanhos; lojas de departamentos, que são lojas de grande porte (com área de venda superior a 4000 m^2) e oferecem várias linhas de produtos e serviços; minilojas de departamento ou magazines, que podem ser vistas como lojas de departamento de linhas limitadas, o que faz com que também se aproximem muito do formato de loja especializada; *category killer*, formato mais comum nos Estados Unidos e que representa uma grande loja especializada em uma linha de produtos; lojas de desconto; lojas de fábricas, que são lojas de propriedade de fabricantes e que comercializam exclusivamente as suas linhas de produtos.

No caso de varejo de serviços, existe uma infinidade de formatos que vai desde salão de beleza até clínicas médicas. Os varejos de serviços podem ser agrupados, segundo Berman (1998), a partir dos seguintes critérios: grau de tangibilidade, grau de

competência, intensidade de mão-de-obra, grau de contato com o cliente e por objetivo (lucro e não-lucro).

Com relação ao varejo sem lojas, explicando melhor cada um de seus formatos, temos: o marketing direto, que é definido como um sistema de marketing interativo entre vendedor e consumidor, que utiliza alguns veículos de comunicação para que o contato entre as partes seja feito, sendo um exemplo disto a compra de produto por catálogo, via correio; a venda direta, que é um sistema que possui como essência o contato direto entre vendedor e consumidor para demonstrações e explicações dos produtos, tendo como exemplo a Natura e a Avon; máquinas automáticas de venda (*vending machines*), onde os consumidores compram produtos/serviços através de máquinas, sendo representadas, por exemplo, pelas máquinas de refrigerantes e cafés situadas em locais de grande fluxo; e o varejo virtual (online), que é conceitualmente realizado na internet e não em lojas físicas e tem apresentado um crescimento exponencial na última década.

Outra classificação existente no varejo é a feita pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para a realização da Pesquisa Anual do Comércio (PAC). As atividades do segmento empresarial do comércio brasileiro, para efeito dessa análise, foram agrupadas em três categorias distintas: (1) comércio atacadista, (2) comércio varejista, e (3) comércio de veículos automotores e motocicletas e de peças. As definições para cada uma destas categorias, segundo IBGE, são:

- Comércio atacadista: revenda de produtos que serão utilizados no processo produtivo. Destinam-se às instituições públicas, revendedores, indústrias, profissionais autônomos, etc.
- Comércio varejista: revenda de produtos novos ou usados destinados, predominantemente, às pessoas físicas para consumo pessoal ou doméstico, independente da natureza e quantidade vendida, com exceção de veículos automotores, motocicletas e de peças.
- Comércio de veículos automotores e motocicletas e de peças: revenda dos produtos que não estão incluídos no comércio varejista, predominantemente, às pessoas físicas para consumo pessoal ou doméstico, independente da natureza e quantidade vendida.

Estabelecidos os vários tipos de classificações e formatos existentes, deve-se, neste momento, posicionar o varejo de moda. Em termos da classificação do varejo de

propriedade, o varejo de moda é constituído, predominantemente, pelo varejo independente, onde o varejista é proprietário de apenas uma loja, mas encontramos no formato de rede e de franquia também. Além disso, no formato sem propriedade, verifica-se a presença do varejo on-line, sendo a categoria com maior crescimento dos últimos dois anos.

Considerando a classificação do IBGE, o foco do nosso estudo está inserido dentro da categoria de comércio varejista e dentro da subcategoria de tecidos, artigos de armarinho, modas e calçados. Dentro desta subcategoria estão todos os formatos colocados acima.

De acordo com a Pesquisa Anual do Comércio (PAC), realizada pelo IBGE, o comércio varejista de moda e complementos, no ano de 2015, teve um faturamento de 101 bilhões de reais, o que representa 7,1% do total apurado pelo comércio varejista.

O varejo de moda também é conhecido como varejo de vestuário, varejo de confecção ou varejo mole. O mix de produtos comercializados pelos diversos estabelecimentos é bastante amplo, incluindo artigos que podem ir desde moda até itens de cama, mesa, banho, acessórios e calçados.

O setor busca passar a mensagem de que a caracterização e as várias formas de se vestir traduzem também o estado de espírito e o estilo de vida das pessoas que os utilizam. Desta forma, a maior parte das empresas busca renovar constantemente suas coleções e promover o consumo constante das mesmas, o que está alinhado com o conceito amplamente utilizado de *Fast Fashion*, tendo como seu principal expoente a empresa Zara.

O varejo de moda brasileiro é caracterizado por apresentar uma baixa concentração. Os dados da última PAC (2015) evidenciam um total de mais de 225 mil empresas presentes neste segmento. As grandes lojas especializadas, representadas por Renner, C&A, Riachuelo, Pernambucanas e Marisa, em bloco, têm 16% do faturamento total do setor, o qual alcançou o montante de R$195 bilhões em 2017, segundo IEMI (2018). Cerca de 70% do volume produzido é consumido pelas classes B e C, que representam 63% da população brasileira. Já a classe A (5% da população) responde por 18% do total. As 150 maiores cidades do país consomem 61% do total. Em termos de distribuição de gastos, o brasileiro possui um gasto de 5% da sua renda com itens de moda e, aproximadamente, 59,2% do consumo nacional de moda corresponde às classes C, D e E.

Além de extremamente pulverizado, como já foi observado anteriormente, o varejo de moda é extremamente segmentado, onde cada empresa busca atuar em determinado nicho de mercado, focada em um público-alvo específico, de forma a identificar e atender melhor às necessidades de seus consumidores e, garantindo, então, maior aderência e competitividade.

A seguir, definiremos de forma sucinta, o mercado de baixa renda, pois o Lojão do Brás, empresa de onde são oriundos os dados de vendas, tem foco no público de classes C e D, que compõe parte do mercado de baixa renda brasileiro.

Conforme Prahalad (2009), o mercado de baixa renda possui características próprias e pressupor que as características são as mesmas dos consumidores das demais classes sociais leva a distorções tanto na criação de novos produtos como na maneira na sua forma de distribuição. Tal fato faz com que as empresas tenham que organizar seus diversos departamentos para atender de forma mais adequada as especificidades deste público.

O autor coloca, ainda, que a capacidade de consumo do público de baixa renda é determinada pelos seguintes aspectos: viabilidade, que diz respeito ao aspecto inovador para atender às necessidades deste público; o acesso, que diz respeito ao desenvolvimento da cadeia de distribuição para entregar o produto, muitas vezes em áreas de difícil acesso; e disponibilidade, que diz respeito aos pontos específicos para garantir a eficiência da disponibilidade, de forma a não impactar o preço da oferta da empresa.

2.2. O Lojão do Brás

De forma mais detalhada, o Lojão iniciou suas atividades em 21 de outubro de 1995, a partir de uma vasta e bem-sucedida experiência de seus proprietários no setor têxtil. A primeira e maior loja da rede está localizada no Largo da Concórdia, no Brás.

A proposta de valor da empresa sempre foi oferecer qualidade, variedade de produtos e um conjunto de serviços que resultassem em uma melhor experiência de compra para o seu consumidor.

Atualmente, o Lojão do Brás é uma rede com 18 pontos-de-venda, localizados nas principais cidades do estado de São Paulo. Apresentou um faturamento estimado de R$ 330 milhões de reais em 2016. Tem como público-alvo a família da classe C e D. A empresa comercializa uma grande diversidade produtos, agrupados em departamentos, subdepartamentos e categorias. Os grandes departamentos são: moda, cama, mesa e banho, moda íntima, moda praia, acessórios, beleza, utilidade doméstica e o departamento diversos. Cerca de 65% das suas vendas estão concentradas no departamento de moda. Os principais concorrentes da rede são Torra-Torra, Caedu, Eskala e Marisa.

2.3. Importância da Previsão de Vendas

Considerando um ambiente de negócios cada vez mais competitivo e em constante mudança, a acuracidade do processo de previsão de vendas torna-se crítico e decisivo para os resultados de uma organização, podendo, inclusive, ser visto como uma vantagem competitiva da companhia dentro do seu segmento de atuação.

Makridakis et al. (1998) apresentaram os pontos que levaram ao crescimento da importância do processo de previsão, entre eles temos: o aumento da complexidade das organizações e do ambiente em que elas estão inseridas, tornando extremamente difícil analisar o impacto dos diversos fatores; aumento da sistematização do processo de tomada de decisão e o desenvolvimento contínuo de diversas técnicas de previsão com as suas respectivas aplicações, juntamente com a sua disseminação em diversos níveis dentro das companhias, não ficando restrita somente para os cargos de diretoria.

Lindberg e Zackrisson (1991) apontam os quatro maiores problemas associados ao uso de previsões no apoio à tomada de decisões: (1) a incerteza sobre o futuro; (2) a escolha do método para gerar a previsão; (3) a qualidade e confiabilidade do conjunto de dados de entrada do método; e (4) a interpretação correta da previsão, isto é, a forma como serão utilizadas as previsões na tomada de decisão. Desta forma, o papel da previsão de demanda é avaliar incertezas e riscos. Segundo Makridakis et al. (1998), a incerteza não pode ser totalmente eliminada e, consequentemente, o futuro pode sempre mudar e ser diferente do previsto.

Thomassey (2010) coloca algumas das restrições e complexidades presentes na indústria da moda, as quais são: o longo processo de manufatura, a efemeridade do aspecto de moda, o impacto da sazonalidade, o impacto de variáveis exógenas exploratórias, a volatilidade das tendências de moda influenciando as demandas do consumidor e a alta variedade de produtos, em termos de cor, tamanho e modelo. Tais aspectos dificultam a realização de melhorias que busquem aumentar a precisão das previsões, o que é imprescindível para ser bem-sucedido no gerenciamento da cadeia de suprimentos e distribuição. Além disso, o autor coloca que a implantação de técnicas avançadas de previsão de vendas permite: obter a redução do efeito chicote, que representa o nível de distorção da percepção da demanda ao longo da cadeia de abastecimento; aumentar a eficiência de produção para o fornecedor; reduzir a ruptura, que implica em perda de vendas; reduzir o nível das remarcações de preços e aumentar a margem de lucro da organização.

2.4. Previsão de Vendas no Varejo de Moda

Conforme Parente (2000), após definido o *mix* de produtos, as empresas varejistas desenvolvem atividades com intuito de garantir o abastecimento adequado para os seus clientes. Para isto, desenvolvem atividades em termos de gestão de estoques e compras, onde a previsão de vendas desempenha um papel-chave para determinar a quantidade a ser comprada. Neste sentido, os erros nas previsões de vendas podem causar faltas ou excessos de alguns produtos, o que acaba gerando graves problemas e impactos financeiros aos varejistas.

Antes de especificarmos como o processo de previsão é feito no varejo de moda, é fundamental apresentarmos alguns aspectos que afetam especificamente a previsão deste segmento do varejo.

Conforme Choi et al. (2014), o processo de produção e distribuição da indústria de moda contempla várias empresas, é subdivido em múltiplas etapas e grande parte dele está inserido nas cadeias de produção globais. Tal complexidade é intensificada, pois o atributo de moda tem se tornado, cada vez mais, uma característica efêmera e os gostos dos consumidores vêm mudando constantemente.

Os mesmos autores colocam também algumas especificidades dos produtos de moda, as quais devem ser consideradas pelos sistemas de previsão: as vendas são muito sazonais; muitas variáveis exógenas afetam o volume de vendas, tais como promoções, poder de compra dos consumidores, entre outras; as vendas dependem das tendências de moda, o que impacta na ausência de histórico de informação para maior parte das previsões; os itens possuem muitas variações, em termos de cor, tamanho e outros atributos.

Outros aspectos que devem ser analisados e estabelecidos quando estamos fazendo a previsão de vendas no varejo de moda são:

- **Horizonte de previsão:** horizonte de longo prazo, geralmente de pelo menos um ano ou de curto prazo, que representam algumas semanas e são importantes para efeito de reposição. Deve-se considerar que a escolha do horizonte implica em escolha de métodos de previsão distintos.
- **Ciclo de vida dos produtos:** no setor de moda temos, normalmente, três tipos de natureza de itens, que são: os itens básicos, os quais são vendidos ao longo de todo o ano; os itens de moda, que são comprados uma única vez, não possuem reposição, estão disponíveis durante um curto espaço de tempo e são vistos como apostas da área comercial; e os itens mais vendidos, que são itens vendidos ao longo de todo o ano com pequenas modificações acompanhando as tendências de moda e com reposições. Em termos de previsão de vendas, faz sentido verificarmos os itens básicos e os itens mais vendidos.
- **Nível de agregação dos produtos:** temos desde o nível maior de agregação, que é a marca, até o menor nível, que é o SKU. Como os dados vão se tornando menos disponíveis quanto maior o nível de granularidade, mais uma vez, é necessário a aplicação de técnicas distintas de previsão. Lembrando que as técnicas tradicionais demandam um maior volume de dados.
- **Sazonalidade:** esta característica é crítica para o varejo de moda e afeta a previsão em diferentes intensidades, dependendo do produto em si. Por exemplo, quando se compara calça jeans com jaquetas, pode-se afirmar que a sazonalidade é um aspecto mais impactante para o último caso.
- **Variáveis exógenas:** uma série de variáveis exógenas afetam ou o tráfego nas lojas ou a decisão de compra do consumidor, como, por exemplo: variáveis macroeconômicas, feriados, promoções, entre outras. No entanto, o impacto destas variáveis é difícil de ser mensurado e, muitas vezes, elas estão altamente

correlacionadas entre si. A seguir, na Figura 3, é apresentado um quadro resumo com estas variáveis exógenas exploratórias, elaborado por Thomassey (2010).

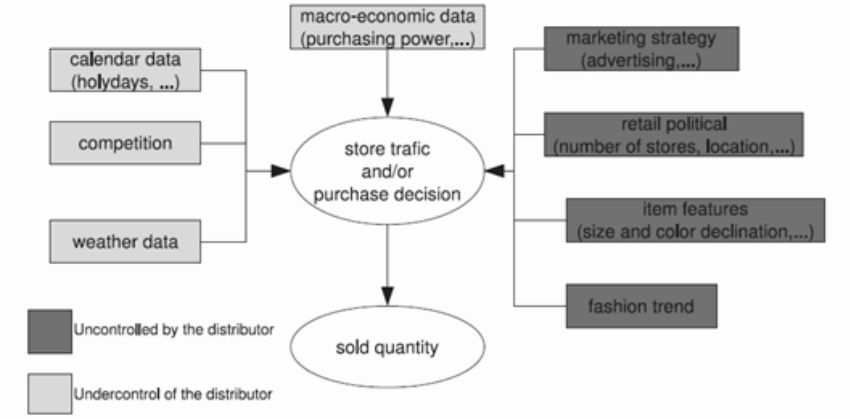


Figura 3 – Variáveis explicativas relacionadas à venda de roupas
Fonte: Thomassey, 2010, p. 473.

Estamos diante de um setor com uma série de peculiaridades e em constante transformação, o que certamente só atrai um número cada vez maior de pesquisadores com o intuito de contribuir para o seu desenvolvimento. Mais do que isto, estamos diante de um mercado cada vez mais competitivo, onde entender a dinâmica do impacto da sensibilidade de preços nos processos de previsão, mais especificamente, por meio de promoções, tem um papel cada vez mais relevante e deve ser contemplada dentro das previsões em si.

Retomando o processo de previsão de forma geral, segundo Hyndman (2013), os passos básicos para a tarefa de previsão são: (1) a definição do problema, onde é estabelecido o escopo do que será previsto e dos envolvidos neste processo, (2) a obtenção da informação, tanto em termos de dados como também a própria experiência das pessoas que já estão atuando neste processo, (3) análise exploratória, com o intuito de fazer uma investigação inicial dos dados, (4) a escolha efetiva dos métodos que serão escolhidos para previsão, sendo realizada a comparação com, pelo menos, dois ou três métodos e sendo este procedimento dependente da disponibilidade dos dados, da força da relação entre as variáveis exploratórias, da variável a ser prevista e da forma como a previsão será utilizada, e (5) uso e avaliação do modelo de previsão por meio da acuracidade. Na Figura 4, temos um quadro resumo deste processo.

Figura 4 – Processo de Previsão
Fonte: adaptado de Hyndman, 2013.

De forma muito semelhante Pellegrini e Fogliatto (2000) colocam que a implantação de um processo de previsão para uma empresa depende de quatro aspectos: (1) a identificação e definição do problema de previsão, (2) a aplicação dos métodos de previsão, (3) a definição dos procedimentos para a escolha dos métodos adequados para o objeto que será previsto, e (4) do suporte organizacional para que os métodos selecionados sejam implantados.

Em relação à definição do problema, temos que o objetivo é conseguir realizar a previsão das vendas mensais em termos de unidades da empresa Lojão do Brás para as categorias de produtos escolhidas, que são representativas para o faturamento da companhia. Sendo o horizonte de previsão a semana seguinte para cada categoria de produto e a seleção do método foi feita com base nos métodos mais utilizados pelo setor de estudo deste trabalho. Em termos de critério de desempate, é considerada a Média dos Erros Percentuais Absolutos (MAPE) para selecionar o método adequado de previsão para a empresa.

Na etapa de obtenção da informação houve grande facilidade de acesso à base de dados da empresa, assim como para a captura da série de dados das variáveis exógenas contempladas neste trabalho. O detalhamento dos dados é apresentado na parte de metodologia, assim como uma análise exploratória dos mesmos.

A próxima etapa consiste na escolha dos modelos de previsão a serem adotados. Antes disso, é interessante apresentar o conjunto dos métodos possíveis para um posterior detalhamento daqueles que forem selecionados. Deve-se destacar que existe uma grande diversidade de métodos e os mesmos podem ser agrupados por diversas abordagens.

Winklhofer et al. (1996) desenvolveram uma estrutura para avaliação das questões que aparecem dentro de um processo de previsão. Mais especificamente, os autores desenvolveram um arcabouço explicitando todos os aspectos que devem ser contemplados no momento de realizar a previsão em si. Esta estrutura também serve

para classificar os diversos estudos realizados na área. Na Figura 5, é possível verificar um quadro com esta estrutura:

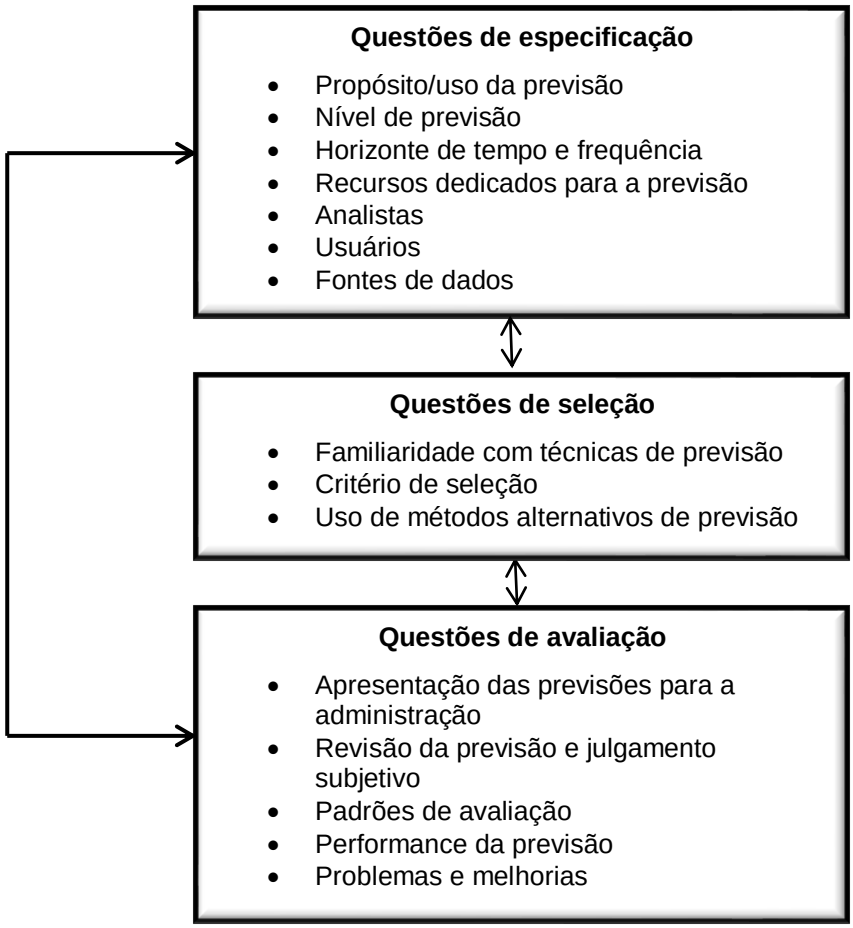


Figura 5 – Framework para análise de métodos de previsão
Fonte: Winklhofer et al., 1996, p. 204.

De forma resumida, o *framework* apresentado na Figura 5 pretende evidenciar todos os pontos que devem ser especificados no momento de efetivação de um processo de previsão, os quais podem ser agrupados em três grandes grupos: questões de desenho, em que é definido o objetivo, o tipo de previsão, os recursos envolvidos, as pessoas que farão parte do projeto e as fontes de dados a serem utilizadas; questões de seleção, em que é estabelecido os métodos a serem utilizados no projeto; e questões de avaliação, em que é determinada a forma de apresentação, revisão e medição do desempenho das previsões construídas.

Makridakis et al. (1998) colocam que as situações nas quais se utilizam técnicas de previsão variam de forma ampla, tendo em vista os diversos horizontes de tempo considerados, os fatores que afetam a variável a ser prevista, o tipo de padrão de data

contemplado e diversos outros aspectos.

Diante de tamanha variação, várias técnicas foram desenvolvidas com o objetivo de atender de forma mais assertiva tais aspectos, as quais podem ser agrupadas em três grandes categorias: a quantitativa, onde existe um volume suficiente de informação quantitativa; a qualitativa, onde há pouco ou nenhum dado quantitativo disponível, existindo somente informação qualitativa; e a de imprevisibilidade, onde não há nem a presença de informação quantitativa e qualitativa. Dentro do grupo quantitativo, temos a subdivisão em séries de tempo, onde se supõe a continuação do padrão histórico, e a subdivisão em pesquisa exploratória, que visa compreender como variáveis exploratórias afetam a variável a ser prevista. Abaixo, na Figura 6, temos um quadro resumo com a categorização dos métodos de previsão expostos acima:

Figura 6 – Categorias de métodos de previsão
Fonte: elaboração própria.

Outra abordagem de agrupamento de métodos de previsão é a colocada por Armstrong (2008). Para o autor, as técnicas devem ser subdivididas, primeiramente, em técnicas de Julgamento, que contempla técnicas com ênfase qualitativa ou subjetiva, e em técnicas Estatísticas, com ênfase quantitativa. A principal diferença destes agrupamentos é que o primeiro é orientado pela opinião de um indivíduo ou de um grupo de indivíduos e, o segundo, pelos dados históricos. O autor defende que a combinação destes dois tipos de técnicas pode levar a previsões mais precisas sobre o futuro.

Considerando a forma de categorização de Armstrong (2008), evidenciada na Figura 7, o foco deste trabalho é a utilização das técnicas estatísticas, o que pressupõe a existência de uma base de dados.

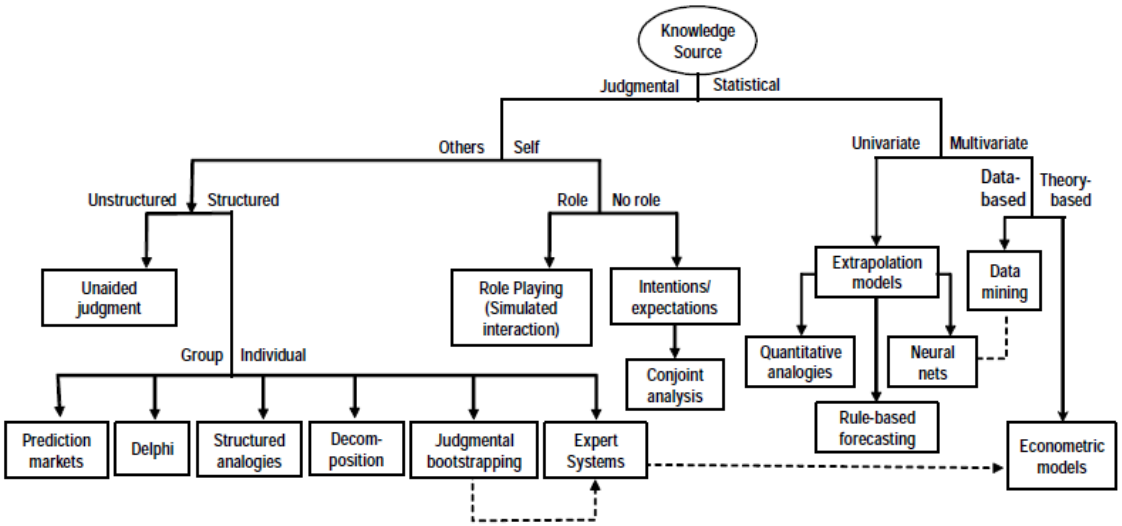


Figura 7 – Categorias de métodos de previsão
Fonte: Armstrong, 2008, p. 2.

Choi et al. (2014) colocam que, no caso da indústria da moda, as técnicas de previsão podem ser agrupadas em técnicas estatísticas, em técnicas usuais e em técnicas avançadas. As técnicas estatísticas clássicas dizem respeito aos modelos de previsão de séries de tempo e aos modelos exploratórios, sendo a técnica de regressão a sua principal representante. As técnicas usuais são métodos customizados pelas empresas, considerando, sobretudo, a complexidade e o custo dos métodos estatísticos e dos métodos avançados. O último agrupamento de métodos são as técnicas avançadas, que contemplam as técnicas mais novas de previsão. A maior parte destas técnicas lida com a não linearidade e pode ser constituída por meio da composição de mais de uma técnica.

A nomenclatura de “métodos avançados” não é a ideal. Certamente, a denominação métodos não lineares seria mais adequada para este agrupamento de métodos. No entanto, será utilizada esta nomenclatura, tendo em vista a referência bibliográfica adotada de Choi et al. (2014). Na Figura 8, pode-se visualizar um quadro ilustrativo com o agrupamento proposto pelos autores.

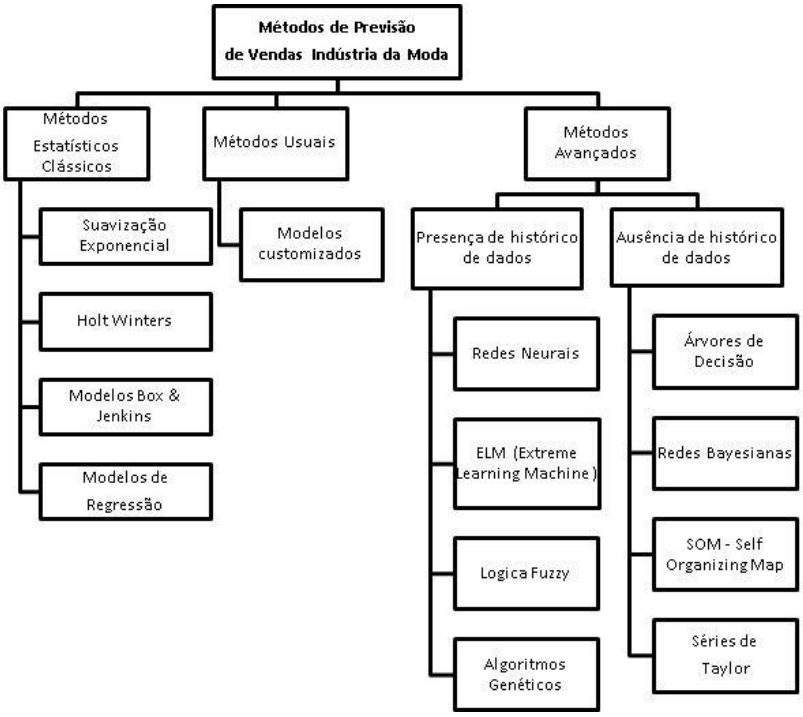


Figura 8 – Categorias de métodos de previsão setor de moda
Fonte: elaboração própria.

Dentro de cada um destes agrupamentos, há subagrupamentos que contém as diversas variações de cada uma das técnicas de previsão, os quais serão explicitados a seguir.

- **Métodos tradicionais:** incluem os métodos de séries temporais como ARIMA, Holt Winter e os modelos de regressão. Entretanto, estes métodos não são implementados de forma eficiente no varejo de moda, sobretudo por estes modelos demandarem um grande volume de histórico de dados e demandarem certa experiência dos operadores;
- **Métodos usuais:** considerando o custo e a complexidade de alguns métodos tradicionais, muitas empresas desenvolvem modelos customizados tendo como base as vendas do último ano, realizando alguns ajustes em termos de sazonalidade e de variáveis exploratórias. Todavia, todas as variáveis não são consideradas e os resultados dependem, novamente, da experiência do operador, o que faz com que, mais uma vez, tenhamos um *gap* de necessidade entre o modelo utilizado e a previsão efetivada.
- **Métodos avançados de previsão de vendas:** este agrupamento deve ser subdividido em dois grupos, tendo como base a disponibilidade de dados. Na presença de histórico, a rede neural é a técnica mais utilizada para previsões

> de curto prazo. A *Extreme Learning Machine* (ELM) é outra técnica também utilizada e tem apresentado resultados superiores aos das redes neurais. *Fuzzy Inference Systems* (FIS), que são modelos com Lógica *Fuzzy*, é outra técnica implementada e que tem como principal vantagem a inclusão dos impactos das variáveis exploratórias. Choi et al. (2014) colocam que foram obtidos ganhos significativos em termos acuracidade para previsões de médio prazo com a FIS, pois esta técnica contempla o efeito de variáveis exógenas. Nos estudos apresentados, as técnicas mais avançadas de predição apresentaram previsões mais precisas do que os modelos tradicionais. Importante destacar que, na ausência de histórico, é necessário que o novo produto, que terá sua previsão realizada, seja *clusterizado*, segundo os grupos criados, com base em suas características e, posteriormente, seja feito um reajuste da previsão após as primeiras semanas de vendas. Neste caso, geralmente, são aplicados os seguintes modelos: redes neurais, redes Bayesianas e árvores de decisão, que lidam melhor com a não linearidade e com relacionamentos mais complexos.

Um aspecto relevante citado por Makridakis et al. (1998) é que a realização de previsões quantitativas pressupõe o atendimento de três requisitos: a informação sobre o passado deve estar disponível, a informação deve ser quantificável em termos de um padrão numérico e, o mais importante deles, a consideração de que pelo menos parte do padrão passado irá se perpetuar no futuro. A última condição é conhecida como suposição de continuidade e pode ser vista como uma premissa básica para utilização de todos os métodos quantitativos de previsão e de vários métodos qualitativos.

A próxima etapa consiste na escolha dos modelos de previsão a serem adotados. Segundo Armstorng (2008), a escolha do método de previsão deve ser feita com base no estágio do ciclo de vida do produto e na disponibilidade dos dados. O autor recomenda a consideração de mais de um método e coloca que é possível evitar a complexidade da seleção simplesmente combinando as previsões.

Com base nas diversas técnicas colocadas acima e utilizando a última abordagem apresentada de categorização de métodos de previsão, o trabalho em questão abordará tanto métodos estatísticos clássicos, por meio dos modelos Box-Jenkins, tendo em vista que esta é uma metodologia muito utilizada e leva em conta somente a variável a ser prevista como própria variável explicativa do modelo, como por

métodos avançados, mais especificamente, por meio das redes neurais.

Dentro dos modelos Box-Jenkins, será detalhado o funcionamento dos modelos: AR, MA, ARMA e ARIMA e SARIMA. Já no caso do agrupamento dos métodos avançados, será feito o detalhamento das redes: MLP e redes neurais recorrentes LSTM. Tanto no caso dos métodos clássicos como no caso dos avançados serão apresentadas as variações dos modelos univariados para modelos multivariados, sendo contempladas então variáveis exógenas.

Após a escolha deste conjunto de técnicas, a próxima etapa é a avaliação do modelo escolhido. Para isto, Passari (2003) coloca que podemos analisar os modelos através dos seguintes critérios: acurácia, custo, habilidade do analista, características desejadas de previsão, características específicas do problema, facilidade do uso, requerimento de dados, disponibilidade de *software*, velocidade, facilidade de interpretação, eficiência, enviesamento e a capacidade de incluir a experiência do gestor.

Ainda segundo Passari (2003), não existe consenso na literatura sobre o critério ou o conjunto de critérios mais importantes para se selecionar uma das técnicas de previsão analisadas. Apesar disso, o critério mais utilizado é o de acuracidade. Posteriormente, são evidenciadas as medidas de desempenho para verificar a acuraria de cada um dos modelos utilizados.

Explicitados os modelos selecionados neste trabalho, a próxima etapa é o detalhamento de cada um deles para que seja possível o seu uso.

2.4.1. Processo de previsão de vendas no Lojão do Brás

O processo de previsão de vendas na empresa Lojão do Brás, atualmente, é feito na periodicidade semanal no horizonte de um ano e usa como *baseline* (referência) as vendas do ano anterior. Em termos de agregação, a previsão é realizada no nível de categoria de produto – por exemplo, camiseta masculina – e, depois, desdobrada em termos da quantidade de modelos que cada loja tem capacidade de receber. Não ocorre a realização de previsões no nível de item (*sku*).

É fundamental ter em mente que o objetivo deste trabalho é a realização de previsões no nível agregado de produto.

Após a consideração do *baseline* das vendas do ano anterior, é verificada a previsão de crescimento da economia para o ano da previsão, por meio do relatório Focus do Banco Central, o qual é ajustado para a expectativa da empresa pela diretoria, considerando o histórico da relação entre o PIB e o crescimento obtido pela empresa ao longo dos anos. Então esta variação é aplicada ao *baseline* para se determinar a previsão final.

Em termos de efeitos comparativos, o modelo atual será considerado como uma das técnicas utilizadas, sendo possível constatar o seu nível de precisão diante das demais técnicas de previsão.

2.5. Principais trabalhos realizados

A realização deste estudo no Brasil é de grande importância, tendo em vista que foram encontrados trabalhos de aplicação de técnicas tradicionais de previsão de vendas para o varejo de moda, porém não temos relatos de aplicação de técnicas mais avançadas, descritas anteriormente, o que é extremante aplicável e pode trazer ganhos diante da grande incerteza em todo o processo de previsão e também, como já observado anteriormente, diante das diversas variáveis exploratórias, que acabam impactando a previsão de vendas em si.

Quando é realizada a verificação para o varejo brasileiro de forma geral, já que não foi localizado nenhum trabalho de aplicação das técnicas mais avançadas de previsão de vendas para o varejo de moda, foram encontrados três trabalhos relevantes de aplicação de redes neurais para previsão de vendas no varejo, os quais são: Felisoni de Angelo et al. (2011), Pasquotto (2010) e Almeida e Passari (2006). No primeiro caso, os autores compararam a previsão de vendas de forma agregada do mercado varejista por meio de duas técnicas: séries temporais ARIMA e modelos de Redes Neurais (RNs) e concluíram que a diferença de precisão entre as duas previsões geradas foi baixa.

Deve-se ressaltar que as RNs, efetivamente, apresentaram resultados de acuracidade superior, porém existe um grau maior de complexidade para definição da arquitetura adequada e há uma série de testes e validações que precisam ser realizadas. Desta forma, observa-se claramente um *trade off* entre um pequeno ganho

de precisão contra um alto investimento de tempo para se estabelecer o modelo apropriado no caso da rede neural.

No trabalho de Pasquotto (2010) é explorada a aplicação de redes neurais recorrentes simples, mais conhecidas como Rede de Elman, na previsão de séries mensais de três séries temporais do varejo de bens e serviços do Brasil, as quais são: produtos farmacêuticos, adubo e tráfego aéreo. Neste trabalho, o autor não obteve maiores ganhos de acuracidade com aplicação deste tipo de técnica.

Já no trabalho de Almeida e Passari (2006), a previsão de vendas é aplicada no nível individual de produto, seguindo, coincidentemente, uma recomendação final deixada no trabalho de Felisoni de Angelo et al. (2011), que analisaram os dados de forma agregada. Interessante notar que os autores, diferentemente do outro trabalho abordado, e o que foi bastante enriquecedor, exploraram a previsão de vendas do produto de forma individual contemplando também o impacto da demanda simultânea de demais produtos relacionados às vendas desse produto, além de considerar também os impactos de promoção e de preço, que fazem parte do composto varejista. Com base na comparação da acuracidade da previsão, por meio da análise do erro quadrático médio e do erro percentual absoluto médio das redes neurais contra as técnicas de modelagem Naive e de regressão linear, o desempenho das redes neurais foi superior no nível individual de produto.

A produção bibliográfica internacional é muito mais ampla e, por isto, foi necessário estabelecer um foco. Sendo assim, a maior parte dos esforços foi para a análise de artigos com uma abordagem voltada para o varejo de moda, com o intuito exatamente de capturar as suas peculiaridades, o que não foi feito para análise do conteúdo nacional, como já visto anteriormente. Além disso, foi verificada uma variedade enorme de aplicação de técnicas não convencionais, passando por Redes Bayesianas, Lógica Fuzzy, Redes Neurais e a própria combinação de mais de uma destas técnicas simultaneamente.

Deve-se ressaltar, desde já, com base em toda a literatura estudada, a importância de se ter uma previsão para o varejo com um nível de precisão crescente, tendo como objetivo principal garantir maior lucratividade para empresas em um ambiente de maior competição global. Mais especificamente, uma previsão mais assertiva garantirá menores níveis de estoques, menores remarcações, menores rupturas, maiores margens e um efeito chicote menor, o que significa uma cadeia de *supply chain* mais integrada e eficiente.

Em 1998, Zhang et al. escreveu um trabalho muito completo sobre redes neurais com foco no processo de previsão. Conforme os autores, uma das principais aplicações da área de Rede Neural artificial é para previsão, sendo que foi em 1964 a primeira aplicação construída com este objetivo. Zhang et al. (1998) coloca que os principais benefícios de uma Rede Neural são: ela consegue aprender por meio de exemplos, pode generalizar a partir de dados amostrais, ela funciona como um aproximador universal, sendo mais genérica e flexível que os métodos tradicionais de estatística, e não é linear. Em termos de limitações, o mesmo autor coloca os seguintes aspectos: para processos lineares com poucos distúrbios, a rede neural não é uma metodologia eficiente, ela funciona como uma caixa preta, sendo difícil de explicar e interpretar; ela também pode sofrer mais frequentemente de problemas de *overffiting*, dado o grande número de parâmetros que demanda; e não há um método claro para determinarmos a melhor estrutura de rede, sendo necessário um processo de tentativa e erro para chegar à rede mais adequada. Pode-se afirmar, então, que as redes neurais são mais apropriadas para tratar grandes bases de dados, problemas não lineares e problemas de séries temporais com mais de uma variável.

Adicionalmente, Zhang et al. (1998) destaca que o ponto crítico de qualquer rede neural é a definição de sua arquitetura, que, mais especificamente, pode ser dividida nos seguintes aspectos: número de *inputs*, número de camadas ocultas e número de *outputs*. De forma geral, a maior parte das redes neurais possui uma camada oculta, em alguns casos sendo necessário ter duas para ser mais eficiente. O número de *inputs*, no caso de séries temporais, corresponde ao número de defasagens necessário para entender o padrão da série e, assim, fazer a previsão. É a decisão mais crítica a ser tomada, sendo que muitos autores fazem uso do *Akaike Information Criteria* (AIC) para definir tal parâmetro. No caso do número de *outputs*, ele é determinado pelo horizonte da previsão. A função de ativação, na maioria dos casos, é a função logística, podendo ser usada a função linear no caso de previsões de séries de tempo. O algoritmo de treinamento mais popular é o de *backpropagation*, que funciona como um gradiente iterativo.

Um último ponto colocado pelos autores acima é a importância da normalização dos dados para garantir um desempenho superior das redes neurais. Este processo é pré-requisito quando se utiliza uma função não linear de ativação e, mesmo quando se utiliza uma função linear, é recomendado, pois o processo de normalização visa

minimizar os problemas oriundos do uso de unidade de dispersões distintas entre as variáveis e reduzir o tempo de treinamento da rede.

Alon et al. (2001) publicaram um dos primeiros *papers* relevantes sobre a questão da previsão de vendas no varejo com redes neurais e fizeram uma comparação entre RNs e métodos tradicionais, considerando a previsão de vendas de forma agregada, assim como vimos no primeiro artigo brasileiro. Com base na previsão mensal dos dados do varejo americano, através da comparação do erro percentual absoluto médio das RNs contra o valor encontrado das técnicas tradicionais – ARIMA, suavização exponencial de Winters e regressão múltipla –, os autores concluíram que as RNs apresentaram as melhores performances. Além disso, os autores realizaram testes em diversos cenários macroeconômicos e perceberam que as redes neurais apresentaram desempenho muito superior frente às técnicas tradicionais de previsão quando a economia passa por um grande período de volatilidade.

Outra conclusão pertinente e que vai ao encontro da usabilidade das RNs, a qual será vista na revisão teórica posteriormente, é que este tipo de técnica consegue capturar as tendências não lineares e os padrões de sazonalidades. Nelson et al. (1999), discordam efetivamente do ponto de vista em relação à sazonalidade, colocando a necessidade de dessazonalização dos dados para garantir a obtenção de redes neurais mais precisas.

Assim como Felisoni de Angelo et al. (2011), Alon et al. (2001) colocam a importância de as empresas avaliarem o *trade off* entre a acuracidade da previsão e os custos e limitações de implantação de cada método. Apesar dos estudos evidenciarem que as RNs são mais precisas na média, elas são menos parcimoniosas e o custo envolvido para se determinar a estrutura mais adequada é muito maior, exigindo um grande nível de esforço.

Adentrando já na especificidade do setor de moda, onde também está presente na literatura com os termos de têxtil e *fashion*, foram encontrados vários trabalhos dos autores Thomassey, Happiette e Castelain, os quais conseguiram em seus artigos apontar várias peculiaridades do setor e criar alguns modelos específicos de previsão de vendas.

Thomassey et al. (2002) colocam que um sistema de previsão de vendas para o setor têxtil no médio prazo (um ano) deve levar em conta os seguintes aspectos: conseguir reagir rapidamente às variações de tendência e sazonalidade, conseguir identificar e suavizar eventos randômicos, conseguir gerar previsões com pouco

histórico de dados e incluir os efeitos das variáveis exploratórias. Diante do alto grau de incerteza e do grande número de parâmetros que afetam a demanda, os autores acreditam que os modelos tradicionais não são adequados e propuseram dois modelos: HFCCX (*Hybrid Forecasting model with Fuzzy estimation of Corrective Coefficients of the Explanatory variables influence*) e AHFCCX (*Automatic Hybrid Forecasting model with Fuzzy estimation of Corrective Coefficients of the Explanatory variables influence*), sendo que ambos os modelos usam lógica Fuzzy para mapear e prever os efeitos das variáveis exploratórias sobre as vendas e a diferença entre os modelos, e que a última proposta não requer a intervenção de um especialista, realizando a previsão de forma automatizada.

Explicado melhor a técnica de previsão adotada pelos autores acima, HFCCX, ela é dividida em três estágios: os dados de vendas são dessazonalizados da influência das variáveis exploratórias, que neste caso são preço, as estações do ano e o período de férias; os dados dessazonalizados são utilizados para prever as vendas considerando o ajuste sazonal; as vendas previstas são sazonalizadas novamente com a influência das variáveis exploratórias correspondentes da próxima estação. Uma das vantagens dos modelos de inferência *Fuzzy* é que ele permite o uso de regras linguísticas, sendo extremamente intuitivo e, além disso, consegue lidar bem com relações não lineares.

A comparação da técnica prevista por meio de erro quadrático médio e do critério de SBIC, que visa encontrar o modelo mais parcimonioso, contra as técnicas clássicas, Holt Winter com sazonalidade, ARIMA e regressão múltipla com variáveis exploratórias, permitiu verificar uma performance superior da técnica AHCFCCX, que é a HFCCX automatizada. No entanto, deve-se destacar a complexidade e o tempo necessário de aprendizado exigido pelo algoritmo genético do sistema Fuzzy quando comparamos com a maior simplicidade das técnicas tradicionais. Thomassey et al. (2002), no artigo “A short and mean-term automatic forecasting system – application to textile logistict”, apresentam o modelo SAMANFIS (*short-term forecasting model by adjustment of mean-term forecast with adaptive neural fuzzy inference system*), que é o modelo para previsão de curto prazo, complementando o primeiro *paper*, que tinha foco no médio prazo. O objetivo do curto prazo é conseguir se adaptar aos acontecimentos reais, tendo como grande limitação um pequeno volume de informações. O modelo proposto consiste também em um sistema de inferência Fuzzy com uma rede neural correspondente, que passa por um processo de treinamento.

A técnica SAMANFIS tem uma acuracidade superior, em termos de erro percentual médio absoluto, que os métodos tradicionais Holt Winter, Arima e Naive.

Após estes trabalhos de Thomassey, Happiette e Castelain, verifica-se uma contribuição muito relevante de Celia et al. (2003), que fizeram um trabalho aplicado para previsão de vendas de moda. O primeiro ponto interessante que os autores ressaltam e que é crítico para os principais modelos de previsão é que: a maior parte das previsões considera os valores passados como base para prever o futuro. Estes autores dividem os modelos de previsão em três categorias: (1) univariados, que usam o tempo como única variável e são citados como exemplos os modelos de média móveis Holt Winters, Séries de Fourier e ARIMA; (2) multivariável, que usam um conjunto de variáveis e tentam estabelecer as relações de causa e efeito do sistema, como exemplo temos regressão múltipla e o ARIMA multivariável (MARIMA); e (3) métodos qualitativos, que incluem pesquisa de mercado Delphi, painel, analogia histórica e *soft computing*, que são métodos que conseguem identificar padrões não lineares, tendo como exemplos Redes Neurais, Lógica *Fuzzy* e algoritmos genéticos.

Estes autores consideram complexa a previsão de vendas na moda, pois existem muitas variáveis endógenas e exógenas envolvidas, tais como preço, cor, clima, efeito de clima etc. Embora uma regressão linear consiga levar em conta as variáveis exógenas colocadas anteriormente, sua aproximação é estritamente linear. Neste sentido, os autores propõem lançar mão de uma técnica não linear como a rede neural para verificar sua performance contra as técnicas clássicas de suavização exponencial e Holt Winters. A análise do coeficiente de determinação R^2 mostrou um desempenho superior das RNs para as diversas categorias de produtos analisadas perante as técnicas tradicionais. No entanto, a correlação entre a venda atual e a sua previsão foi baixa, indicando um possível problema de *overfitting*, que está diretamente ligado a um problema de treinamento no modelo.

Até o momento, pelo que já foi apresentado, os modelos neurais vêm apresentando melhores resultados dos que as técnicas tradicionais. Apesar disto, Chu e Zhang (2003) fazem um artigo comparando os modelos lineares e não lineares para previsão de vendas no varejo de forma agregada e tocam na questão da dessazonalização dos dados como crítica para performance das redes neurais. Os autores colocam que, historicamente, modelar e prever dados com sazonalidade é um dos maiores esforços de pesquisas, que os métodos heurísticos têm tentado superar nas últimas décadas.

Também ressaltam que vários estudos foram conduzidos comparando modelos tradicionais e redes neurais e existe um misto de conclusões sobre a flexibilidade dos métodos não lineares, que é realmente um ponto compensatório em relação à simplicidade da previsão linear. Posteriormente, é colocada a dificuldade de tratar os dados sazonais nos dois modelos. Diante disto, surge a proposta do artigo que é: verificar qual tipo de modelo é mais adequado para situações de sazonalidade. Comparando a acuracidade de um modelo SARIMA, através do erro quadrático médio, do erro percentual médio absoluto e do erro médio absoluto, com os modelos de redes neurais com dados sazonalizados e dessazonalizados, o trabalho conclui que o melhor modelo com base nos critérios de erro é o de redes neurais com dados dessazonalizados, confirmando o que foi evidenciando no modelo de Nelson et al. (1999).

Deve-se ressaltar que não existe um consenso sobre a dessazonalização dos dados para a utilização da técnica de previsão de redes neurais. Muitos pesquisadores e estudiosos advogam que uma das competências das redes neurais é conseguir identificar claramente os ciclos e padrões de sazonalidade presentes em uma série temporal, não sendo necessário eliminá-los da série para que a modelagem seja realizada.

Depois de verificada esta questão da sazonalidade e do seu impacto nos modelos de redes neurais, pode-se retomar a evolução dos trabalhos do tema de previsão de vendas no varejo de moda ao longo do tempo.

Segundo Thomassey (2010), para se manterem competitivas, as empresas vêm investindo em sofisticados sistemas de informação e logística e, sobretudo, em sistemas de previsão mais precisos e confiáveis. Segundo o autor, é fundamental que os sistemas de previsão para o setor levem em conta os seguintes pontos: sazonalidade das vendas, as diversas variáveis exógenas que impactam as vendas, a dependência das tendências de moda, o que impacta na ausência do histórico de dados, a existência de várias opções de cores, tamanhos e de atributos *fashion*. Atualmente, conforme colocado por Thomassey, as empresas, em sua grande maioria, não utilizam os métodos clássicos de previsão e sim métodos próprios, tendo como base para previsão as vendas passadas e a incorporação de algumas variáveis. Neste caso, observa-se que um número limitado de variáveis é considerado e não são consideradas as técnicas mais modernas de previsão, que garantem uma maior acuracidade.

O autor acredita que a utilização de lógica Fuzzy, algoritmos genéticos e RNs é adequada para a indústria de moda, dada as suas peculiaridades. Este ponto já foi detalhado nos artigos do mesmo autor apresentados anteriormente. Para realização deste tipo de previsão é necessário, no entanto, um histórico mínimo de dados, fazendo com que seja possível a realização da previsão apenas no nível maior de agregação de categoria produto e não em níveis menores, tais como: item, cor e tamanho.

Para previsão no nível de item, que foi a grande colaboração deste artigo, o sistema deve conseguir trabalhar com as seguintes características: ausência de histórico de vendas (substituição total dos itens), um número grande de itens com histórico, pertencentes a coleções equivalentes, e presença dos critérios descritivos. Desta forma, o objetivo passa a ser relacionar os critérios descritivos dos itens que serão previstos com as descrições de produtos antigos, tornando possível encontrar produtos equivalentes, o que permitirá realizar a previsão de venda neste nível de granularidade. Neste caso, é recomendada a utilização de outras técnicas de *machine learning*, tais como modelos de redes Bayesianas ou árvores de decisão para fazer a classificação do item a ser previsto no *cluster* adequado.

Com base na simulação realizada para os 20 itens do artigo em análise, Thomassey (2010) concluiu que seria necessária a reestruturação de toda a cadeia de abastecimento para que fosse possível capturar totalmente o benefício de previsões mais confiáveis.

Wong e Guo (2010) apresentaram de forma sucinta em seu trabalho a evolução, as últimas aplicações e uma sugestão de modelo de previsão para ser utilizado para previsão de vendas no varejo de moda. Os autores consideraram os modelos de Thomassey muito complexos de serem utilizados. Diante disto, os mesmos autores propõem a utilização de um sistema inteligente híbrido (HIS) para superar as limitações dos modelos com redes neurais. Mais especificamente, é feito um aprimoramento da determinação dos pesos da rede neural, o que vai ao encontro das necessidades do setor, que tem como características apresentar grande volatilidade, ciclos de vida curtos dos produtos e mudanças constantes nas necessidades dos consumidores. Considerando os testes mensais, trimestrais e anuais, o modelo HIS apresentou previsões mais precisas que os modelos ARIMA, rede neural e ELM, tendo como parâmetro o erro quadrático médio na maior parte dos testes realizados, indicando então ser o modelo mais apropriado. Menos no caso anual, no qual não é possível

realizar tal afirmação, já que o trabalho é para vários intervalos de tempo, tais como: semanal, mensal e anual.

Aksoy et al. (2012) verificaram que, dada à alta complexidade do setor de moda, sobretudo devido à grande quantidade de artigos e falta de dados históricos, é sugerida a utilização de um modelo ANFIS para previsão de vendas na indústria, que combina a capacidade de raciocínio de um sistema de lógica *Fuzzy* com a capacidade de aprendizado do sistema de rede neural, retomando novamente as ideias de Thomassey. Um ponto que é salientado por estes autores é a questão das cadeias de produção global, que estão em forte evolução no setor de moda, e a mudança mais frequente nos hábitos de consumo, trazendo mais volatilidade e incerteza para o processo de previsão da demanda. Ademais, o aumento da competição levou à previsão de demanda a ser estudada por vários pesquisadores.

Diante disto, foi percebido que a combinação de mais de um método resultava em um desempenho superior, em termos de precisão, do que a utilização de métodos individuais. Mas existia a ausência do desenvolvimento efetivo de um modelo *Neuro Fuzzy*. É fundamental perceber, de antemão, que este modelo é uma evolução do modelo de Thomassey, Happiette e Castelain.

Outro aspecto colocado pelos autores é que o fato dos produtos serem substituídos ao longo das coleções faz com que o histórico dos dados seja perdido e, por isto, os métodos quantitativos devem ser descartados como metodologia de previsão. Surge, então, a necessidade de um modelo mais adequado e o modelo ANFIS parece suprir uma série de exigências para previsões do setor estudado, incorporando critérios qualitativos e quantitativos. A partir dos critérios de erro quadrático médio, de erro percentual médio absoluto e de erro médio absoluto, o modelo híbrido proposto pelos autores apresentou uma assertividade superior ao modelo de rede neural puro.

Com base em Xia et al. (2012), foi possível constatar a aplicação de mais um método híbrido: Extreme Learning Machine com Métricas Adaptativas (AD-ELM). O objetivo das métricas adaptativas é reduzir o problema de determinação de tendência e reduzir o efeito de *overfitting*, isto é, quando há uma aderência muito grande no conjunto de dados de treino e que deixa de existir no conjunto de dados de teste. Os resultados experimentais gerados pela aplicação de testes em vários conjuntos de dados e, comparado com os modelos de redes neurais, AR, Extreme Machine Learning, utilizando como forma de avaliação de precisão o erro quadrático médio e o erro

percentual absoluto médio, evidenciaram que o modelo sugerido AD-ELM consegue realizar previsões com erros muitos menores que os outros modelos, sendo a melhor escolha para previsão de vendas no varejo de moda.

Segundo Choi et al. (2014), a utilização das redes Bayesianas é feita de forma similar do modelo aplicado por Thomassey (2010), onde, a partir das caraterísticas de um item, fazemos a sua previsão com base no histórico de vendas existente de um item similar.

Do ponto de vista da utilização de redes neurais para previsão de séries temporais, três trabalhos de destaque, que podem ser aplicados para qualquer setor, são o de Dorffner (1996), o de Gers et al. (2001) e o Längkvist et al. (2014). Dorfnner (1996) apresenta um resumo de todos os tipos de redes neurais que podem ser aplicados na modelagem de séries de tempo e coloca que o grande benefício da utilização das redes neurais MLP é a possibilidade de aproximar funções não lineares, o que permite criar aplicações mais poderosas.

No trabalho feito por Gers et al. (2001), os autores apresentam testes da aplicação dos modelos de rede neural recorrente LSTM versus a utilização de redes neurais Multilayer Perceptron (MLP). Dada a grande promessa das redes LSTM conseguirem modelar a dependência temporal para as observações de uma série, os autores quiseram efetivamente constatar a superioridade deste tipo de rede. Neste sentido, a conclusão obtida foi que uma MLP de janela fixa teve um desempenho melhor que um modelo LSTM simples tipo AR e recomendaram usar este método somente no caso em que a técnica de MLP de janela fixa não apresentasse bons resultados.

Por fim, no trabalho realizado por Längkvist et al. (2014), é evidenciada uma revisão de aprendizagem profunda para modelagem de séries temporais. Os autores destacam que o principal problema na modelagem de séries de tempo é a consideração do tempo de forma estática. Desta forma, devem ser usadas técnicas que permitam a captura das dependências tanto de longo como de curto prazo, sendo o LSTM uma das técnicas possíveis, a qual consegue memorizar mais de cem períodos.

Diante dos trabalhos apresentados anteriormente, pode-se afirmar que a dissertação em questão tem um caráter bastante inovador no contexto nacional, já que não há trabalhos com a aplicação de redes neurais recorrentes LSTM para nenhum campo de estudo do varejo, e no caso específico do setor de varejo de moda, não foram verificados trabalhos sequer com a aplicação de redes neurais multicamadas (MLP).

Toda a dissertação será desenvolvida usando referências e métodos extraídos deste capítulo.

Após apresentada esta revisão bibliográfica, a qual serve como fio condutor para os próximos passos deste trabalho, será exposta a revisão teórica das técnicas de previsão que utilizaremos efetivamente para a realização da previsão de vendas das séries temporais desta dissertação.

3. Revisão Teórica

Nesta etapa do trabalho, é exposta a uma revisão dos principais conceitos e modelos teóricos quantitativos utilizados para a previsão de séries temporais. Sendo assim, inicialmente é apresentada uma descrição de características presentes nas séries temporais e, depois, a descrição dos métodos utilizados.

3.1. Características das Séries Temporais

Na primeira parte deste capítulo são apresentadas características relevantes para o entendimento das séries temporais.

3.1.1. Definição

Segundo Morettin e Toloi (2006), uma série temporal é qualquer conjunto de observações ordenadas no tempo, sendo que é esta ordenação cronológica que faz surgir o efeito de autocorrelação entres os dados da série. Conforme Gujarati (2006), esses dados podem ser coletados em intervalos regulares, que podem ser de forma diária, semanal, mensal, trimestral, semestral, anual ou ainda em intervalos extremamente curtos de forma quase contínua, a exemplo das cotações das ações.

Ao se analisar uma série temporal, a partir de Morettin e Toloi (2006), os principais objetivos são:

- Investigar o mecanismo gerador da série temporal;
- Criar previsões de valores futuros da série de curto ou longo prazo;
- Descrever graficamente o comportamento da série verificando a existência de tendências, ciclos e variações sazonais, construção de histogramas e diagramas de dispersão, etc.;
- Verificar periodicidades relevantes através de análise espectral;

- Modelagem do fenômeno sobre consideração;
- Obtenção de conclusões em termos estatísticos;
- Avaliação da adequação do modelo em termos de previsão.

Em termos de caracterização, os mesmos autores colocam que as séries podem ser de ordem univariada, baseada somente em uma série histórica, ou multivariada, baseada em mais de uma série histórica. Em relação ao tipo de dado, a série pode ser contínua, onde existe um intervalo contínuo de dados, ou discreta, quando há um conjunto de dados registrado em tempos específicos.

3.1.2. Componentes

Gujarati (2006) coloca que é fundamental determinar os componentes básicos para identificar um padrão de comportamento da série, permitindo uma melhor compreensão e, assim, tornando possível a geração de previsões.

Os principais tipos de componentes que podem estar presentes dentro de uma série temporal são:

- Tendência: determina o comportamento de longo prazo da série, indicando se ela sobe, desce ou permanece estável. Deve-se destacar que a tendência pode ser linear ou apresentar formas polinomiais.
- Ciclos: são padrões que se repetem ao longo da série em períodos superiores há um ano.
- Sazonalidade: são padrões que se repetem ao longo da série dentro de um período de um ano.
- Termo errático: são ocorrências irregulares decorrentes de fatores aleatórios.
- Nível: é a constante.

Fava (2000) constata que, até o início da década de 1970, a maior parte dos trabalhos sobre séries temporais tinha como objetivo decompor a série nos elementos vistos anteriormente, considerando o mecanismo gerador da série de tempo como sendo de natureza determinista.

3.1.3. Estocasticidade

Estocasticidade é um conceito que está ligado diretamente ao campo probabilístico. Neste caso, mais especificamente, o que se quer salientar é que o processo gerador de uma série temporal não é determinístico, ele pode assumir diversos valores.

Conforme Bueno (2008), sabendo que Ω representa um conjunto amostral e τ, os subconjuntos de Ω, cuja probabilidade associada para cada um deles é *P*, temos que: considerando (Ω, τ, P) um espaço de probabilidades e *Z* um conjunto de índices de números reais, definindo a função $y(.,.)$ por $y(.,.): S\ x\ Z \rightarrow R$, a sequência ordenada de variáveis aleatórias $\{y(.,t), t \in Z\}$ é chamada processo estocástico.

De acordo com Gujarati (2006), um processo é estocástico ou aleatório quando se tem um conjunto de variáveis aleatórias ordenadas no tempo.

Enders (2009) define que os valores de uma série estocástica são obtidos por meio de uma função de distribuição de probabilidade em cada instante do tempo e a série obtida reflete apenas um conjunto de realizações possíveis dentro do processo estocástico. Adicionalmente, isto se torna possível pela presença da autocorrelação entre os termos da série.

3.1.4. Estacionariedade

A partir de Bueno (2008), a constatação de estacionariedade é que permitirá a realização de inferências estatísticas dos parâmetros estimados de um modelo, onde temos um processo estocástico. De forma simplificada, uma série temporal é fracamente estacionária quando suas características estatísticas – média, variância e covariância – são constantes ao longo do tempo. Assim, matematicamente, para uma série $y(t)$, onde $\{y_t, t \in Z\}$ e Z = $\{0, \pm 1, \pm 2, \dots\}$ apresentar estacionariedade fraca, ele deve atender as seguintes condições:

Média: $E[y_t] = \mu$

Variância: $Var(y_t) = E[(y_t - \mu)^2] = \gamma = \sigma^2$

Covariância: $\gamma_k = E[(y_t - \mu)(y_{t-k} - \mu)]$

Assim uma série temporal é estacionária se sua média, variância e covariância permanecem constantes, independente do período que foram analisadas, ou seja, elas não variam no decorrer do tempo.

Ao longo deste trabalho não é utilizado o conceito de estacionariedade estrita, que é uma propriedade mais forte de distribuição idêntica ao longo da série.

Outro conceito fundamental, quando se trata de séries temporais estocásticas, é a de ruído branco, pois o termo de erro presente em toda série deve possuir este comportamento. De Oliveira e Favero (2003) colocam que uma sequência ε_t é considerada um ruído branco se para cada período t existir média zero, variância constante e não existir correlação com qualquer realização da própria série. Abaixo, seguem as condições:

Média zero: $E[\varepsilon_t] = E[\varepsilon_{t-1}] = \cdots = 0$

Variância constante: $E[\varepsilon_t^2] = E[\varepsilon_{t-1}^2] = \cdots = \sigma^2$

Covariância nula: $E[\varepsilon_t, \varepsilon_{t-s}] = E[\varepsilon_{t-j}, \varepsilon_{t-j-s}] = \ldots = 0$

Em resumo, um ruído branco é, simultaneamente, temporalmente homogêneo, estacionário e sem memória.

3.1.5. Homecedasticidade e Heterocedasticidade

Segundo Gujarati (2006), uma série temporal pode ser classificada mediante o comportamento da sua variância, a qual pode ser apresentada de formas diferentes em

relação às dispersões, seja com homocedasticidade ou com heterocedasticidade, fator de suma importância para a escolha de modelos que a representem.

Para o mesmo autor, a variância de u_i (erro) para cada X_i (variável explicativa), ou seja, a variância condicional de u_i é um número positivo constante, no caso de presença de homocedasticidade. Em outras palavras, significa que as populações Y correspondentes aos valores de X têm a mesma variância ou ainda que a variação em torno da linha de regressão ou linha das relações médias entre X e Y é a mesma para todos os X. Este conceito está evidenciado na equação abaixo:

$$Var[u_i|\, X_i] = E[u_i - E\,(u_i\,|X_i)]^2 = E\,[u_i^2|X_i]\; =\; \sigma^2$$

Por outro lado, segundo Maia (2017), o conceito de heterocedasticidade é o oposto ao colocado acima. Mais especificamente, a variância dos erros será diferente para cada valor condicional de X_i, como pode ser visto na equação abaixo.

$$Var[u_i|\, X_i] =\;\; \sigma_i^2$$

Maia (2017) também coloca que as principais causas da heterocedasticidade são:

- Natureza das variáveis: algumas variáveis apresentam tipicamente tendência à heterocedasticidade;
- Valores extremos: a presença de valores extremos na amostra pode ampliar significativamente a variabilidade em um determinado ponto do ajuste;
- Falhas na especificação do modelo: a omissão de variáveis importantes pode levar à heterocedasticidade;
- Transformação incorreta das variáveis: a transformação incorreta das variáveis, por meio, por exemplo, de alteração da forma funcional, pode levar à heterocedasticidade.

Conforme Wooldridge (2015), a presença de heterocedasticidade, embora não cause viés ou inconsistência dos parâmetros, invalida os erros-padrão, as estatísticas t e as estatísticas F. Sendo assim, os estimadores deixam de ser eficientes.

3.1.6. FAC e FACP

Bueno (2008) coloca que a função de autocorrelação (FAC) é o gráfico da autocorrelação contra a defasagem. A partir da análise desta função, é possível identificar a ordem *q* de um processo de média móvel (MA), o qual é apresentado posteriormente. Já a função de autocorrelação parcial (FACP), que é a função de autocorrelação parcial, é o processo pelo qual se eliminam as correlações implícitas entre duas observações. Por esta função, é possível identificar a ordem *p* de um processo autorregressivo (AR).

Vasconcellos e Alves (2000) reforçam a importância da FAC e da FACP na determinação de quais os filtros AR, I (integração) e MA compõem o processo gerador da série e as suas respectivas ordens.

A FAC de uma série y_t pode ser definida por:

$$\rho_k = \frac{\sum_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

ou

$$\rho_k = \mathrm{Cov}[y_t, y_{t-k}]/Var[y_t] = \gamma_k/\gamma_0$$

Sendo que γ_k é a autocovariância da k-ésima de ordem y_t:

$$\gamma_k = E[(y_t - \mu)(y_{t-k} - \mu)], \qquad k \in Z$$

Necessariamente, o coeficiente de autocorrelação ρ_k varia entre -1 e 1. Além disso, os coeficientes seguem, aproximadamente, uma distribuição normal com média zero e variância $1/n$.

No caso da autocorrelação parcial, de ordem k, representado por ϕ_{kk}, ela é determinado pelo último coeficiente de cada uma das autorregressões:

$$y_t = \phi_{k1}\, y_{t-1} + \phi_{k2}\, y_{t-2} + \cdots + \phi_{kk-1}\, y_{t-k-1} + \phi_{kk}\, y_{t-k} + \varepsilon_t$$

3.1.7. Linearidade

Linearidade é um conceito importante neste trabalho, pois as redes neurais são conhecidas por conseguirem modelar com maior adequação séries que apresentam não linearidades, o que não é possível com os métodos clássicos, que são utilizados como forma de comparação com os modelos de redes neurais. Simplificadamente, não linear significa que a saída não pode ser reproduzida a partir de uma combinação linear das entradas.

A linearidade é discutida aqui como a forma da relação das observações com as suas respectivas observações passadas. No caso de uma relação linear, conseguimos escrever por meio de uma equação de reta que a observação y_t é função de suas observações anteriores:

$$y_t = \alpha y_{t-1} + \beta y_{t-2} + \cdots$$

Caso não seja possível escrever da maneira acima, estamos diante de uma relação não linear. É possível afirmar, de modo geral, que uma série temporal econômica ou financeira é não linear quando responde de forma diferente a choques grandes ou pequenos.

3.1.8. Enfoques de Séries Temporais

Conforme Morettin e Toloi (2006), para a análise de séries temporais há dois enfoques predominantes: os modelos paramétricos e os modelos não paramétricos. No caso dos modelos paramétricos, a análise ocorre no campo temporal, isto é, com um número finito de parâmetros. Exemplos destes modelos são: modelo AR, ARMA, ARIMA e SARIMA. No caso dos modelos não paramétricos, a análise se dá no domínio da frequência e podem apresentar um número infinito de parâmetros. Como exemplo deste caso temos as RNAs.

Os autores ressaltam também que, independentemente do tipo de modelo, os modelos escolhidos devem ser simples e parcimoniosos, ou seja, devem ser especificados com o menor número de parâmetros possível.

3.1.9. Transformações

Para que seja possível construir modelos de séries temporais, muitas vezes, é necessário realizar algumas transformações previamente, tornando viável a aplicação dos diversos modelos que discutiremos a seguir.

Gujarati (2006) ressalta que, na maioria dos procedimentos de análise estatística de séries temporais, supõe-se que estas séries sejam estacionárias. Caso a série não seja estacionária, devem-se aplicar as seguintes transformações: logaritmo, diferenciação e diferenciação sazonal.

Utiliza-se a transformação logarítmica quando a estacionariedade está ligada de certa maneira à variância da série. No final da modelagem é imprescindível realizar a transformação inversa.

A técnica mais utilizada é, na verdade, a diferenciação, que consiste em tirar a primeira diferença da variável em questão y_t: $\Delta y_{t.}$ Este processo deve ser repetido até se atingir a estacionariedade. Abaixo, temos respectivamente a equação da primeira e da n-ésima diferença:

$$\Delta y_t = y_t - y_{t-1} \qquad \Delta^d y_t = \Delta^{d-1} y_t - \Delta^{d-1} y_{t-1}$$

Quando autocorrelações muito grandes forem observadas em *lags* espaçados *n* períodos adiante e com periodicidade, pode-se lançar mão da diferenciação sazonal para ajustar a série. Abaixo, as equações de diferenciação sazonal são apresentadas:

$$\Delta_S^1 y_t = \Delta_S y_t = y_t - y_{t-s}$$
$$\Delta_S^2 y_t = \Delta_S y_t - \Delta_S y_{t-s}$$
...
$$\Delta_S^D y_t = \Delta_S^{D-1} y_t - \Delta_S^{D-1} y_{t-s}$$

No caso das redes neurais, um procedimento muito utilizado é a normalização. Isto significa que os dados utilizados para as previsões, muito provavelmente, terão que ser padronizados antes do treinamento da rede. Este processo é válido para todos os modelos de redes neurais artificiais. Quando um modelo é treinado com dados com escalas diferentes, com um amplo intervalo de valores, isso pode tornar o treinamento mais lento e até impedir a convergência da rede. Sendo assim, existem dois procedimentos para ajustar a série neste caso: a normalização ou a padronização.

A normalização é um redimensionamento dos dados do intervalo original para que todos os valores fiquem dentro do intervalo de zero e um. A normalização exige que sejam estimadas com precisão valores observáveis mínimos e máximos.

A padronização de um conjunto de dados é feita por meio da reavaliação da distribuição de valores, fazendo com que a média de observação dos valores seja 0 e o desvio padrão seja 1. Como a normalização, a padronização pode ser útil e até mesmo necessária em aprendizado e máquina quando seus dados têm valores de entrada com diferentes escalas. A padronização assume que as observações apresentam uma distribuição gaussiana.

Conforme Vellasco (2007), as entradas de uma rede neural devem ser variáveis informativas e sem correlação. Do contrário, podem atrapalhar a rede na tarefa de aprendizagem. Sobre os dados de entrada, é útil realizar uma normalização dos padrões. Neste sentido, as variáveis de entrada devem ser pré-processadas de modo que seu valor médio calculado sobre todo o conjunto de treinamento seja próximo de zero, ou seja, pequeno quanto ao todo do conjunto, ajudando assim na velocidade do treinamento do modelo e na sua convergência.

Mais especificamente, nesta dissertação, foi utilizada a seguinte forma de normalização para as séries estudadas, sendo i cada elemento da série, pertencente ao conjunto Z.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

3.1.10. Testes estatísticos

Neste item, é especificada uma série de testes estatísticos, os quais são utilizados para verificar aspectos da série temporal, que são fundamentais para aplicação dos métodos escolhidos.

Testes de raiz unitária

O primeiro ponto a ser analisado nos estudos de séries temporais é a verificação se a série possui estacionariedade, para isso a análise gráfica consegue dar um indício desta característica, mas não é conclusiva. Segundo Bueno (2008), a verificação visual de uma série dificilmente permite distingui-la como sendo de tendência estocástica ou tendência determinística, podendo gerar então conclusões equivocadas.

Como a análise gráfica não é o suficiente para determinar a estacionariedade da série, existem alguns testes estatísticos para verificar tal condição, os quais são: *Dickey-Fuller* aumentado (ADF), Phillips-Pherron (PP) e Kwiatkowski–Phillips–Schmidt–Shin (KPSS).

ADF

O teste ADF é uma adaptação do teste *Dickey-Fuller* (DF). Bueno (2008) destaca que o problema do teste DF é a consideração do erro como ruído branco, o que geralmente não ocorre em diversas séries temporais e pode levar a distorções no poder do teste. Então, fez-se um ajuste no teste, introduzindo a quantidade de variáveis autorregressivas necessárias, de maneira que o teste de resíduos não rejeite a hipótese de que há um ruído branco.

De forma mais específica, o teste DF consiste em regredir a série y_t contra seus valores defasados de um período y_{t-1}. As hipóteses a serem analisadas no teste são:

H_0: $\rho = 0$ – série não estacionária

H_1: $\rho \neq 0$ – série estacionária

$$\Delta y_t = y_t - y_{t-1}$$

O próximo passo consiste em testar a significância estatística do parâmetro de y_{t-1} por meio da comparação da estatística t do teste com os valores críticos da tabela criada por *Dickey-Fuller*, considerando três possíveis modelos: equação com intercepto, equação com intercepto e com tendência, e equação sem intercepto e sem tendência.

No caso do teste ADF, temos as mesmas hipóteses do teste DF a serem testadas:

H_0: $\rho = 0$ – série não estacionária

H_1: $\rho \neq 0$ – série estacionária

A principal diferença é a contemplação da presença de autocorrelação entre os resíduos. Logo, podemos expressar o teste ADF da seguinte forma:

$$\Delta y_t = \beta_0 y_t + \delta t + \beta_1 y_{t-1} + \eta_1 \Delta y_{t-1} + \eta_2 \Delta y_{t-2} + \cdots + \eta_p \Delta y_{t-p-1} + \varepsilon_t$$

Neste caso, também será testada a significância estatística do parâmetro de y_{t-1} por meio da comparação da estatística *t* do teste com os valores críticos da tabela criada por *Dickey-Fuller,* ajustada para o teste aumentado.

Phillips-Perron (PP)

O teste PP, segundo Bueno (2008), tem como principal objetivo fazer uma correção não paramétrica ao teste DF, possibilitando que seja consistente mesmo que existam variáveis defasadas pendentes e correlação serial nos erros. Desta forma, o

teste PP torna desnecessária a especificação de um modelo com ordem suficientemente autorregressivo para eliminar a correlação serial dos resíduos.

H_0: $\rho = 0$ – série não estacionária
H_1: $\rho \neq 0$ – série estacionária

Existem três especificações possíveis para o teste:

AR (1) sem *drift*: $y_t = \rho y_{t-1} + \varepsilon_t$
AR (1) com *drift*: $y_t = \alpha_0 + \rho y_{t-1} + \varepsilon_t$
AR (1) com *drift* e tendência linear: $y_t = \alpha_0 + \beta_t + \rho y_{t-1} + \varepsilon_t$

O teste utiliza uma regressão base dada por:

$$\Delta y_t = \beta_1 + \beta_2 t + \delta y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta y_{t-i} + \varepsilon_t$$

E a estatística Z é calculada da seguinte forma:

$$Z = n\hat{\delta}_n - \frac{n^2\hat{\sigma}^2}{2s_n^2}\left(\hat{\lambda}_n^2 - \hat{\gamma}_{0,n}\right),$$

Em que:

$$\hat{\gamma}_{j,n} = \frac{1}{n}\sum_{i=1+j}^{n} r_i\, r_{i-j},$$

$$\lambda_n^2 = \hat{\gamma}_{0,n} + 2\sum_{j=1}^{q}\left(1 - \frac{j}{q+1}\right)\hat{\gamma}_{j,n},$$

$$s_n^2 = \frac{1}{n-k}\sum_{i=1}^{n} r_i^2.$$

A estatística Z é comparada à tabela de valores críticos e assim é verificada a presença de raiz unitária.

Kwiatkwiski-Phillips-Schimidt-Shin (KPSS)

Para realizar a confirmação dos resultados verificados no teste ADF e PP, podemos realizar o teste KPSS. Conforme Bueno (2008), o KPSS é um teste alternativo para verificação da presença de raiz unitária em uma série, já que o teste ADF é considerado de baixo poder. Isto significa que o teste ADF apresenta uma alta probabilidade de cometer o erro tipo II, que é de não rejeitar a hipótese nula quando ela é falsa.

O teste KPSS, desenvolvido em 1992, apresenta uma mudança nas hipóteses a serem testadas. A hipótese nula é que a série é estacionária, assim temos:

H_0: $\rho = 0$ – série é estacionária
H_1: $\rho \neq 0$ – série não é estacionária

O teste consiste no seguinte modelo:

$$y_t = \delta t + \zeta t + \varepsilon_t,$$
$$\zeta_t = \zeta_{t-1} + u_t,$$

Em que: $u_t \sim$ i.i.d. (0, σ2 u) e ε_t é um processo estacionário.
A estatística do teste é determinada por:

$$LM = \frac{\sum_{t=1}^{T} S_t^2}{\sigma_e^2}$$

Em que: S_t^2 é a soma dos resíduos de uma regressão de y_t sobre um intercepto e uma tendência e σ_e^2 é a soma dos quadrados dos resíduos da regressão dividido por T.

O valor calculado é comparado com os valores críticos da tabela de estatística do teste para verificar a sua significância.

Testes de Ljung Box

Na análise de séries temporais assume-se que os erros sejam um ruído branco, isto é, a séries dos erros tem que apresentar média zero, variância constante e não apresentar correlação serial.

Sendo assim, após a estimativa de um modelo com base na metodologia Box-Jenkins, os resíduos devem ser analisados. Para que o modelo seja adequado, os resíduos não podem estar autocorrelacionados, o que significa não apresentar memória. Caso isto não ocorra, o modelo deve ser descartado e outro modelo deve ser especificado.

Para verificar a autocorrelação dos resíduos é aplicado o teste de Ljung Box. Ele é um teste de Box-Pierce adaptado a amostras pequenas para valores elevados da ordem de autocorrelação. O teste de Ljung Box testa a hipótese nula de ausência de autocorrelação mediante a comparação da estatística do teste com o valor correspondente em uma distribuição χ^2. A estatística Q do teste é dada por:

$$Q(K) = \mathrm{n}(\mathrm{n}+2)\sum_{k=1}^{K}\frac{r_k^2}{n-k}$$

Em que: *Q*(*K*) é a distribuição qui-quadrado (χ^2) com K graus de liberdade, n é o tamanho da amostra. Sendo que as hipóteses são definidas como:

H_0: os resíduos são ruídos branco

H_1: os resíduos não são ruídos branco

Testes de Normalidade

Ainda na análise dos resíduos, deve-se verificar se os mesmos, além de não poderem apresentar nenhuma memória, são normalmente distribuídos, o que é característico de um processo de ruído branco. Neste trabalho, é utilizado o teste de Shapiro Wilk, apesar da existência de outros testes. A estatística do teste é a seguinte:

$$W = \frac{b^2}{s^2} = (\sum_{i=1}^{n} a_i\, y_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y_i})^2)$$

Sendo y_t a variável observada e a_i os coeficientes tabelados na tabela de Shapiro Wik. A estatística do teste também pode ser avaliada pelo *p-value*, o qual, se for muito baixo, leva a rejeição da hipótese nula de normalidade para a variável analisada.

Teste de Heterocedasticidade

Em relação aos resíduos, deve-se também verificar se eles são homocedásticos. Para isto, utilizamos o teste ARCH-LM, o qual identifica sinais de heterocedasticidade condicional. Conforme Aiube (2013), este teste foi proposto por Engle em 1982. A volatilidade condicional será constante se todos os α_i do modelo ARCH(q) forem nulos, conforme equação abaixo:

$$h_t = \bar{\omega} + \sum_{i=1}^{q} \alpha_i y_{t-i}^2$$

Deve-se fazer uma regressão, onde:

$$\hat{e}_t^2 = \bar{\omega} + \sum_{i=1}^{q} \alpha_i \hat{e}_{t-i}^2 + \nu_t$$

Sendo que e_t são os resíduos estimados da regressão. Neste teste as hipóteses são:

H_0*:* $\alpha_1 = \alpha_2 = \alpha_q = 0$ - sem efeito ARCH

H_1: presença do efeito ARCH

Conforme Bueno (2008) pode-se mostrar que:

$ARCH - LM_h = T \times R^2$ distribuído assintoticamente com uma distribuição qui-quadrado (χ^2) sob a hipótese nula. Rejeita-se a nula se o valor calculado exceder o valor tabelado.

Critérios de comparação dos modelos

Diante da possibilidade de serem produzidas várias alternativas de modelos, conforme a metodologia Box-Jenkins, é importante ter em mente uma métrica que possa determinar o número ideal de parâmetros para o modelo e, logo, tornar possível a seleção do modelo mais apropriado. Segundo Bueno (2008), a ideia do critério de informação é minimizar uma função baseada nos resíduos, penalizada pelo número de regressores. O melhor modelo será o mais parcimonioso, satisfazendo o fato dos resíduos terem os menores valores possíveis. Conforme o autor, há três principais critérios de informação: a estatística de Schwarz (BIC), a estatística de Akaike (AIC) e a estatística Hanna-Quinn (HC).

Neste trabalho será adotado o critério de Akaike, o qual é determinado pela seguinte equação:

$$AIC(p,q) = \ln \sigma^2 + n\frac{2}{T}$$

3.2. Métodos Estatísticos Clássicos

Os métodos estatísticos clássicos podem ser divididos em lineares e não lineares. A não linearidade, neste caso, está diretamente ligada ao tratamento da heterocedasticidade, o que só passou a ser estudado a partir da década de 1980 por Engle.

3.2.1. Métodos Estatísticos Clássicos Lineares

Estabelecidos os conceitos iniciais para análise de uma série temporal, torna-se possível especificar os métodos que são usados para modelar as séries deste trabalho. Neste item são apresentados os modelos paramétricos e lineares, representados pelos modelos: AR, MA, ARMA, ARIMA, SARIMA, ARMAX e SARIMAX. O caractere X que aparece nos dois últimos modelos refere-se à inclusão de variáveis exógenas.

Em 1976, surgiu a metodologia de Box-Jenkins, que pode ser vista como um marco divisor para a área de estudo de séries temporais e popularizou o uso dos modelos ARIMA (*autoregressive integrated moving moverage*). Esta metodologia tem como objetivo propor e ajustar modelos lineares, podendo ser estacionários ou não, a uma determinada série de tempo. Ela será adotada no caso dos métodos estatísticos clássicos para este trabalho, tendo em vista o seu amplo uso até os dias de hoje.

Conforme a metodologia de Box-Jenkins (1976), atualizada em 1994, com intuito principal da realização de previsões, é possível prever os valores futuros de uma série por meio de seus valores passados e presentes, explorando a correlação serial existente entre os termos da série. Os componentes existentes em um modelo ARIMA (*p*,*q*,*d*) são: o componente autorregressivo (AR - *p*), o componente de integração (I - *d*), quando a série não é estacionária, e o componente de médias móveis (MA - *q*). Conforme De Oliveira (2003), a equação a seguir define a forma como esta metodologia realiza o processo de estimativa de uma série:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + L + \alpha_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_t + L + \beta_q \varepsilon_{t-q},$$

Em que:

- α_0: constante do modelo estimado;
- α_1 até α_q: parâmetros que ajustam os valores passados de y_t;
- β_1 até β_q : parâmetros de ajuste da série em termos de choques passados;
- ε_t: ruído branco;
- Demais termos ε: sequência de choques aleatórios e independentes uns dos outros.

Segundo Enders (2009), uma grande limitação do uso dos modelos ARIMA é que ele pressupõe uma relação de linearidade entre as variáveis contempladas no modelo, o que nem sempre é adequado. Outro pré-requisito para a utilização da metodologia é a condição de estacionariedade.

Para se determinar um modelo para uma série temporal a partir da metodologia Box-Jenkins, conforme Gujarati (2006), os seguintes passos devem ser seguidos:

- **Identificação:** descobrir as ordens *p* e *q* do modelo, que dizem respeito, respectivamente, a ordem autorregressiva (*p*) e a ordem de médias móveis (*q*). A definição destes parâmetros é feita pela análise dos gráficos da função de autocorrelação (FAC ou ACF) e da função de autocorrelação parcial (FACP ou PACF).
- **Estimação:** este passo consiste em estimar os parâmetros dos termos autorregressivos e de médias móveis;
- **Verificação:** deve-se analisar se os resíduos estimados não rejeitam a hipótese nula de que seja um ruído branco. Caso ocorra rejeição, deve-se retomar o primeiro passo de identificação de um modelo mais apropriado;
- **Previsão:** no último passo é realizada a previsão em si, porém é fundamental verificar o potencial de previsão do modelo.

Métodos de médias móveis – MA(*q*)

O modelo de médias móveis é um modelo dependente dos seus erros defasados e é representado por MA(*q*), sendo equivalente a um ARIMA (0,0,*q*). Abaixo, evidenciamos este processo:

$$y_t = \mu + \sum_{j=0}^{q} \theta_j \, \varepsilon_{t-j}, \theta_0 = 1$$

Em que:

θ_q : corresponde ao parâmetro do modelo MA de ordem *q*;

E_t: representa os choques, isto é, as diferenças entre os valores reais e os previstos da série em cada instante *t*.

A série y_t pode ser definida como uma combinação do ruído branco μ com os erros dos períodos anteriores não correlacionados. Adicionalmente, o processo de médias móveis satisfaz as condições de estacionariedade: média nula, variância constante e sem a presença de autocorrelação temporal. A análise da FAC e da FACP nos permite afirmar que, no caso do processo de médias móveis, ocorre uma truncagem da FAC, que é onde se determina a ordem MA (*q*) da série e ocorre um decaimento da FACP.

Modelo autorregressivo – AR(*p*)

O processo autorregressivo é aquele que depende dos seus próprios valores passados. A seguir é apresentado um processo definido de ordem *p*. É fundamental ter em mente que a condição de estacionariedade, sendo ela forte ou fraca, é que $|\phi| < 1$, do contrário verifica-se um processo explosivo. Ainda, conforme Bueno (2008), se as raízes da polinomial $(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p)$ estiverem fora do círculo unitário ou, equivalentemente, se as raízes da polinomial $(\lambda^p - \phi_1 \lambda^{p-1} - \phi_2 \lambda^{p-2} - \phi_p)$ estiverem dentro do círculo unitário, o processo será fracamente estacionário e, desta forma, poderá ser representado por um processo MA. Abaixo, temos a representação geral do processo AR(*p*).

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t = c + \sum_{j=1}^{p} \phi_j y_{t-j} + \varepsilon_t,$$

Sendo que:

y_t : corresponde à observação da série temporal no tempo *t*;

ϕ_j : corresponde ao parâmetro do modelo AR de ordem *p*;

ε_t : representa o resíduo, ou seja, o que não pode ser explicado pelo modelo.

Assim como no processo de médias móveis, é através da análise das funções FAC e FACP que conseguimos identificar a ordem do modelo. Todavia, no caso de um processo AR (*p*), a FAC decai lentamente e na FACP é que ocorre a truncagem da série

no *lag p*. De forma geral, um modelo AR(*p*) é uma regressão linear da série contra seus valores defasados.

Modelo ARMA (*p*, *q*)

O processo do modelo ARMA, que é o processo autorregressivo de médias móveis, nada mais é que a combinação dos processos AR (*p*) e MA (*q*). A generalização do processo ARMA (*p*,*q*) pode ser escrita da seguinte forma:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i (y_{t-i} - \mu) + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j} = \mu + \vartheta(L)\, \varepsilon_t$$

Em que ϕ_i são parâmetros autorregressivos, θ_j são parâmetros de médias móveis e μ é considerado um processo puramente aleatório com média zero e variância constante.

Mais uma vez estamos diante de uma série estacionária com média zero e variância constante. A análise das funções FAC e FACP nos permitirá constatar o modelo mais adequado. No caso de um modelo ARMA (*p*,*q*), a FAC apresenta um decaimento exponencial após a defasagem *q* e a FACP também evidencia um decaimento exponencial após a defasagem *p*. Pode-se dizer, conforme Bueno (2008), que os modelos ARMA (*p*,*q*) são mais difíceis de serem identificados. O autor apresentou um quadro resumo para identificação dos modelos AR(*p*), MA(*q*) e ARMA (*p*,*q*), o qual está evidenciado na Tabela 1:

Tabela 1– Identificação modelos AR(*p*), MA(*q*) e ARMA (*p*,*q*)

Modelos	FAC	FACP
AR(*p*)	Decai	Truncada na defasagem *p*
MA(*q*)	Truncada na defasagem *p*	Decai
ARMA(*p* ,*q*)	Decai se *j* > *q*	Decai se *j* > *p*

Fonte: Bueno, 2008, p. 47.

Modelo ARIMA (*p*, *d*, *q*)

Os modelos ARMA (*p*, *q*) tem como requisito apresentar séries estacionárias. No entanto, no mundo real, temos muito mais contato com séries que não são estacionárias, que precisam passar pelo processo de diferenciação para que se alcance o ruído branco. Os modelos ARIMA são equivalentes aos modelos ARMA, mas com a ausência de estacionariedade. O parâmetro *d* determina exatamente o número de diferenciações necessárias para que se consiga determinar a estacionariedade. Segundo Enders (2009), a maior parte das séries alcança a estacionariedade até a segunda diferenciação.

Desta forma, se y_t torna-se estacionária após passar por d diferenças, e se a série z_t que resulta destas diferenciações for modelada por um processo ARMA (*p*,*q*), pode-se afirmar que a série temporal y_t é descrita por meio de um modelo ARIMA(*p*,*d*,*q*). A generalização do modelo ARIMA encontra-se abaixo:

$$z_t = \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + \ldots + \varphi_p z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$

$$z_t = \Delta^d y_t$$

Modelo SARIMA (*p*, *d*, *q*) x (*P*, *D*, *Q*)s

Algumas séries temporais apresentam um componente sazonal relevante, o que significa que o componente sazonal de período *s* pode ser determinado por meio dos seus valores passados, tendo em vista um mesmo período. Importante considerar que, neste tipo de modelagem, a sazonalidade é tratada como um processo estocástico e não determinístico, sendo que no último são usadas variáveis *dummies* para modelagem.

A sazonalidade pode ser modelada tanto por seus valores defasados, representados pelos termos de ordem *P*, como também pelas suas médias móveis, representados pelos termos de ordem *Q*. No caso da necessidade da diferenciação,

aparecerão termos *D*. Abaixo, pode-se verificar a generalização dos modelos SARMA (*P*,*Q*), sem a necessidade de diferenciação:

$$y_t = \phi_1 y_{t-s} + \phi_2 y_{t-2s} + \ldots + \phi_P y_{t-P_s} + \varepsilon_t - \theta_1 \varepsilon_{t-s} - \theta_2 \varepsilon_{t-2s} - \theta_Q \varepsilon_{t-Q_S}$$

A determinação dos modelos SARIMA segue os mesmos passos da modelagem Box-Jenkins, que foram especificados para os modelos ARIMA. A diferença é que passa a ser necessária também a verificação das defasagens sazonais.

Modelo ARIMAX (*p*, *d*, *q*)

Conforme Camelo et al. (2017), o ARIMAX é um modelo derivado do ARIMA. A diferença entre os dois é que o modelo ARIMAX possui, além dos parâmetros autorregressivos e de médias móveis, possíveis entradas exógenas. O modelo ARIMAX pode ser entendido como a combinação dos modelos Auto Regressivo AR(*p*), Integrado (*d*), Média-Móvel MA (*q*) e Exógeno X (*r*), e que pode então ser simbolizado por ARIMAX (*p*, *d*, *q*, *r*). Abaixo, segue a equação matemática para representá-lo:

$$y_t = \rho + \sum_{i=1}^{p} \beta_i \, y_{t-i} + \sum_{j=i}^{r} \omega_j \, w_j + \sum_{j=1}^{q} (\theta_j \varepsilon_{t-j}) + \varepsilon_{t,}$$

Em que:

y_t: variável dependente no tempo *t*;

ρ: é uma constante;

y_{t-i}: é a variável dependente defasada por *i* passos de tempo;

β_i: é o coeficiente de y_{t-i};

p: é o número máximo de intervalos de tempo;

w_j: representa as variáveis exógenas

ω_j: representa os coeficientes das variáveis exógenas;

r: é o número máximo de variáveis exógenas;

θ_j: é o coeficiente do termo de ε_{t-j};

ε_t: é a componente de erro do modelo, com $\varepsilon_t \sim N(o,\sigma^2)$.

O modelo ARIMAX é um modelo multivariável, tendo em vista que a série temporal é explicada pelos seus próprios valores defasados e por outras variáveis.

Existem três pressupostos, conforme Kohn (1979), que devem ser cumpridos para a aplicação do modelo ARIMAX: (1) as variáveis que compõem o modelo devem ser originadas de uma população multivariada, tendo seus resíduos a média zero; (2) as variáveis exógenas são estacionárias; e (3) a variável dependente é estacionária. É fundamental que as variáveis exógenas estejam correlacionadas com a variável dependente que se deseja explicar.

Para se determinar como será a configuração ideal do modelo, será utilizado o processo de *backward elimination*, que consiste em iniciar o modelo com o conjunto completo de variáveis e depois são retiradas, uma a uma, as variáveis que não contribuem para a compreensão da variável dependente.

Conforme Guimarães (2006), os passos do *backward elimination* são: (1) realizar a regressão com todas as variáveis independentes contempladas, (2) calcular o valor *p* de cada variável, e (3) comparar o valor do *p-value* de cada variável calculada com o nível de confiança (α) estipulado para o estudo. Caso o *p-value* da variável seja superior ao nível de significância, a variável deve ser removida e o modelo deve ser repetido até que fiquem somente variáveis com *p-value* menor que α.

Modelo SARIMAX (*p*, *d*, *q*) x (*P*, *D*, *Q*)s

O modelo SARIMAX (*p*,*d*,*q*) x (*P*,*D*,*Q*)s é um modelo bastante similar ao modelo ARIMAX, com a única diferença que o componente sazonal, assim como ocorre no modelo SARIMA, no SARIMAX a sazonalidade é tratada como um processo estocástico e não determinístico.

Desta forma, a sazonalidade pode ser modelada tanto por seus valores defasados, representados pelos termos de ordem *P*, como também pelas suas médias móveis, representados pelos termos de ordem *Q*. Sendo necessária a diferenciação, teremos os termos *D*. O modelo SARIMAX pode ser entendido como a combinação dos modelos auto regressivo AR(*p*), integrado (*d*), média-móvel MA(*q*), Exógeno X (*r*) autorregressivo sazonal SAR(*P*), integrado sazonal (*D*) e média-móvel sazonal SMA(*Q*), podendo então ser simbolizado por SARIMAX (*p*, *d*, *q*, *r*) x (*P*, *D*, *Q*)s.

Previsão Inocente (*Naive*)

A previsão de inocente é um método extremamente simples, pois consiste em adotar os valores futuros da série como iguais ao valor atual: $y_t(h) = y_t$, para todo *h*>0. Geralmente, esta técnica é utilizada para fazer comparação com técnicas de predição mais elaboradas, visando verificar o ganho de performance e acuracidade. A previsão Naive é definida por:

$$y_{t+1} = y_t$$

Sendo que y_{t+1} é a previsão feita em *t* para o dia *t* +1 e y_t é o valor observado no dia *t*.

3.2.2. Métodos Estatísticos Clássicos Não Lineares

Conforme Morettin e Toloi (2006), os modelos lineares não são adequados para descrever o comportamento de séries temporais que apresentam variância condicional evoluindo no tempo ou simplesmente volatilidade. Porém, os modelos não lineares, assim considerados pois são não lineares na variância, a exemplo do ARCH, GARCH e suas variantes, além de modelos de volatilidade estocástica, são modelos mais apropriados para capturar e representar a volatilidade, que é a variância condicional de uma variável, comumente um retorno. De forma geral, pode-se afirmar que uma série temporal econômica ou financeira é não linear quando responde de forma diferente a choques grandes ou pequenos. Até o momento, considerou-se que a volatilidade é constante ao longo do tempo.

Os modelos ARCH e GARCH são muito utilizados para modelar a volatilidade das séries financeiras. Conforme será visto nos testes da série deste trabalho, por meio do teste ARCH-LM, é constatada a presença de heterocedasticidade e, então, de elementos ARCH. Por isto, a seguir, são especificados os modelos ARCH e GARCH.

Modelo ARCH

O modelo ARCH, apresentado em Engle (1982), foi o primeiro a ser proposto com o objetivo de explicar a dinâmica da variação dos retornos dos ativos financeiros. Neste modelo, a variância condicional do erro é modelada conforme um modelo autorregressivo, segundo as expressões abaixo:

$$y_t = \phi_1 y_{t-1} + \varepsilon_t ,$$

$$\varepsilon_t = \nu_t \sqrt{h_t} ,$$

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 .$$

Em que ϕ, α_0 e α_i são parâmetros a serem estimados, considerando que ε_t e ϑ_t são normalmente distribuídos.

Conforme Morettin e Toloi (2006), o modelo ARCH(*q*) considera que a variância condicional do erro é determinada pela defasagem do erro ao quadrado, sendo que *q* é o número de defasagens dos erros ao quadrado. Devido ao grande número de defasagens que, geralmente, é necessário, um modelo mais geral foi proposto, o modelo GARCH, o qual está detalhado a seguir.

Modelo GARCH

A partir do trabalho inicial de Engle, o artigo de Bollerslev, de 1986, amplia a classe dos modelos ARCH, por meio de uma estrutura mais flexível, a qual foi denominada modelos GARCH (*generalized* ARCH).

Como em muitas aplicações é requerido que o modelo ARCH apresente muitos *lags* (*q*) ou defasagens, os modelos GARCH surgem para tornar possível escrever a volatilidade com menos parâmetros. O modelo GARCH (*p*, *q*) pode, então, ser representado por:

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i}$$

Em que *p* >= 0 e *q* > 0 são as estruturas de defasagens e α_0, α_i *e* β_i são os parâmetros a serem estimados.

No modelo GARCH (*p, q*), a variância condicional de ε_t, σ^2 depende do quadrado dos resíduos nos *p* períodos anteriores e da variância condicional nos *q* períodos anteriores.

3.3. Métodos Avançados

Neste capítulo são evidenciados os métodos avançados, que dizem respeito às redes neurais, tanto no modelo de múltiplas camadas como no modelo de redes neurais recorrentes LSTM. Antes, porém, são explicitados conceitos importantes das redes neurais.

3.3.1. Redes Neurais, Aprendizagem de Máquina e Aprendizagem Profunda

Aprendizagem Profunda é uma subárea de Aprendizagem de Máquina, a qual investiga técnicas para simular o comportamento do cérebro humano em tarefas como reconhecimento visual, reconhecimento de fala e processamento de linguagem natural. Algoritmos de aprendizagem profunda objetivam produzir representações hierárquicas de alto nível dos dados de entrada por meio de camadas de processamento sequencial em uma rede neural artificial.

Métodos de Aprendizado Profundo (*Deep Learning*) são, atualmente, o estado da arte em muitos problemas possíveis de se resolver via aprendizado de máquina, em particular problemas de classificação. No entanto, ainda há pouco entendimento de como esses métodos funcionam, porque funcionam e quais as limitações envolvidas ao utilizá-los.

O aprendizado de máquina (*Machine Learning*) é uma das áreas da inteligência artificial (IA) que tem como objetivo o desenvolvimento de técnicas computacionais

sobre o aprendizado e, assim, construir sistemas capazes de aprender de maneira automática. Mais especificamente, o aprendizado de máquina é um método de análise de dados que busca automatizar o desenvolvimento de modelos analíticos, o que se dá por meio de algoritmos.

A aprendizagem de máquina se faz com base na indução, onde a inferência é realizada com base em um conjunto de problemas. As redes neurais representam um dos algoritmos mais utilizados da aprendizagem de máquina. Outros algoritmos são: regressão, árvore de decisão, *k-means*, entre outros.

Com base na Figura 9 é possível entender a relação entre IA, aprendizagem de máquina e aprendizagem profunda. Assim, a aprendizagem profunda é um dos campos da aprendizagem de máquinas que, por sua vez, é uma das áreas de IA, que pode ser definida como a capacidade de uma máquina imitar o comportamento humano inteligente. Mais especificamente, a IA é uma área de estudos da computação que se interessa pelo estudo e criação de sistemas que possam exibir um comportamento inteligente e realizar tarefas complexas com um nível de competência, que é equivalente ou superior ao de um especialista humano.

Para a Universidade de Stanford, o aprendizado de máquinas é como a ciência consegue fazer com que os computadores atuem de maneira específica, sem que sejam explicitamente programados para fazê-lo. Isto é feito por meio de algoritmos, que conseguem analisar dados, aprender com os mesmos e fazer uma previsão.

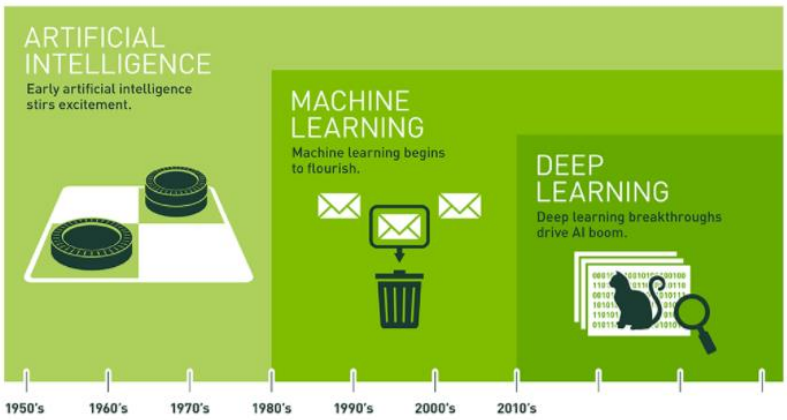


Figura 9 – Relação entre *AI, Machine Learning e Deep Learning*
Fonte: Nvidia, 2018.

O último conceito é o de aprendizagem profunda, que é uma das formas de aprendizado de máquina que tenta imitar a estrutura de cérebro humano, sendo representada por alguns algoritmos específicos de *machine learning*. As redes neurais

mais complexas e densas, com várias camadas intermediárias são um dos algoritmos de aprendizagem profunda.

3.3.2. Conceitos

As redes neurais recebem esta denominação exatamente por tentarem imitar o comportamento biológico do cérebro humano. Mais especificamente, o cérebro humano tem as seguintes características: processamento altamente complexo, ser não linear e paralelo, o que o possibilita realizar um conjunto de atividades com uma eficiência superior quando comparada com os computadores. A capacidade de aprendizado é outra característica essencial do sistema nervoso humano, permitindo que ele se adapte diante de uma série mudanças.

Antes de se falar sobre as redes neurais artificiais, é interessante detalhar, de forma breve, o funcionamento de um neurônio biológico, que facilitará a compreensão do neurônio artificial. Abaixo, temos a Figura 10, que é uma ilustração do neurônio biológico:

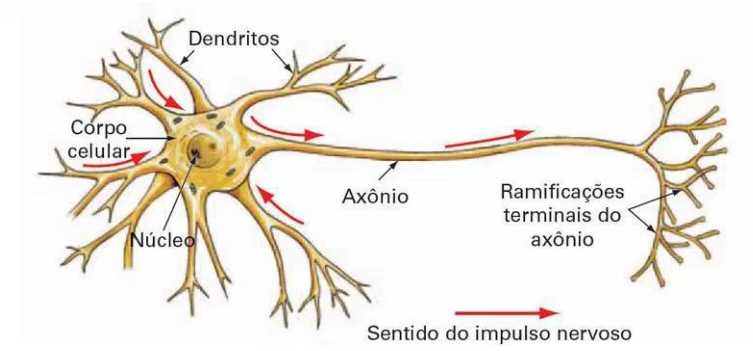


Figura 10 – Representação simplificada do neurônio biológico
Fonte: Data Science Academy.

Conforme Braga et al. (2016), a principal célula do sistema nervoso central é o neurônio, o qual é responsável por conduzir impulsos ao longo de todo o sistema. Este elemento biológico pode ser dividido três partes: os dendritos, o corpo celular e o axônio. A função dos dendritos consiste em captar os estímulos oriundos de outros neurônios conectores ou neurônios sensitivos. Já o corpo celular é responsável por processar todas as informações originadas dos dendritos com o objetivo de produzir um potencial de ativação, que determinará se o neurônio poderá disparar ou não um impulso elétrico

por meio do axônio. Por fim, é responsabilidade do axônio conduzir os impulsos elétricos para outros neurônios conectores. Como é apresentado na próxima etapa deste trabalho, o neurônio artificial emula as partes verificadas do neurônio biológico.

Formalmente, segundo Haykin (2001): “Uma rede neural é um processador maciço e paralelamente distribuído, constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos:

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.
2. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido”.

O autor coloca ainda que os principais benefícios da utilização de uma rede neural são: a não linearidade; mapeamento de entrada-saída, propriedade que está ligada com a questão de aprendizagem; adaptabilidade; reposta a evidências, característica que está vinculada com o reconhecimento de padrões e com o grau de confiança desta classificação; informação contextual, que é a forma como a informação é considerada pela estrutura construída e pela ativação das partes; tolerância a falhas; implementação em VLSI (*very-large-scale-integration*), em paralelo, garantindo uma alto desempenho e a uniformidade de análise do projeto, pois os neurônios representam um dos componentes comuns em qualquer RNs, possibilitando uma integração e uma modulação.

Haykin (2001) coloca ainda que foi demonstrado que determinados tipos de RNs podem aproximar qualquer função contínua a qualquer nível de precisão. Assim, pode-se dizer que as RNs são formas funcionais mais flexíveis do que os métodos estatísticos tradicionais. Desde já, vale a ressalva de que, apesar de os modelos tradicionais serem lineares e menos flexíveis, são mais fáceis de serem entendidos e implementados.

Algumas desvantagens das RNs são: dificuldade de justificar o comportamento das redes, podendo ser consideradas “caixas-pretas”; nas RNs que utilizam o algoritmo de *backpropagation* o tempo de treinamento pode ser muito longo; dificuldade de determinar a topologia mais adequada para a RN, considerando os objetivos de convergência e generalização e, dependendo do tipo RN, pode ocorrer a necessidade de um grande volume de dados para a realização do seu treinamento.

Geralmente, as RNs são utilizadas para a resolução de problemas complexos, em que não se conhece exatamente como se desenvolve o relacionamento entre as variáveis envolvidas. Mas, dada a existência de um grande volume dados, é possível aprender e entender tal relacionamento e comportamento, não sendo o objetivo das RNs entender o relacionamentos entre as variáveis envolvidas no problema.

Segundo Fernandes et al. (1995), as redes neurais possuem vários pontos de contato com os modelos estatísticos e econométricos tradicionais, mas com diferença em suas terminologias. A seguir, na Tabela 2, é apresentado um quadro resumo que equipara tais terminologias:

Tabela 2 – Dicionário de Redes Neurais – Econometria

Redes Neurais	Estatística
pesos	parâmetros
conjunto de treinamento	amostra
entradas	variáveis exógenas
saídas	variáveis endógenas
retropropagação	aproximação estocástica
treinamento ou aprendizado	estimação
sinal de entrada	valor das variáveis exógenas
sinal de saída	valor estimado
alvo	valor da variável endógena

Fonte: Fernandes et al., 1995, p. 5.

3.3.3. Histórico

Os primeiros estudos de modelos sobre as redes neurais artificiais surgiram nos anos 1940. Apesar da intensa evolução até os dias de hoje, o estudo dos primeiros modelos e de sua evolução possibilita um entendimento mais claro e completo das redes neurais. A seguir, temos na Figura 11, uma ilustração, que evidencia a evolução das redes neurais.

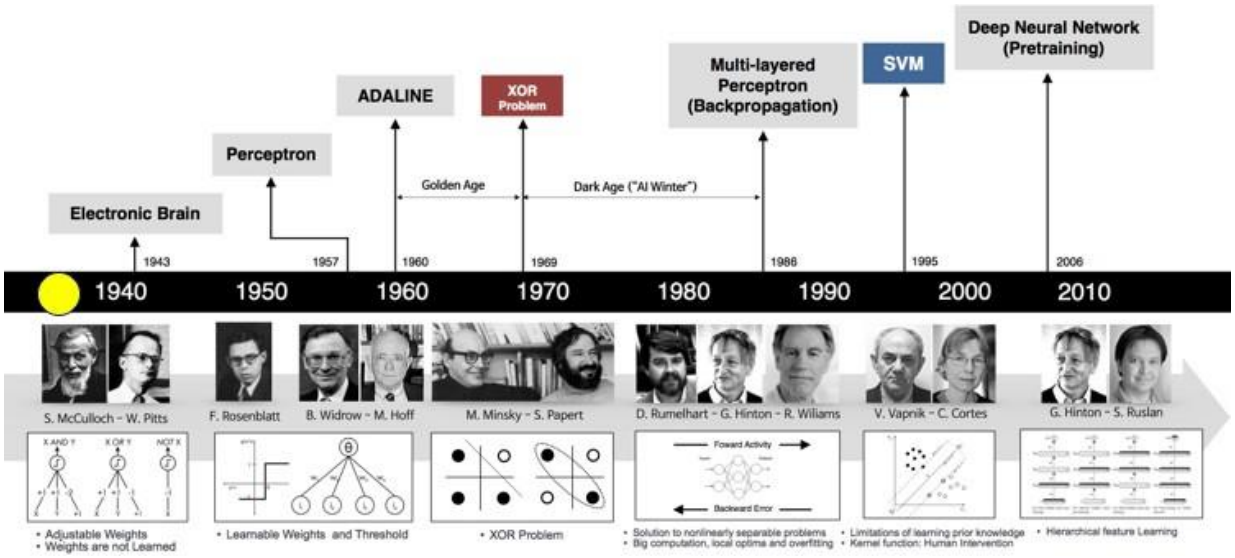


Figura 11 – Marcos no desenvolvimento das Redes Neurais
Fonte: Data Science Academy.

das redes neurais, o livro *The Organization of Behavior*, que influenciou vários modelos de RNAs de destaque na atualidade.

Frank Rosenblatt criou o Perceptron em 1959, um marco para as RNAs e que é fundamental para este campo de estudo até hoje. Contemporâneo ao Percetron surgiu o modelo Adeline (*Adaptive Linear Element*), criado por Bernard Widrow, em 1962. Ambos os modelos são baseados na aprendizagem supervisionada por correção de erros.

Minsky e Papert realizaram um trabalho em 1969, onde constataram que os modelos de redes neurais baseados no Perceptron não eram capazes de aprender uma simples função lógica do tipo “ou exclusiva” (XOR = Exclusive Or). Este trabalho afetou significativamente os estudos das redes neurais, o que fez com que ao longo da década de 1970 não fosse apresentado nenhum avanço expressivo nesta área do conhecimento.

No entanto, na década de 1980, surgiram novos modelos que deram um novo impulso às redes neurais. Mais especificamente, em 1982, surgia o modelo de Hopfield, o qual era baseado em um tipo de rede diferente dos modelos baseados no Perceptron, sendo uma rede com conexões recorrentes com aprendizagem não supervisionada.

Avanços importantes ocorreram na década de 1980, sobretudo devido ao desenvolvimento dos computadores mais velozes, o que até os dias de hoje são fundamentais para a evolução deste campo de estudo.

O modelo de Kohonen surgiu também na década de 1980 e é um modelo diferenciado, pois permite o aprendizado competitivo com uma auto-organização da rede neural, criando os chamados “mapas de atributos auto organizáveis”.

Outro fato relevante foi o surgimento do *Multilayer Perceptron* – MLP, onde o novo algoritmo de aprendizado chamado *backpropagation* conseguiu resolver grande parte dos problemas do Perceptron, inclusive o problema do XOR. Este modelo foi desenvolvido por diferentes pesquisadores quase ao mesmo tempo, como D. Parker e D. Rumelhart, mas foi Rumelhart e Hinton que tornaram este algoritmo famoso com a sua obra “*Parallel Distributed Processing – PDP*”.

O último modelo de destaque, neste período, foi o modelo ART (Adaptive Ressonance Theroy), criado por Gail Carpenter e Stephen Grossberg. Este modelo possui um aprendizado do tipo não supervisionado, criando protótipos (*clusters*) dos padrões aprendidos. O modelo ART teve diversas versões posteriores, entre elas versões do tipo semi-supervisionado e com uso de conceitos da lógica nebulosa (Fuzzy-ART). Nesta mesma época, também foi evidenciado o algoritmo Support Vector Machines (SVM).

Pode-se verificar, pelos fatos mencionados acima, que os estudos sobre as redes neurais passaram por uma grande revolução, a partir dos anos 1980, e esta área de estudos tem se destacado tanto pelas promissoras características apresentadas pelos modelos de redes neurais propostos, como também pelas condições tecnológicas atuais de implementação, as quais permitem desenvolver arrojadas implementações de arquiteturas neurais paralelas em *hardwares* dedicados, obtendo assim performances superiore destes sistemas, quando comparamos com a estrutura convencional. Mais especificamente, isto se dá pela utilização das *graphics processing unit* (GPU), que são as placas de vídeo. A evolução dos estudos das redes neurais são as redes neurais profundas, que possuem várias camadas intermediárias.

3.3.4. O neurônio artificial de uma rede neural

Uma rede neural é composta por uma série de neurônios, os quais estão interligados entre si, influenciando uns aos outros e formando um sistema amplo, o qual é capaz de armazenar conhecimento através de exemplos apresentados e, desta forma,

tornando possível realizar inferências para novos dados. Na Figura 12 pode-se verificar o modelo de um neurônio artificial, segundo Haykin (2011, p. 38):

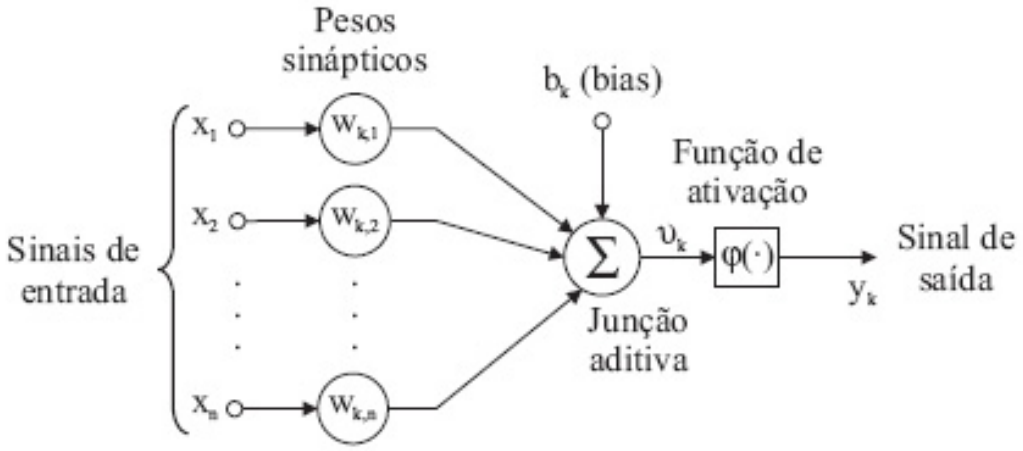


Figura 12 – Rede Neural
Fonte: HAYKIN, 2001, p. 38.

Sendo que:

x_n: são as entradas;
ω_{kn}: são os pesos da rede;
b_k: é o termo polarizador, também conhecido como bias;
ϑ_k: é a combinação linear dos pesos, o que representa a soma ponderada dos pesos pelas entradas acrescentada do *bias*;
$\varphi(.)$: é a função de ativação;
y_k: representa a saída do neurônio.

As equações matemáticas do neurônio são:

$$\vartheta_k = \sum_{n=0}^{n} x_n \, \omega_{kn},$$

$$y_k = \varphi(\vartheta_k).$$

Pode-se afirmar então que, dentro de uma rede neural, os parâmetros a serem estimados são os pesos e o polarizador (*bias*). Considerando que cada neurônio recebe a soma ponderada de todas as entradas, então, o polarizador deverá ter uma entrada fixa +1 ou -1.

É imprescindível fazer uma distinção entre parâmetros e hiperparâmetros dentro do contexto de rede neural. Um parâmetro do modelo é uma variável de configuração interna do modelo, sendo que seu valor é estimado a partir de dados e tendo como principais características: ser requerido pelo modelo para realizar previsões, não precisar ser configurado manualmente e ser salvo como modelo aprendido. No caso das redes neurais, os parâmetros do modelo são: os pesos e o *bias*.

Por outro lado, um hiperparâmetro de um modelo é uma configuração externa ao modelo, sendo que seu valor não pode ser estimado por meio de dados e tendo como principais características: ser utilizado para ajudar a estimar os parâmetros do modelo, ser especificado pelo desenvolvedor do modelo e, geralmente, ser configurado por heurística. Alguns exemplos de hiperparâmetros de redes neurais são: a taxa de aprendizado, o *momentum* e o otimizador, os quais serão detalhados posteriormente.

O *bias*, também conhecido como limiar de ativação, tem como objetivo mover os valores em uma determinada direção. Conforme Silva et al. (2010), o limiar de ativação é uma variável que especifica qual será o patamar adequado para que o resultado produzido pelo combinador linear possa produzir um valor de disparo em direção à saída do neurônio. É fundamental a presença do *bias*, pois a sua ausência faria com que, para entradas nulas, a função de ativação resultasse sempre em um mesmo valor.

O processo dentro de uma rede neural pode ser sucintamente explicado da seguinte maneira: a tarefa realizada por um nó começa com o recebimento dos sinais, originados de outros nós pelas conexões de entrada, ponderados pelos respectivos pesos das conexões e sendo que o seu somatório irá corresponder ao total do sinal de entrada do nó. Posteriormente, o limiar de ativação do neurônio é adicionado a este total, o qual geralmente tem valor negativo. Sinais fracos de entrada serão inibidos por este limiar. Em seguida, é aplicada uma função ao sinal de entrada, que computa um valor y correspondente ao sinal de saída do nó e que é transmitido, como entrada, aos outros nós, aos quais o primeiro está conectado. Esta função é conhecida como função de ativação.

Em termos de representação, uma rede neural é um grafo de funções matemáticas. Lembrando que a teoria dos grafos tem como objetivo estudar as relações entre os objetos de um determinado conjunto. O grafo é composto por nós e arestas. Os nós em cada camada (exceto os nós na camada de entrada) executam funções matemáticas usando entradas de nós das camadas anteriores. As arestas do grafo representam as conexões entre os nós, por onde os valores fluem de uma camada para

a próxima. Essas arestas também podem realizar operações nos valores que fluem ao longo delas, como multiplicar por pesos, adicionar *bias*, entre outras operações matemáticas.

3.3.5. Funções de ativação

Segundo Braga et al. (2016), a função de ativação é responsável por gerar a saída *y* do neurônio através da ponderação dos valores dos pesos e das entradas. É fundamental ter em mente que a escolha da função de ativação pode contribuir de forma decisiva para que a rede consiga captar todas as não linearidades dos dados. Neste sentido, o propósito da função de ativação é introduzir a não linearidade na rede neural.

Uma função de ativação serve também como um limite (*threshold*), alternativamente denominado classificação ou partição. O objetivo de uma função de ativação em um contexto de aprendizagem profunda é garantir que a representação no espaço de entrada seja mapeada para um espaço diferente na saída.

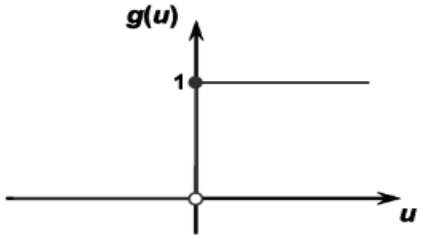
Existem diversas funções de ativação, as quais, segundo Silva et al. (2010) podem ser agrupadas em funções parcialmente diferenciáveis e funções totalmente diferenciáveis, com base no seu domínio de definição. Para as funções que serão apresentadas a seguir, u será o valor de entrada para função de ativação e g é a função de ativação.

Iniciaremos apresentando as principais funções parcialmente diferenciáveis:

Função degrau

Utilizada no modelo proposto por McCulloch e Pits (1943), o resultado da utilização desta função assumirá valores positivos quando o potencial de ativação do neurônio for maior ou igual a zero, do contrário receberá o valor nulo. Abaixo temos as equações matemáticas e a representação gráfica.

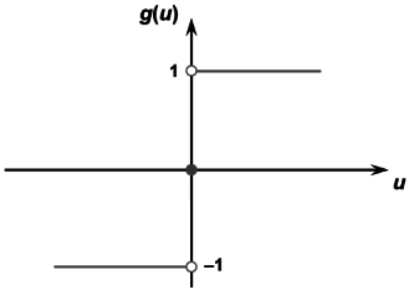
$$g(u) = \begin{cases} 1, & \text{se } u \geq 0 \\ 0, & \text{se } u < 0 \end{cases}$$



Função degrau bipolar ou função sinal

Neste caso, o resultado produzido por esta função assumirá valores unitários positivos quando o potencial de ativação do neurônio for maior que zero, receberá valor nulo quando o potencial for nulo, e valores unitários negativos, quando o potencial for menor que zero. A seguir temos as equações matemáticas e a representação gráfica.

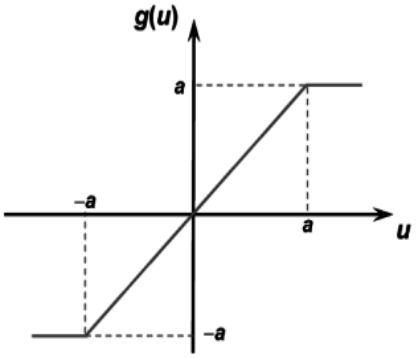
$$g(u) = \begin{cases} 1, & se\ u > 0 \\ 0, & se\ u = 0 \\ -1, & se\ u < 0 \end{cases}$$



Função rampa simétrica

Nesta função, os valores retornados são iguais aos próprios valores dos potenciais de ativação, quando estes são definidos no intervalo [-*a*,*a*], restringindo-se aos valores limites em caso contrário. Abaixo temos as equações matemáticas e a representação gráfica.

$$\mathrm{g}(u) = \begin{cases} a, \ se \ u > a \\ u, \ se \ -a \leq u \geq a \\ -a, \ se \ u < -a \end{cases}$$

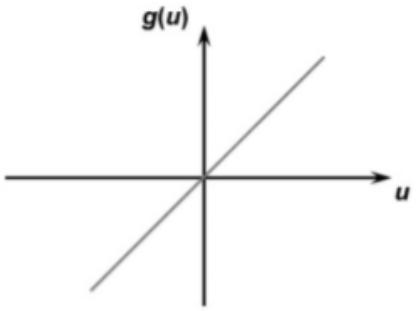


Agora passarão a serem analisadas as funções totalmente diferenciáveis, o que significa que suas derivadas de primeira ordem existem e são determinadas em todos os pontos do seu domínio.

Função linear

Geralmente é utilizada em neurônios de camada de saída. Conforme Vellasco (2007), esta função é utilizada em detrimento das outras, quando o objetivo é eliminar os efeitos de saturação presente em funções sigmoides e hiperbólicas. De forma simplificada, o efeito saturação ocorre quando a rede fornece como saída o resultado nulo. A função de ativação linear produz resultados de saídas idênticos aos valores de potencial de ativação u. A seguir, temos as equações matemáticas e a representação gráfica.

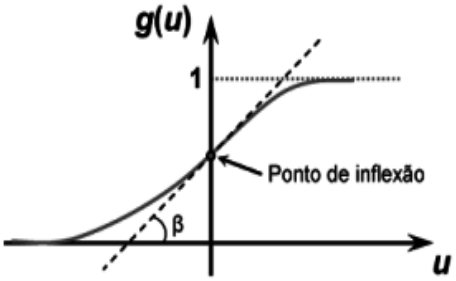
$$\mathrm{g}(u) = u$$



Função logística ou *logsig* ou sigmoide

O resultado de saída produzido pela utilização da função logística terá valores reais entre zero e um. A função logística apresenta somente valores positivo e o formato de um “S”. Abaixo temos sua expressão matemática e seu gráfico.

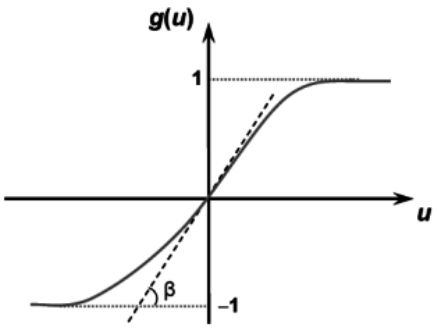
$$g(u) = \frac{1}{1 - e^{-\beta+\mu}}$$



Função tangente hiperbólica ou *tansig* ou *tanh*

Já no caso da função tangente hiperbólica, o resultado de saída da função sempre assumirá valores reais entre -1 e 1. Abaixo temos sua expressão matemática e seu gráfico, onde β é uma constante associada ao nível de inclinação da função logística.

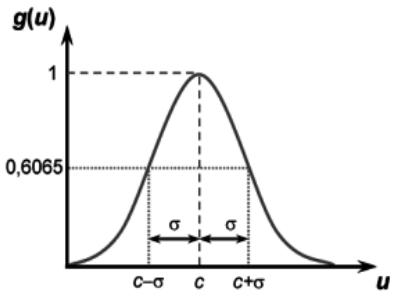
$$g(u) = \frac{1 - e^{-\beta+\mu}}{1 + e^{-\beta+\mu}}$$



Função gaussiana

Nesta função de ativação, a saída do neurônio produzirá resultados iguais para aqueles valores de potencial de ativação u, os quais estejam posicionados a uma mesma distância de seu centro, sendo que sua curva é simétrica em relação a este. A seguir temos sua expressão matemática, sendo *c* um parâmetro que define o centro da função gaussiana e σ o desvio padrão. Abaixo, temos seu gráfico.

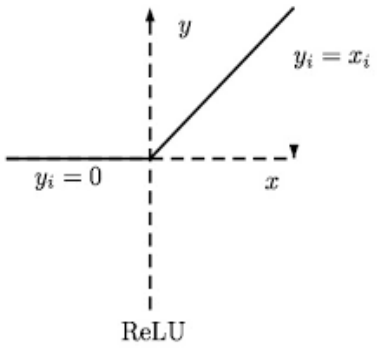
$$\mathrm{g}(u) = e^{\frac{-(u+C)^2}{2+\sigma^2}}$$



Historicamente e como vimos até o momento, as redes neurais utilizaram as seguintes funções de ativação: função hiperbólica tangente, sigmoide ou linear. No entanto, segundo Ponti e Costa (2017), para *deep learning* a função retificadora linear (ReLU) tem sido mais utilizada, pois a mesma facilita o processo de treinamento deste tipo de redes, mais densas e complexas. Dahl et al. (2013) também destacam as melhorias obtidas nas redes profundas com o uso da ReLU e da técnica de regularização *dropout*.

A melhoria do treinamento com a ReLU ocorre porque as funções sigmoidais saturam a partir de um determinado ponto, enquanto que a ReLU é função identidade para valores positivos. Adicionalmente, ao se calcular a derivada da ReLU, seu gradiente terá sempre uma direção não nula, o que não ocorre nas funções sigmoidais e evita o problema de dissipação de gradiente, que será discutido depois. Deve-se ressaltar que a ReLU só é utilizada em camadas ocultas. Abaixo, pode ser vista a equação matemática da ReLU e seu gráfico.

$$g(u) = \max(0, x)$$



Outras funções de ativação usadas em *deep learning* atualmente são: CUBE, ELU, HARDSIGMOID, HARDTANH, IDENTITY, LEAKYRELU, RATIONALTANH, RELU, RRELU, SOFTMAX, SOFTPLUS, SOFTSIGN.

Após a apresentação de todas estas funções de ativação e dos benefícios do uso da ReLU, é importante destacar que a escolha de determinada função de ativação está intrinsecamente ligada às propriedade do problema estudado. Não há uma regra clara para determinar esta decisão. Contudo, uma escolha mais adequada contribuirá para tornar possível a convergência fácil e rápida da rede neural.

Para este trabalho são utilizadas as funções de ativação sigmoide e a ReLU.

3.3.6. Etapas de modelagem da rede neural

Segundo Vellasco (2007), as etapas de modelagem de uma rede neural são:

1) Treinamento e aprendizado: que é obtido pelo ambiente gerador de dados;
2) Associação: que consiste no reconhecimento de padrões distintos;
3) Generalização: que está ligada com a capacidade de a rede reconhecer com êxito o ambiente, que origina os dados, e não necessariamente os dados utilizados no treinamento.

A mesma autora coloca que a escolha das variáveis de entrada é crítica, pois é necessário que existam dados suficientes e representativos para que o aprendizado seja feito pela RNA. O conhecimento é passado para rede através de um algoritmo de treinamento e o aprendizado é transformado e armazenado nos pesos.

O aprendizado apenas se torna viável com apresentação do conjunto dos dados com uma determinada frequência, que é denominada época. Uma época significa uma apresentação completa de todo o conjunto de treinamento. O aprendizado ocorre até que os pesos e o *bias* se estabilizem por meio de uma função custo, permitindo que ocorra a convergência para um valor mínimo e o objetivo pretendido seja alcançado.

3.3.7. Arquiteturas de redes

Conforme Silva et al. (2010), a arquitetura de rede neural estabelece a forma como os diversos neurônios serão dispostos entre si. Os arranjos possíveis são estruturados por meio do direcionamento das conexões sinápticas dos neurônios. Necessariamente, uma arquitetura de redes é composta pelos seguintes tipos de camadas: de entrada, onde ocorre o recebimento das informações; intermediárias, que são as camadas responsáveis por extrair as características associadas ao processo a ser inferido; e a de saída, que é a camada responsável pela apresentação dos resultados da rede. Sendo que cada camada é composta por vários neurônios, com exceção da camada de entrada.

Deve-se ressaltar que a arquitetura de rede está diretamente ligada ao tipo de algoritmo de aprendizagem a ser utilizado, aspecto discutido na próxima parte.

É fundamental destacar o conceito de topologia neste momento, que diz respeito às diferentes formas de composições que uma determinada arquitetura de rede neural pode apresentar. Mais especificamente, diz respeito à definição dos seguintes aspectos: a forma como os neurônios estão organizados e interconectados, o que pressupõe a definição do número de camadas, o número de neurônios, que compõem cada camada e o tipo de conexão entre os neurônios. Adicionalmente, deve-se considerar o valor da taxa de aprendizado, o tipo de função de ativação, a presença ou não do termo *momentum*, a forma de otimização, entre outros.

Em termos de topologia, não existe uma formalização clara e estabelecida para determinar quantas camadas, neurônios, função de ativação e otimizador a ser utilizado em uma rede neural. Neste caso, é função do cientista construir várias redes com diferentes parâmetros e avaliar a rede com melhor desempenho. Apesar disto, é possível seguir referências de outros trabalhos para estabelecer estes pontos.

Segundo Zhang et al. (1998), para evitar problemas de *overfitting*, alguns pesquisadores apresentaram regras empíricas para determinar o número de neurônios na camada intermediária e elas são: 2n+1, 2n, n e n/2, sendo n o número de *inputs* da camada de entrada. O autor destacou ainda que vários estudos de RNAs, em que o número de neurônios da camada oculta era igual ao número de entradas da camada de entrada, reportaram resultados melhores em termos de previsão.

Existem basicamente três tipos principais de arquiteturas de RNAs: as redes diretas ou não recorrentes (*feed foward networks*), as redes recorrentes (*feed backward networks*) e redes competitivas.

Nas redes não recorrentes ou diretas, não ocorre retroalimentação da rede por meio das suas próprias saídas ou das camadas ocultas para as suas entradas. Em outras palavras, não ocorre a formação de ciclos nas conexões entre os neurônios e não há conexões ligando um neurônio de uma camada a outro de uma camada anterior ou de uma mesma camada. Nestas redes, os neurônios são dispostos em camadas distintas e ordenados sequencialmente.

Além disso, nas redes não recorrentes, o fluxo da informação é sempre da camada de entrada para a camada de saída e estas redes podem ser de camada única, como o *Perceptron* e ou de múltiplas camadas (MLP). Nestas redes, cada neurônio de uma camada é conectado com todos os neurônios da camada seguinte e não há conexões entre neurônios de uma mesma camada.

Já no caso das redes recorrentes, existem ciclos nas conexões, o que torna o treinamento mais complexo, mas que são mais adequadas a determinados tipos de problemas, como a modelagem de sistemas que variam no tempo, conforme Pommeranzenbaum (2014). Em outras palavras, em uma rede recorrente, a saída de um neurônio é aplicada como entrada no próprio neurônio e/ou em outros neurônios de camadas anteriores.

Este tipo de rede são redes mais gerais e sua estrutura, segundo Vellasco (2007), não é obrigatoriamente organizada em camadas. Ainda segundo a mesma autora, o fato destas redes possuírem realimentação faz com que a resposta aos estímulos dados seja dinâmica. Então, a inclusão de uma nova entrada faz com que uma nova saída seja calculada, a qual realimenta e modifica a própria camada de entrada, gerando uma memória. Exemplos deste tipo de rede são: rede de Elman, de Jordan e LSTM.

Por fim, as redes competitivas são redes neurais, onde os neurônios estão divididos em duas camadas: a camada de entrada e a camada de saída, chamada de grade. Na última camada, os neurônios são colocados para competir entre si, com base no nível de similaridade entre o padrão de entrada e a grade de neurônios. Então, apenas o neurônio vencedor será ativado a cada iteração. A rede de Kohonen é um tipo de rede competitiva.

Os avanços nesta área de estudo fez com que ocorresse uma multiplicação dos tipos de redes neurais com diversas formas de arquitetura, sendo o algoritmo de *backpropagation* o mais utilizado para treinamento da rede. Alguns tipos de redes são: redes MLP (multicamadas), tendo como principais algoritmos de treinamento o *Backpropagation*, *Quickprop* e *Resilient Propagation*; redes RBF (redes de função basial); redes recorrentes, onde temos como exemplos de configuração a rede de Elman, Jordan, *Time-Delay Neural Network* e *Time-Lagged Feedforward Neural Network* (TLFN); *Deep learning neural networks*; Redes não supervisionadas, como a de Kohonen e ART; métodos híbridos, composto pelos modelos Neurofuzzy e as RNs Genéticas.

Em termos de Deep *Learning*, destacam-se os seguintes tipos de arquitetura: *autoencoders*, as *Deep Belief Networks* (DBNs), *Generative Adversarial Networks* (GANS), as redes convulacionais e as redes neurais recorrentes, por meio das *Long Short Term Memory* (LSTM) e das *Gate Recurrent Units* (GRU).

A seguir, na Figura 13, é apresentado um quadro completo das possíveis arquiteturas de redes neurais do instituto Asimov, que é uma empresa focada em inteligência artificial.

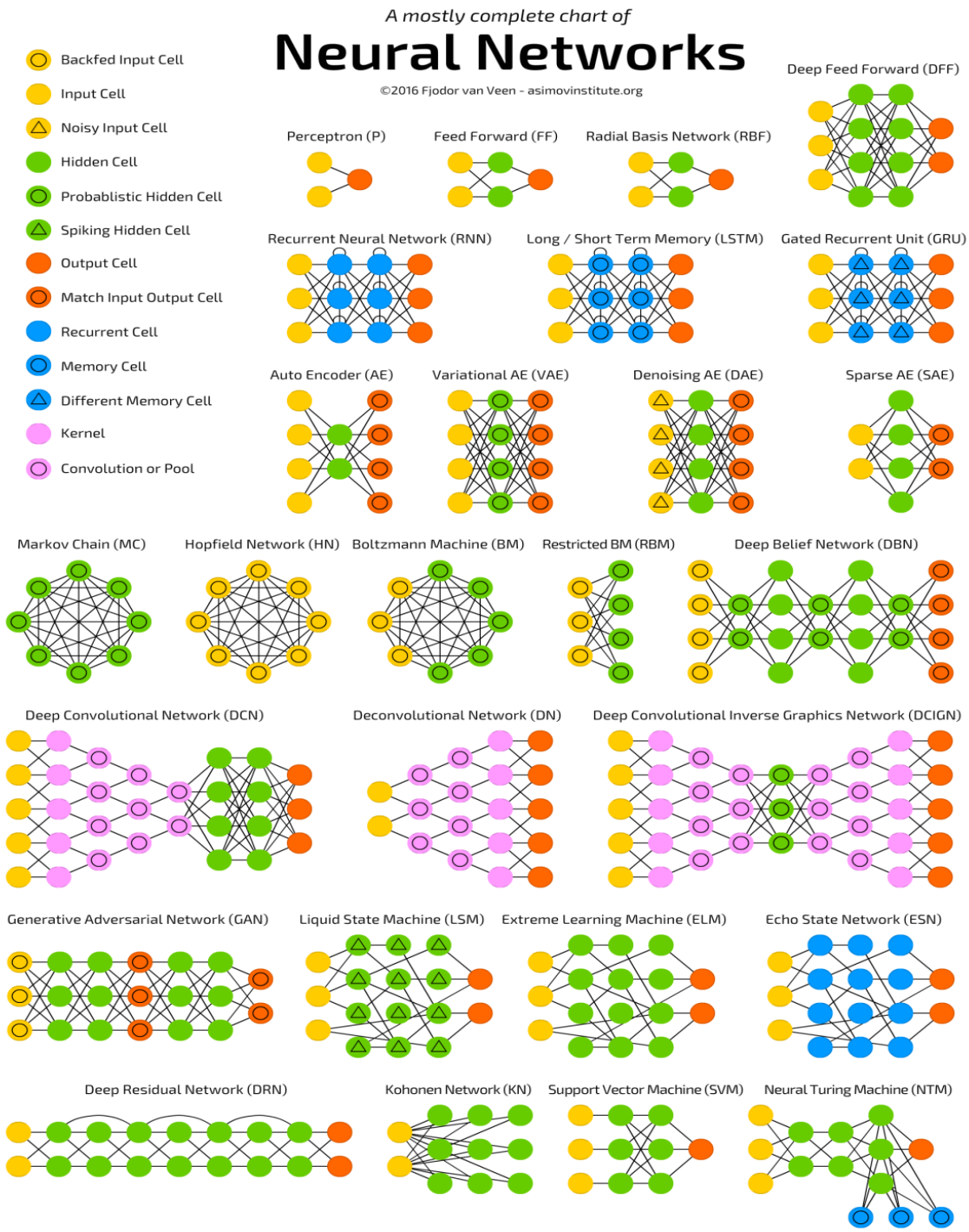


Figura 13 – Quadro de Redes Neurais
Fonte: ASIMOVE Institute.

A seguir serão detalhadas as redes neurais utilizadas neste trabalho.

3.3.8. Rede *Multilayer Perceptron* (MLP)

A rede MLP é uma generalização da rede de camada única alimentada para frente. Segundo Braga et al. (2016), as redes de única camada têm a limitação de solucionar somente problemas com características lineares. No entanto, sabemos que a maior parte dos problemas reais apresentam não-linearidades, por isso se faz necessário as redes MLP.

Conforme Felisoni de Angelo et al. (2011), o modelo de Rede Neural mais utilizado para previsão de séries temporais é o MLP, que é uma extensão do Perceptron proposto por Rosenblatt, sendo composto de várias camadas de neurônios. Neste caso, os sinais de entrada são transmitidos da camada de entrada para a camada de saída, por meio de camadas intermediárias, e não há retroalimentação dos sinais intermediários ou de saída. Traçando um paralelo com a estatística, cada nó de entrada seria uma variável independente e cada nó de saída representa uma variável dependente.

Uma rede MLP é composta por, no mínimo, três camadas: de entrada, oculta, que contém neurônios que processam o problema a ser modelado, e a de saída. A função das múltiplas camadas de uma rede MLP é transformar o problema descrito pelo conjunto de dado no espaço de entrada em uma representação que possa ser consumida pela camada de saída da rede. Com exceção da camada de entrada, as demais camadas são formadas por neurônios.

O algoritmo de *backpropagation* é o mais utilizado para o treinamento da rede MLP, o qual será detalhado posteriormente. Deve-se destacar que é a existência da camada oculta, não existente na rede Perceptron, que permite a resolução de situações onde os dados não são linearmente separáveis. A partir de Braga et al. (2016), os nós da camada de entrada são determinados pela complexidade do problema, o número de neurônios da camada de saída são definidos pela dimensionalidade da resposta desejada e, a grande maioria dos problemas práticos, raramente necessita de mais de uma camada intermediária. Deve-se ter em mente que o aumento do número de camadas escondidas e de neurônios pode gerar uma rede muito complexa, a qual poderá apresentar problemas de *overfitting* e dificuldade de generalização. Este problema não é exclusivo da rede neural MLP, pode ser verificado em qualquer tipo de rede neural. Por outro lado, uma rede neural com poucas camadas e neurônios gera

uma rede pequena, que não conseguirá modelar o problema em estudo. Estamos diante das questões de convergência e generalização.

Braga et al. (2016) colocam que a definição do número de neurônios em cada camada é crítica para o desempenho da rede. Na verdade, a definição do número de neurônios é o aspecto mais importante no aprendizado das redes neurais. No entanto, como já foi colocado anteriormente, não existe uma regra geral que determine de forma precisa o número adequado de neurônios. Os autores colocam que o problema de dimensionamento envolve um ajuste entre a complexidade do modelo neural e a complexidade do problema a ser resolvido. Assim deve-se buscar a minimização do erro quadrático, juntamente com a minimização da complexidade da rede em si.

Segundo Vellasco (2007), existem três características básicas das redes MLP: (1) o modelo de cada neurônio da rede inclui uma função de ativação não-linear, caso não existisse a não linearidade, não faria sentido utilizar uma rede MLP, com a presença da camada oculta; (2) a rede necessariamente contém uma ou mais camadas de neurônios ocultos, que permitem que a rede possa aprender tarefas mais complexas; (3) a rede apresenta um alto grau de conectividade a partir dos seus pesos. Além disso, a autora menciona como principal deficiência, a sua alta complexidade e conectividade, o que torna difícil a sua compreensão teórica. Na Figura 14 é evidenciada uma representação de uma rede MLP.

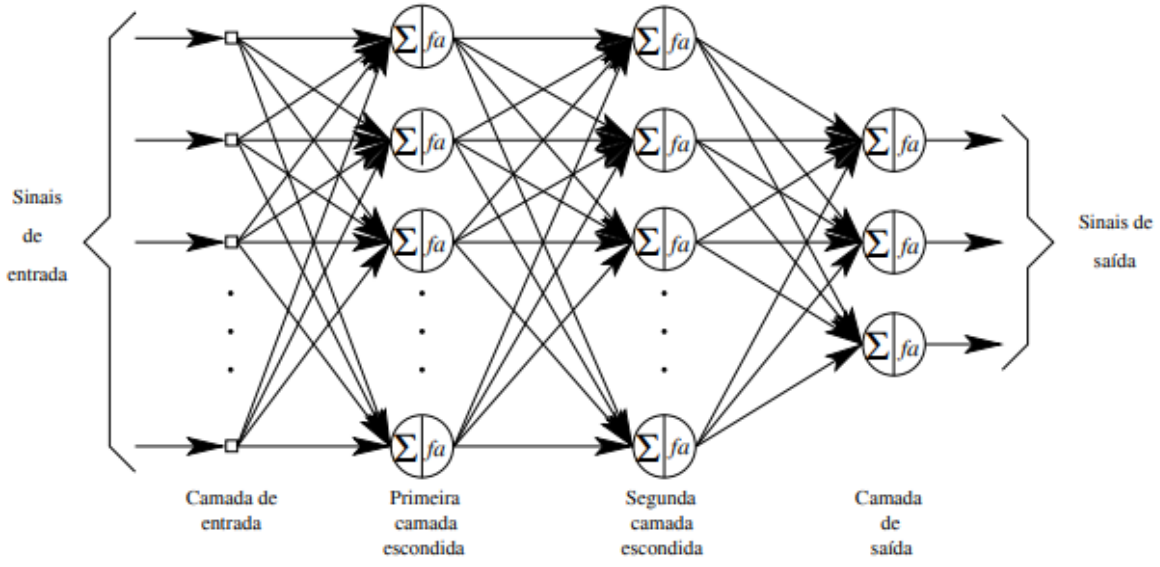


Figura 14 – Rede Neural MLP
Fonte: Adaptado de HAYKIN, 2001, p. 186.

Uma MLP deve passar por duas etapas: a primeira de treinamento, onde é separada uma parte dos dados, conhecida como *insampl*e, para especificação do modelo. Posteriormente, passa-se à segunda fase, que é a de execução, onde a previsão é realizada efetivamente, considerando o restante dos dados não utilizados (*outsample*).

A obtenção da diferença entre os valores conhecidos e os computados mantém relação direta com o benefício de aprendizagem das redes neurais, onde o resultado computado pela rede é sistematicamente avaliado e a diferença em relação ao resultado desejado determina alterações na especificação da rede. O ajuste dos pesos mantém associação direta com a fisiologia do cérebro que, de forma extremamente simplificada, remete a uma aproximação ou afastamento das ligações (sinapses) entre neurônios biológicos. Quanto maior o peso tanto mais estreita é a ligação.

Em resumo, pelos neurônios de entrada inclui-se informação na rede. O neurônio de saída transmite as respostas para rede. Os neurônios transmitem a informação através dos dados de entrada e pelas conexões determinadas, a partir de um processo de aprendizagem. Os pesos de conexão permitem aos neurônios ponderar a influência das variáveis preditivas, dos dados de entrada na rede e na resposta que ela gera. Uma variável preditiva com forte influência na previsão a ser obtida fará com que os pesos de conexão entre a variável e os neurônios a ela conectados sejam maiores.

3.3.9. Rede *Focused Time Delay Neural Network* (FTDNN)

Segundo Haykin (2001), as redes *Focused Time Delay Neural Network* (FTDNN) são redes MLP, onde o aspecto temporal é incluído na rede através da camada de entrada, por meio da introdução de várias defasagens da variável que se visa prever. Deve-se considerar que as entradas precisam ser submetidas sequencialmente à rede. Sendo assim, estamos incluindo uma memória de curto prazo na camada de entrada da rede.

$$y_t = f(x_{t-1}, x_{t-2}, \ldots, x_{t-p})$$

A consideração das defasagens torna possível à rede identificar em que ponto ela está de um determinado intervalo de tempo, possibilitando a resolução de

problemas, onde o tempo é uma variável relevante. Na Figura 15, é possível verificar este tipo de rede neural.

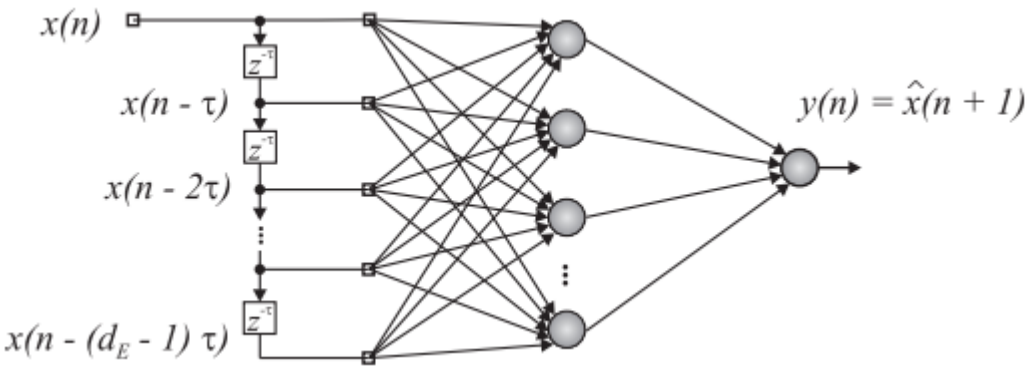


Figura 15 – Arquitetura genérica de rede FTDNN
Fonte: Menezes Junior, 2006, p. 64.

Em que $x(n)$ é um vetor que contém de elementos da série contados a partir do elemento atual $x(n)$, separados um do outro por t unidades de tempo, y é a saída do neurônio e x é a variável escalar predita baseada em atrasos no tempo.

Neste caso, o processo realizado claramente é a inclusão um modelo AR, visto anteriormente, dentro da especificação da rede neural, com a vantagem da identificação dos *lags* pela própria rede, por meio da consideração dos pesos, os quais são ajustados pela Regra de Delta.

Nesta dissertação é testada a MLP no formato FTDNN e com uma adaptação para incluir variáveis exógenas.

3.3.10. Rede Neural Recorrente

Conforme Silva et al. (2010), as redes neurais recorrentes (RNNs) ou realimentadas são redes em que as saídas dos neurônios são realimentadas como sinais de entrada para outros neurônios. De forma simplificada, as saídas tornam-se entradas, realimentando a rede. Tal fato faz com estas redes possam ser utilizadas em sistemas que apresentem variação ao longo do tempo, o que é característico, por exemplo, de séries temporais. Segundo Haykin (2001), a existência de realimentação

gera um grande impacto na rede tanto em termos de desempenho como de treinamento. Na Figura 16, verifica-se um exemplo de rede neural recorrente.

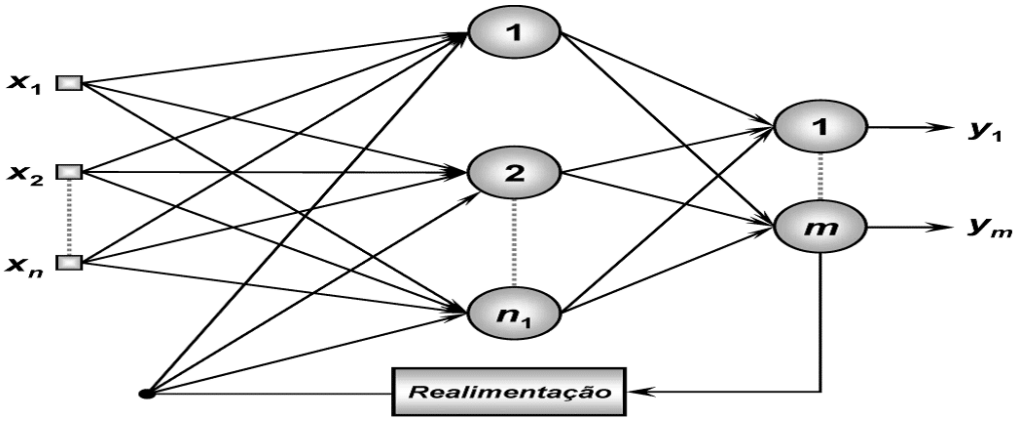


Figura 16 – Rede recorrente
Fonte: Silva et al., 2010, p. 49.

Braga et al. (2016) colocam que a computação convencional não lida de forma satisfatória com as tarefas de reconhecimento de padrão dinâmico, como é o caso da previsão de séries temporais, então é preciso a utilização de estruturas mais complexas, capazes de representar o tempo e apresentar memória. Desde já, é importante destacar que o algoritmo de treinamento também precisa ser ajustado e o mais utilizado é o *backpropagation through de time* (BPTT).

Segundo Olah (2015), as redes recorrentes apresentam *loops* que permitem que a informação persista ao longo da rede. Para o autor, uma rede neural recorrente pode ser vista como múltiplas cópias da mesma rede, cada uma passando uma mensagem para o seu sucessor. A seguir, na Figura 17, é mostrada a ilustração da rede recorrente fechada e da mesma desdobrada.

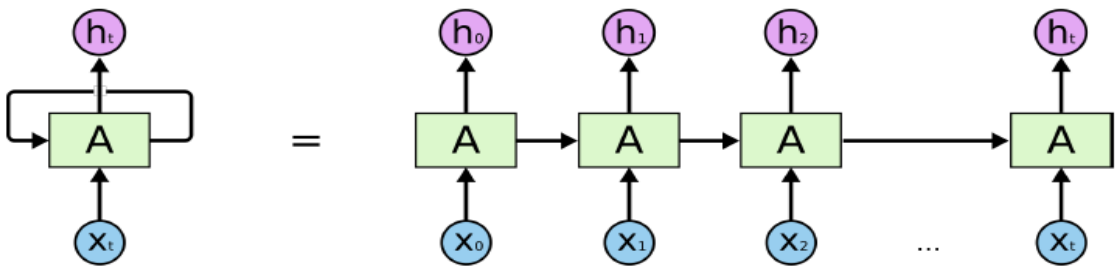


Figura 17 - Rede recorrente desdobrada
Fonte: Olah, 2015.

A partir da ilustração anterior, é possível constatar que as redes neurais recorrentes possuem a natureza semelhante a uma cadeia, o que está estritamente ligado com processos sequenciais. Diante desta característica, nos últimos anos, pesquisadores obtiveram bastante êxito aplicando RNNs em uma série de problemas reais, tais como: reconhecimento da fala, modelagem de idiomas, tradução, legendas e até mesmo previsões de séries temporais. Em todos estes problemas, a questão da sequência e da memória são relevantes. O tipo de RNN que obteve o maior êxito nestes casos foi a *Long Short Memory* (LSTM) pela capacidade de lidar melhor com as dependências de longo prazo e que será detalhada posteriormente.

Como já foi explicitada anteriormente, a principal vantagem das RNNs é que elas conseguem conectar informações anteriores à tarefa atual. Em outras palavras, significa que, para o entendimento do estágio atual, é necessária determinada quantidade de informação passada.

As RNNs devem ser vistas como um tipo especial de rede neural projetada para problemas sequenciais. Comparando com uma rede MLP, uma RNN pode ser pensada como uma MLP com *loops*. Mais especificamente, em uma dada camada, cada neurônio pode passar seu sinal para frente e para o lado. A RNN tem sua recomendação de uso quando os dados a serem modelados fazem parte de uma sequência e esta ordem é importante.

No caso de uma série temporal, o objetivo é prever o próximo valor de acordo com valores anteriores. Assim, a entrada para a RNN, em cada intervalo de tempo, é o valor atual, bem como um vetor estado com os dados do período anterior, que é a memória.

Olah (2015) apresenta um exemplo bem interessante considerando um modelo de linguagem que busca prever a próxima palavra com base nas anteriores. Se o objetivo é tentar prever a última palavra em “as nuvens estão no céu”, não é necessário de mais contexto – é bem óbvio que a próxima palavra será o céu. Nesses casos, onde o espaço entre as informações relevantes e o local que é necessário é pequeno, as RNNs podem aprender a usar a informação passada.

No entanto, há casos em que é necessário mais contexto. Por exemplo, tentar prever a última palavra no texto “Eu cresci na França ... Eu falo francês com fluência”. Informações recentes sugerem que a próxima palavra é provavelmente o nome de uma linguagem, mas se o objetivo for restringir qual idioma, é fundamental ter o contexto da França, mais adiante. Adicionalmente, é possível que a diferença entre as informações

relevantes e o ponto em que o contexto está seja muito grande, tendo em vista os três pontinhos da frase.

Assim, à medida que esse *gap* aumenta, as RNNs convencionais não conseguem aprender a conectar a informação. Este aspecto foi abordado por Hochreiterem (1991) na sua tese de doutorado. Mas as LSTM, que são um tipo especial de RNN conseguem lidar com este tipo de situação de dependência de longo prazo de forma adequada.

3.3.11. Rede *Long Short Term Memory* (LSTM)

A rede neural LSTM (*Long Short Term Memory* – rede de memória de longo prazo) é uma rede neural do tipo recorrente, sendo assim sofre retroalimentação. Foi desenvolvida, em 1997, por Sepp Hochreiter e Jürgen Schmidhuber e foi aprimorada, em 2000, pela equipe de Felix Gers, com a inclusão do portão de esquecimento. As redes de memória de longo prazo são um tipo especial de rede neural recorrente, as quais são capazes de aprender dependências de longo prazo, o que não é possível nas redes neurais MLP tradicionais. Uma desvantagem da rede LSTM é que ela precisa de um volume expressivo de dados e alto poder de computação para alcançar um bom desempenho. A partir de Greff, Srivastava, Koutnik, Steunebrink, Schmidhuber (2017), as redes LSTM se tornaram os modelos de estado da arte para vários problemas de aprendizagem de máquina.

Normalmente, este tipo de rede neural é utilizado para classificar, processar e prever séries temporais, em que não se sabe de antemão e de forma clara o efeito de memória da própria série. É uma arquitetura recomendada para redes onde existe a persistência de memória. Quando se traça um paralelo com o funcionamento do cérebro humano, pode-se dizer que não começamos a pensar do zero a cada momento, a nossa compreensão tem como base experiências anteriores e é este tipo de memória que a rede LSTM replica com várias limitações. Em relação à previsão de séries temporais, o objetivo é modelar séries por meio de atrasos temporais, onde não se sabe exatamente quantos atrasos são relevantes para modelar a série e a duração desta persistência.

Atualmente, este tipo de rede é muito aplicado para a compreensão de texto. A partir de 2016, as principais empresas de tecnologia, Google, Apple e Amazon,

passaram a utilizar este tipo de rede neural para uma série de produtos, sobretudo para os produtos de reconhecimento de voz, tais como Google Tradutor, Siri e Alexa. Outras aplicações da rede LSTM são: controle de robôs, previsão de séries temporais, composição de música etc.

A principal diferença de uma rede recorrente LSTM para os demais tipos de redes recorrentes é que ela não usa uma função de ativação dentro dos seus componentes recorrentes, o que faz com que o valor armazenado não seja modificado de forma interativa. Adicionalmente, outra vantagem do uso deste tipo de rede, conforme Gomes et al. (2017), é que ela evita problemas de propagação do erro ao longo de toda a estrutura, o que é muito comum nos outros tipos de redes neurais recorrentes. Conforme Chung et al. (2014), outros tipos de redes neurais recorrentes têm dificuldade de capturar as dependências de longo prazo, porque os gradientes tendem a desaparecer ou a explodir.

Conforme Hochreiter e Schmidhuber (1997), as redes recorrentes possuem a dificuldade de aprender a armazenar informações para longos intervalos de tempo. Hochreiter et al. (2001) colocam de maneira mais específica que RNNs padrões treinadas com o algoritmo de gradiente estocástico descendente têm dificuldade de treinar redes com mais de 10 fases de tempo de dependências, devido ao problema de dissipação do gradiente.

Isto ocorre, sobretudo, pela dificuldade de fazer o fluxo do erro fluir ao longo do gradiente. A rede LSTM tem como principal vantagem truncar o gradiente, o que permite que erro flua por meio do carrossel de erro constante, que é um dos elementos do bloco de memória. A nova estrutura do LSTM permite que este tipo de rede resolva tarefas mais complexas e de longa duração, que não tinham sido resolvidos por algoritmos de redes recorrentes anteriores, tendo como principal justificativa o fato dos gradientes não explodirem ou se dissiparem.

A descida do gradiente, que é a forma como se atinge o objetivo de minimizar o erro da função custo e se ajustam os pesos para que o valor previsto do modelo se aproxime do real é o algoritmo mais utilizado para treinamentos de redes neurais. No entanto, este algoritmo pode ter problemas de explosão ou dissipação. Os métodos baseados em gradientes aprendem o valor de um parâmetro ao entender como uma pequena alteração no valor do parâmetro afetará a saída da rede. Se uma alteração no valor do parâmetro provoca uma alteração muito pequena na saída da rede, a rede simplesmente não consegue aprender o parâmetro efetivamente, levando ao problema

de dissipação do gradiente. A escolha da função de ativação, juntamente com o otimizador, geralmente aqueles que são baseados em gradientes, podem levar a este problema.

R2RT (2016) fazem uma analogia com o mundo real e evidenciam o princípio que está por trás da rede LSTM. Assim, os autores argumentam que, para garantir a integridade das mensagens no mundo real, é preciso anotá-las. Escrever é uma mudança incremental que pode ser aditiva ou subtrativa ou que pode permanecer inalterada. No caso das LSTMs, tudo é escrito e se considera que nenhuma interferência de outras unidades de estado ou entradas externas carrega seu estado anterior para frente, ou seja, isso significa que qualquer mudança de estado é incremental, de modo que:

$$s_{t+1} = s_t + \Delta s_{t+1}$$

Segundo Hochreiter e Schmidhuber (1997) e Gers et al. (2000), a topologia de um neurônio de uma LSTM é baseada em portões, os quais controlam a quantidade de informação que entra em cada unidade, a quantidade de informação que é memorizada e a quantidade de informação que deve passar para as próximas unidades. Esta rede é treinada usando BPTT e supera o problema de dissipação do gradiente. Ao invés de neurônios, as redes LSTM possuem blocos de memória conectados através de camadas. Deve-se ressaltar que Gers et al. (2000) identificaram uma fraqueza das redes LSTM na versão original, que é a dificuldade de aprendizado deste tipo de redes para séries longas, que não são segmentadas em subsequências de treinamento. Para isto, os autores propõem a inclusão dos portões de esquecimento, que conseguem redefinir o conteúdo da célula de memória, eliminando o conteúdo que não é mais necessário. Na Figura 18, temos a ilustração da unidade básica da rede LSTM.

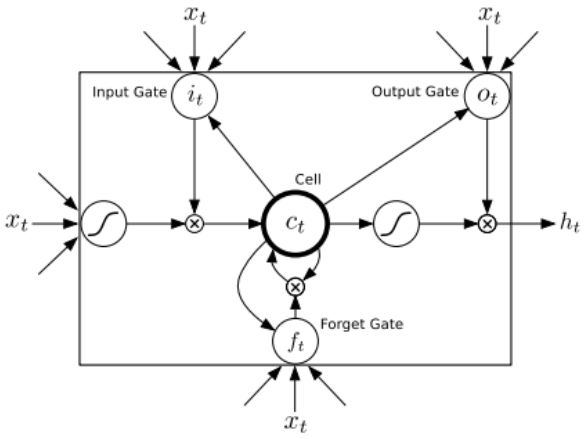


Figura 18 – Célula de memória LSTM
Fonte: Graves, 2014, p. 2.

Quando é realizada a comparação do bloco de memória com o neurônio padrão, um bloco de memória possui componentes, que o tornam mais inteligente do que um neurônio clássico, e possui um componente de memória para armazenar sequências recentes. Cada bloco contém portões que gerenciam o estado e a saída do bloco. Um bloco trabalha com uma sequência de entrada e cada portão dentro de um bloco usa as unidades de ativação para controlar se elas são acionadas ou não, fazendo com que a mudança de estado e a adição de informações fluam através do bloco condicional. Existem três tipos de portões dentro de uma unidade:

- **Portão de entrada:** determina os valores da entrada para atualizar o estado da memória;
- **Portão de esquecimento:** determina quais informações jogar fora do bloco;
- **Portão de saída:** determina o que produzir com base na entrada e na memória do bloco.

Cada neurônio opera como uma máquina de miniestado, sendo que os portões das unidades têm pesos que são aprendidos durante o procedimento de treinamento.

Todas as redes neurais recorrentes têm a forma de uma cadeia de módulos repetitivos de rede neural. Em RNNs padrão, este módulo de repetição terá uma estrutura muito simples, como uma única camada *tanh*.

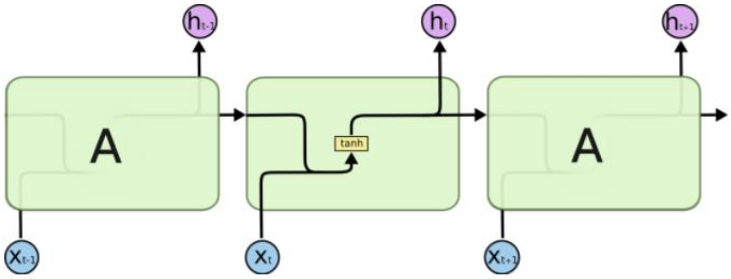


Figura 19 – Módulo de repetição de uma RNN com uma camada escondida
Fonte: Olah, 2015.

A seguir é evidenciado o funcionamento de uma LSTM, conforme Olah (2015). Assim, primeiramente é fundamental destacar que o módulo de repetição em uma RNN padrão ou também chamada de vanila contém uma única camada. As LSTMs também possuem essa estrutura de cadeia, mas o módulo de repetição tem uma estrutura diferente. Em vez de ter uma única camada de rede neural, existem quatro camadas, que interagem de forma complementar, com intuito de obter melhores resultados.

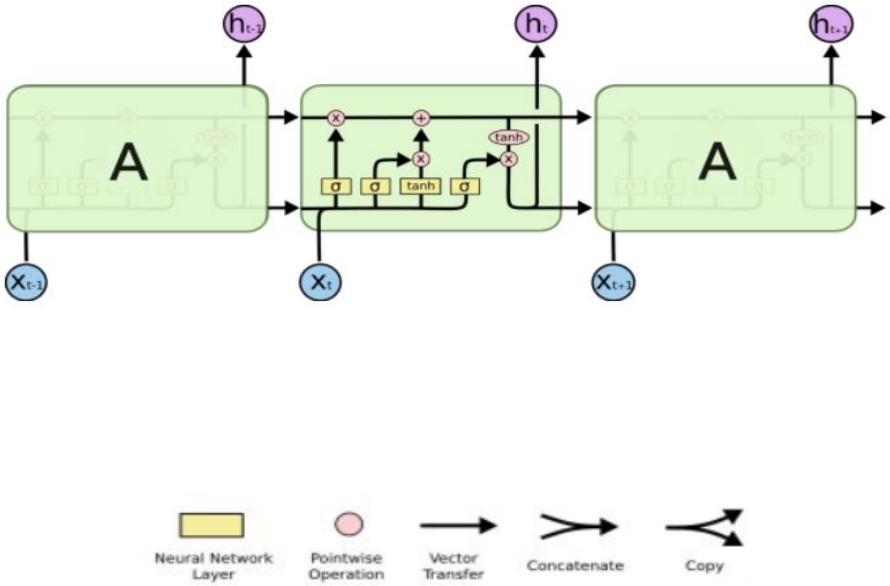


Figura 20 – Módulo de repetição de um LSTM com 4 camadas
Fonte: Olah, 2015.

Na Figura 20, cada linha carrega um vetor inteiro, desde a saída de um nó até as entradas de outros. Os círculos cor-de-rosa representam operações pontuais, como a adição de vetores, enquanto as caixas amarelas são camadas de rede neural. As linhas de fusão denotam concatenação, enquanto uma linha de bifurcação denota que o seu conteúdo vai ser copiado e as cópias irão para lugares distintos.

Uma parte muito importante das LSTMs é o estado da célula, a qual é a linha horizontal que passa pela parte superior do diagrama. O estado celular funciona como uma correia transportadora. Ele corre diretamente pela cadeia inteira, com apenas algumas pequenas interações lineares. A informação apenas passa ao longo dela e de maneira inalterada.

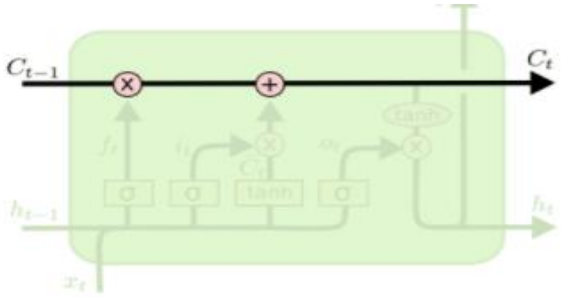
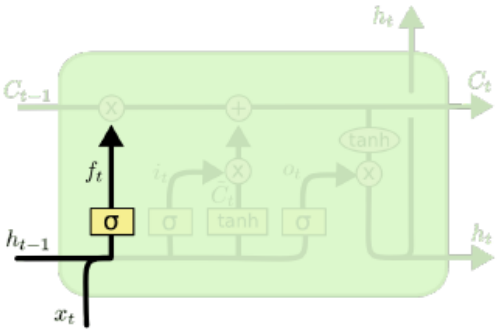


Figura 21 – Ideia principal LSTM
Fonte: Olah, 2015.

O LSTM tem a capacidade de remover ou adicionar informações ao estado celular por meio dos portões, que controlam o fluxo de informação ao longo da rede. Eles são compostos por uma camada de rede neural sigmoide e uma operação de multiplicação de pontos. A camada sigmoide produz números entre zero e um, descrevendo quanto de cada componente deve ser deixado passar. Um valor igual a zero significa “não deixar passar nada”, enquanto um valor de um significa “deixar passar tudo”. Um LSTM tem três desses portões para proteger e controlar o estado celular.

O portão de esquecimento tem a função de decidir quais as informações que serão afastadas do estado da célula. Esta camada recebe h_{t-1} e X_t e emite um output entre 0 e 1 para cada número na célula de estado C_{t-1} . O valor um representa “manter completamente isso” enquanto um valor zero representa “esquecer”.

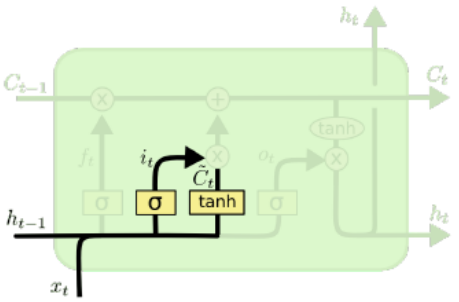
A seguir, será apresentado um modelo de linguagem tentando prever a próxima palavra com base em todas as anteriores. Neste problema, a célula de estado pode incluir o gênero do assunto atual, para que os pronomes corretos possam ser usados. Neste sentido, quando é apresentado um novo assunto, o gênero do assunto antigo deve ser esquecido.



$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f)$$

Figura 22 – Passo-a-passo LSTM – 1º passo
Fonte: Olah, 2015.

O segundo passo consiste em escolher quais novas informações serão armazenadas na célula estado, o que é subdividido em duas tarefas: (1) atualização das informações de entrada pelo portão de entrada, e (2) por meio de uma função *tanh*, é criado um vetor de novos valores candidatos, C_t , que podem ser adicionados ao estado. No exemplo de modelo de linguagem, o objetivo era adicionar o gênero do novo assunto à célula de estado para substituir o antigo, que tinha que ser esquecido.

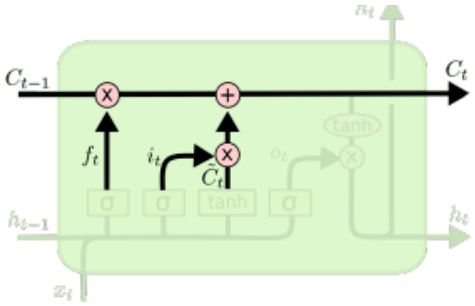


$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i)$$

$$\widetilde{C}_t = \tanh(W_C \times [h_{t-1}, x_t] + b_C)$$

Figura 23 – Passo-a-passo LSTM – 2º passo
Fonte: Olah, 2015.

A próxima etapa consiste em atualizar a célula de estado anterior, C_{t-1}, na nova célula de estado C_t. Nas fases anteriores foram escolhidas as informações relevantes, agora é necessário gravá-las. Para isto, multiplica-se o estado antigo por f_t, eliminando as informações que foram escolhidas para serem esquecidas. Então, se adiciona $i_t \times C_t$. Estes são os novos valores candidatos escolhidos para serem atualizados como novo estado. No caso do modelo de linguagem, é neste momento em que a informação sobre o gênero do sujeito antigo é eliminada e são adicionadas as novas informações, conforme as escolhas das etapas anteriores.

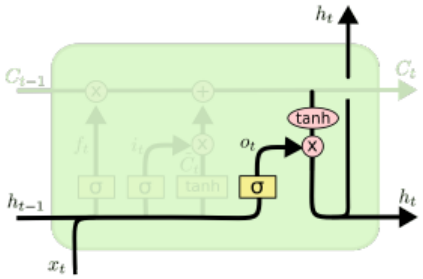


$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t}$$

Figura 24 – Passo-a-passo LSTM – 3º passo
Fonte: Olah, 2015.

Finalmente, chega-se na determinação do que será gerado como saída, que tem como base o valor da célula estado, mas será uma versão filtrada. Sendo assim, primeiramente, é executada uma função sigmoide, a qual decide quais partes da célula de estado serão usadas no *output*. Em seguida, coloca-se a célula de estado por meio da função de ativação *tanh* e realiza a multiplicação pela saída do portão sigmoide, garantindo que permaneçam somente as informações escolhidas.

No exemplo em questão, uma vez que somente foi visto o assunto, pode-se querer exibir informações relevantes para um verbo, caso este seja o tipo de palavra que venha depois. Especificamente, pode-se produzir se o sujeito é no singular ou plural, para que assim seja possível saber em que forma um verbo deve ser conjugado, mais uma vez, se for o que vem a seguir.



$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \times \tanh(C_t)$$

Figura 25 – Passo-a-passo LSTM – 4º passo
Fonte: Olah, 2015.

A estrutura de LSTM evidenciada na Figura 26 pode receber alguns ajustes, dependendo do objetivo de modelagem e da complexidade do problema:

- LSTMs *Stateless*: quando se redefine o estado inicial de uma LSTM a cada novo lote. Isto pode ser desejado quando não houver a continuidade

em determinadas partes de uma sequência de previsão. O padrão é a LSTM *stateful*, em que a previsão depende da ativação da célula de memória no período anterior;

- *Stacked* LSTM: o modelo LSTM original é composto por uma única camada escondida, pode-se fazer um ajuste para incluir múltiplas camadas, tendo como possíveis benefícios uma necessidade menor de neurônios nas camadas de treinamento e uma redução do tempo de treinamento.

Em termos de evolução dos modelos LSTM, deve-se destacar que, após a proposição inicial, em 1995, deste modedo, surgiram muitas variações do mesmo, tais como: Gers e Schmidhuber (2000) incluíram as conexões *peephole*, que permitem que os portões se comuniquem com o estado da célula, a GRU modelada por Cho (2014), que combina os portões de esquecimento e de entrada em um único portão de atualização, entre outras.

3.3.12. Redes neurais e séries temporais

Nesta etapa do trabalho, são evidenciados alguns aspectos das arquiteturas das redes neurais, considerando objetivo de realizar previsões para séries temporais.

Conforme Brownlee (2017), as redes neurais recorrentes têm a capacidade de conseguir aprender o contexto temporal das sequências das entradas, podendo realizar melhores previsões. Diferentemente do que ocorre nos modelos estatíscos tradicionais de séries temporais, no caso das redes recorrentes LSTM, não se faz mais necessário especificar as quantidades de entradas defasadas que serão necessárias para realizar a previsão, isto é, a própria rede recorrente consegue aprender a dependência temporal.

No caso da previsão de séries temporais, a principal dificuldade está na identificação da dependência temporal entre as observações. Brownlee (2017) coloca que a previsão de séries temporais tem sido feita por métodos lineares, como a modelagem ARIMA, porque são facilmente compreendidos e eficazes para a maioria dos problemas. No entanto, tais métodos apresentam as seguintes limitações: apenas trabalham com séries completas, em que não faltam dados; supõem a existência de

uma relação linear; a dependência temporal é fixa e tem que ser especificada; observa-se uma maior eficiência somente para previsões de curto prazo, ficando as previsões de longo prazo prejudicadas.

Quando se passa a utilizar as redes neurais, os ganhos para previsão de séries temporais são: a possibilidade de lidar com dados incompletos e não consideração da suposição de linearidade. Dorffner (1996) destaca que a capacidade das redes neurais conseguir aproximar funções arbitrárias não lineares promete a construção de aplicações mais poderosas. Em outras palavras, isto significa que as redes neurais conseguem suportar um número arbitrário tanto de entradas como de saída, permitindo entradas multivariadas e uma previsão de vários passos à frente.

Apesar desta vantagem, o autor destaca alguns problemas como a necessidade da grande quantidade de dados, a necessidade de uma avaliação mais cuidadosa e as limitações de aprendizado.

Conforme Brownlee (2017), as redes neurais diretas (*feedfoward*) apresentam uma ótima capacidade, mas ainda têm a limitação da necessidade de especificação a priori do número de defasagens temporais para o modelo. Na maioria das vezes, esta dependência não é claramente conhecida e sua descoberta é feita por meio de análises de tentativas e erros e através de uma janela fixa.

Diante disto, uma rede mais apropriada seria a rede recorrente de memória de curto e longo prazo (LSTM), pois elas adicionam o tratamento da ordem entre as observações da série, que está sendo modelada. A sequência passa a ser uma nova dimensão para a função que é aproximada. Simplificadamente, não está sendo feito mais o mapeamento de entrada para saídas de forma isolada. Na verdade, neste caso, a rede neural é capaz de aprender uma função para as entradas ao longo do tempo determinando uma saída. As LSTMs conseguem aprender o contexto necessário para fazer previsões em problemas de previsão de séries temporais, em vez de ter esse contexto pré-especificado e corrigido.

A principal promessa das redes recorrentes é que a dependência temporal dos dados de entradas pode ser aprendida, não sendo necessário especificar uma janela fixa. Adicionalmente, muitos autores colocam que uma boa prática para modelagem com redes neurais para séries temporais, como visto em Nelson et al. (1999) é o ajuste para eliminação da tendência e da sazonalidade. No caso das redes recorrentes, existe uma corrente de estudiosos que colocam que estas redes são capazes de aprender tais estruturas sistemáticas, não sendo necessária a sua eliminação. Esta suposição foi

considerada para os modelos que serão vistos no próximo capítulo, mas é um tratamento que pode ser feito, considerando uma abordagem comparativa, em trabalhos futuros. Assim, não foi contemplado um tratamento nos dados para eliminar os componentes de tendência e de sazonalidade.

Gers, Eck e Schmidhuber (2001) apresentaram um *paper* exatamente para verificar a capacidade das LSTMs conseguirem resolver limitações das redes neurais, que utilizam janelas de tempo de tamanho fixo. Os autores afirmam, inicialmente, que os problemas de previsão de séries temporais univariáveis são realmente mais simples do que os tipos de problemas, tradicionalmente, usados para demonstrar as capacidades das LSTMs. Neste *paper* foram avaliadas duas séries, considerando uma abordagem de autorregressão (AR), onde para modelar um passo à frente, deve-se considerar algumas defasagens passadas, que é a abordagem comum da estatística clássica. No caso da LSTM foi exposta uma observação de cada vez, diferentemente do caso da MLP, onde foi especificado um conjunto de variáveis defasadas como entradas.

As conclusões do estudo comparando a abordagem LSTM com a MLP de janela fixa foram:

- Os resultados para rede LSTM foram piores do que a rede MLP;
- O aumento do número de blocos de memória não melhorou significativamente os resultados;
- As LSTMs podem não ser adequadas às formulações de tipo AR da previsão de séries temporais, sobretudo quando as observações atrasadas são próximas do tempo previsto;
- A rede LSTM conseguiu aprender comportamentos cíclicos e a sazonalidade.

Portanto, no trabalho de Gers, Eck e Schmidhuber (2001), uma rede MLP baseada em janela de tempo superou a abordagem de LSTM-AR, considerando somente algumas entradas recentes. Neste caso, a grande vantagem da LSTM, que é aprender a lembrar eventos únicos por períodos muito longos e desconhecidos, não gerou previsões com maior acuracidade neste trabalho. Então, os resultados sugerem usar a rede recorrente LSTM somente em tarefas, em que as abordagens tradicionais baseadas em janela de tempo não consigam obter resultados satisfatórios.

Brownlee (2017) sugere que a utilização de modelos com mais camadas e com um treinamento por um número maior de épocas poderia fazer com que o modelo LSTM superasse o modelo MLP de janela fixa, visto anteriormente.

3.3.13. Tipos de treinamento

O processo de treinamento é o responsável pela realização do ajuste de pesos da rede neural, o qual pode ser supervisionado ou não supervisionado. Segundo Haykin (2001), o supervisionado ocorre por meio do auxílio de um professor, que nada mais é do que a presença do conjunto de dados de saída, não somente os dados de entrada, para que a rede consiga se ajustar ao padrão desejado. Sendo assim, a função do algoritmo de treinamento supervisionado é realizar os ajustes internos de peso da rede para que ela consiga gerar dados de saída o mais próximo possível da saída esperada. Para isto, calcula-se uma função erro, que deve ser minimizada pelo processo de treinamento.

Braga et al. (2016) colocam que o exemplo mais típico de aprendizado supervisionado é o aprendizado por correção de erros, onde o objetivo é minimizar o erro da resposta do modelo em relação à saída real. Outro tipo de aprendizado supervisionado é o aprendizado por reforço.

Já no caso do aprendizado não supervisionado, não temos a presença dos dados de saída para realizar os ajustes necessários. Neste sentido, segundo Haykin (2001), a rede tem que desenvolver habilidades internas para codificar as características de entrada e criar automaticamente classes, sendo o critério de vizinhança determinante para tal processo. Exemplos deste tipo de aprendizado são: aprendizado Hebbiano e por competição.

Neste trabalho, o foco será a aprendizagem supervisionada por correção de erros. Na Tabela 3, temos um quadro resumo dos tipos de regras de aprendizagem:

Tabela 3 - Tipos de regra de aprendizagem

Tipos de Regra de Aprendizagem	Descrição
Aprendizagem por Correção de Erro (Regra Delta)	Este tipo de regra é utilizada no treinamento supervisionado, onde o ajuste dos pesos se dá por meio do erro, o qual é obtido pelo cálculo da diferença ente o valor de saída da rede e o valor real em um ciclo de treinamento. O erro vai reduzindo gradualmente conforme a evolução das épocas.
Aprendizagem Hebbiana	É a regra baseada no postulado na aprendizagem de Hebb, que diz que quando dois neurônios em ambos o lados de uma sinapse são ativados sincronamente e ao mesmo tempo, logo a força desta sinapse é seletivamente aumentada. Este processo de treinamento é feito localmente, onde o ajuste do peso das conexões é realizado com base nas atividades dos neurônios
Aprendizagem de Boltzmann	É um método de aprendizagem estocástica, onde os neurônios são estocáticos, podendo apresentar dois estados possíveis: ligado (+1) e desligado (-1) e são divididos em dois grupos funcionais: presos e livres, os quais são responsáveis pela interação com o ambiente e pela explicação das restrições advindas dos padrões de entrada do ambiente. Neste tipo de regra de aprendizagem, os neurônios possuem conexões bidirecionais.
Aprendizagem Competitiva	Neste modelo de aprendizagem, os neurônios são obrigados a competir entre si e apenas um será ativo, em uma determinada iteração, o vencedor, isto é, o que apresentar maior similaridade em o padrão de entrada. Assim, todos os pesos dos neurônios próximos ao neurônio vencedor terão seus valores ajustados.

Fonte: elaboração própria.

3.3.14. Algoritmos de treinamento

Nesta parte do trabalho são evidenciados os algoritmos de treinamento para redes neurais. O processo de treinamento é como a rede neural aprende os pesos de forma que melhor represente os dados do modelo. Neste sentido, o principal objetivo é que a rede faça previsões que se aproximem o máximo possível dos valores reais. Será detalhado o algoritmo de *backpropagation* para redes não recorrentes e o algoritmo *backpropagation through the time* para as redes recorrentes.

Antes de realizar a especificação exatamente do funcionamento de cada um destes algoritmos, é importante destacar a importância do Adeline, idealizado Widrow e Hoff, em 1960, e o seu algoritmo de treinamento: a Regra de Delta, que é a base para o *backpropagation*. A Regra de Delta é também conhecida como algoritmo LMS (*last mean square*) ou método do gradiente descendente.

Na Regra de Delta utiliza-se a função de soma dos erros quadrados, para verificar a diferença entre o valor previsto pelo modelo e o valor real:

$$E = \frac{1}{2} \sum \mu \sum j \left[y_j^{\mu} - \hat{y}_j^{\mu} \right]^2$$

O valor previsto pela rede neural depende dos pesos e é definido por:

$$\hat{y}_j^{\mu} = f\left(\sum_i w_{ij}\, x_i^{\mu}\right)$$

Assim temos que:

$$\mathrm{E} = \frac{1}{2} \sum \mu \sum j \left[y_j^{\mu} - f\left(\sum_i w_{ij}\, x_i^{\mu}\right)\right]^2$$

O objetivo aqui é atingir o menor valor em termos de erro previsão por meio dos pesos. Para isto, é utilizado o método iterativo de descida do gradiente. Na matemática, o vetor gradiente é um vetor que indica o sentido e a direção na qual, por deslocamento a partir do ponto especificado, é possível atingir o maior incremento possível no valor de uma grandeza. O gradiente é a derivada generalizada para funções com mais de uma variável. Em cada etapa do processo iterativo é feito o cálculo do erro e do gradiente e estes valores são utilizados para determinar quanto deve ser a variação dos pesos para a realização do ajuste necessário.

De forma mais detalhada, a descida do gradiente funciona tendo como primeiro passo o cálculo da inclinação do plano no ponto atual, que inclui o cálculo das derivadas parciais da perda em relação a todos os parâmetros, que é o gradiente em si. Em seguida, utiliza-se o gradiente para modificar os pesos, de modo que a próxima passagem direta através da rede mova os pesos para valores que levem para a saída mais baixa na superfície. Em suma, o método de descida do gradiente é a mudança de peso sináptico proporcional à sua contribuição para o erro calculado. Na Figura 26, é evidenciada uma imagem que representa a descida do gradiente.

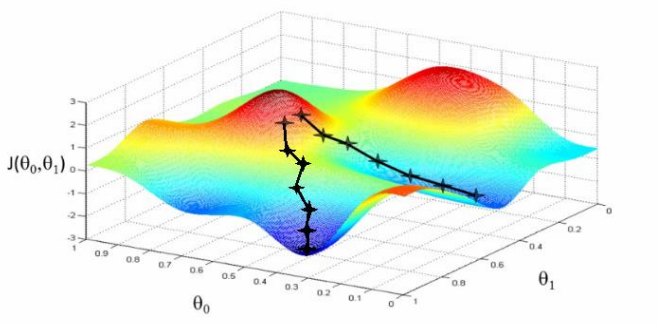


Figura 26 – Ilustração gradiente
Fonte: Data Science Academy.

Deve-se destacar que este processo pode levar a um erro extremamente baixo, que não necessariamente será o menor. Neste caso, chega-se em um mínimo local. Mas existem alguns procedimentos para minimizar este tipo de problema.

Backpropagation

Segundo Braga et al. (2016), o algoritmo de redes MLP mais utilizado é o *backpropagation*, que utiliza o mecanismo de correção de erros para ajuste dos pesos da rede. Este treinamento é composto por duas fases: a propagação para frente (*foward)* e a retropropagação (*backward)*. Na fase de propagação, a partir de um dado de entrada, o sinal é propagado ao longo da rede até gerar uma determinada saída e o seu respectivo erro. Já na fase retro propagação, com base na saída desejada e na obtida ocorre a atualização dos pesos. Neste caso, o erro é corrigido em cada camada, ajustando-se os pesos no sentido inverso. Na Figura 27, temos uma imagem que apresenta o processo de *backpropagation*:

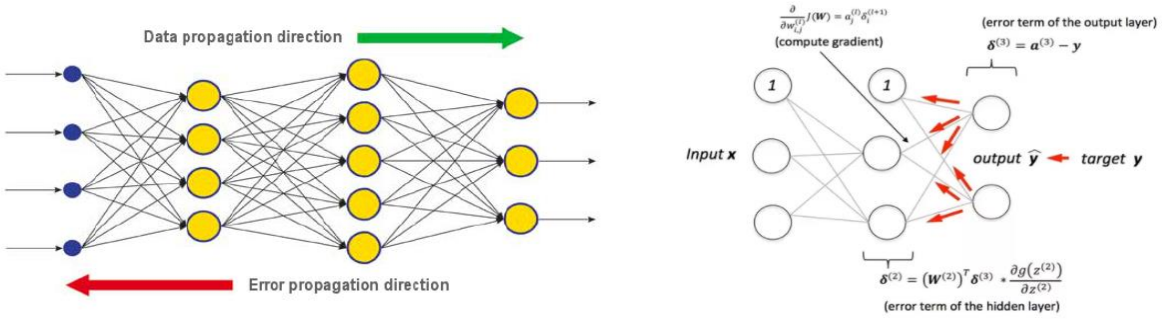


Figura 27 – Processo *backpropagation*
Fonte: Data Science Academy.

Segundo Vellasco (2007), o algoritmo de retropropagação pode ser implementado em cinco fases:

- Apresentação do padrão, que é a saída real;
- Propagação do sinal para frente e realização do cálculo da saída para todos os neurônios;
- Cálculo dos deltas para todos os neurônios de saída;

- Retropropagação do sinal de erro, o qual é distinto do sinal funcional, pois parte do neurônio de saída para realização do cálculo dos deltas das camadas ocultas;
- Atualização dos pesos da rede e apresentação do novo padrão para que o processo recomece.

Como já foi colocado anteriormente, as redes que utilizam *backpropagation* trabalham com uma variação da Regra de Delta apropriada para redes multicamadas: a Regra de Delta Generalizada. A regra de Delta padrão nada mais é do que a utilização do gradiente sobre uma função de minimização do erro quadrático. Por outro lado, a Regra de Delta Generalizada é apenas uma adaptação para que seja possível aplicar este processo em redes multicamadas.

O gradiente pode ser definido como a derivada da função de erro no valor atual do peso. E a função de erro mede a diferença da saída da rede neural da saída esperada. Cada peso, que é a conexão entre dois neurônios, tem um gradiente que é a inclinação da função de erro. Logo, o cálculo do gradiente da função de erro permite que o método de treinamento determine se deve aumentar ou diminuir o peso. Consequentemente, este ajuste busca diminuir o erro da rede neural.

A seguir, temos as equações que evidenciam a Regra de Delta e o processo de *backpropagation* que tem como objetivo calcular o valor de mudança do peso para cada peso da RN.

Fator de variação dos pesos:

$$\Delta\omega_{ij} = -\eta \frac{\partial E}{\partial \omega_{ij}}$$

Em que:

$\Delta\omega_{ij}$: variação do peso do neurônio *j* da conexão *i*;

η; taxa de aprendizado;

E: somatório do erro quadrático médio total.

Generalização do erro:

$$E = \frac{1}{2}\sum_{i=1}^{k}(t_i - s_i)^2$$

Em que:

t_i: saída real;

s_i: saída gerada.

Assim, temos pela regra da cadeia que:

$$\Delta\omega_{ij} = -\eta \frac{\partial E}{\partial \omega_{ij}} = -\eta \frac{\partial E}{\partial net_j} . \frac{\partial net_j}{\partial \omega_{ij}}$$

Em que:

$$net_j = \sum s_i . \omega_{ij} + \theta_j \qquad \frac{\partial net_j}{\partial \omega_{ij}} = S_i$$

Considerando:

$$e_j = -\frac{\partial E}{\partial net_j}$$

Temos que:

$$\Delta\omega_{ji} = \eta . s_i . e_j$$

E o ajuste do peso será de:

$$\omega_{ij}^{t+1} = \omega_{ij}^{t} + \Delta\omega_{ij}^{t+1} = \omega_{ij}^{t} + \eta . s_i . e_j$$

No caso de múltiplas camadas, este processo ocorre desde a saída, passando pelas camadas escondidas e chegando ao ajuste dos pesos das entradas.

Deve-se destacar que, muitas vezes, o tempo de aprendizado pode ser muito alto com a utilização do *backpropagation* e, além disso, pode-se chegar a um mínimo local e não global da função. A inclusão da taxa de aprendizado e do termo *momentum* podem ajudar a solucionar estes possíveis problemas.

A taxa de aprendizado (η) é o principal parâmetro que controla a atualização dos pesos e está compreendido no intervalo [0,1]. É o parâmetro que determina com que velocidade a rede neural aprenderá. Caso o valor deste parâmetro seja muito alto, pode-se passar do ponto onde ocorre a minimização do erro. Por outro lado, uma taxa de aprendizado muito pequena leva a uma trajetória suave e de pequenas mudanças nos pesos a cada iteração, alongando o tempo de treinamento da rede. Segundo Haykin (2001), a taxa de aprendizado é uma constante positiva que desempenha, na prática, um papel importante para o desempenho da aprendizagem por correção de erro e influencia de forma crítica a convergência do processo iterativo de treinamento.

Já o termo *momentum* (α), segundo Braga et al. (2016), é um parâmetro que pondera o quanto as matrizes sinápticas foram alteradas entre duas interações anteriores e sucessivas. Em outras palavras, é um termo que tem como objetivo guardar o efeito das mudanças passadas dos pesos na direção do movimento atual dos pesos, com objetivo de assegurar a convergência. Este recurso permite um aumento da taxa de aprendizado sem gerar oscilações durante a execução do algoritmo de *backpropagation*. De forma simplificada, o *momentum* funciona como impulso, adicionando uma força na direção em que um peso estava se movendo.

Outro aspecto importante dentro do processo *backpropagation* é a forma como ocorre a correção dos pesos em si, que pode ser de quatro formas: alteração padrão (on-line), por lote (*batch*), por *batch size* ou *mini-bath training*. No ajuste padrão, a alteração dos pesos é realizada sempre que um novo par de entrada e saída é apresentado à rede, gerando um erro que provoca uma correção individual na sequência. Já na alteração por lote, todos os pares do conjunto de dados de treinamento são apresentados à rede, o que gera um erro para todos o lote e é, somente, a partir deste que ocorre a atualização dos pesos. O ajuste deixa de ser individual e passa a ser feito por uma média. No caso do *batch size*, a atualização dos pesos ocorre na soma de algum tamanho do lote de elementos do conjunto de treinamento, o qual deve ser previamente definido. E, no caso do *mini batch training*, em que a atualização ocorre conforme o tamanho do lote, mas que essencialmente é muito pequeno.

A última definição a ser explicitada é a definição de época. Uma rede neural tem seu aprendizado fundamentado na apresentação dos dados à rede. Sendo assim, cada vez que é apresentado todo o conjunto de dados à rede, é realizada uma época de treinamento. Ao iniciar um processo de repetição de apresentação dos dados,

começamos outra época. O número de épocas é um dos hiperparâmetros, que devem ser especificados para o processo de treinamento de uma rede neural.

Backpropagation through the time **(BPTT)**

Segundo Braga et al. (2016) existem duas maneiras para treinar uma rede recorrente e que não envolvem o uso de aproximação no cálculo dos gradientes, e elas são: o *backropagatiom through time* (BPTT) e as redes recorrentes de tempo real (RTRL). Neste trabalho, o foco é o BPTT, que é uma adaptação do tradicional *backpropagation*. Os autores colocam que este método apresenta como principal desvantagem o fato que nenhum aprendizado é feito de forma a alcançar toda a sequência de fases de tempo, mesmo assim, é a técnica mais amplamente utilizada para redes neurais recorrentes.

No caso de uma MLP padrão, utiliza-se o *backpropagation*, onde o erro da rede é retropropagado e usado para atualizar os pesos da rede. Não é possível utilizar o algoritmo desta forma em uma RNN, pois o gradiente pode explodir ou se degradar. A alternativa é o BPTT, algoritmo em que, dada a inclusão da sequência de entrada, os gradientes são calculados e, posteriormente, alterados para manter a restrição de igualdade da matriz de pesos em cada camada oculta.

Conforme Haykin (2001), o BPTT considera que a operação temporal de uma rede recorrente pode ser desdobrada em uma rede alimentada à frente, onde sua arquitetura é acrescida de uma camada a cada período adicionado.

As etapas de treinamento do BPTT são:

1. Apresentar a sequência de etapas de tempo em termos de pares de entradas e saídas;
2. Fazer o desdobramento da rede;
3. Fazer a retropropagação do erro e atualizar os pesos;
4. Repetir até a convergência.

O algoritmo BPTT demanda, proporcionalmente, mais capacidade computacional à medida que aumenta o número de etapas de tempo. Se as sequências de entrada são compostas por muitas etapas de tempo, será necessário um número elevado de derivadas para que ocorra atualização dos pesos. Tal fato pode trazer à tona

os problemas de explosão e/ou dissipação do gradiente. Adicionalmente, a aprendizagem se torna extremamente lenta. Algumas alternativas para este problema são: o BPTT por época e o BPTT truncado.

3.3.15. Otimizadores de algoritmos

Depois de especificado os algoritmos de treinamentos que serão utilizados neste trabalho, é importante apresentar os otimizadores, os quais podem ser incluídos nos algoritmos de treinamentos, já que a escolha do mesmo impacta diretamente no tempo de treinamento da rede neural. Alguns dos pontos negativos do gradiente descendente são: ele é extremamente sensível à taxa de aprendizado e ao *momentum*; a taxa de aprendizado deve ser extremamente pequena para que a RN seja treinada de forma precisa, o *momentum* deve ser grande o suficiente para superar os mínimos locais, mas pequeno para não promover oscilações no processo de treinamento; é mantida uma única taxa de aprendizado e/ou *momentum* ao longo do treinamento, o que não é interessante, e a mesma vale para todos os pesos.

Além do gradiente descendente e do gradiente descendente estocástico, temos os seguintes otimizadores, que são métodos adaptativos de taxa de aprendizado: Adadelt, Adagrad, Adam, Nesterovs, Rmsprop, Conjugate gradient, Hessian Free, entre outros.

Ruder (2016) apresenta uma análise dos principais otimizadores e de seus desempenhos. O autor coloca que, em termos de escolha, se os dados de entrada forem escassos serão obtidos resultados melhores através da utilização de um dos métodos adaptativos de taxa de aprendizado, que inclusive é o caso deste trabalho.

Em relação ao SGD, o autor coloca que este otimizador consegue encontrar um mínimo, mas pode demorar significativamente mais do que com outros otimizadores, é mais dependente de uma programação robusta de inicialização e pode ficar preso nos mínimos locais. Portanto, se o objetivo é alcançar uma convergência rápida no treinamento de uma rede neural profunda ou complexa, mais uma vez, os métodos de taxa de aprendizado adaptativa devem ser escolhidos.

Com base nos estudos feitos pelo autor, os otimizadores Adagrad, Adadelta, RMSprop e Adam são os mais adequados e fornecem a melhor convergência, conforme a Figura 28.

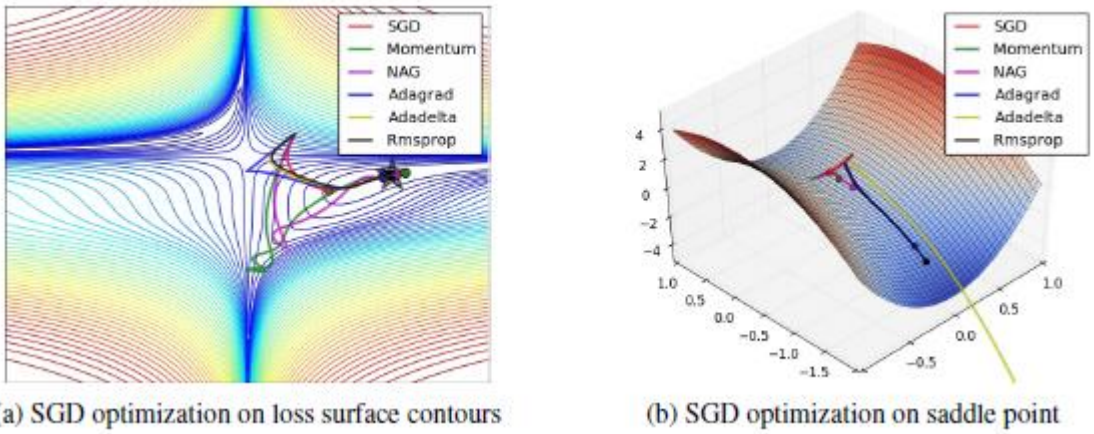


Figura 28 – Desempenho otimizadores
Fonte: Ruder, 2016, p. 10.

Neste trabalho serão utilizados o *Stochastic Gradient Descendent* (SGD) pela sua ampla utilização, apesar do seu desempenho inferior, o Adam e o RMSprop, que são métodos de taxa de aprendizado adaptativa. Importante destacar que tanto o Adam como o RMSProp são extensões do SGD. Abaixo, segue o detalhamento de cada um deles.

Gradiente descendente estocástico (SGD)

O SGD é o mais popular dos otimizadores de *backpropagation*. É uma versão do gradiente descendente (GD), onde são utilizadas amostras aleatórias do conjunto de treinamento. A descida estocástica do gradiente seleciona um elemento do conjunto de treinamento aleatoriamente e, posteriormente, calcula o gradiente e realiza a atualização de peso. Este processo continua até que o erro atinja um nível aceitável. Em comparação ao gradiente descendente, pode-se dizer que o processo de treinamento e de convergência é mais rápido, mais eficiente computacionalmente e diminui a superexposição, já que somente parte do conjunto de dados de treinamento é utilizada.

Neste trabalho, é usada a opção de aplicação de Nesterov, que é uma característica que pode ser incluída no SGD, visando reduzir o risco de escolha de um mini lote ruim no processo de treinamento.

RMSProp

O método RMSProp é um método de adaptação da taxa de aprendizado criado por Geoffrey Hinton.

Este otimizador mantém as taxas de aprendizado por parâmetro, as quais são adaptadas com base na média das magnitudes recentes dos gradientes para o peso. Ao invés de ter apenas uma taxa global de aprendizado, passa-se a ter um vetor de taxas de aprendizado para cada parâmetro treinável. É atualizado iterativamente com uma média em execução de magnitudes de quadrados de gradientes anteriores. As mudanças nos pesos durante o treinamento agora não são puramente na direção do gradiente, mas sim na direção da divisão elementar do gradiente por este vetor, que está sendo mantido.

Desta forma, este algoritmo funciona bem em problemas com treinamento padrão (on-line) e não estacionários.

Adam

Adam foi apresentado por Kingma e Ba (2105). Adam é diferente da descida estocástica do gradiente clássica. O SGD mantém uma taxa de aprendizado única para todas as atualizações de peso e a taxa de aprendizado não muda durante o treinamento. Já no Adam uma taxa de aprendizado é considerada para cada peso da rede e adaptada separadamente à medida que a aprendizagem é implementada.

Este otimizador possui as seguintes vantagens: simplicidade na utilização, computacionalmente mais eficiente, baixa requisição de memória, adequado para problemas que são grandes em termos de dados e/ou parâmetros, adequado para

problemas não estacionários e adequado para problemas com gradientes ruidosos ou escassos.

Kingma e Ba (2105) descrevem Adam como um otimizador que alia as vantagens de duas outras extensões do SGD: o Adagrad, um otimizador que mantém uma taxa de aprendizado por parâmetro, melhorando o desempenho em problemas com gradientes escassos, e o RMSProp, que também mantém as taxas de aprendizado por parâmetro, as quais são adaptadas com base na média das magnitudes recentes dos gradientes para o peso. Além disso, ao invés de adaptar as taxas de aprendizado dos parâmetros com base no primeiro momento médio (a média) como faz o RMSProp, o Adam também faz uso da média dos segundos momentos dos gradientes (a variância).

Deve-se ressaltar que, tanto para o Adam como para o RMSprop, são estabelecidas taxas de aprendizado iniciais.

3.3.16. Regularização

Quando se fala de regularização, é importante explicitar dois conceitos relevantes: (1) a convergência, que é a capacidade da rede de aprender todos os padrões do conjunto de treinamento, o que é diretamente afetado pelo tamanho da rede, e (2) a generalização, que é a capacidade de um modelo responder de maneira adequada aos novos padrões apresentados, os quais compõem o conjunto de dados testes. Pode-se afirmar que a capacidade de generalização de um modelo é o aspecto mais relevante para qualquer técnica preditiva, incluindo as redes neurais.

Posto isto, chega-se à questão de regularização. A regularização é qualquer alteração feita no algoritmo de aprendizagem com o objetivo de aumentar a generalização e evitar o super ajustamento (*overfitting)*. Adicionalmente, a utilização de regularização permite reduzir o nível de esforço computacional e possíveis problemas de multicolinearidade. Verifica-se o *overfitting* no caso em que o modelo se adapta de forma perfeita aos dados utilizados em seu treinamento, porém, apresentando uma baixa taxa de acerto quando confrontado com os dados do conjunto de teste.

A regularização é uma das técnicas que pode impedir este problema. Existem várias formas de regularização, entre as mais utilizadas estão: L1, L2 e *dropout*.

A regularização L1 (Lasso) e L2 (Ridge) são algoritmos de regressão que funcionam incluindo uma penalidade de peso ao treinamento da rede neural, o que leva a rede neural a manter pesos com baixos valores e são adicionados como parte do algoritmo de *backpropagation*. O L1 vai forçar para que os pesos apresentem uma distribuição Laplace, já o L2 faz uma aproximação para que os pesos apresentem uma distribuição Gaussiana.

O L1 adiciona uma penalidade ao valor absoluto na magnitude do coeficiente com o objetivo de levar uma parte deles ao valor zero, o que faz com que sejam eliminados da rede, aumentando a esparsidade da mesma. Já a L2 executa uma penalidade equivalente ao quadrado do peso dos coeficientes.

O *dropout* é uma técnica de regularização criada por Geofrey Hinton, em 2012, sendo a mais utilizada atualmente, que desativa alguns dos neurônios com uma probabilidade *p*. A desativação de um neurônio significa mudar o valor de saída para 0. Os neurônios que sofreram *dropout* têm os parâmetros reajustados, multiplicados por *p* (que é a probabilidade). Pode-se afirmar que o *dropout* tem um efeito similar ao de se fazer uma média de todos os possíveis modelos da rede neural, que utilizam um subconjunto dos parâmetros disponíveis na camada afetada pela regularização.

Neste trabalho não será aplicada nenhuma técnica de regularização, o que pode ser contemplado como uma melhoria para trabalhos futuros.

3.3.17. Capacidade computacional e redes neurais

Cientistas de dados, tanto na indústria como na academia, vêm utilizando GPUs para a aprendizagem de máquinas a fim de fazerem melhorias inovadoras em uma variedade de aplicações, tais como: classificação de imagem, análise de vídeo, reconhecimento de fala e processamento de linguagem natural.

Apesar da aprendizagem por máquinas já existir há décadas, dois fatos recentes provocaram um uso generalizado e intensivo da aprendizagem por máquinas: a alta disponibilidade de grandes quantidades de dados para treinamento e a computação paralela presente nas GPUs (placas de vídeos), em que vários cálculos são realizados simultaneamente, operando sob o princípio de que grandes problemas geralmente podem ser divididos em problemas menores, que então são resolvidos em paralelo.

Segundo Jungsuk e Iane (2014), com GPUs o conteúdo pré-gravado de fala ou multimídia pode ser transcrito muito mais rapidamente. Comparando com a implementação da CPU, pode-se realizar um reconhecimento até 22 vezes mais rápido.

Conforme MSV (2017), as GPUs – as placas de vídeos que tinham forte apelo no mundo jogos – ganharam bastante destaque na área de aprendizado profundo, uma das áreas da aprendizagem de máquinas. Isto ocorreu, porque este tipo de modelagem depende de cálculos matemáticos e estatísticos sofisticados. As redes convolucionais e recorrentes são algumas das aplicações modernas da aprendizagem profunda e, em muitas funções, imitam o cérebro humano com uma estreita semelhança com a neurociência.

As CPUs tradicionais são projetadas para abordar os cálculos em ordem sequencial, o que significa que cada operação matemática terá que aguardar a conclusão anterior. Por outro lado, uma GPU apresenta um processador com milhares de núcleos capazes de realizar milhões de operações matemáticas em paralelo, de forma simultânea. Há uma relação estreita entre renderização gráfica e aprendizagem profunda. Ambos os cenários lidam com muitas operações de multiplicação de matriz por segundo. Para colocar as coisas em perspectiva, as últimas GPUs da Nvidia vêm com 3.584 núcleos, enquanto as CPUs do servidor de ponta da Intel podem ter no máximo 28 núcleos.

O crescimento da GPU trouxe a Nvidia para o centro das atenções. Enquanto o mercado de PC e de servidores tradicional testemunha um declínio, o mercado de GPU está em célere expansão, ainda mais considerando o cenário de *cloud computing* e a sua utilização por meio de serviço.

Em resposta ao forte crescimento do GPU, o Google lançou o chip TPU (*Tensor Processing Unit*) em 2016, e já está na segunda geração em 2017. O TPU foi projetado para o aprendizado de máquina adaptado para o TensorFlow. Em alguns casos o TPU já superou as GPUs em termos de velocidade. Além disso, o TPU é mais eficiente em termos energéticos.

Enfim, pelo que foi colocado acima, o desenvolvimento de *hardware* tem potencializado e impulsionado a área de aprendizagem de máquina e de aprendizagem profunda.

3.4. Avaliação das Técnicas de Previsão

Conforme Passari (2003), para se avaliar métodos de previsão existem diferentes critérios, tais como: acurácia, custo, habilidade do analista, características desejadas de previsão, características específicas do problema, facilidade do uso, requerimento dos dados, disponibilidade de *software*, velocidade, facilidade de interpretação, eficiência, viés e capacidade de incorporar a experiência do gestor.

Com base na revisão sistemática da literatura, a acurácia é o critério mais utilizado para fazer a avaliação dos métodos de previsão. Segundo Aczel (1993 apud Passari, 2003, p. 24), as principais medidas de desempenho são:

- ME: *Mean Error* (erro médio)

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(Y_t - \hat{Y}_t\right)$$

- MSE: *Mean squared error* (erro quadrático médio)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_t - \hat{Y}_t\right)^2$$

- MAE: *Mean absolute error* (erro absoluto médio)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_t - \hat{Y}_t\right|$$

- RMSE: *Root mean squared error* (raiz do erro quadrático médio)

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_t - \hat{Y}_t\right)^2\right]^{\frac{1}{2}}$$

- MPE: *Mean percentual error* (erro percentual médio)

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_t - \hat{Y}_t}{Y_t}\right]$$

- MAPE: *Mean absolute percentual error* (média dos erros percentuais absolutos)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|$$

Passari (2003) destaca que a escolha da medida de desempenho da acurácia deve ser feita com base nas características da série a ser prevista e no funcionamento de cada uma das métricas estipuladas anteriormente. O MSE é amplamente utilizado, mas afetado pela presença de *outliers*. Caso a série possua *outliers*, é recomendado utilizar o MAPE.

Como as séries desta dissertação possuem vários *outliers*, o que será visto no próximo capítulo, a partir da análise descritiva de cada série, será adotado o MAPE como medida de verificação da acuracidade.

4. Metodologia

Inicialmente, é fundamental enquadrar este trabalho dentro das classificações de metodologia de pesquisa. Conforme Gil (1994), podemos considerar que o método científico aplicado foi o método dedutivo, que é um método racionalista, o qual considera a razão como única maneira de se determinar o conhecimento verdadeiro. Neste sentido, é utilizado um raciocínio descendente, da análise mais geral para particular, até se determinar a conclusão.

Em termos de classificação de pesquisa, este trabalho apresenta vários tipos de classificações, onde cada uma delas atende um ponto de vista específico. Assim, ainda conforme Gil (1994), as classificações possíveis são: conforme a natureza, a abordagem, o objetivo e os procedimentos técnicos.

Sob o ponto de vista da natureza, é uma pesquisa aplicada, pois gera conhecimentos para aplicações práticas objetivando alcançar a solução do problema de previsão de vendas das empresas varejistas e, mais especificamente, do Lojão do Brás.

Já em relação à abordagem, é uma pesquisa quantitativa, pois considera que tudo é quantificável e foi aplicado um conjunto de métodos matemáticos e estatísticos para testar e fundamentar de forma quantitativa o relacionamento entre a variável resposta e as variáveis explicativas, com o intuito de se determinar o método de previsão de venda com melhor acuracidade.

Com relação ao objetivo, é uma pesquisa experimental, pois se determinou o método de previsão que é mais adequado e preciso para a previsão de vendas de cada uma das categorias de produtos de moda contempladas.

Por último, sobre o ponto de vista dos procedimentos técnicos, este trabalho pode ser considerado uma pesquisa tanto bibliográfica, pois foi elaborada, em sua grande parte, com base em material já publicado, como também uma pesquisa experimental, pois foram aplicadas diversas técnicas de previsão, visando configurá-las de forma a garantir o melhor desempenho em cada uma delas e, assim, se determinar o método com melhor acuracidade.

Após o enquadramento deste trabalho nas classificações de metodologia de pesquisa, devem-se explicitar os dados que serão utilizados, juntamente como o seu processamento e o detalhamento de como cada técnica de previsão de vendas será

utilizada, incluindo a especificação do *software* aplicado, já que cada modelo utilizou um *software* específico.

4.2. Definição do problema de previsão

Neste item está especificado o escopo do problema de previsão tratado na presente dissertação. Sendo assim, temos:

Setor:

A empresa, cujos dados são analisados, pertence ao setor do varejo de moda e comercializa uma série de produtos, passando por moda masculina, feminina e infantil. Além disso, a empresa também comercializa outros produtos, como moda lar, moda íntima, moda bebê, moda praia, acessórios e maquiagem, categorias que não são analisadas neste trabalho. Especificamente, a análise é restrita a algumas categorias de produtos de moda.

Variável a ser prevista:

A variável a ser prevista são as vendas semanais – em termos de quantidade e em unidades – de quatro categorias de produtos: camiseta de homem, camiseta de menino, blusa de mulher e blusa de menina. Estas categorias foram escolhidas devido a sua alta representativa, aproximadamente 35% das vendas totais do departamento de moda, dentre as 70 categorias existentes.

Horizonte de previsão:

Previsão da quantidade vendida em unidades para cada categoria de produto para a próxima semana. Todos os dados foram considerados na granularidade semanal, pois é nesta frequência que a empresa estudada utiliza e, coincidentemente, foi a forma verificada nos trabalhos de Thomassey, autor com vários artigos sobre o assunto abordado nesta dissertação.

Objetivo:

Conseguir obter previsões com o maior nível de acuracidade, tendo como critério o MAPE (média dos erros percentuais absolutos).

4.3. Definição dos métodos de previsão

Para que sejam realizadas as previsões, é fundamental estabelecer quais são os métodos utilizados neste trabalho, tendo em vista que o principal objetivo do mesmo é a realização da avaliação de modelos quantitativos de previsão para séries temporais por meio dos métodos estatísticos clássicos, do método atual utilizado pela empresa e dos métodos avançados. Na Figura 29, temos uma ilustração, que evidencia o conjunto de modelos construídos.

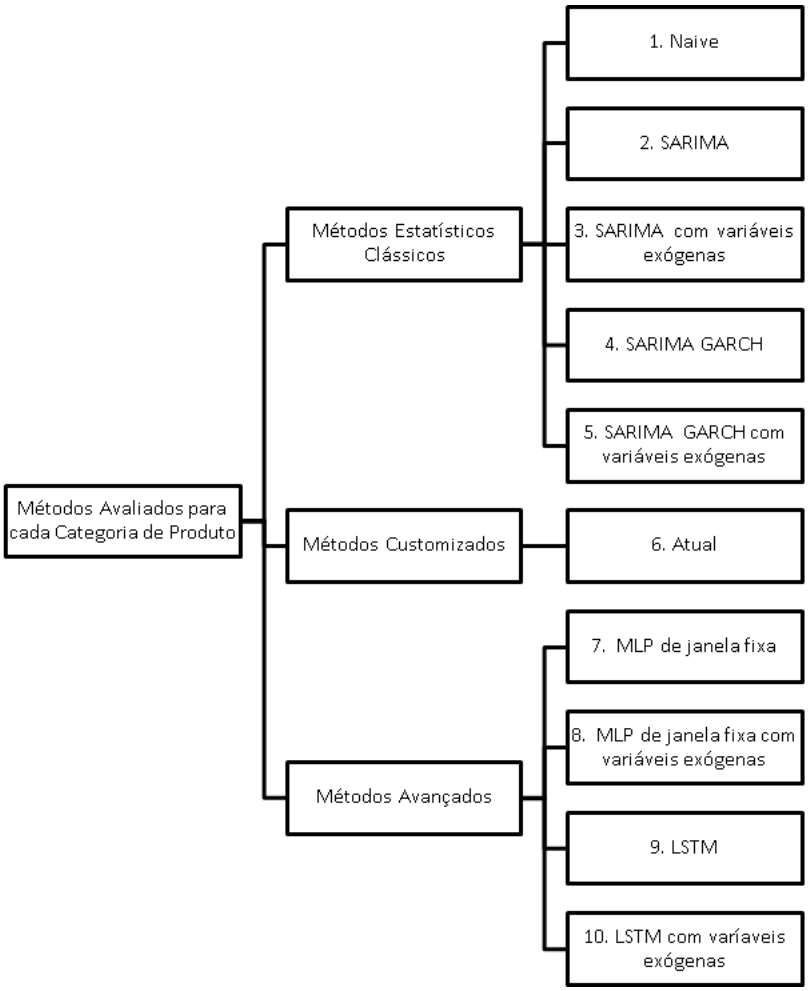


Figura 29 – Métodos de previsão utilizados
Fonte: elaboração própria.

Para todos os modelos, com exceção do Naive e do modelo atual, que é o modelo vigente utilizado pela empresa, temos uma variação do mesmo modelo com a inclusão das variáveis exógenas, já que o último objetivo do trabalho é exatamente a constatação do impacto das variáveis externas.

A partir da construção de todos estes métodos de previsão, torna-se possível alcançar os demais objetivos deste trabalho, os quais são: verificar qual o método mais adequado para cada série, incluindo a comparação com o método atual usado.

Deve-se ressaltar que, para cada um dos métodos escolhidos, foi feita mais de uma alternativa de modelo.

4.4. Base de dados

Para formulação deste trabalho são utilizadas tanto bases primárias como bases secundárias de dados. Os dados de natureza primária, capturados na sua forma original, são representados pelas quantidades e pelos preços médios, os quais foram corrigidos pelo Índice de Preços ao Consumidor (IPCA).

Já os dados de natureza secundária dizem respeito aos dados que foram coletados e manipulados por outras pessoas com outros fins e são representados por todas as variáveis exógenas do trabalho, tais como: temperaturas médias registradas, índice de precipitação médio, taxa de desemprego na região metropolitana de São Paulo, índice de confiança do consumidor, índice de volume de vendas no comércio varejista de moda de São Paulo, rendimento médio real efetivo das pessoas ocupadas, massa salarial ampliada disponível corrigida pelo IPCA e o saldo da carteira de crédito a pessoas físicas em relação ao PIB.

A escolha das varáveis exógenas tem como base dois trabalhos: o estudo da SBVC de 2016 e o trabalho de Moraes e Silva (2015). No estudo da SBVC de 2016, verifica-se a existência de quatro elementos determinantes para o crescimento e desenvolvimento do setor de varejo, os quais são: renda, emprego, crédito e confiança. No trabalho de Moraes e Silva (2015), os autores verificaram que as vendas no varejo são condicionadas por três fatores macroeconômicos, os quais são: taxa de juros, taxa de desemprego e renda média real. Com base na importância destes fatores foram selecionadas as estatísticas mais relevantes para cada um deles.

A seguir, é apresentada a Tabela 4 com o resumo destas variáveis com a sua fonte e a sua tipificação. A maior parte das variáveis foi extraída do banco de dados de séries temporais do Banco Central, o que não necessariamente significa que ele é o órgão responsável pela geração dos dados em si.

Tabela 4 – Variáveis exógenas

Código	Descrição	Unidade	Fonte	Periodicidade	Código Fonte
TMED	Temperatura média semanal	Graus Celsius	Somar Meteorologia	Diária	
PREC	Nível de precipitação médio semanal	Milímetros	Somar Meteorologia	Diária	
ICC	Índice de confiança do consumidor	Índice	Banco Central	Mensal	4393
TAXA_DESEMPREGO	Taxa de desempego da região metropolitana de SP - SEADE/PED	Participação%	IPEA DATA	Mensal	
REND_REAL	Rendimento médio real efetivo do trabalhador no Brasil	R$	Banco Central	Mensal	10790
MSAD_REAL	Massa Salarial Ampliada Disponível PNADC	R$	Banco Central	Mensal	22080
CRED_PF	Saldo de crédito disponível a pessoas físicas como percentual do PIB	Participação%	Banco Central	Mensal	20624
PC_VEST	Índice com base fixa em 2014 que mede o nível de atividade do comércio varejista de vestuário de SP	Índice	IBGE	Mensal	3418

Fonte: elaboração própria.

Além disso, foram incluídas também algumas variáveis *dummies* para controlar o efeito dos principais eventos do calendário varejista, responsáveis por gerar picos de vendas, tais como: dia das Mães, dia dos Namorados, dia dos Pais, dia das Crianças e Natal. Foi inserida, por fim, uma *dummy* para o mês de dezembro, já que este mês representa 25% das vendas totais da empresa.

Deve-se salientar o fato de parte das variáveis exógenas terem uma periodicidade mensal, que foi adaptada para semanal, permitindo assim a sua utilização nos modelos de predição. A adaptação foi feita repetindo o valor verificado em um determinado mês para as suas respectivas semanas, não foi realizado nenhum tratamento de gradação nos valores. É fundamental ter em mente que tal ponto pode impactar a relevância das variáveis exógenas para explicar as séries de produtos escolhidas.

As categorias de produtos analisadas possuem séries de vendas semanais, compreendendo as semanas de janeiro de 2007 a dezembro de 2016, e representam a quantidade de unidades vendidas para cada semana. Quando se utiliza um modelo preditivo, o foco não está nos dados passados, mas sim em utilizar esses dados para tentar criar um modelo, que possibilite de alguma forma prever o futuro.

Assim, a forma mais simples de avaliar este modelo é reservar parte dos dados para construir o modelo e a outra parte para validá-lo. Esta divisão é conhecida como *insample* (treino) e *outsample* (teste). Verifica-se, então, que será empregado o princípio da indução, onde são obtidas conclusões genéricas a partir de um conjunto particular de dados, que são os dados de treinamento. Depois é feita a avaliação do modelo no conjunto de dados teste.

Neste sentido, o período de treinamento compreenderá as semanas de janeiro de 2007 a dezembro de 2014. As alternativas de cada modelo com o melhor desempenho na fase de treinamento foram escolhidas e testadas nas semanas do período de janeiro de 2015 a dezembro de 2016, o período de teste. Importante ressaltar que, para a avaliação comparativa final entre os modelos, foi considerada a medida de acurácia MAPE na fase teste de cada um deles e para cada categoria de produto.

As variáveis exógenas que foram escolhidas são consideradas variáveis determinantes pelo setor para entendimento tanto da sazonalidade, por meio da temperatura e das chuvas, como do cenário econômico do país. A partir delas é possível compreender o nível de crédito, confiança, renda e emprego da economia.

Com exceção da temperatura e do nível de precipitação, as demais variáveis exógenas apresentam defasagens em sua mensuração. Tal procedimento foi adotado, pois o intuito é realizar um experimento em condições extremamente próximas ao real. Optou-se, então, por defasar as variáveis de acordo com a disponibilidade das mesmas. Especificamente, as seguintes variáveis exógenas foram defasadas em dois meses: índice de confiança do consumidor, índice de volume de vendas no comércio varejista de moda de São Paulo, rendimento médio real efetivo das pessoas ocupadas, massa salarial ampliada disponível corrigida pelo IPCA, taxa de desemprego na região metropolitana de São Paulo e o saldo da carteira de crédito a pessoas físicas em relação ao PIB%.

Abaixo, é evidenciado um detalhamento maior de cada uma destas variáveis.

Vendas semanais por categoria de produto

Como o principal objetivo é fornecer informações de previsões de vendas para os sistemas de compras da empresa Lojão do Brás, a análise recai sobre o histórico dos últimos 10 anos em semanas da quantidade vendida das principais categorias de produtos da empresa do departamento de moda, que são cerca de quatro categorias, as quais representam em torno de 35% das vendas totais do departamento de moda. Exemplificando, camiseta masculina adulta é uma categoria de produto. No total serão utilizadas 524 observações para cada categoria de produto.

Um aspecto importante que merece ser destacado é o grau de granularidade da previsão que se visa realizar, a qual será desenvolvida no nível de agrupamento de produto, dentro de cada categoria, e não no nível de *SKU*.

O conceito de categoria utilizado pela empresa é a junção do subdepartamento com o produto. Dentro de categoria temos um conjunto de itens e, por conseguinte, de SKUs, que seriam as cores e tamanhos de cada item. Na Figura 30, é apresentada a hierarquia de departamentos de produtos do Lojão do Brás:

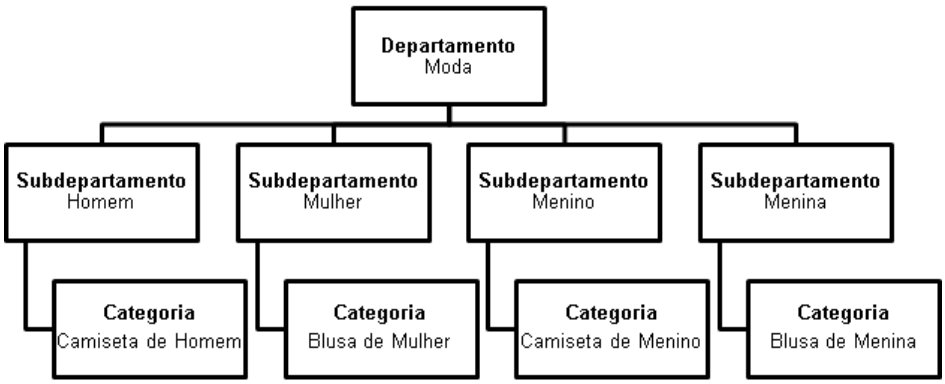


Figura 30 – Hierarquia de departamentos de produtos do Lojão do Brás
Fonte: elaboração própria.

Especificamente, as quatro categorias de produtos que são previstas nesta dissertação são: camiseta de homem, blusa de mulher, camiseta de menino e blusa de menina.

Preço médio categoria de produto

Considerando que o público-alvo da empresa Lojão do Brás é extremamente sensível a preço, a inclusão desta variável tem como objetivo capturar impactos nas

quantidades vendidas, a partir de alterações nos preços, o que pode ser efeito de promoções, liquidações, etc. Atualmente, a empresa possui o dado diário, assim é feita uma média semanal para inclusão desta informação no modelo. Para que a informação fosse comparável, foi feita a correção dos preços pelo IPCA, tendo como base o ano de 2012.

Temperatura e nível de precipitação médios registrados

O Lojão do Brás compra informação meteorológica junto a Somar Meteorologia e, assim, possui as condições registradas em termos de temperatura e nível de precipitação na região onde atua. Para os modelos que foram previstos, como os dados recebidos são diários, mais uma vez, foi feito o cálculo da média semanal tanto para temperatura como para o nível de precipitação. A temperatura afeta, sobretudo, as vendas de produtos de inverno. Já o nível de precipitação afeta o nível das vendas em geral, devido à empresa ter o foco em lojas de rua e não de *shopping*, como outros varejistas.

Calendário varejista (*dummies*)

Foi criada uma *dummy* para cada data relevante do calendário varejista, tendo em vista a relevância de controlar o efeito de cada uma destas datas, as quais afetam com diferentes intensidades a quantidade vendida: Dia das Mães, Dia dos Namorados, Dia dos Pais, Dia das Crianças e Natal.

Mês de dezembro

Além das *dummies* do calendário varejista, foi incluída uma *dummy* somente para o mês de dezembro, visto que este mês representa 25% das vendas totais anuais da empresa.

Índice de confiança do consumidor

É utilizado o índice da Federação do Comércio de Bens, Serviços e Turismo do Estado de São Paulo (FecomercioSP) de confiança do consumidor com periodicidade mensal. Sabe-se que quando o consumidor está satisfeito e otimista em relação ao futuro, ele tende a gastar mais e é este aspecto que se almeja incluir no modelo.

Crédito disponível a pessoas físicas como participação do PIB

Para determinação do consumo, o nível de crédito disponível na economia é um fator fundamental. Além disso, conforme ocorre um aumento do valor agregado de um bem, ocorre um aumento de necessidade de crédito por parte dos consumidores para conseguir adquiri-lo. Este indicador é divulgado mensalmente.

Massa salarial ampliada disponível real

Para o varejo, é importante verificar o tamanho da massa salarial disponível na economia, pois é o indicador que reflete, de forma conjunta, o aumento do número de empregos e o nível de empregos, evidenciando então a soma de todos os salários pagos. A informação disponibilizada pelo Banco Central foi corrigida pelo IPCA para ser comparável ao longo dos períodos analisados. Esta série possui periodicidade de divulgação mensal.

Rendimento médio real efetivo do trabalhador

O rendimento médio efetivo do trabalhador é uma informação pesquisada mensalmente pelo IBGE, através da Pesquisa Nacional por Amostra de Domicílios (PNAD), e é fundamental a consideração da sua evolução, pois um aumento de renda pode levar a um aumento de consumo, o que também precisa ser controlado.

Taxa de desemprego na região metropolitana de SP

Um maior nível de desemprego, certamente, afeta o nível de consumo, sendo assim levamos em conta a taxa de desocupação medida pela Pesquisa de Emprego e Desemprego (PED), da Fundação Sistema Estadual de Análise de Dados (SEADE) de São Paulo, com periodicidade de divulgação mensal.

Volume de vendas da pesquisa mensal do comércio para categoria de moda em SP

A pesquisa do comércio feita pelo IBGE funciona como um termômetro de como está evoluindo a economia para este setor. Neste sentido, o objetivo da inclusão desta variável é exatamente conseguir controlar este efeito, já que melhorias neste indicador podem indicar aumentos nas quantidades comercializadas. O índice de volume de vendas para comércio varejista de tecidos, moda e calçados no estado de São Paulo foi o escolhido, pois é o tipo de atividade mais diretamente relacionado ao tipo de atividade desenvolvida pela empresa. Este índice compara o nível de volume de venda do mês com a média mensal obtida no ano de 2014.

Como já foi colocado anteriormente, foi necessária a inclusão de parte das variáveis exógenas com defasagens, devido a sua falta de disponibilidade no momento de previsão das séries temporais. Também foram incluídas defasagens da própria variável predita para a construção de todos os modelos.

Nos modelos estatísticos clássicos, com exceção do modelo Naive, faz parte de metodologia Box-Jenkins a identificação do número de defasagens necessárias. No caso dos métodos avançados de redes neurais, foram consideradas 52 semanas de defasagens com o objetivo de que, na fase de treinamento, o modelo conseguisse capturar a sazonalidade. No caso do modelo Naive, é considerada somente uma defasagem e, por fim, no modelo atual a referência é a defasagem da mesma semana do ano anterior.

Depois de estabelecida as bases de dados usadas ao longo deste trabalho, o próximo passo é apresentar uma análise univariada e descritiva de cada série.

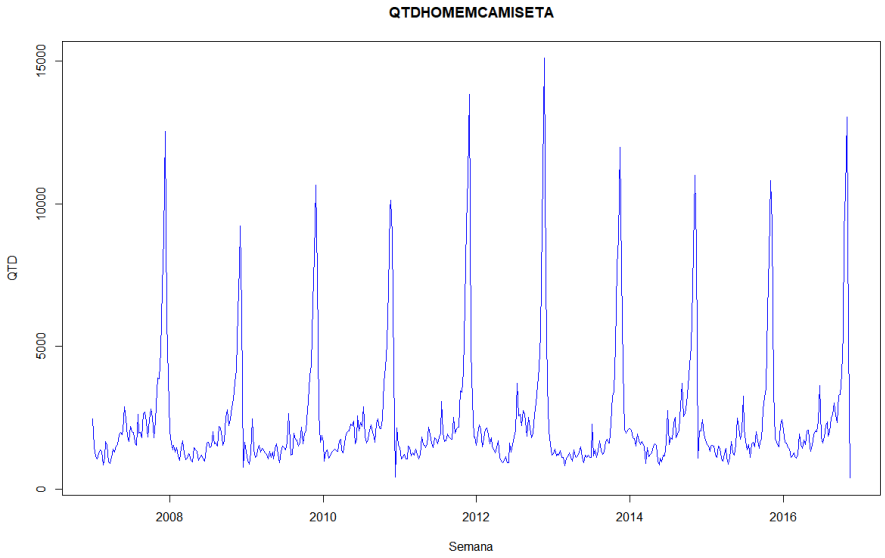
4.5. Análise univariada das séries temporais

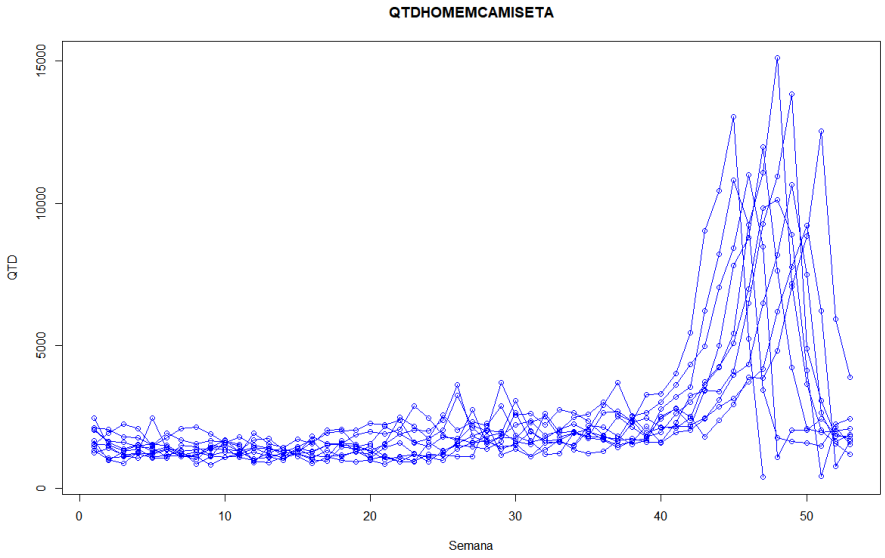
Nesta parte do trabalho, faremos uma análise univariada das quatro séries de tempo que são previstas. Sendo assim, busca-se fazer uma análise exploratória de cada uma das séries de produtos a serem modeladas, as quais são variáveis quantitativas discretas, com intuito de verificar a sua forma de distribuição, por meio da estatística descritiva.

Camiseta de Homem

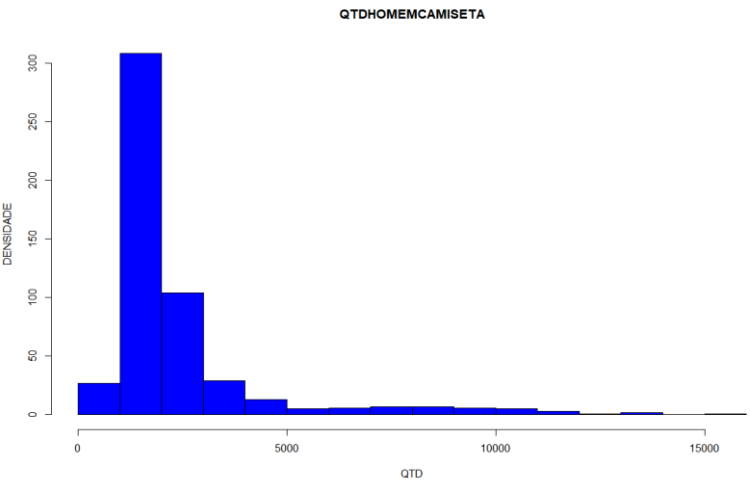
A primeira série analisada é a série de camiseta de adulto masculino. Pode-se perceber claramente um padrão sazonal, tendo as últimas semanas de cada ano, que representam os meses de novembro e dezembro, um peso maior na distribuição. Verifica-se que a média é superior à mediana, o que caracteriza uma distribuição assimétrica à esquerda. Além disso, a média é superior à medida do terceiro quartil, o que é um indício da presença de *outliers* na série, que acabam elevando o valor da média.Estatísticamente, os valores considerados *outliers* são aqueles menores que o valor do primeiro quartil menos a diferença entre o primeiro e terceiro quartil multiplicado por 1,5 ou maiores que o terceiro quartil adicionada a mesma diferença entre quartis informada anteriormente.

A série apresenta uma grande dispersão, com coeficiente de variação de 90% e cerca de 11% dos dados da série são *outliers*, conforme a definição estatística, ultrapassando o limite superior. Devido à presença destes valores extremos, a média é extremamente alta, próxima do valor do terceiro quartil.





MEDIDAS DESCRITIVAS	QTDHOMEMCAMISETA
Q1	1359
Mediana	1695
Média	2413
Q3	2386
Erro padrão	92
Modo	1647
Desvio padrão	2172
Variância da amostra	4717278
Curtose	10
Assimetria	3
Intervalo	14717
Mínimo	377
Máximo	15094
Soma	1389785
Contagem	576
Coeficiente de Variação	90%
#Outliers	63
%Outliers/Contagem	10,9%



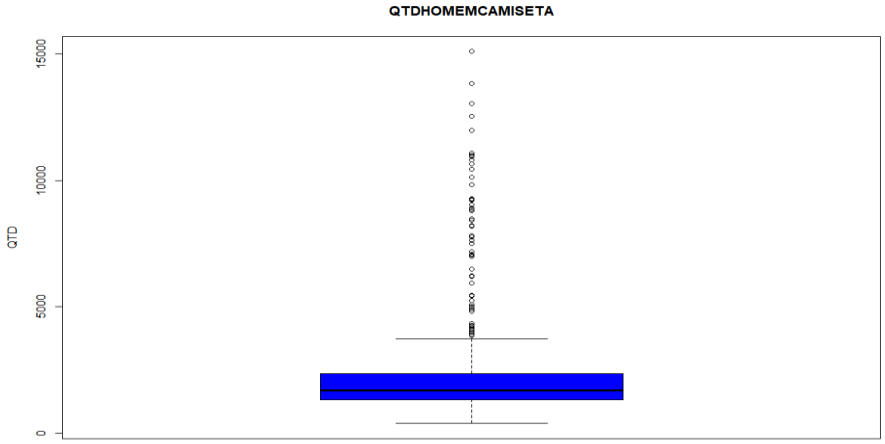
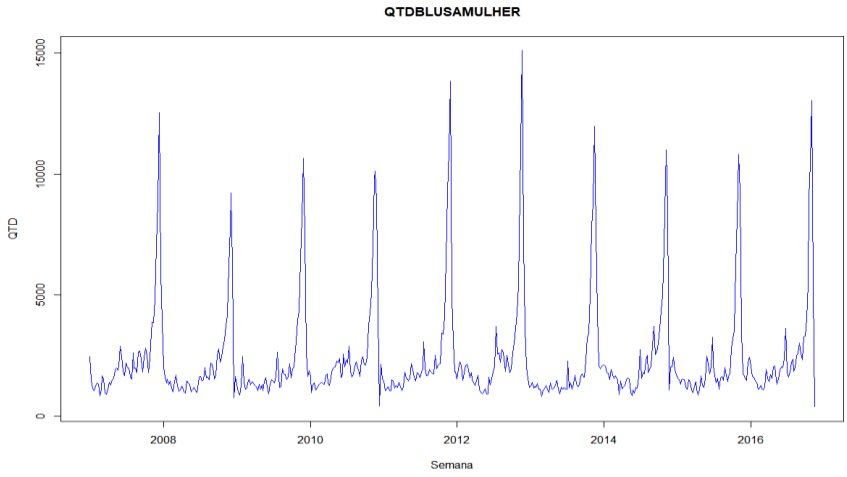
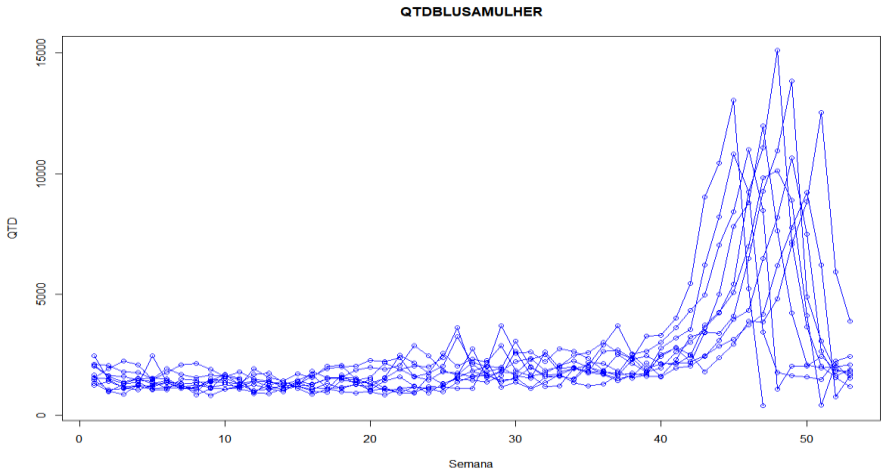


Figura 31 – Análise univariada série camiseta de homem
Fonte: elaboração própria.

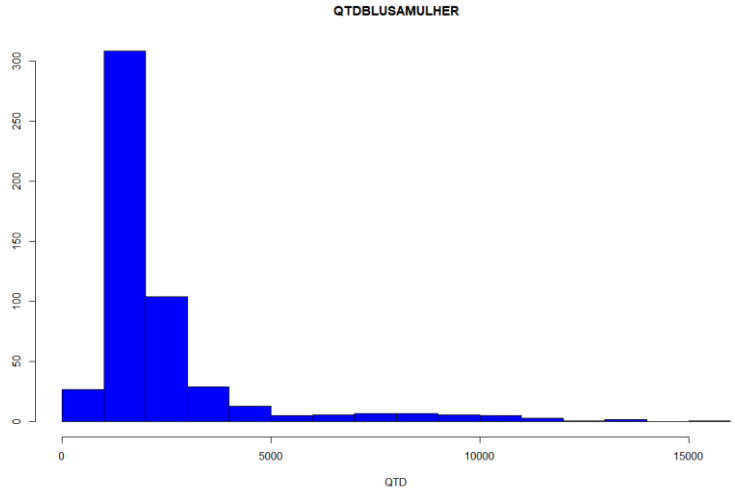
Blusa de Mulher

A série de blusa do adulto feminino é uma série com padrão sazonal similar ao de camiseta do adulto masculino, evidenciando maior volume de vendas nos últimos meses do ano, mais especificamente em novembro e dezembro. Em termos de medidas descritivas, a série tem um nível de concentração menor, quando é feita a comparação com a série de camiseta de homem, mas continua com um padrão de assimetria à esquerda, já que a média é superior à mediana. A partir da análise das medidas de dispersão, verifica-se que esta série é a que possui a maior dispersão, já que apresenta os maiores valores para as medidas de variância e intervalo. Além disso, seu coeficiente de variação é de 78%. Cerca de 10,5% dos dados podem ser considerados *outliers.*





MEDIDAS DESCRITIVAS	QTDMULHERBLUSA
Q1	2429
Mediana	3267
Média	4384
Q3	4804
Erro padrão	97
Modo	3041
Desvio padrão	3402
Variância da amostra	11575571
Curtose	7
Assimetria	3
Intervalo	24227
Mínimo	878
Máximo	25105
Soma	2524946
Contagem	576
Coeficiente de Variação	78%
#Outliers	60
%Outliers/Contagem	10,4%



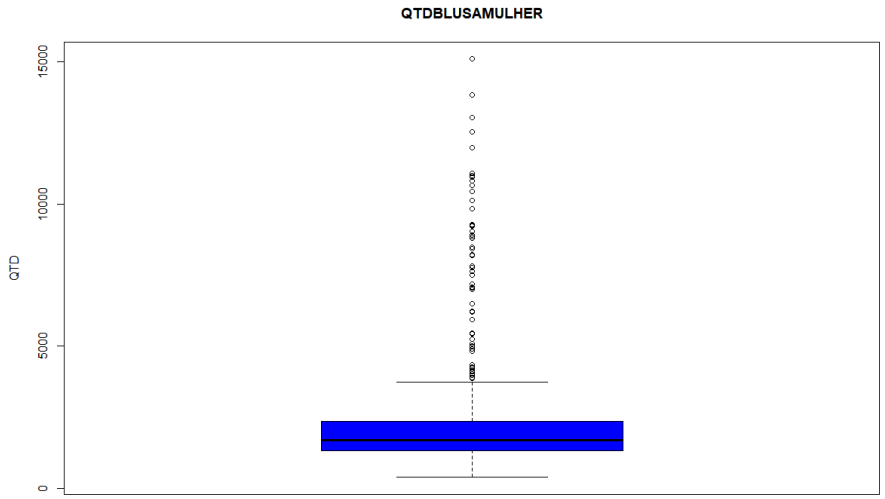
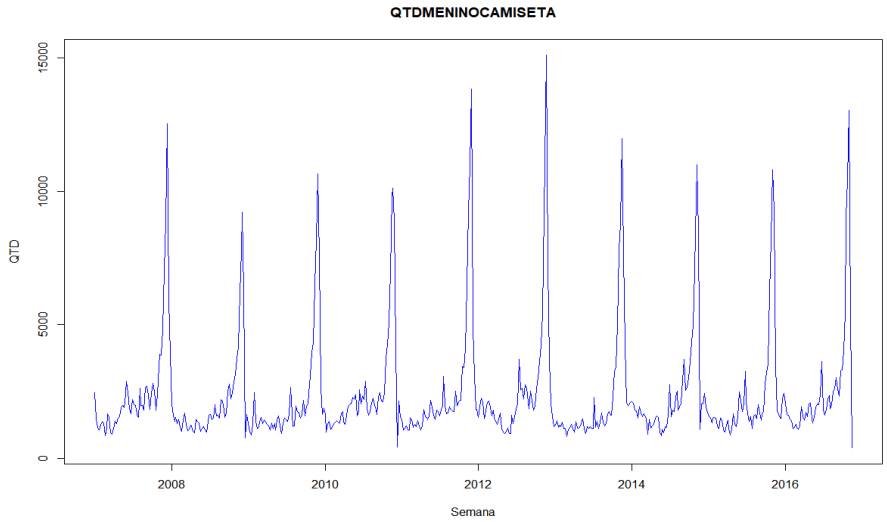
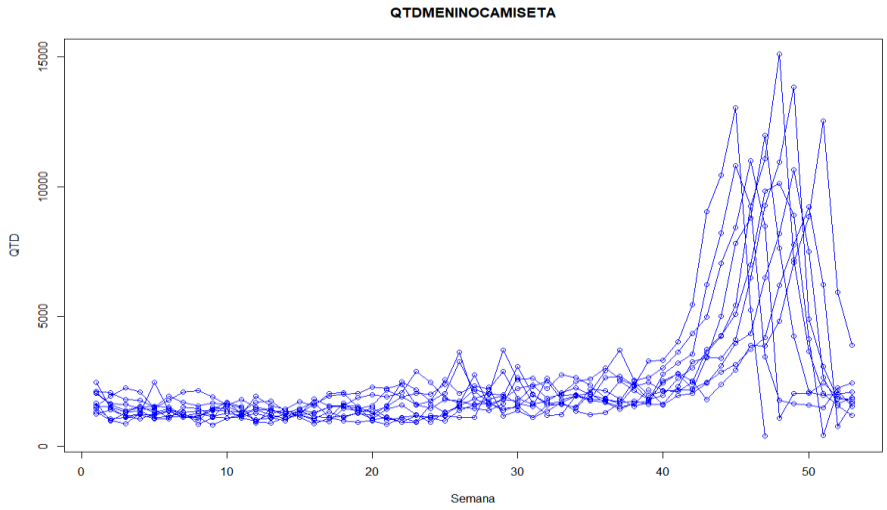


Figura 32– Análise univariada série blusa de mulher
Fonte: elaboração própria.

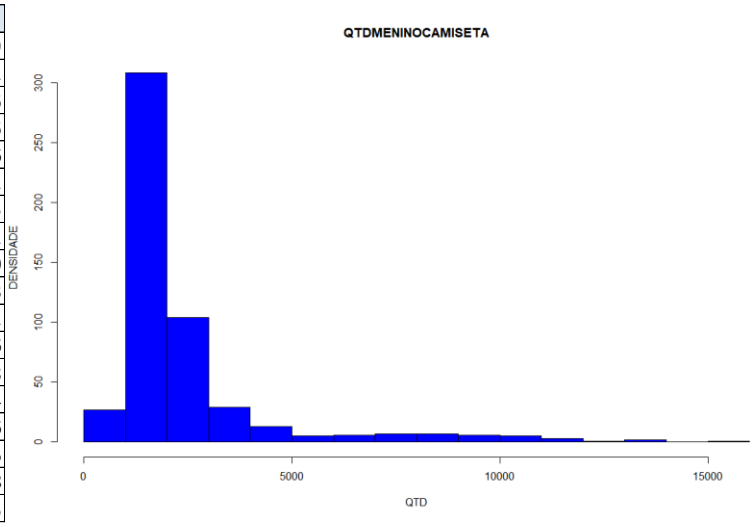
Camiseta de Menino

A série de quantidade de camiseta de menino tem um comportamento muito similar à da série de camiseta de homem, apresentando forte característica sazonal, com maior peso para as semanas dos meses de novembro e dezembro. É possível constatar a presença de uma série com distribuição assimétrica à esquerda, já que a média é superior à mediana. Outro ponto relevante é verificar que a média é superior ao terceiro quartil, indicando a existência de *outliers*, que acabam elevando o valor da média. Adicionalmente, a série se caracteriza por apresentar uma alta dispersão, com um coeficiente de variação de 89% e com cerca de 12% dos seus dados podendo ser considerado como *outlier*s, conforme a definição estatística.





MEDIDAS DESCRITIVAS	QTDMENINOCAMISETA
Q1	1439
Mediana	1821
Média	2476
Q3	2336
Erro padrão	96
Modo	1802
Desvio padrão	2205
Variância da amostra	4862622
Curtose	9
Assimetria	3
Intervalo	13452
Mínimo	266
Máximo	13718
Soma	1426384
Contagem	576
Coeficiente de Variação	89%
#Outliers	68
%Outliers/Contagem	11,8%



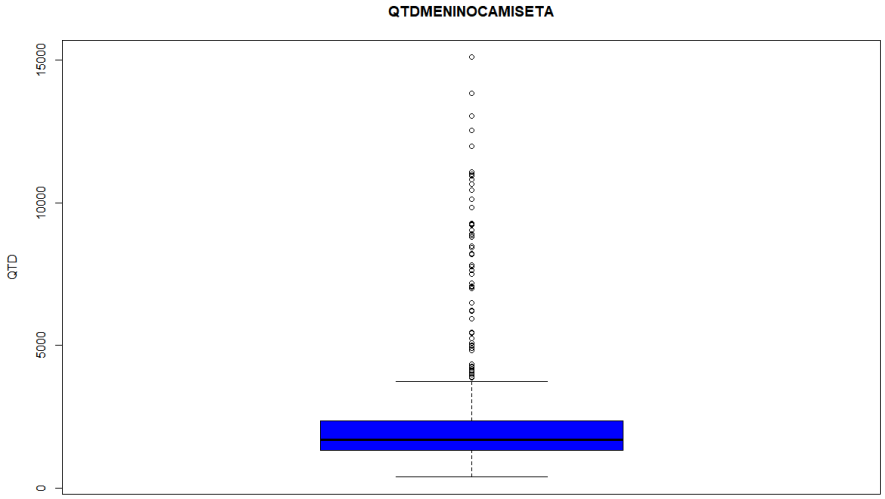
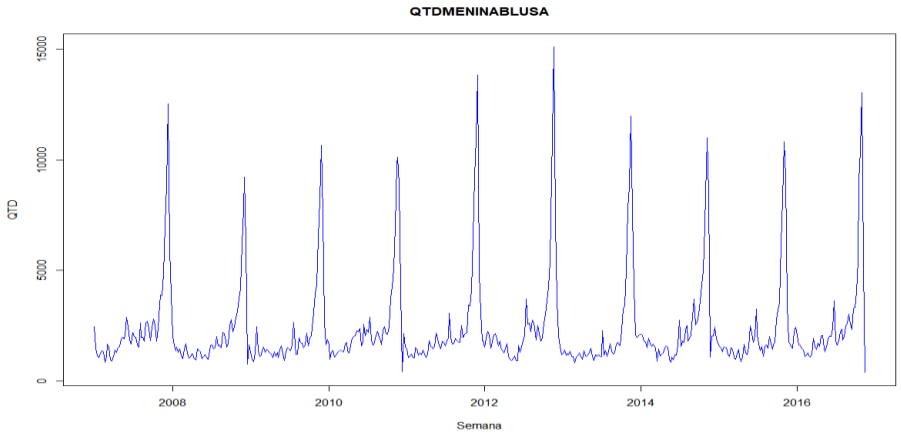


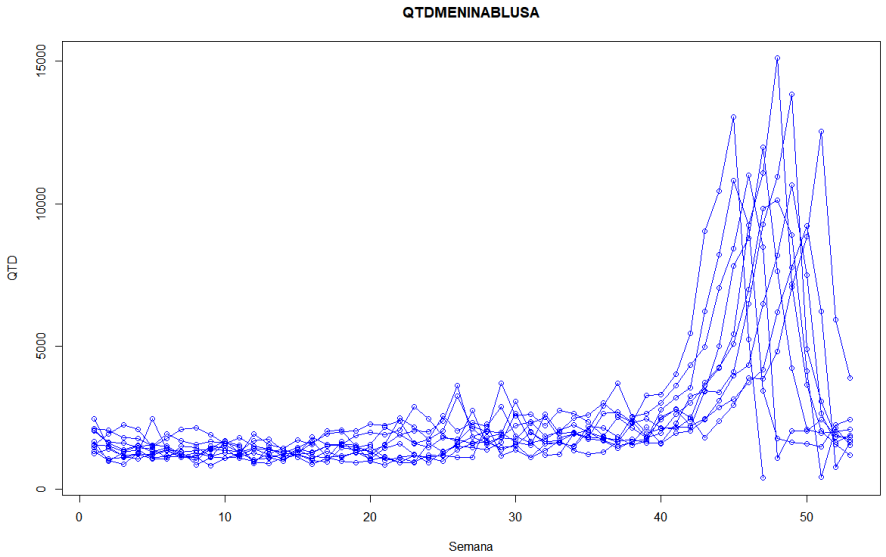
Figura 33 – Análise univariada série camiseta de menino
Fonte: Elaborado pelo autor

Blusa de Menina

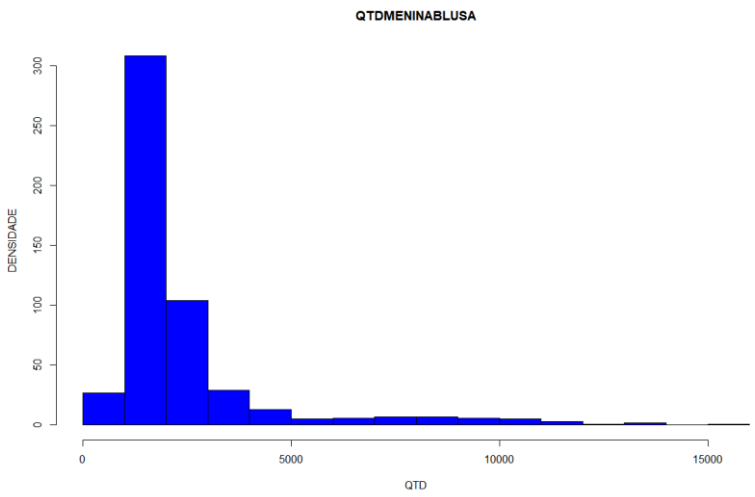
A série de blusa de menina apresenta um padrão sazonal mais concentrado nas últimas semanas do mês de dezembro, não tendo um pico tão forte em novembro, como ocorreu com as demais séries estudadas. A série em questão evidencia uma assimetria negativa, com a média muito superior à mediana. Neste caso, em termos de medidas descritivas, verifica-se que o valor da medida do terceiro quartil é muito próximo da média, o que é um indício da presença de *outliers*. A análise das medidas de dispersão permite verificar que a série possui uma alta dispersão, com o segundo maior intervalo

dentre as séries analisadas, com alta variância, coeficiente de variação de 90%, e cerca de 10,4% dos dados da série são considerados *outliers*.





MEDIDAS DESCRITIVAS	QTDMENINABLUSA
Q1	1478
Mediana	1983
Média	2671
Q3	2690
Erro padrão	94
Modo	2427
Desvio padrão	2401
Variância da amostra	5764389
Curtose	15
Assimetria	3
Intervalo	21301
Mínimo	160
Máximo	21461
Soma	1538724
Contagem	576
Coeficiente de Variação	90%
#Outliers	60
%Outliers/Contagem	10,4%



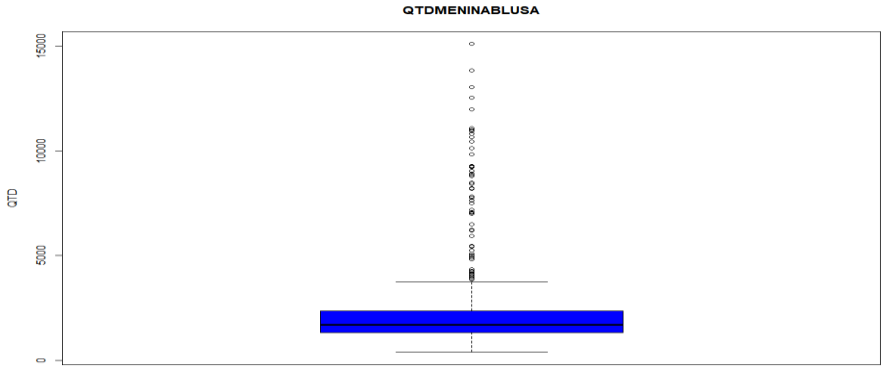


Figura 34 – Análise univariada série blusa de menina
Fonte: elaboração própria.

Em resumo, a análise univariada de todas as séries de produtos, que foram utilizadas neste trabalho, permite afirmar que todas as séries possuem uma forte característica sazonal, sendo que as semanas de novembro e dezembro são que apresentam maior relevância em termos de quantidade vendida para cada uma delas, possuem uma distribuição assimétrica à esquerda, já que a média é superior à mediana e possuem uma alta dispersão. Em todas as séries foi verificada a presença de *outliers*, conforme a definição estatística formal. No entanto, estas observações representam, em sua grande maioria, as quantidades vendidas nas semanas de dezembro, que é um padrão esperado e que não pode ser eliminado da série para a sua efetiva compreensão. Tal ressalva é fundamental, já que observações outliers, geralmente, são eliminadas das séries no momento da especificação do modelo de predição. Na Figura 35, é evidenciado um gráfico com as quatro séries em conjunto, o que torna possível verificar alguns dos pontos mencionados acima.

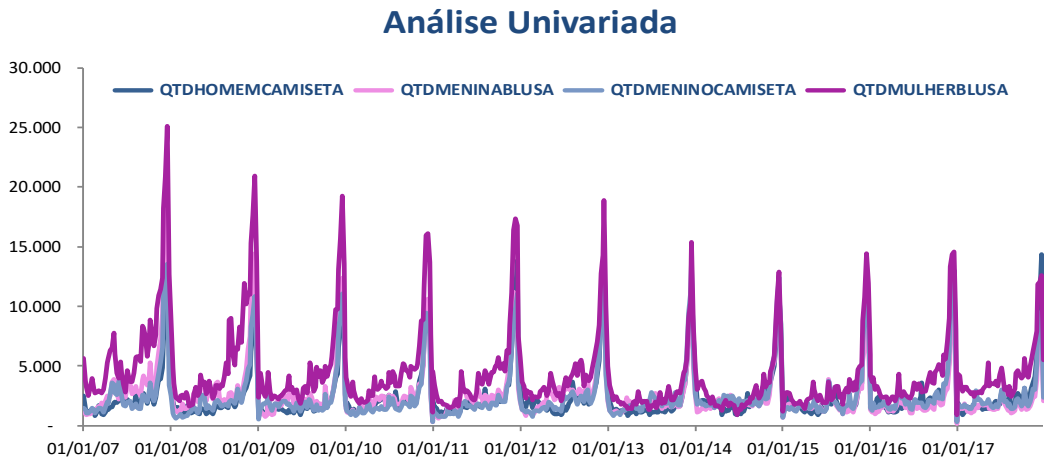


Figura 35 – Análise univariada em conjunto
Fonte: Elaborado pelo autor

4.6. Especificação dos métodos de previsão

Conforme já colocado anteriormente, para cada um dos métodos de previsão foi desenvolvido um conjunto de alternativas de modelos, seguindo o referencial teórico apresentado.

É importante destacar que cada um destes métodos pode ser visto como um algoritmo, cujo objetivo é encontrar o modelo mais eficiente em termos de acuracidade, o que não necessariamente garante que seja o melhor. Diante deste fato, os métodos usados são heurísticos, pois são algoritmos exploratórios que conduzem à resolução do problema em questão, por meio de aproximações, que não necessariamente permitem chegar à solução exata e definitiva e sim na melhor possível.

É por isto que para cada um dos métodos serão desenvolvidos vários modelos buscando encontrar aquele que seja mais preciso. Além disso, não existe uma forma exata de definir quais os parâmetros exatos para se chegar ao modelo ideal, faz parte do papel do analista testar um conjunto de possibilidades para alcançar a solução mais eficiente.

Todos os métodos usam a própria variável a ser prevista na sua forma defasada como varíavel de *input* para a realização da previsão em si. A principal diferença entre os métodos utilizados é quantidade de defasagens usadas na especificação de cada um deles. Especificamente, todos os métodos utilizam a quantidade vendida para uma determinda categoria em semanas passadas para prever na semana corrente a quantidade que será vendida. No caso dos métodos estatísticos clássicos, existe uma análise prévia das defasagens que serão utilizadas, o que não ocorre nos métodos avançados, visto que foi considerado um número amplo de defasagens, que tem suas respectivas importâncias determinadas pelos pesos calculados pela rede neural.

No caso dos métodos estatísticos clássicos foi utilizada a modelagem Box-Jenkins para identificar os termos autorregressivos e de médias móveis necessários para cada um dos modelos, mais especificamente, esta informação é obtida pela análise da FAC e da FACP. Para cada categoria de produtos foi desenvolvido sete modelos no caso do método SARIMA, quatro modelos no método SARIMA com variáveis exógenas, quatro modelos no método SARIMA GARCH e oito modelos no método SARIMA GARCH com exógenas.

Já no caso dos modelos avançados, foram construídos 126 modelos para cada um dos métodos: MLP, MLP com exógenas, LSTM e LSTM com exógenas, testando os seguintes hiperparâmetros, que fazem parte da especificação da topologia da rede:

- **Número de defasagens de entrada**: 52;
- **Número de camadas escondidas**: 1;
- **Número de camadas de saída**: 1;
- **Número de neurônios na camada escondida**: 6,13,26,52,104,156,208;
- **Taxa de aprendizado**: 0.001, 0.01, 0.1;
- **Funções de ativação da camada escondida**: Sigmoide e RELU;
- **Otimizadores**: SGD, ADM e RMSProp.

Assim, no total, para cada uma das séries de categorias de produto, incluindo a alternativa Naive e modelo atual, foram produzidos 529 modelos, tendo em vista os dez métodos de previsão escolhidos. Abaixo, na Tabela 5, é evidenciado um quadro resumo especificando as alternativas testadas e o *software* utilizado para cada uma delas.

Tabela 5 – Alternativas de modelos para cada método de previsão

Método	Modelos Testados	Software
1. Naive	Avaliada 1 alternativa, que considera a última defasagem como próxima previsão	R
2. SARIMA	Avaliadas 7 alternativas, por meio da implantação da metodologia Box-Jenkins	R
3. SARIMA com exógenas	Avaliadas 4 alternativas, por meio da implantação da metodologia Box-Jenkins	R
4. SARIMA GARCH	Avaliadas 4 alternativas, por meio da implantação da metodologia Box-Jenkins	R
5. SARIMA GARCH com exógenas	Avaliadas 4 alternativas, por meio da implantação da metodologia Box-Jenkins	R
6. Atual	Avaliada 1 alternativa, onde o modelo considera o valor do ano anterior com baseline e ajusta por um fator	R
7. MLP	Avaliadas 126 alternativas com base na variação dos seguintes parâmetros: • Número de defasagens de entrada: 52; • Número de camadas escondidas: 1; • Número de camadas de saída: 1; • Neurônios na camada escondida: 6,13,26,52,104,156,208; • Taxa de aprendizado: 0.001, 0.01, 0.1; • Funções de ativação da camada escondida: Sigmoid e RELU; • Otimizadores: SGD, ADM e RMSProp.	Python
8. MLP com exógenas	Avaliadas 126 alternativas com base na variação dos seguintes parâmetros: • Número de defasagens de entrada: 52; • Número de camadas escondidas: 1; • Número de camadas de saída: 1; • Neurônios na camada escondida: 6,13,26,52,104,156,208; • Taxa de aprendizado: 0.001, 0.01, 0.1; • Funções de ativação da camada escondida: Sigmoid e RELU; • Otimizadores: SGD, ADM e RMSProp.	Python
9. LSTM	Avaliadas 126 alternativas com base na variação dos seguintes parâmetros: • Número de defasagens de entrada: 52; • Número de camadas escondidas: 1; • Número de camadas de saída: 1; • Blocos de memória: 6,13,26,52,104,156,208; • Taxa de aprendizado: 0.001, 0.01, 0.1; • Funções de ativação da camada escondida: Sigmoid e RELU; • Otimizadores: SGD, ADM e RMSProp.	Python
10. LSTM com exógenas	Avaliadas 126 alternativas com base na variação dos seguintes parâmetros: • Número de defasagens de entrada: 52; • Número de camadas escondidas: 1; • Número de camadas de saída: 1; • Blocos de memória: 6,13,26,52,104,156,208; • Taxa de aprendizado: 0.001, 0.01, 0.1; • Funções de ativação da camada escondida: Sigmoid e RELU; • Otimizadores: SGD, ADM e RMSProp.	Python

Fonte: elaboração própria.

4.7. *Softwares* e ferramentas de modelagem

Para a manipulação, o tratamento e a produção dos modelos foram utilizados os *softwares* R e Python. O R é uma linguagem de programação de computadores muito utilizada para tratamento de dados estatísticos e gráficos. Pesquisas recentes mostram que a popularidade do R tem aumentado substancialmente nos últimos anos. O objetivo era utilizar somente o R, porém para que os modelos de redes neurais, sobretudo para que o LSTM obtivesse um desempenho maior, em termos de velocidade, se fez necessário o uso do Python. Diante deste fato, os modelos avançados foram processados no Python.

Ross Ihaka e Robert Gentleman criaram a linguagem *open source* R, em 1995. O objetivo era desenvolver uma linguagem que oferecesse uma maneira melhor e mais amigável de fazer análise de dados, estatísticas e modelos gráficos. Inicialmente, o R foi usado, principalmente, em meios acadêmicos e pesquisas, mas ultimamente o R tem se disseminado por diversos setores, inclusive nas corporações. Isso faz do R uma das línguas estatísticas de crescimento mais rápido no mundo.

Uma das principais vantagens do R é a sua enorme comunidade, que oferece suporte através de listas de endereços e uma extensa documentação. Há também o CRAN, que é um enorme repositório de pacotes R, tornando a linguagem extensível e em que os usuários também podem contribuir. Esses pacotes são uma coleção de funções R, que facilitam o acesso imediato às mais recentes técnicas e funcionalidades sem a necessidade de desenvolver tudo a partir do zero.

O Python foi criado por Guido Van Rossem, em 1991, tendo como principal objetivo criar uma linguagem que apresentasse alta produtividade e legibilidade do código. Pode ser considerada uma linguagem flexível e altamente adequada para se construir algo novo, tendo em vista seu foco em legibilidade, simplicidade e sua baixa curva de aprendizagem.

Semelhante ao R, o Python também é extensível e possui vários pacotes. O Python possui uma grande comunidade, mas está um pouco mais dispersa, já que é uma linguagem de uso geral. No entanto, a utilização do Python para ciência de dados está crescendo rapidamente, chegando a ultrapassar o R. Vale ressaltar que diferentemente do R, o Python é uma linguagem de programação interpretada de propósito geral e não específico.

Até 2016, o R era a linguagem com maior adoção e foi ultrapassada em 2017 pelo Python, como pode ser observado na Figura 36.

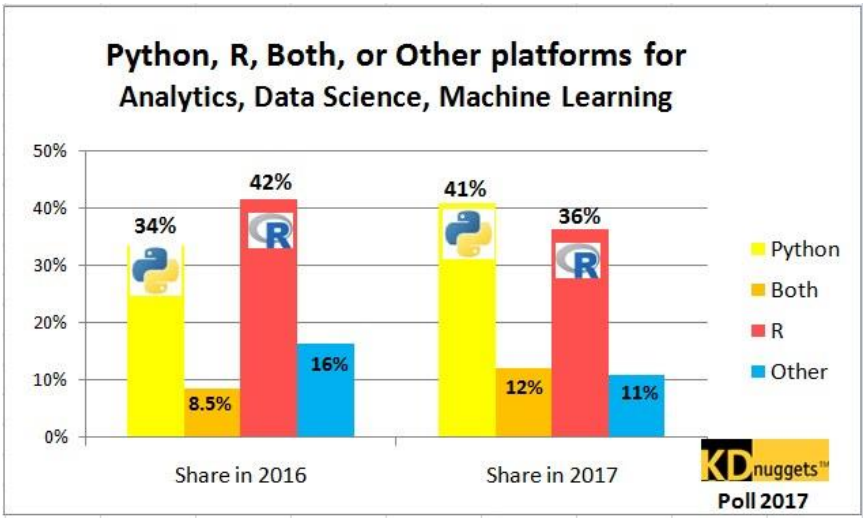


Figura 36 – Evolução ferramentas de *data science* e *machine learning*
Fonte: Kdnuggets, 2017.

Por fim, devem-se destacar duas bibliotecas importantes, instaladas juntamente com o Python, para que os modelos avançados, que são modelos de aprendizagem profunda, fossem construídos: o Tensorflow e o Keras.

O Tensorflow é uma biblioteca Python para computação numérica criada e lançada pelo Google. É uma biblioteca com código aberto para aprendizado de máquina, que pode ser utilizada em diversos problemas.

Já o Keras é uma biblioteca enxuta de Python, que pode ser executada sobre o Theano ou Tensorflow e foi desenvolvida para tornar a implementação de modelos de aprendizagem profunda mais rápida e fácil, alavancando o ambiente de pesquisa e desenvolvimento deste campo de conhecimento. Assim como o Tensorflow, o Keras também foi desenvolvido pela Google. As características do Keras são: modularidade, extensibilidade e legibilidade.

4.8. Critério de avaliação dos modelos de previsão

Os resultados das previsões de cada um dos modelos e para cada categoria de produtos foram avaliados em dois momentos. A primeira avaliação ocorreu dentro do

período de treinamento, que compreendeu as semanas dos anos de 2007 a 2014. Assim, para cada método, foi verificada a acuracidade pelo MAPE das diversas alternativas de modelos construídos para este período. A partir disto, em um segundo momento, foi eleita a alternativa de modelo mais precisa de cada método e foi constatada a sua respectiva acuracidade, agora para o conjunto de teste, que compreende as semanas dos anos de 2015 e 2016. Este procedimento tornou possível a efetiva verificação do método mais preciso para cada uma das quatro séries temporais estudadas.

5. Desenvolvimento dos modelos e análise comparativa

Esta é a parte prática desta dissertação, pois aqui é evidenciado cada um dos dez métodos de previsão escolhidos com as suas respectivas alternativas de modelos. Cada um dos métodos foi desenvolvido sobre o mesmo conjunto de dados de treinamento e de teste. Primeiramente, é evidenciado o resultado de cada método para cada série na fase de treinamento, posteriormente, é apresentado o resultado na fase de teste e, por fim, é feita uma análise comparativa para cada série com o objetivo de verificar qual método obteve a melhor acuracidade na fase de teste.

5.1. Método 1 – Naive

O modelo Naive é o método com a premissa mais simples entre os avaliados. Ele considera que a previsão nada mais é do que o último valor observado. Abaixo, temos os resultados alcançados para cada série na fase de treinamento, juntamente com o respectivo gráfico comparando a série real com a prevista.

$$y_t = y_{t-1}$$

Série 1 – Camiseta de Homem

Como temos somente uma alternativa de modelo neste caso, a medida de acuracidade MAPE foi de 26,56%, o que significa que as previsões geradas apresentam um erro médio percentual absoluto de 27%.

Tabela 6 – Resumo dos resultados do modelo 1 para série 1 – Camiseta de Homem

Modelos	MAPE Treino
Naive	**26,56%**

Fonte: elaboração própria.

Na Figura 37, é apresentado o gráfico, onde se verifica a aderência do modelo previsto contra os dados reais e o MAPE do conjunto de dados teste.

Modelo Naive - QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
Naive	26,56%	33,90%

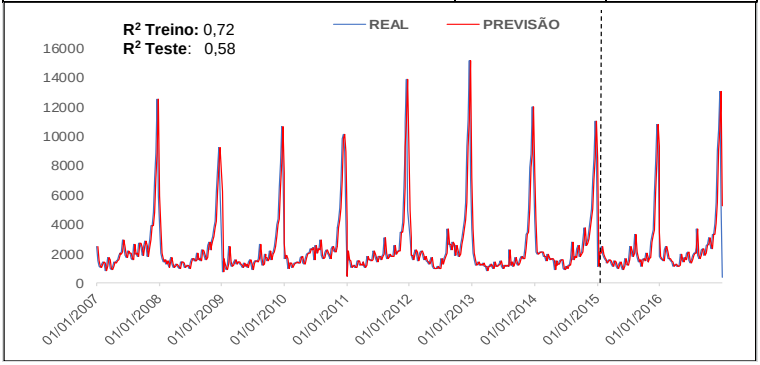


Figura 37 – Gráfico com resultados melhor alternativa método 1 para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Série 2 – Blusa de Mulher

No caso do modelo Naive, existe somente uma alternativa de modelo, a qual apresentou um MAPE de 23,03% para a série de blusa de mulher, indicando que o erro médio percentual absoluto é de 23%.

Tabela 7 – Resumo dos resultados do modelo 1 para série 2 – Blusa de Mulher

Modelos	MAPE Treino
Naive	**23,03%**

Fonte: elaboração própria.

Na Figura 38, temos o gráfico com os dados reais e os dados previstos. Além disso, também é apresentado o MAPE no conjunto de teste, que foi de 25,24%.

Modelo Naive - QTDMULHERBLUSA	MAPE Treino	MAPE Teste
Naive	23,03%	25,24%

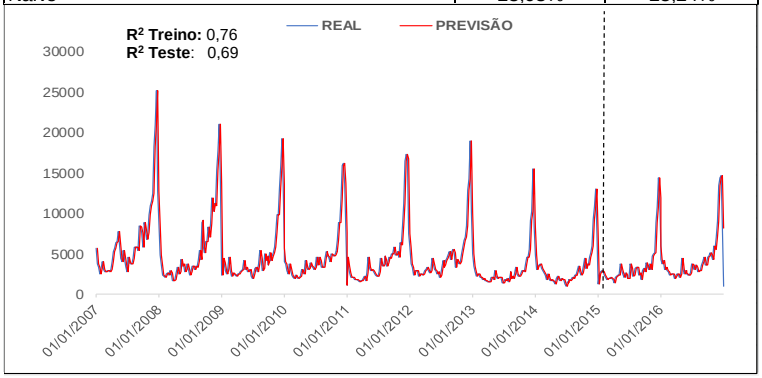


Figura 38 – Gráfico com resultados melhor alternativa método 1 para a série 2– Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

No caso da série temporal de camiseta de menino, foi obtido um MAPE no conjunto de treino de 29,14%, o pior resultado entre todas as séries temporais testadas para este método.

Tabela 8 – Resumo dos resultados do modelo 1 para série 3 – Camiseta de Menino

Modelos	MAPE Treino
Naive	**29,14%**

Fonte: elaboração própria.

Abaixo, na Figura 39, é possível comparar os dados reais e os dados previstos para a série em análise, a qual apresentou uma piora considerável na medida de acuracidade MAPE no conjunto de dados testes, que passou de 29,14% para 36,28%.

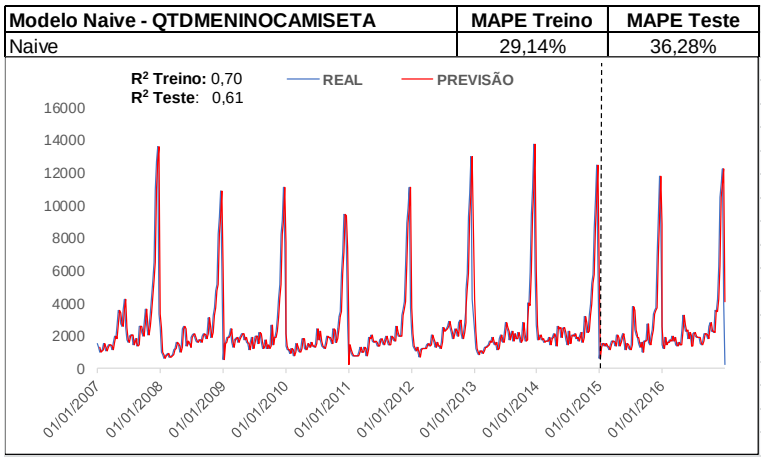


Figura 39 – Gráfico com resultados melhor alternativa método 1 para a série 3 – Camiseta de Menino
Fonte: elaboração própria

Série 4 – Blusa de Menina

Na série de blusa de menina para o método Naive, em que a previsão é sempre igual ao último valor observado, obteve-se um MAPE de 28,58%.

Tabela 9 – Resumo dos resultados do modelo 1 para série 4 – Blusa de Menina

Modelos	MAPE Treino
Naive	**28,58%**

Fonte: elaboração própria.

Assim como ocorreu na série de camiseta de menino, a série de blusa de menina também apresentou uma piora significativa na execução do método Naive no conjunto de dados testes, o que fez com que o MAPE saísse de cerca de 30% para 39%.

Modelo Naive - QTDMENINABLUSA	MAPE Treino	MAPE Teste
Naive	29,58%	39,09%

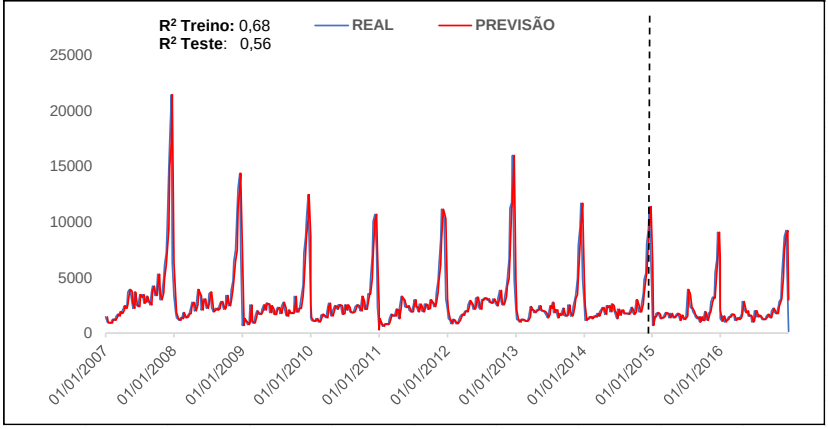


Figura 40 – Gráfico com resultados melhor alternativa método 1 para a série 4 – Blusa de Menina
Fonte: elaboração própria.

5.2. Método 2 – SARIMA

Com base na metodologia de Box-Jenkins apresentada no capítulo de revisão teórica, são desenvolvidos os seguintes passos para identificação do modelo mais apropriado para cada série: (1) verificação da estacionariedade da série e determinação da ordem de integração necessária, caso a série não seja estacionária; (2) identificação, que consiste em verificar qual modelo melhor descreve o comportamento da série, o que é feito com base na análise das funções de autocorrelações (FAC) e das funções de auto correlações parciais (FACP); (3) estimação, onde é realizada a estimação dos parâmetros em si; e (4) verificação, onde é feita a análise dos resíduos, visando identificar a adequação do modelos para os dados em questão.

Importante considerar que o método SARIMA nada mais é do que a previsão da sazonalidade com a modelagem Box-Jenkins. Logo, não é feita a inclusão de variáveis *dummies* para modelar a sazonalidade.

Série 1 – Camiseta de Homem

A análise dos testes de raiz unitária ADF, PP e KPSS, considerando as três modalidades de teste, com intercepto, como tendência e intercepto e sem tendência e sem intercepto, torna possível afirmar que a série é estacionária, a partir da comparação da estatística do teste com os valores tabelados para diferentes níveis de significância. O tipo sem tendência e sem intercepto não é possível de ser feito no teste KPSS. A seguir, é apresentada a Tabela 10 com o resumo dos testes.

Tabela 10 – Resumo dos testes de raiz unitária para a série 1 – Camiseta de Homem

	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**
T calculado	- 8,21	- 6,84	0,09	- 8,23	- 6,87	0,02	- 3,21	- 4,31	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,97	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

Fonte: elaboração própria.

O próximo passo consiste em fazer a análise da FAC e da FACP para identificar os termos AR e MA que serão necessários para construir os possíveis modelos SARIMA dessa série.

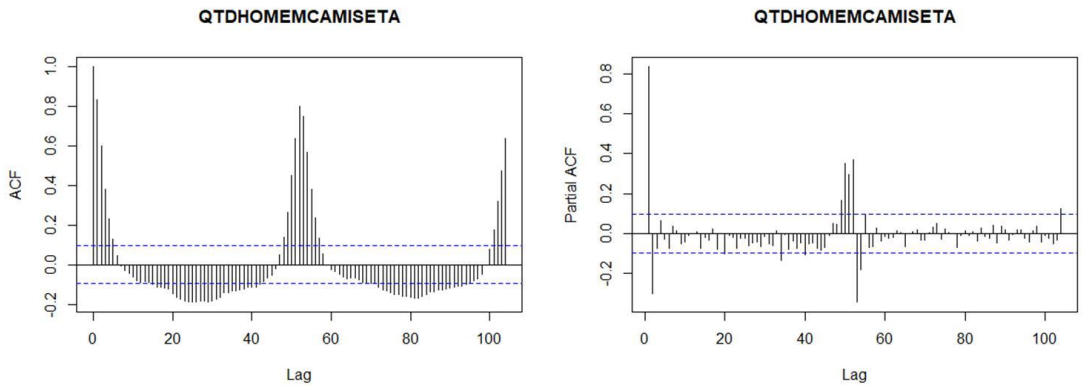


Figura 41 – Gráfico FAC e FACP da série 1 – Camiseta de Homem
Fonte: elaboração própria.

A partir da análise do FAC e do FACP, é possível constatar que se está diante de um modelo muito próximo de um AR (1), dado o decaimento da FAC e do alto valor do primeiro *lag* no FACP. Adicionalmente, a análise dos gráficos de autocorrelação permite verificar um forte padrão sazonal. Diante disto, os modelos considerados foram:

- SARIMA $(1,0,1)(0,1,0)_{52}$
- SARIMA $(1,0,0)(0,1,0)_{52}$

- SARIMA $(1,0,0)(1,0,0)_{52}$
- SARIMA $(1,0,0)(1,1,0)_{52}$
- SARIMA $(1,0,1)(1,1,0)_{52}$
- SARIMA $(1,0,2)(0,1,0)_{52}$
- SARIMA $(1,0,2)(1,1,0)_{52}$

Depois da identificação dos modelos, é fundamental fazer a análise dos resíduos de cada um deles para verificar se os mesmos não estão correlacionados. Além disso, deve-se verificar se eles são homocedásticos e normalmente distribuídos. Não foi constatada a existência de correlação em cada uma das alternativas contempladas, o que possibilita a sua utilização.

No entanto, os resíduos dos modelos analisados são heterocedásticos, através do teste ARCH-LM, o que levará a utilização da abordagem GARCH. Também não são normais, por meio do teste de Shapiro.

Abaixo, na Tabela 11, temos o quadro resumo com a medida MAPE e o critério de AIC de cada um dos modelos na fase de treinamento. Deve-se ressaltar que será utilizado o MAPE para a escolha do modelo mais adequado. O critério de AIC é o parâmetro de referência para determinar o modelo mais parcimonioso.

Tabela 11 – Resumo dos resultados do modelo 2 para série 1 – Camiseta de Homem

Modelos	MAPE Treino	AIC
SARIMA$(1,0,1)(0,1,0)_{52}$	**18,89%**	**5.924,94**
SARIMA$(1,0,0)(0,1,0)_{52}$	18,90%	5.922,95
SARIMA$(1,0,0)(1,0,0)_{52}$	20,03%	6.789,98
SARIMA$(1,0,0)(1,1,0)_{52}$	19,93%	5.924,91
SARIMA$(1,0,1)(1,1,0)_{52}$	18,92%	5.926,90
SARIMA$(1,0,2)(0,1,0)_{52}$	19,66%	5.922,75
SARIMA$(1,0,2)(1,1,0)_{52}$	19,71%	5.924,67

Fonte: elaboração própria.

Os modelos evidenciam MAPEs muito próximos, mas o modelo com melhor desempenho é: SARIMA$(1,0,1)(0,1,0)_{52,}$ que apresentou MAPE de 18,89% e um dos critérios de AIC mais baixos. A seguir, na Figura 43, é apresentada a análise dos resíduos deste modelo, juntamente com a verificação da homocedasticidade e da normalidade.

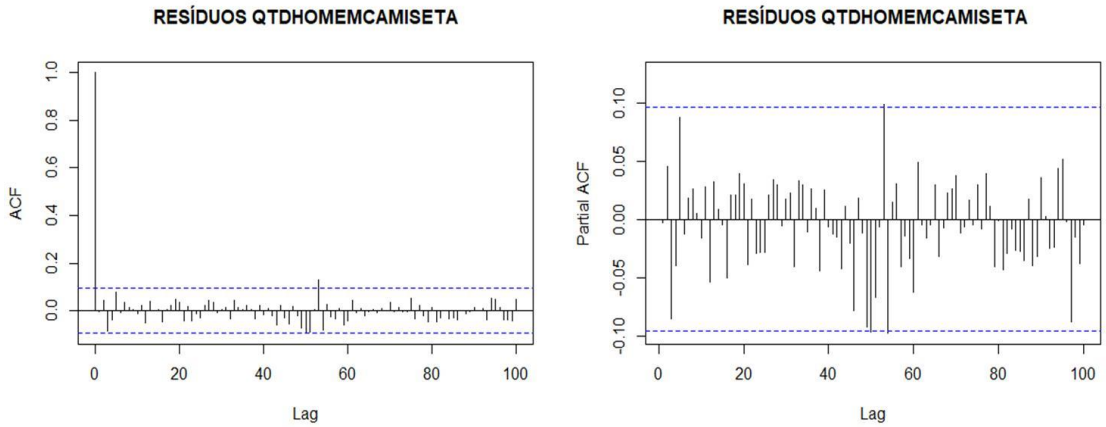


Figura 42 – Gráfico FAC e FACP dos resíduos do modelo SARIMA $(1,0,1)(0,1,0)_{52}$ para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Tabela 12 – Resumo dos testes de correlação, homocedasticidade e normalidade dos resíduos do modelo SARIMA $(1,0,1)(0,1,0)_{52}$ para série 1 – Camiseta de Homem

Testes Resíduos	Teste Autocorrelação Ljung-Box	Teste Heterocedasticidade LM teste	Teste Normalidade Shapiro teste
p-value	0,995	-	-

Fonte: elaboração própria.

Por fim, deve-se evidenciar na Figura 43, o gráfico com os dados previstos contra os dados reais, juntamente com o MAPE no conjunto de dados do teste. Diferentemente do que ocorreu no método Naive, no método SARIMA, o MAPE do conjunto de teste piorou significativamente.

Modelo SARIMA - QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
SARIMA(1,0,1)(0,1,0)52	18,89%	46,27%

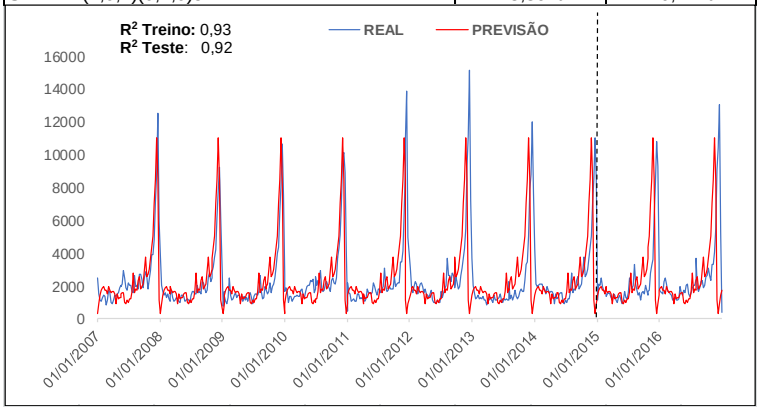


Figura 43 – Gráfico com resultados melhor alternativa método 2 para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Série 2 – Blusa de Mulher

Para a utilização da metodologia Box-Jenkins, um pré-requisito é que a série seja estacionária. Para a série de blusa de mulher e, por meio dos testes de raiz unitária ADF, PP e KPSS, pode-se afirmar que a série é estacionária nas três modalidades testadas, as quais são: com intercepto, com intercepto e com tendência e sem intercepto e sem tendência. Na Tabela 13, é apresentado um resumo com os resultados destes testes.

Tabela 13 – Resumo dos testes de raiz unitária para a série 2 – Blusa de Mulher

	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**
T calculado	-6,98	-6,58	0,34	-7,36	-6,65	0,07	-2,86	-3,57	-
T crítico - 1%	-3,44	-3,44	0,74	-3,97	-3,97	0,22	-2,57	-2,57	-
T crítico - 5%	-2,87	-2,87	0,46	-3,42	-3,42	0,15	-1,94	-1,94	-
T crítico - 10%	-2,57	-2,57	0,35	-3,13	-3,13	0,12	-1,62	-1,62	-

Fonte: elaboração própria.

O primeiro passo da metodologia Box-Jenkins é a identificação dos termos autorregressivos e de médias móveis necessários para modelar a série em questão, o que é feito a partir da verificação da função de autocorrelação e de autocorrelação parcial.

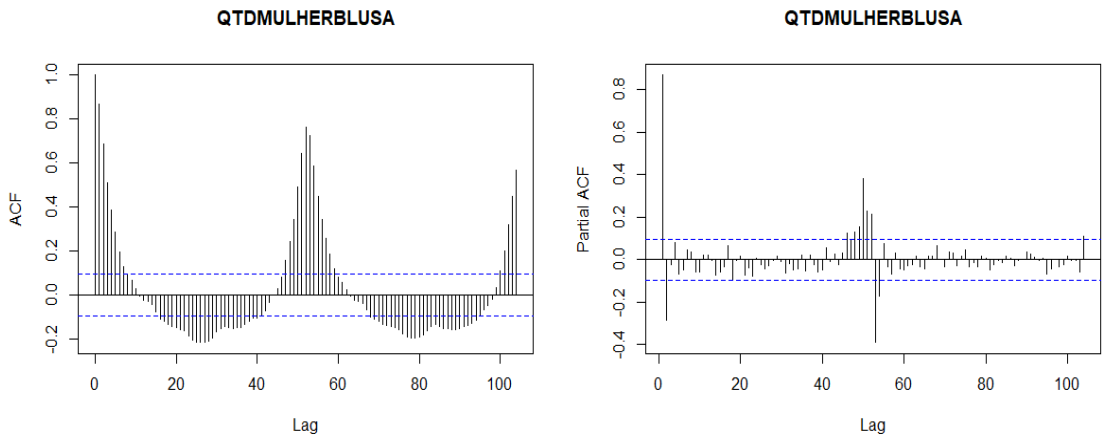


Figura 44 – Gráfico FAC e FACP da série 2 – Blusa de Mulher
Fonte: elaboração própria.

Com base na análise dos gráficos da Figura 45, pode-se constatar um decaimento exponencial da FAC, o que é característico de um modelo AR. Adicionalmente, percebe-se a relevância do termo AR(1), com menor intensidade de um AR(2) e um claro aspecto sazonal ao longo da série, o que se intensifica nas última

semanas do ano. Diante destes pontos, foram sugeridas as seguintes alternativas de modelos para avaliação:

- SARIMA $(2,0,2)(1,1,0)_{52}$
- SARIMA $(1,0,0)(1,1,0)_{52}$
- SARIMA $(1,0,1)(1,1,0)_{52}$
- SARIMA $(1,0,2)(0,1,0)_{52}$
- SARIMA $(1,0,2)(1,1,0)_{52}$
- SARIMA $(2,0,2)(0,1,0)_{52}$
- SARIMA $(3,0,2)(1,1,0)_{52}$

O próximo passo da modelagem Box-Jenkins é a análise dos resíduos, para verificar se eles não são autocorrelacionados, homocedásticos e normalmente distribuídos. Baseada na análise da FAC e da FACP dos resíduos e, juntamente, com a análise da Ljung-Box, verificou-se que não há autocorrelação dos resíduos em todas as alternativas testadas. No entanto, evidenciou-se a presença de heterocedasticidade, através do teste ARCH-LM, e de não normalidade dos resíduos, por meio do teste de Shapiro.

A solução do problema de heterocedasticidade é feita pela aplicação da abordagem GARCH, próximo método a sert desenvolvido. A seguir, na Tabela 14, é apresentado um quadro resumo com o resultado do MAPE e do critério de AIC de cada uma das alternativas de modelos testadas. Com base no MAPE, o melhor modelo SARIMA é o SARIMA (2,0,2)(1,1,0)52, o qual também apresentou o menor critério AIC, que é o desejado para se determinar o modelo mais parcimonioso.

Tabela 14 – Resumo dos resultados do modelo 2 para série 2 – Blusa de Mulher

Modelos	MAPE Treino	AIC
SARIMA$(2,0,2)(1,1,0)_{52}$	**18,98%**	**6.262,05**
SARIMA$(1,0,0)(1,1,0)_{52}$	19,40%	6.262,90
SARIMA$(1,0,1)(1,1,0)_{52}$	19,20%	6.262,33
SARIMA$(1,0,2)(0,1,0)_{52}$	19,56%	6.265,84
SARIMA$(1,0,2)(1,1,0)_{52}$	19,06%	6.264,12
SARIMA$(2,0,2)(0,1,0)_{52}$	19,50%	6.265,85
SARIMA$(3,0,2)(1,1,0)_{52}$	18,99%	6.263,62

Fonte: elaboração própria.

Mesmo já tendo sido esclarecida a inexistência de autocorrelação dos resíduos, é importante evidenciar os testes de autocorrelação, heterocedasticidade e normalidade dos resíduos da alternativa de modelo escolhida. O *p-value* do teste de autocorrelação foi extremamente alto, bem superior ao nível de significância de 5%, o que impede de rejeitar a hipótese de não existência de autocorrelação. Por outro lado, os resíduos são heterocedásticos e não são normalmente distribuídos.

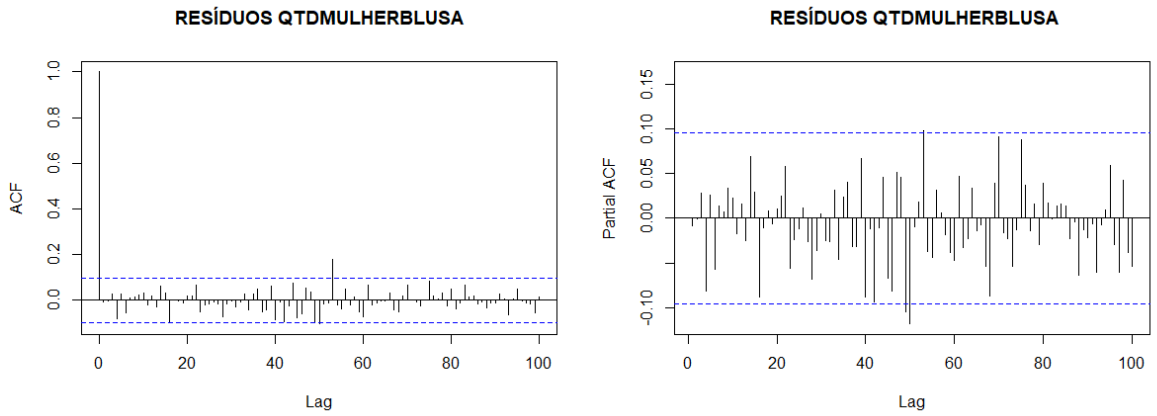


Figura 45 – Gráfico FAC e FACP dos resíduos do modelo SARIMA $(2,0,2)(1,1,0)_{52}$ para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Tabela 15 – Resumo dos testes de correlação, homocedasticidade e normalidade dos resíduos do modelo SARIMA $(2,0,2)(1,1,0)_{52}$ para série 2 – Blusa de Mulher

Testes Resíduos	Teste Autocorrelação Ljung-Box	Teste Heterocedasticidade LM teste	Teste Normalidade Shapiro teste
p-value	0,491	-	-

Fonte: elaboração própria.

Na Figura 46, pode-se confrontar os dados reais e as previsões feitas pela alternativa de modelo escolhida. Também é possível visualizar o resultado do MAPE da aplicação do modelo selecionado para o conjunto de dados de teste, que foi de 44,67%, valor muito maior daquele verificado no conjunto dados de treino.

Modelo SARIMA - QTDMULHERBLUSA	MAPE Treino	MAPE Teste
SARIMA(2,0,2)(1,1,0)$_{52}$	18,98%	44,67%

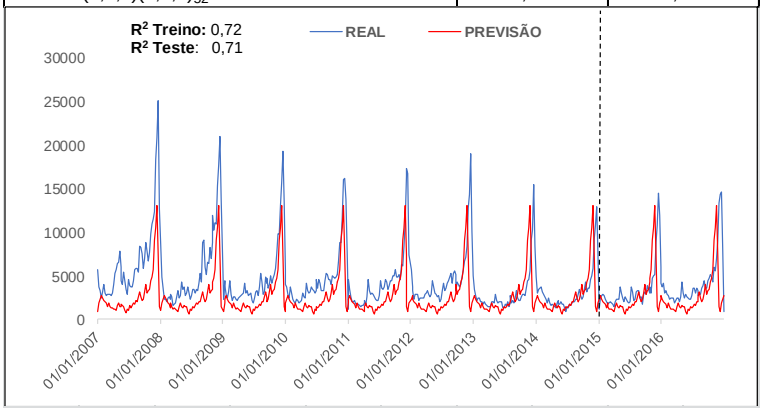


Figura 46 – Gráfico com resultados melhor alternativa método 2 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

A análise da estacionariedade para a série de camiseta de menino, uma condição para o uso do método em evidência, através dos testes de raiz unitária ADF, PP e KPSS e em três variações possíveis, possibilita afirmar, assim como ocorreu com as outras séries, que a série de camiseta de menino é estacionária.

Tabela 16 – Resumo dos testes de raiz unitária para a série 3 – Camiseta de Menino

	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**
T calculado	- 8,98	- 6,49	0,06	- 8,98	- 6,48	0,02	- 3,43	- 4,43	-
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,97	0,22	- 2,57	- 2,57	-
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	-
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	-

Fonte: elaboração própria.

Seguindo a sequência de metodologia Box-Jenkins, o primeiro passo é a identificação das alternativas de modelos apropriados, com base na análise dos gráficos FAC e FACP, os quais tornam possível verificar um decaimento exponencial da FAC, indicando fortemente um modelo AR. Além disso, percebe-se uma truncagem de primeira ordem da PACF, reforçando a presença de um modelo AR(1).

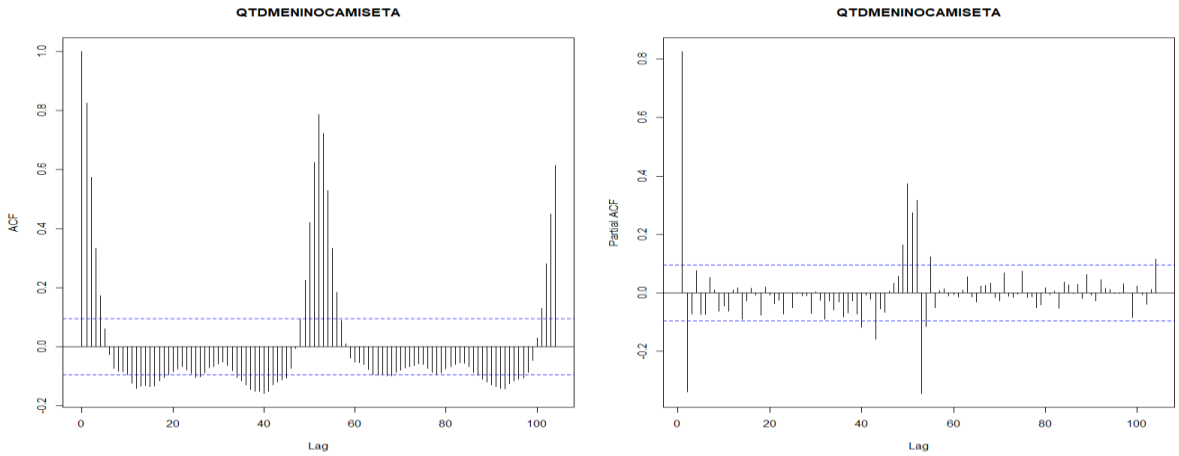


Figura 47 – Gráfico FAC e FACP da série 3 – Camiseta de Menino
Fonte: elaboração própria.

Diante desta análise, foram testadas algumas alternativas de modelos. Em parte delas, foram incluídos alguns termos de média móvel com o objetivo de encontrar modelos mais assertivos:

- SARIMA $(1,0,0)(0,1,0)_{52}$
- SARIMA $(1,0,0)(1,0,0)_{52}$
- SARIMA $(1,0,0)(1,1,0)_{52}$
- SARIMA $(1,0,2)(0,1,0)_{52}$
- SARIMA $(1,0,2)(1,1,0)_{52}$
- SARIMA $(2,0,2)(0,1,0)_{52}$
- SARIMA $(2,0,2)(1,1,0)_{52}$

Considerando os modelos identificados acima, o próximo passo é realizar a verificação dos resíduos, visando examinar a adequação das alternativas escolhidas. Pode-se afirmar que todas as alternativas acima não apresentaram autocorrelação nos resíduos. Todavia, verificou-se a presença de heterocedasticidade, que será tratado através da modelagem GARCH, e distribuição não normal para os mesmos.

Como todas as alternativas são viáveis e tendo em vista a especificação do MAPE como métrica de acuracidade, a opção selecionada diante do conjunto de possibilidades existentes foi o modelo SARIMA $(1,0,0)(0,1,0)_{52}$, por ela ter apresentado o menor MAPE. O mesmo modelo também evidenciou o menor critério de AIC, indicando ser o mais parcimonioso. Na Tabela 17, pode-se visualizar um resumo com os resultados de cada modelo.

Tabela 17 – Resumo dos resultados do modelo 2 para série 3 – Camiseta de Menino

Modelos	MAPE Treino	AIC
SARIMA(1,0,0)(0,1,0)$_{52}$	**21,21%**	**5.903,35**
SARIMA(1,0,0)(1,0,0)$_{52}$	22,67%	6.791,00
SARIMA(1,0,0)(1,1,0)$_{52}$	21,26%	5.905,03
SARIMA(1,0,2)(0,1,0)$_{52}$	21,79%	5.903,25
SARIMA(1,0,2)(1,1,0)$_{52}$	21,86%	5.905,20
SARIMA(2,0,2)(0,1,0)$_{52}$	21,77%	5.905,23
SARIMA(2,0,2)(1,1,0)$_{52}$	21,85%	5.907,30

Fonte: elaboração própria.

Cabe apresentar a análise dos resíduos da alternativa escolhida, a qual mostrou resíduos bem-comportados, sem a presença de autocorrelação. Mais uma vez, verificou-se a presença da heterocedasticidade e a falta de normalidade dos resíduos da alternativa escolhida, tendo em vista os baixos valores obtidos nos testes para avaliar tais características, levando à rejeição da hipótese nula em ambos os casos. Na Figura 48, temos a análise dos resíduos, e na Tabela 18 o quadro resumo com os resultados dos testes dos mesmos.

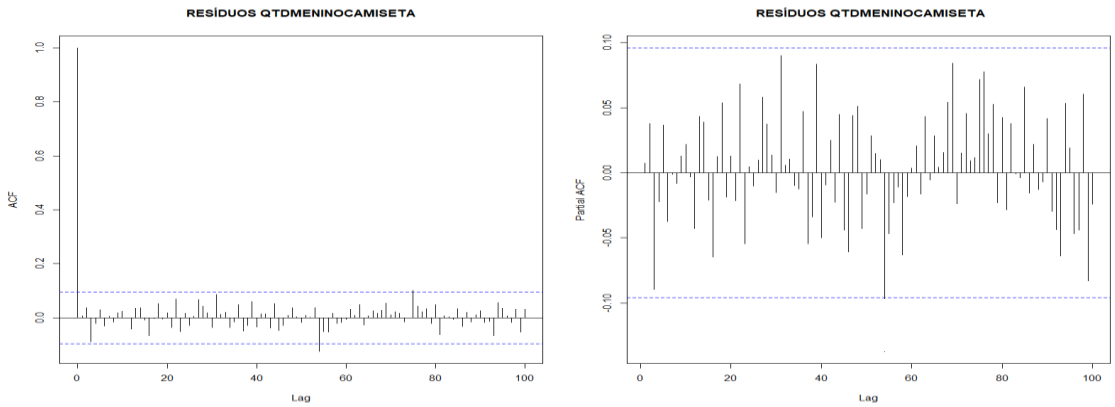


Figura 48 – Gráfico FAC e FACP dos resíduos do modelo SARIMA (1,0,0)(0,1,0)$_{52}$ para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Tabela 18 – Resumo dos testes de correlação, homocedasticidade e normalidade dos resíduos do modelo SARIMA (1,0,0)(0,1,0)$_{52}$ para série 3 – Camiseta de Menino

Testes Resíduos	Teste Autocorrelação Ljung-Box	Teste Heterocedasticidade LM teste	Teste Normalidade Shapiro teste
p-value	0,990	-	-

Fonte: elaboração própria.

Por fim, é importante verificar a acuracidade do modelo selecionado para o conjunto de dados teste, que foi de 51,06%, muito menor do que no conjunto de dados treino, o que é esperado. No Figura 49, é possível comparar os dados reais e as previsões feitas para a série de camiseta de menino com base no modelo SARIMA (1,0,0)(0,1,0)52.

Modelo SARIMA - QTDMENINOCAMISETA	MAPE Treino	MAPE Teste
SARIMA$(1,0,0)(0,1,0)_{52}$	21,21%	51,06%

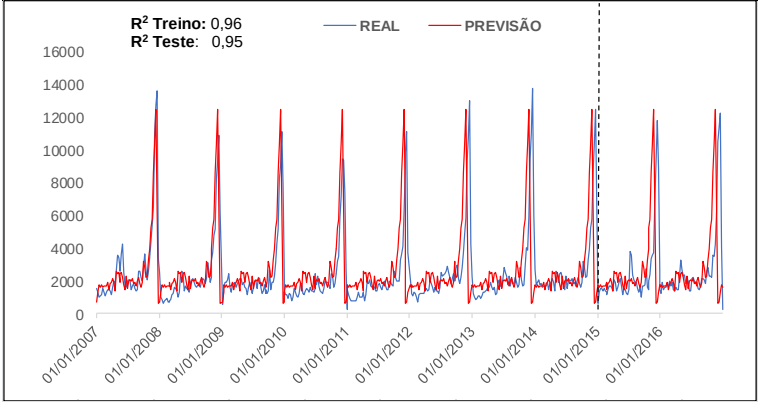


Figura 49 – Gráfico com resultados da melhor alternativa método 2 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

Mais uma vez, se faz necessário verificar a estacionariedade da série, que é uma exigência da metodologia Box-Jenkins. A partir da aplicação dos testes de raiz unitária: ADF, PP e KPSSS nas três variações a seguir, é possível constatar que a série de blusa de menina também é estacionária. Adiante, temos a Tabela 19 com o quadro resumo com as estatísticas calculadas e os valores críticos das tabelas de cada teste para cada nível de significância.

Tabela 19 – Resumo dos testes de raiz unitária para a série 2 – Blusa de Menina

	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**
T calculado	- 7,73	- 6,59	0,35	- 7,33	- 6,47	0,02	- 1,72	- 6,47	-
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,97	0,22	- 2,57	- 3,97	-
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 3,42	-
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 3,13	-

Fonte: elaboração própria.

A próxima etapa é a realização da identificação do modelo ou modelos apropriados para a série em questão, o que é feito com base na análise das funções FAC e FACP. Assim como ocorreu com as outras séries, percebe-se claramente uma

forte presença de sazonalidade, o que é intensificado nas últimas semanas do ano. Observa-se um decaimento exponencial da FAC e uma truncagem de segunda ordem na FACP, sugerindo uma maior adequação de um modelo AR(2).

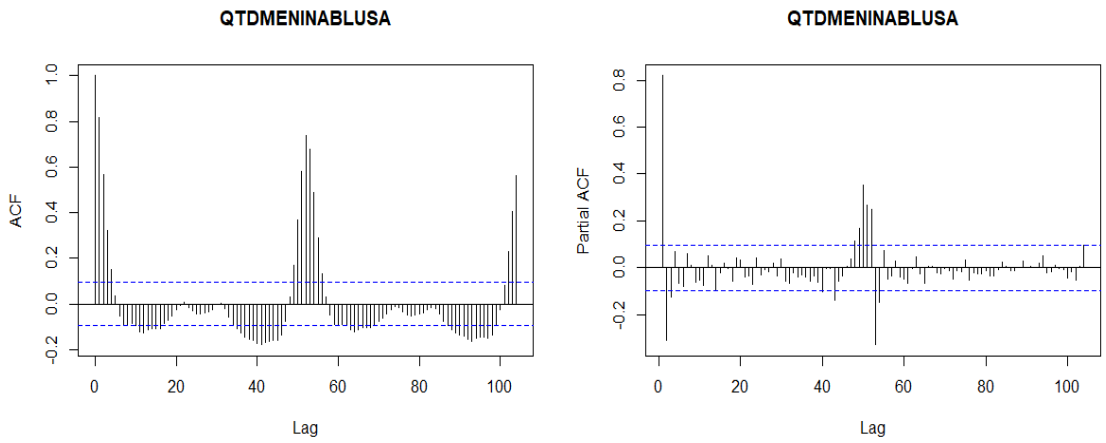


Figura 50 – Gráfico FAC e FACP da série 4 – Blusa de Menina
Fonte: elaboração própria.

Com base na análise dos gráficos da Figura 51 foram sugeridas algumas alternativas de modelos para moldar a série em análise. Dentre as sugestões, foram incluídos alguns termos de média móvel com o objetivo de se encontrar modelos mais aderentes. Neste sentido, as alternativas consideradas são:

- SARIMA $(2,0,3)(0,1,0)_{52}$
- SARIMA $(1,0,1)(1,1,0)_{52}$
- SARIMA $(2,0,2)(1,1,0)_{52}$
- SARIMA $(1,0,0)(1,1,0)_{52}$
- SARIMA $(1,0,2)(0,1,0)_{52}$
- SARIMA $(2,0,2)(0,1,0)_{52}$
- SARIMA $(3,0,2)(1,1,0)_{52}$

Depois de estabelecer os modelos, passa-se para a etapa de validação, que ocorre através da análise dos resíduos, a qual, para todos os modelos especificados acima, não identificou a presença de autocorrelação entre os termos. Contudo, verificou-se que os resíduos são heterocedásticos, o que significa que os estimadores deixam de ser eficientes, mas continuam sendo válidos, assim como não são normalmente distribuídos.

Como todas alternativas de modelos verificadas geraram resíduos bem-comportados, a escolha do modelo mais apropriado foi aquele que apresentou o menor MAPE no conjunto de dados treino, indicando maior acuracidade. Neste caso foi o modelo SARIMA (2,0,3)(0,1,0)52, o qual apresentou uma MAPE de 20% e se mostrou ser a alternativa mais parcimoniosa, já que é o modelo com o menor critério de AIC.

Tabela 20 – Resumo dos resultados do modelo 2 para série 4 – Blusa de Menina

Modelos	MAPE Treino	AIC
SARIMA(2,0,3)(0,1,0)52	**20,02%**	**6.030,67**
SARIMA(1,0,1)(1,1,0)52	20,29%	6.039,62
SARIMA(2,0,2)(1,1,0)52	21,02%	6.030,67
SARIMA(1,0,0)(1,1,0)52	20,49%	6.040,46
SARIMA(1,0,2)(0,1,0)52	21,42%	6.040,57
SARIMA(2,0,2)(0,1,0)52	20,93%	6.041,37
SARIMA(3,0,2)(1,1,0)52	20,59%	6.036,16

Fonte: elaboração própria.

Da mesma forma como foi feito para as outras séries, é importante apresentar os resultados da análise dos resíduos da alternativa selecionada, visando comprovar sua adequação. Logo, a FAC e a FACP dos resíduos, juntamente com o teste de Ljung-Box, evidenciam a não existência de autocorrelação. Já o teste ARCH-LM permite afirmar que os resíduos são heterocedásticos e, o teste de normalidade de Shapiro, que os mesmos não são normalmente distribuídos.

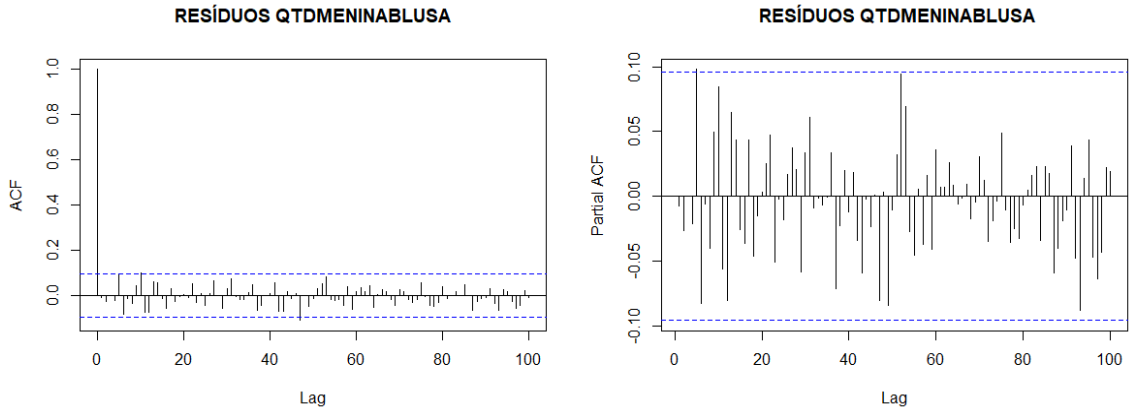


Figura 51 – Gráfico FAC e FACP dos resíduos do modelo SARIMA(2,0,3)$(0,1,0)_{52}$ para a série 4 – Blusa de Menina

Fonte: elaboração própria.

Tabela 21 – Resumo dos testes de correlação, homocedasticidade e normalidade dos resíduos do modelo SARIMA(2,0,3)(0,1,0)$_{52}$ para série 4 – Blusa de Menina

Testes Resíduos	Teste Autocorrelação Ljung-Box	Teste Heterocedasticidade LM teste	Teste Normalidade Shapiro teste
p-value	0,889	-	-

Fonte: elaboração própria.

Por fim, cabe destacar que a aplicação do modelo escolhido para o conjunto de dados testes da série de blusa de menina gerou um MAPE de 64,91%, que foi o pior resultado obtido neste tipo de conjunto de dados para todas as séries analisadas. A série de blusa de menina também foi a que apresentou a pior evolução do MAPE, mais especificamente, do conjunto de treino para o de teste entre as séries estudadas. Na Figura 52, é possível ver a comparação entre os dados reais e os previstos para esta série para a alternativa de modelo selecionada.

Modelo SARIMA - QTDMENINABLUSA	MAPE Treino	MAPE Teste
SARIMA(2,0,3)(0,1,0)$_{52}$	20,02%	64,91%

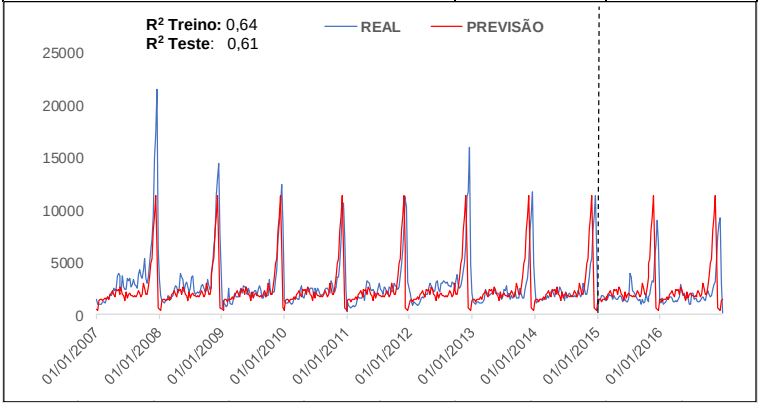


Figura 52 – Gráfico com resultados da melhor alternativa método 2 para a série 4 – Blusa de Menina
Fonte: elaboração própria.

5.3. Método 3 – SARIMA com exógenas

O próximo método é o mesmo método SARIMA com a consideração das variáveis exógenas, ou seja, o modelo passa a ser multivariado.

Diante disto, antes de realizar a análise para cada série, é necessário verificar a estacionariedade das variáveis exógenas, pois este é um pré-requisito para a utilização da metodologia Box-Jenkins. A seguir, são apresentados os resultados dos testes de raiz unitária: ADF, PP e KPSS, considerando as três modalidades possíveis: com

intercepto, como tendência e intercepto e sem tendência e intercepto de cada uma das variáveis exógenas

Tabela 22 – Resumo dos testes de raiz unitária para as variáveis exógenas

TMED	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e sem intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 8,98	- 11,57	0,02	- 9,00	- 11,56	0,02	- 0,54	- 0,64	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,97	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

ICC	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 0,75	0,76	1,71	- 1,68	- 1,68	0,57	- 0,70	- 0,70	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,97	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,56	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

PREC	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 6,97	- 21,47	0,08	- 7,00	- 21,47	0,03	- 2,97	- 17,80	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,42	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,41	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

MSAD_REAL	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 1,65	- 1,50	3,09	- 1,20	- 0,95	0,68	- 3,60	- 8,14	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,98	- 3,97	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

CRED_PF	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 3,45	- 3,99	3,05	- 0,85	- 0,18	0,60	2,75	5,82	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,98	- 3,97	0,22	- 2,56	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,00	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

TAXA_DESEMPREGO	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 1,16	- 1,16	0,79	- 1,61	- 1,45	0,68	- 0,01	0,13	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,98	- 3,98	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

PC_VEST	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 4,92	- 7,05	0,38	- 5,01	- 7,03	0,36	- 0,45	- 1,37	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,98	- 3,98	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	- 0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

PMHOMEMCAMISETA	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 3,88	- 4,09	0,71	- 4,92	- 5,20	0,45	- 0,49	- 0,58	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,97	0,33	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,47	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

PMMENINOCAMISETA	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 4,89	- 4,73	0,63	- 5,02	- 4,85	0,44	- 0,49	- 0,48	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,97	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

PMMULHERBLUSA	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 4,46	- 4,18	1,61	- 6,59	- 5,27	0,29	- 0,45	- 0,56	
T crítico - 1%	- 3,44	- 3,44	0,74	- 3,97	- 3,97	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,35	- 3,13	- 3,13	0,12	- 1,62	- 1,62	

PMMENINABLUSA	Teste com intercepto			Teste com tendência e intercepto			Teste sem tendência e intercepto		
ESTATÍSTICAS	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
T calculado	- 5,44	- 5,64	0,92	- 5,78	- 6,02	0,33	- 0,49	- 0,57	
T crítico - 1%	- 3,44	- 3,44	- 0,74	- 3,98	- 3,97	0,22	- 2,57	- 2,57	
T crítico - 5%	- 2,87	- 2,87	- 0,46	- 3,42	- 3,42	0,15	- 1,94	- 1,94	
T crítico - 10%	- 2,57	- 2,57	0,45	- 3,13	3,13	0,12	- 1,62	- 1,62	

Fonte: elaboração própria.

A análise da Tabela 22, com base na comparação da estatística *t* calculada contra as estatísticas *ts* tabeladas para os níveis de significância de 1%, 5% e 10% para cada uma das variáveis exógenas, torna possível afirmar que as variáveis TMED (temperatura média), PREC (precipitação), PC_VEST (índice de volume vendas do

varejo de moda), CRED_PF (saldo disponível de crédito a pessoas físicas com participação do PIB%), PMHOMEMCAMISETA (preço médio para as camisetas de homem), PMMULHERBLUSA (preço médio para as blusas de mulher), PMMENINABLUSA (preço médio para as blusas de menina) e PMMENINOCAMISETA (preço médio para camisetas de menino) são estacionárias.

Já as variáveis ICC (índice de confiança do consumidor), TAXA_DESEMPREGO (taxa de desemprego na região de metropolitana de SP) e a MSAD_REAL (massa salarial disponível real) não são estacionárias, o que gerou a necessidade de ser feita a primeira diferença para que tais séries se tornassem estacionárias.

Em relação às variáveis *dummies*, não é necessário realizar o teste de estacionariedade, pois elas sempre são estacionárias.

Deve-se destacar que, em alguns casos, o resultado do teste do KPSS foi diferente do teste ADF e do teste PP. Neste caso, o teste ADF foi considerado referência para a confirmação da estacionariedade.

O próximo passo da metodologia Box-Jenkins é a identificação do modelo, o qual tem o mesmo resultado do modelo SARIMA do segundo método. A mudança que ocorre é a constatação das variáveis exógenas que são significantes para modelar cada série em questão. Esta verificação se faz por meio da comparação do *p-value* de cada variável com o nível de significância estabelecido. Para cada uma das séries estudadas foram selecionadas as duas melhores alternativas SARIMA e aplicada duas variações de consideração de variáveis exógenas, gerando no total quatro possíveis modelos para cada série no terceiro método.

Série 1 – Camiseta de Homem

No caso da série da camiseta de homem, as variáveis exógenas significativas e que devem permanecer no modelo são: DUMMY_NATAL, PC_VEST e PMHOMEMCAMISETA.

No segundo método foram elencadas sete alternativas, já neste método consideraram-se os dois melhores métodos em termos de MAPE do segundo método com as variações contemplando todas as variáveis exógenas e somente as variáveis exógenas significativas, totalizando as quatro alternativas abaixo:

- SARIMA(1,0,0)(0,1,0)52 XREG(significativas)
- SARIMA(1,0,1)(0,1,0)52 XREG(todas)

- SARIMA(1,0,1)(0,1,0)52 XREG(todas)
- SARIMA(1,0,1)(0,1,0)52 XREG(significativas)

Os quatro modelos avaliados apresentaram resíduos não correlacionados, homocedásticos e não normais. Diante do fato dos resíduos terem um bom comportamento, definiu-se como modelo mais adequado aquele com menor MAPE, que também foi o mais parcimonioso, devido ao menor critério AIC. Então, o modelo escolhido foi o SARIMA (1,0,0)(0,1,0)52 XREG(significativas) com MAPE no conjunto de treino de 18,78%.

Tabela 23 – Resumo dos resultados do modelo 3 para série 1 – Camiseta de Homem

Modelos	MAPE Treino	AIC
SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas)	**18,78%**	**5.883,03**
SARIMAX(1,0,1)(0,1,0)$_{52}$ XREG(todas)	19,26%	5.895,47
SARIMA(1,0,1)(0,1,0)$_{52}$ XREG(todas)	19,20%	5.897,46
SARIMA(1,0,1)(0,1,0)$_{52}$ XREG(significativas)	18,81%	5.885,02

Fonte: elaboração própria.

A Tabela 23 evidencia que o modelo com melhor MAPE é o SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas). Em relação ao segundo método, eliminou-se o termo MA e houve uma pequena melhora do valor do MAPE de treinamento. Assim como ocorreu no segundo método, os resíduos continuam sendo não autocorrelacionados e não normais. Porém, passaram a ser homocedásticos.

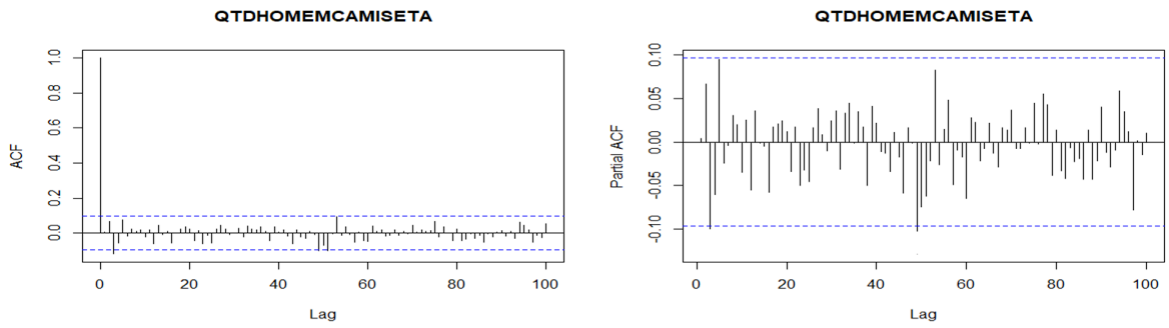


Figura 53 – Gráfico FAC e FACP dos resíduos do modelo SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas) para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Tabela 24 – Resumo dos testes de correlação, homocedasticidade e normalidade dos resíduos do modelo SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas) para série 1 – Camiseta de Homem

Testes Resíduos	Teste Autocorrelação Ljung-Box	Teste Heterocedasticidade LM teste	Teste Normalidade Shapiro teste
p-value	0,990	1,00	-

Fonte: elaboração própria.

Como último passo de cada método feito, é importante verificar o gráfico dos dados previsto contra os dados reais, assim como o MAPE do conjunto de dados de teste, o que é possível na Figura 54. Pode-se afirmar que o modelo SARIMA com as exógenas gerou previsões mais precisas tanto para o conjunto de treinamento como para o conjunto teste, quando comparado com o método SARIMA, atingindo um MAPE de teste de 44,62%.

Modelo SARIMA com exógenas - QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas)	18,78%	44,62%

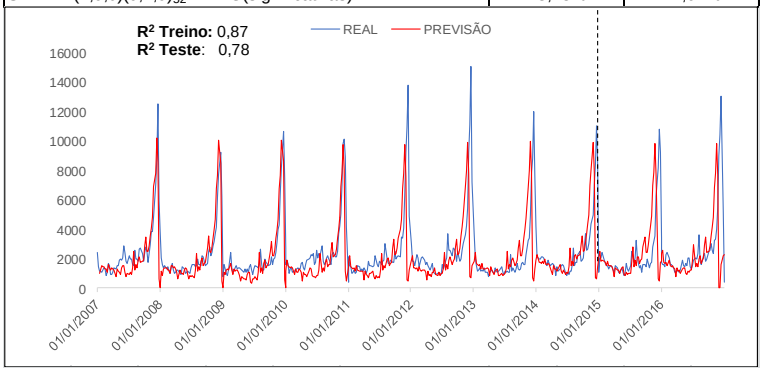


Figura 54 – Gráfico com resultados da melhor alternativa método 3 para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Série 2 – Blusa de Mulher

Para a série de blusa de mulher, as variáveis exógenas significativas e que devem ser contempladas em possíveis modelos são: DUMMY_NATAL, DUMMY_DEZ, CRED_PF e PMMULHERBLUSA.

Na implementação do terceiro método foram consideradas as duas melhores alternativas em termos de MAPE do segundo método, onde se levou em conta sete modelos. Nestas duas alternativas foram feitas duas variações: (1) com todas as variáveis exógenas e (2) somente as variáveis exógenas significativas. Abaixo, temos as quatro alternativas que são analisadas:

- SARIMA (2,0,2)(1,1,0)$_{52}$ XREG(significativas)
- SARIMA (2,0,2)(1,1,0)$_{52}$ XREG(todas)
- SARIMA (3,0,2)(1,1,0)$_{52}$ XREG(todas)
- SARIMA (3,0,2)(1,1,0)$_{52}$ XREG(significativas)

Seguindo a metodologia de Box-Jenkins, o próximo passo é a validação dos modelos selecionados, por meio da análise de resíduos. Todas as alternativas acima apresentaram resíduos bem-comportados e não normais. Contudo, os modelos avaliados passaram a ser homocedásticos, diferentemente do que ocorreu no método SARIMA sem exógenas.

Tal fato nos leva ao próximo passo, que é a seleção do modelo mais apropriado, o que é feito com base no modelo com menor MAPE, que foi o SARIMA $(2,0,2)(1,1,0)_{52}$ XREG(significativas), com MAPE de 17,79% e com o segundo menor critério de AIC. Na Tabela 25, temos um quadro resumo dos resultados em termos de MAPE e critério de AIC para cada alternativa.

Tabela 25 – Resumo dos resultados do modelo 3 para série 2 – Blusa de Mulher

Modelos	MAPE Treino	AIC
SARIMA(2,0,2)(1,1,0)$_{52}$ XREG(significativas)	**17,79%**	**6.191,92**
SARIMA(2,0,2)(1,1,0)$_{52}$ XREG(todas)	17,93%	6.205,98
SARIMA(3,0,2)(1,1,0)$_{52}$ XREG(todas)	18,04%	6.206,16
SARIMA(3,0,2)(1,1,0)$_{52}$ XREG(significativas)	17,97%	6.191,58

Fonte: elaboração própria.

Como foi feito no segundo método, é importante tangibilizar a análise dos resíduos do modelo escolhido, comprovando a ausência de correlação entre os termos, o que pode ser visto pelos gráficos da FAC e da FACP.

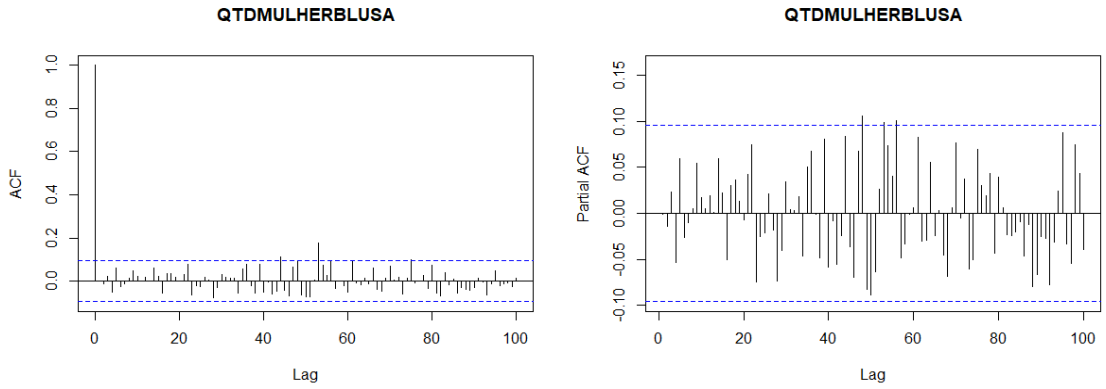


Figura 55 – Gráfico FAC e FACP dos resíduos do modelo SARIMA(2,0,2)(1,1,0)$_{52}$ XREG(significativas) para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

A ausência da autocorrelação também pode ser comprovada pelo resultado do teste Ljung-Box. A presença de homocedasticidade e a falta de normalidade podem ser constatados pelos resultados dos testes respectivamente ARCH-LM e de Shapiro.

Tabela 26 – Resumo dos testes de correlação, homocedasticidade e normalidade dos resíduos do modelo SARIMA(2,0,2)10,1,0)$_{52}$ XREG(significativas) para série 2 – Blusa de Mulher

Testes Resíduos	Teste Autocorrelação Ljung-Box	Teste Heterocedasticidade LM teste	Teste Normalidade Shapiro teste
p-value	0,151	1,00	-

Fonte: elaboração própria.

O resultado do MAPE do modelo selecionado para a série de blusa de mulher no conjunto de dados testes foi de 53,87%, bem inferior ao resultado obtido no conjunto de dados treino de 17,79%. Na Figura 56, podem-se verificar os dados reais e os previstos pelo modelo escolhido.

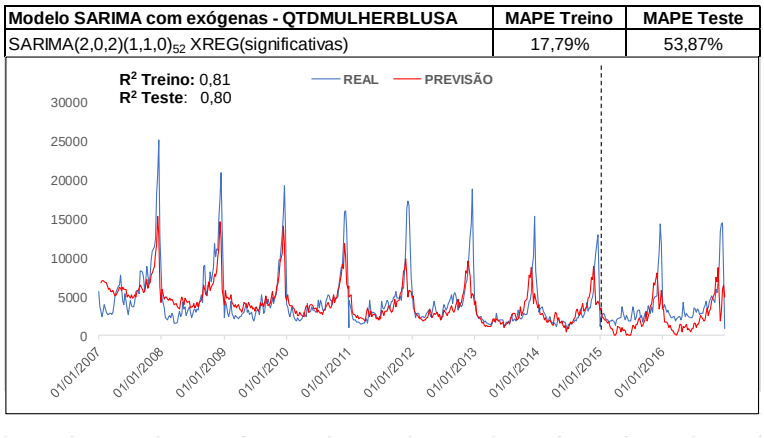


Figura 56 – Gráfico com resultados da melhor alternativa método 3 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

No caso da série de camiseta de menino, as variáveis exógenas significativas, que apresentaram um *p-value* baixo frente ao nível de significância de 5%, foram as seguintes: DUMMY_DEZ, MSAD_REAL e PMMENINOCAMISETA.

Como foi feito para as demais séries, o terceiro método é uma extensão do segundo método, pois a partir do último selecionaram-se as duas alternativas de modelos com maior acuracidade MAPE e se aplicou as variações de consideração de todas as variáveis exógenas e somente as variáveis exógenas significativas, gerando assim quatro modelos possíveis:

- SARIMA (1,0,0)(0,1,0)52 XREG(significativas)
- SARIMA (1,0,0)(0,1,0)52 XREG(todas)
- SARIMA (1,0,0)(1,1,0)52 XREG(todas)

- SARIMA (1,0,0)(1,1,0)52 XREG(significativas)

Os modelos, acima, evidenciaram resíduos bem-comportados, ou seja, não correlacionados, homocedásticos e não normalmente distribuídos. Como todas as alternativas são válidas, conforme a metodologia Box-Jenkins, optou-se pela opção com o menor MAPE e menor critério de AIC, que foi o modelo SARIMA $(1,0,0)(0,1,0)_{52}$ XREG(significativas), o qual apresentou 19,55% de MAPE no conjunto de treino. A seguir, temos a Tabela 27 com o resumo dos resultados de cada opção.

Tabela 27 – Resumo dos resultados do modelo 3 para série 3 – Camiseta de Menino

Modelos	MAPE Treino	AIC
SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas)	**19,55%**	**5.853,77**
SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(todas)	19,86%	5.869,62
SARIMA(1,0,0)(1,1,0)$_{52}$ XREG(todas)	20,03%	5.871,40
SARIMA(1,0,0)(1,1,0)$_{52}$ XREG(significativas)	19,81%	5.855,03

Fonte: elaboração própria.

É importante evidenciar a análise dos resíduos do modelo adotado, justificando assim a sua validação. Logo, com base nos gráficos FAC e FACP dos resíduos, é possível afirmar que não há autocorrelação.

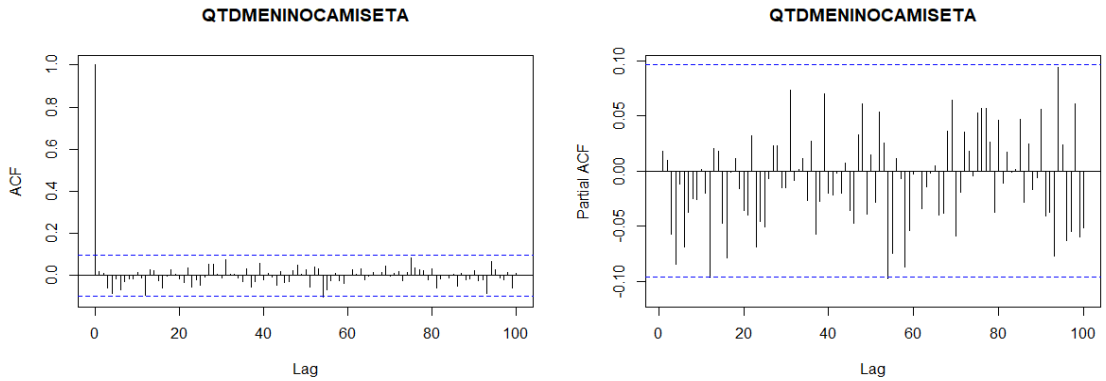


Figura 57 – Gráfico FAC e FACP dos resíduos do modelo SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas) para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

O teste de Ljung-Box também comprova a ausência de autocorrelação, já que a hipótese nula não é rejeitada. O teste de ARCH-LM comprova a presença de homocedasticidade nos resíduos do modelo selecionado e o teste Shapiro constata a falta de normalidade nos mesmos.

Tabela 28 – Resumo dos testes de correlação, homocedasticidade e normalidade dos resíduos do modelo SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas) para série 3 – Camiseta de Menino

Testes Resíduos	Teste Autocorrelação Ljung-Box	Teste Heterocedasticidade LM teste	Teste Normalidade Shapiro teste
p-value	0,997	1,00	-

Fonte: elaboração própria.

Por fim, verificou-se que a aplicação do modelo escolhido para os dados testes gerou um MAPE de 54,62%, bem menos preciso do que o MAPE de 19,55% do conjunto de treino. Abaixo, na Figura 58, pode-se fazer a comparação dos dados reais contra o previsto para a série de camiseta de menino.

Modelo SARIMA com exógenas - QTDMENINOCAMISETA	MAPE Treino	MAPE Teste
SARIMA(1,0,0)(0,1,0)$_{52}$ XREG(significativas)	19,55%	54,62%

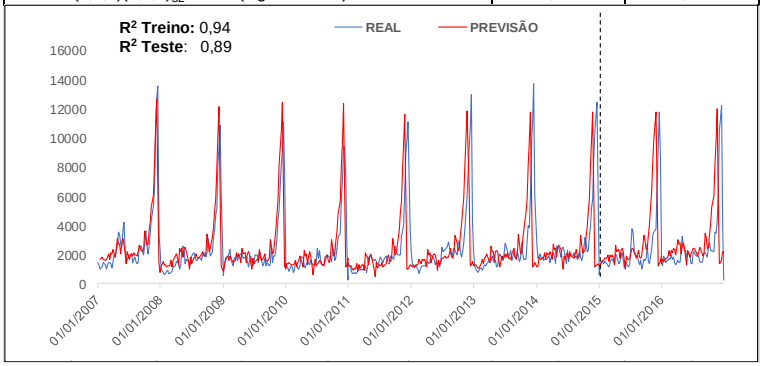


Figura 58 – Gráfico com resultados da melhor alternativa método 3 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

Para consideração de variáveis exógenas em um modelo multivariável, é fundamental que as variáveis exógenas sejam significativas. Para a série de blusa de menina, as variáveis significativas são somente a DUMMY_DEZ e PMMENINABLUSA. Dentre as séries estudadas, esta foi a série que apresentou a menor quantidade de variáveis exógenas significativas.

Como já foi feito com as outras séries, neste caso também foi selecionado os melhores modelos do segundo método e aplicada as variações de inclusão de todas as exógenas e de inclusão somente das variáveis exógenas significativas, totalizando quatro alternativas possíveis, conforme abaixo:

- SARIMA (2,0,3)(0,1,0)52 XREG(significativas)
- SARIMA (2,0,3)(0,1,0)52 XREG(todas)
- SARIMA (1,0,1)(1,1,0)52 XREG(todas)
- SARIMA (1,0,1)(1,1,0)52 XREG(significativas)

Na metodologia Box-Jenkins, após a identificação e estimação do modelo, passe-se à fase de verificação, onde os resíduos são analisados e o objetivo é encontrar resíduos não autocorrelacionados, homocedásticos e normalmente distribuídos. Os quatro modelos acima apresentaram resíduos bem-comportados, homocedásticos e não normais.

Diante disto, a escolha do modelo mais apropriado foi feita com base no modelo que obteve o menor MAPE, isto é, que apresentou a maior precisão no conjunto de dados de treino, que foi a opção SARIMA (2,0,3)(0,1,0)$_{52}$ XREG(significativas), a qual evidenciou o segundo melhor critério de AIC.

Tabela 29 – Resumo dos resultados do modelo 3 para série 4 – Blusa de Menino

Modelos	MAPE Treino	AIC
SARIMA(2,0,3)(0,1,0)$_{52}$ XREG(significativas)	**19,52%**	**6.014,02**
SARIMA(2,0,3)(0,1,0)$_{52}$ XREG(todas)	19,67%	5.994,07
SARIMA(1,0,1)(1,1,0)$_{52}$ XREG(todas)	21,32%	6.892,54
SARIMA(1,0,1)(1,1,0)$_{52}$ XREG(significativas)	19,99%	6.016,92

Fonte: elaboração própria.

Na Figura 59, são apresentados os gráficos da FAC e da FACP dos resíduos do modelo selecionado, juntamente com os resultados do teste Ljung-Box, ARCH-LM e de Shapiro. Neste sentido, na Figura 60, é possível constatar a falta de autocorrelação dos resíduos, mostrando que eles são bem-comportados.

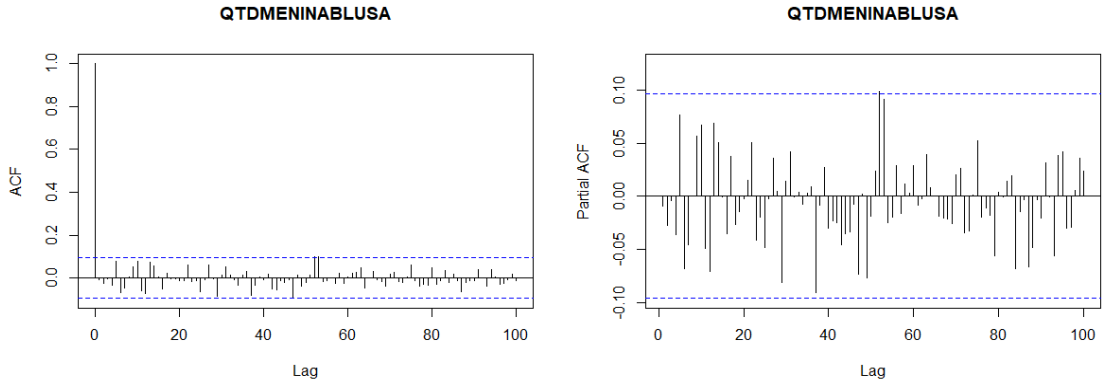


Figura 59 – Gráfico FAC e FACP dos resíduos do modelo SARIMA(2,0,3)(0,1,0)$_{52}$ XREG(significativas) para a série 4 – Blusa de Menina
Fonte: elaboração própria.

O teste de Ljung-Box ratifica a ausência de autocorrelação, o teste ARCH-LM comprova a presença de homocedasticidade e, o de Shapiro, a falta de normalidade, tendo em vista um nível de significância de 5%.

Tabela 30 – Resumo dos testes de correlação, homocedasticidade e normalidade dos resíduos do modelo SARIMA(2,0,3)(0,1,0)$_{52}$ XREG(significativas) para série 4 – Blusa de Menina

Testes Resíduos	Teste Autocorrelação Ljung-Box	Teste Heterocedasticidade LM teste	Teste Normalidade Shapiro teste
p-value	0,985	1,00	-

Fonte: elaboração própria.

O modelo adotado para a série de blusa de menina apresentou resultado ruins em termos de MAPE para o conjunto de dados de teste, mais especificamente um erro percentual médio absoluto de cerca de 70%, o qual foi muito maior do que o valor encontrado no conjunto de dados treino. Na Figura 60, temos o gráfico comparativo com os dados reais e os dados previstos do modelo escolhido.

Modelo SARIMA com exógenas - QTDBLUSAMENINA	MAPE Treino	MAPE Teste
SARIMA(2,0,3)(0,1,0)$_{52}$ XREG(significativas)	19,52%	70,45%

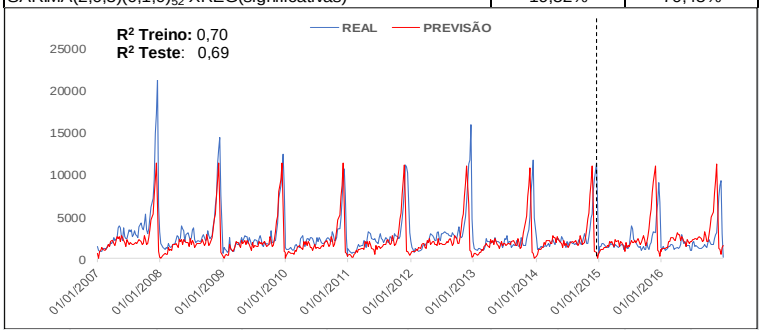


Figura 60– Gráfico com resultados da melhor alternativa método 3 para a série 4 – Blusa de Menina
Fonte: elaboração própria

5.4. Método 4 – SARIMA GARCH

A modelagem GARCH é um método não linear, que tem como principal objetivo tratar a presença de heterocedasticidade condicional, constatada no segundo método. A presença da alta sazonalidade no mês de dezembro pode justificar a maior parte da heterocedasticidade presente nas séries. Verificada a existência da heterocedasticidade condicional é fundamental a especificação de um modelo ARCH

ou GARCH, lembrando que o GARCH é uma modelagem mais generalista e o uso desta abordagem tem como objetivo garantir que valores dos coeficientes dos AR e dos MA sejam ajustados e apresentados da forma adequada.

A identificação da ordem de um modelo GARCH é complexa. O procedimento recomendado é a realização de vários modelos de baixa ordem, verificando a significância dos termos dos quadrados dos resíduos nos *p* períodos anteriores, e dos termos da variância condicional nos *q* períodos anteriores. Neste sentido, devem ser consideradas apenas as sugestões, onde os termos *p* e *q* são significativos.

O mesmo procedimento adotado no terceiro método foi incorporado no método em análise. Assim, foram escolhidas as melhores alternativas de modelos do segundo método, nas quais foram aplicados alguns modelos GARCH, totalizando quatro alternativas de modelos. Em todas as séries estudadas, não foi necessário mais de um termo *p* e/ou *q*. Quando se tentou incluir mais termos, constatou-se que eles não eram significativos. Abaixo, apresenta-se o resultado obtido para cada série.

Série 1 – Camiseta de Homem

Na especificação de um modelo SARIMA GARCH, seguindo a metodologia Box-Jenkins, deve-se realizar a identificação dos termos autorregressivos e de médias móveis, por meio da análise da FAC e da FACP, o que já foi feito no segundo método. O próximo passo é a modelagem da heterocedasticidade condicional, por meio dos modelos GARCH, tendo em vista que a mesma foi verificada também no segundo método.

Para a série temporal camiseta de homem foram analisadas as seguintes alternativas de modelos GARCH, com base na seleção dos melhores modelos em termos de MAPE do segundo método:

- SARIMA $(1,0,1)(0,1,0)_{52}$ GARCH(1,0)
- SARIMA $(1,0,0)(0,1,0)_{52}$ GARCH(1,1)
- SARIMA $(1,0,0)(0,1,0)_{52}$ GARCH(1,0)
- SARIMA $(1,0,1)(0,1,0)_{52}$ GARCH(1,1)

Com base na Tabela 31 verifica-se que as alternativas de modelos propostas possuem medidas de acuracidade muito próximas. Porém, a alternativa mais precisa foi o modelo: SARIMA $(1,0,1)(0,1,0)_{52}$ GARCH(1,0).

Tabela 31 – Resumo dos resultados do modelo 4 para série 1 – Camiseta de Homem

Modelos	MAPE Treino
SARIMA$(1,0,1)(0,1,0)_{52}$ GARCH(1,0)	**18,88%**
SARIMA$(1,0,0)(0,1,0)_{52}$ GARCH(1,1)	18,90%
SARIMA$(1,0,0)(0,1,0)_{52}$ GARCH(1,0)	18,92%
SARIMA$(1,0,1)(0,1,0)_{52}$ GARCH(1,1)	18,90%

Fonte: elaboração própria.

Por último, deve-se evidenciar o gráfico com os dados previstos contra os dados reais, juntamente com o MAPE no conjunto de dados de teste. Verifica-se que os resultados obtidos foram muito próximos do método SARIMA, mais especificamente, apresentam uma pequena piora na aplicação do teste. Logo, a aplicação da modelagem GARCH não contribuiu para melhorar as previsões em si, somente contribuiu para solucionar o problema de heterocedastidade condicional da variância do erro.

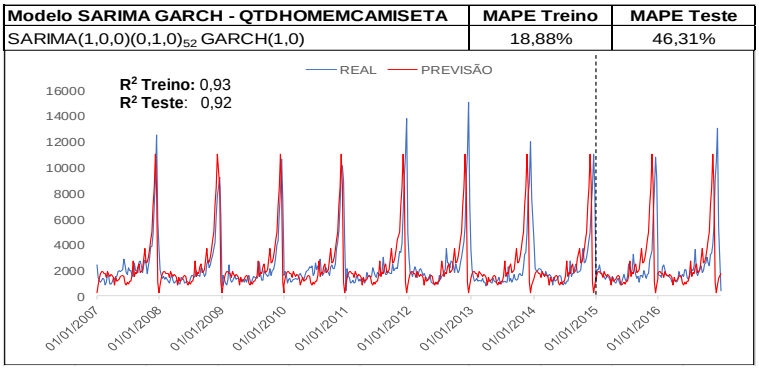


Figura 61 – Gráfico com resultados da melhor alternativa método 4 para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Série 2 – Blusa de Mulher

Constatado o problema de heterocedasticidade na previsão por meio do método SARIMA, uma abordagem para solucionar este problema é o GARCH, que modela a heterocedastidade condicional. Para a especificação de um modelo SARIMA GARCH, deve-se seguir também a metodologia Box-Jenkins. Considerando que os passos iniciais desta metodologia já foram realizados no segundo método, fica pendente somente a parte do GARCH.

Para determinar o modelo GARCH mais adequado selecionamos as duas melhores alternativas de modelos SARIMA desta série e, como visto na revisão teórica, geralmente são necessários poucos termos GARCH para conseguir modelar a heterocedastidade condicional. A seguir, são apresentados os modelos considerados:

- SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,1)
- SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,0)
- SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,1)
- SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,0)

Das alternativas acima, o modelo com maior acuracidade MAPE no conjunto de dados treino é o SARIMA (3,0,2)(1,1,0)$_{52}$ GARCH(1,1) com 18,96%. Vale ressaltar que os MAPEs das demais possibilidades foram extremamente próximos do modelo selecionado, conforme pode ser verificado na Tabela 32.

Tabela 32 – Resumo dos resultados do modelo 4 para série 2 – Blusa de Mulher

Modelos	MAPE Treino
SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,1)	**18,96%**
SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,0)	18,99%
SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,1)	18,97%
SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,0)	18,99%

Fonte: elaboração própria.

No conjunto de dados teste, o MAPE obtido com o modelo selecionado foi de 45,32%, apresentando uma piora significativa em relação ao conjunto de dados treino. Na Figura 62, é feita a comparação dos dados reais contra os dados previstos. A aplicação da abordagem GARCH não melhorou a acuracidade da previsão da série, assim como ocorreu na série de camiseta de homem. Apenas o problema da heterocedasticidade foi solucionado.

Modelo SARIMA GARCH - QTDMULHERBLUSA	MAPE Treino	MAPE Teste
SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,1)	18,96%	45,32%

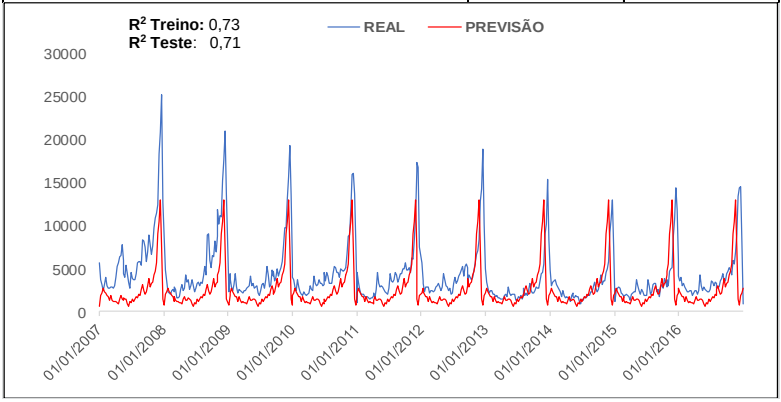


Figura 62 – Gráfico com resultados da melhor alternativa método 4 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

Na série de camiseta de menino, no segundo método, também foi verificado o problema de heterocedasticidade, sendo a abordagem GARCH uma possível solução para este problema. Neste sentido, procedeu-se com a aplicação da metodologia Box-Jenkins, a qual já teve seus passos aplicados no segundo método.

Diante disto, foram escolhidas as duas alternativas de modelos com melhor acuracidade MAPE e considerada duas variações de GARCH para cada uma delas. Não foram contempladas outras possibilidades, pois os termos GARCH não foram significativos e, na maioria das vezes, são necessários poucos termos *p* (autorregressivo) e *q* (resíduos) para modelar a heterocedasticidade condicional. A seguir, temos as opções de modelos contempladas:

- SARIMA (1,0,0)(0,1,0)$_{52}$ GARCH(1,0)
- SARIMA (1,0,0)(0,1,0)$_{52}$ GARCH(1,1)
- SARIMA (1,0,0)(1,1,0)$_{52}$ GARCH(1,0)
- SARIMA (1,0,0)(1,1,0)$_{52}$ GARCH(1,1)

A partir de tais alternativas, o modelo selecionado foi SARIMA (1,0,0)(0,1,0)$_{52}$ GARCH(1,0) por ter apresentado a melhor métrica de precisão com 21,34% no conjunto de dados treino, sendo que as demais alternativas apresentaram valores muito próximos.

Tabela 33 – Resumo dos resultados do modelo 4 para série 3 – Camiseta de Menino

Modelos	MAPE Treino
SARIMA(1,0,1)(0,1,0)$_{52}$ GARCH(1,0)	**18,88%**
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,1)	18,90%
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,0)	18,92%
SARIMA(1,0,1)(0,1,0)$_{52}$ GARCH(1,1)	18,90%

Fonte: elaboração própria.

A aplicação do modelo escolhido para o conjunto de dados teste gerou uma MAPE de 51,60%, que é um erro percentual absoluto médio superior ao verificado no conjunto dados treino e maior que o encontrado no segundo método. Sendo assim, a acuracidade da previsão não apresentou melhoras, somente o problema da heterocedasticidade condicional foi resolvido. A seguir, na Figura 63, é possível observar os dados reais e os dados previstos para a série de camiseta de menino, considerando o modelo adotado.

Modelo SARIMA GARCH - QTDMENINOCAMISETA	MAPE Treino	MAPE Teste
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,0)	21,34%	51,60%

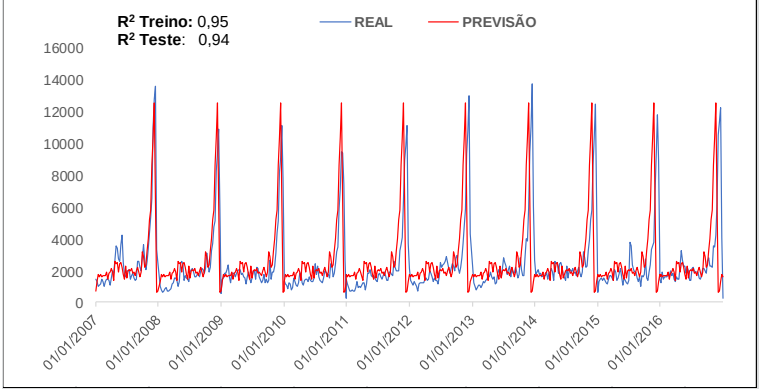


Figura 63 – Gráfico com resultados da melhor alternativa método 4 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

A última série a ser analisada para o método SARIMA GARCH é a série de blusa de menina, a qual também apresentou a característica de heterocedasticidade no método 2. O mesmo procedimento adotado para as outras séries foi aplicado para a série em questão, em conformidade com a metodologia Box-Jenkins.

Desta forma, foram selecionadas as duas melhores alternativas verificadas no segundo método e aplicada algumas possíveis variações para modelar a

heterocedasticidade condicional desta série, o que geralmente é feito com poucos termos p e q. Os possíveis modelos são:

- SARIMA $(1,0,1)(1,1,0)_{52}$ GARCH(1,0)
- SARIMA $(2,0,3)(0,1,0)_{52}$ GARCH(1,1)
- SARIMA $(2,0,3)(0,1,0)_{52}$ GARCH(1,0)
- SARIMA $(1,0,1)(1,1,0)_{52}$ GARCH(1,1)

Com base nas medidas MAPEs evidenciadas na Tabela resumo 34, o modelo escolhido foi o SARIMA$(1,0,1)(1,1,0)_{52}$ GARCH(1,0) com MAPE de 20,24%. A última alternativa apresentou o mesmo valor do MAPE que a opção selecionada. No entanto, optou-se pelo modelo mais parcimonioso, ou seja, com uma quantidade menor de termos na sua especificação.

Tabela 34 – Resumo dos resultados do modelo 4 para série 4 – Blusa de Menina

Modelos	MAPE Treino
SARIMA(1,0,1)(1,1,0)52 GARCH(1,0)	**20,24%**
SARIMA(2,0,3)(0,1,0)52 GARCH(1,1)	20,99%
SARIMA(2,0,3)(0,1,0)52 GARCH(1,0)	21,01%
SARIMA(1,0,1)(1,1,0)52 GARCH(1,1)	20,24%

Fonte: elaboração própria.

No conjunto de dados teste, o modelo escolhido apresentou um MAPE de 64,22%, que é um erro muito maior do que o constatado no conjunto de dados treino. Porém, diferentemente das outras séries, na série de blusa de menina a modelagem GARCH conseguiu solucionar a questão de heterocedasticidade e fazer previsões mais precisas no período de teste. Esta foi a única série em que nos dados testes ocorreram ganhos de precisão com a modelagem GARCH. Abaixo, na Figura 64, temos os dados previstos e os dados reais, tornando possível a realização de uma comparação efetiva.

Modelo SARIMA GARCH - QTDMENINABLUSA	MAPE Treino	MAPE Teste
SARIMA$(1,0,1)(1,1,0)_{52}$ GARCH(1,0)	20,24%	64,22%

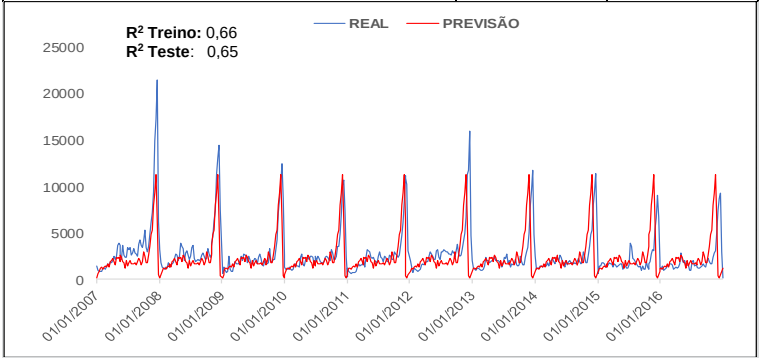


Figura 64– Gráfico com resultados da melhor alternativa método 4 para a série 4 – Blusa de Menina

Fonte: elaboração própria.

5.5. Método 5 – SARIMA GARCH com exógenas

O último método do grupo de métodos estatísticos clássicos efetivado é a inclusão das variáveis exógenas no método SARIMA GARCH. O intuito é verificar os resultados que podem ser obtidos com base em um modelo multivariado, em outras palavras, verificar o impacto das variáveis exógenas.

Para todas as séries analisadas foram considerados os modelos SARIMA GARCH evidenciados no quinto método e, para cada uma das alternativas de modelo proposta, são consideradas duas variações, totalizando oito alternativas: (1) considerando todas as variáveis exógenas e (2) somente são consideradas as variáveis exógenas significativas. Deve-se ressaltar que para cada série temporal existe um conjunto de variáveis externas significativas, que são as mesmas do terceiro método. A seguir é feita a avaliação deste método para cada uma das séries estudadas neste trabalho.

Série 1 – Camiseta de Homem

Considerando os modelos SARIMA GARCH verificados no método anterior, foram especificadas oito alternativas da seguinte maneira: quatro alternativas com todas as variáveis e quatro alternativas somente com as variáveis significativas. No caso da série temporal de camiseta de homem as variáveis significativas são: DUMMY_NATAL, PC_VEST e PMHOMEMCAMISETA. Abaixo, seguem as especificações dos modelos:

- SARIMA (1,0,0)$(0,1,0)_{52}$ GARCH(1,0) XREG(significativas)
- SARIMA (1,0,1)$(0,1,0)_{52}$ GARCH(1,0) XREG(todas)
- SARIMA (1,0,0)$(0,1,0)_{52}$ GARCH(1,1) XREG(todas)
- SARIMA (1,0,0)$(0,1,0)_{52}$ GARCH(1,0) XREG(todas)
- SARIMA (1,0,1)$(0,1,0)_{52}$ GARCH(1,1) XREG(todas)
- SARIMA (1,0,1)$(0,1,0)_{52}$ GARCH(1,0) XREG(significativas)
- SARIMA (1,0,0)$(0,1,0)_{52}$ GARCH(1,1) XREG(significativas)
- SARIMA (1,0,1)$(0,1,0)_{52}$ GARCH(1,1) XREG(significativas)

O modelo que apresentou a maior precisão, em termos de MAPE, apresentou o valor de 18,77% e foi o mesmo encontrado no quarto método com a diferença da inclusão das variáveis exógenas significativas:

Tabela 35 – Resumo dos resultados do modelo 5 para série 1 – Camiseta de Homem

Modelos	MAPE Treino
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,0) XREG(significativas)	**18,77%**
SARIMA(1,0,1)(0,1,0)$_{52}$ GARCH(1,0) XREG(todas)	19,33%
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,1) XREG(todas)	19,31%
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,0) XREG(todas)	19,30%
SARIMA(1,0,1)(0,1,0)$_{52}$ GARCH(1,1) XREG(todas)	19,33%
SARIMA(1,0,1)(0,1,0)$_{52}$ GARCH(1,0) XREG(significativas)	18,80%
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,1) XREG(significativas)	18,78%
SARIMA(1,0,1)(0,1,0)$_{52}$ GARCH(1,1) XREG(significativas)	18,81%

Fonte: elaboração própria.

Ocorreu uma pequena melhora na acuracidade quando se compara com o modelo encontrado no método SARIMA GARCH.

O último ponto a ser verificado é o desempenho da melhor alternativa encontrada para o conjunto de dados testes e as suas previsões em relação aos dados reais. Neste caso, obteve-se um MAPE de teste de 44,55%. Abaixo, segue a Figura 65 com estas informações.

Modelo SARIMAX GARCH - QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,0)	18,77%	44,55%

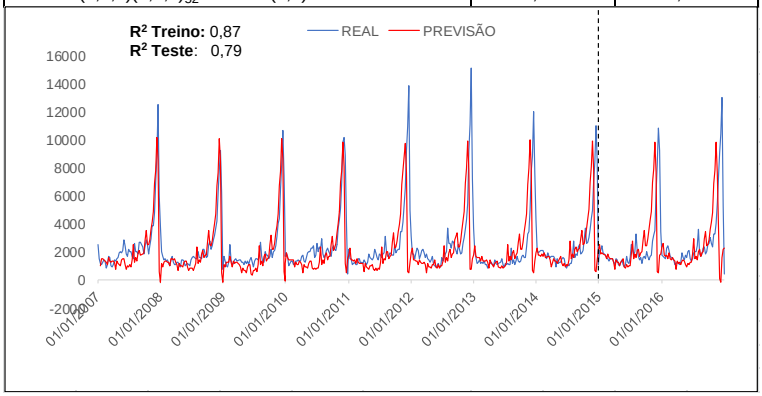


Figura 65 – Gráfico com resultados da melhor alternativa método 5 para a série 1 – Camiseta de Homem

Fonte: elaboração própria.

Série 2 – Blusa de Mulher

O quinto método é uma extensão do quarto método, tendo em vista que a única mudança é a inclusão das variáveis exógenas, as quais são as mesmas contempladas no terceiro método de previsão. No quarto método, foram consideradas quatro alternativas de modelo. Para cada uma delas foi considerada a variação de modelo com todas as variáveis exógenas e somente com as variáveis significativas, totalizando oito modelos para o método corrente, conforme pode ser visto abaixo:

- SARIMA (2,0,2)(1,1,0)$_{52}$ GARCH(1,0) XREG(significativas)
- SARIMA (3,0,2)(1,1,0)$_{52}$ GARCH(1,1) XREG(todas)
- SARIMA (2,0,2)(1,1,0)$_{52}$ GARCH(1,0) XREG(todas)
- SARIMA (2,0,2)(1,1,0)$_{52}$ GARCH(1,1) XREG(todas)
- SARIMA (3,0,2)(1,1,0)$_{52}$ GARCH(1,0) XREG(todas)
- SARIMA (3,0,2)(1,1,0)$_{52}$ GARCH(1,1) XREG(significativas)
- SARIMA (2,0,2)(1,1,0)$_{52}$ GARCH(1,1) XREG(significativas)
- SARIMA (3,0,2)(1,1,0)$_{52}$ GARCH(1,0) XREG(significativas)

Das oito alternativas analisadas, a que apresentou melhor MAPE, no conjunto de dados de treino, com 17,80% de erro, foi a SARIMA (2,0,2)(1,1,0)52 GARCH(1,0) XREG(significativas). Este valor também foi obtido com a sétima alternativa, com dois termos de GARCH, que não foi o modelo adotado por ser menos parcimonioso. Deve-se ressaltar que o modelo escolhido contemplou somente as variáveis exógenas significativas.

Tabela 36 – Resumo dos resultados do modelo 5 para série 2 – Blusa de Mulher

Modelos	MAPE Treino
SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,0) XREG(significativas)	**17,80%**
SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,1) XREG(todas)	18,00%
SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,0) XREG(todas)	17,93%
SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,1) XREG(todas)	17,93%
SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,0) XREG(todas)	18,01%
SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,1) XREG(significativas)	18,01%
SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,1) XREG(significativas)	17,80%
SARIMA(3,0,2)(1,1,0)$_{52}$ GARCH(1,0) XREG(significativas)	17,94%

Fonte: elaboração própria.

A inclusão das variáveis exógenas melhorou a acuracidade da previsão somente no conjunto de dados de treino, quando realizada a comparação com o resultado obtido no quarto método. No conjunto de dados teste houve uma piora do MAPE, o qual passou a ser de 53,18%. Logo, a inclusão das variáveis exógenas não contribuiu para criar um modelo mais assertivo. A seguir, é evidenciada a Figura 66 com os dados reais e os dados previstos com base no modelo escolhido.

Modelo SARIMAX GARCH - QTDMULHERBLUSA	MAPE Treino	MAPE Teste
SARIMA(2,0,2)(1,1,0)$_{52}$ GARCH(1,0)	17,80%	53,18%

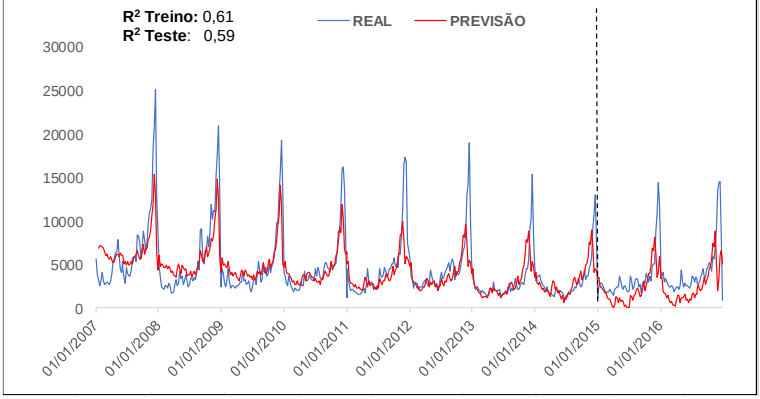


Figura 66 – Gráfico com resultados da melhor alternativa método 5 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

Para a série de camiseta de menino também foi realizado o mesmo procedimento feito para as séries de camiseta de homem e blusa de mulher para o quinto método. Mais especificamente, foram incluídas as variáveis exógenas no quarto método. Para isto, considerou-se duas variações para cada um dos modelos verificados no método anterior: com todas as variáveis exógenas e apenas as variáveis exógenas significativas, chegando ao total de oito modelos possíveis, os quais são:

- SARIMA (1,0,0)(0,1,0)$_{52}$ GARCH(1,1) XREG(significativas)
- SARIMA (1,0,0)(0,1,0)$_{52}$ GARCH(1,0) XREG(todas)
- SARIMA (1,0,0)(0,1,0)$_{52}$ GARCH(1,1) XREG(todas)
- SARIMA (1,0,0)(1,1,0)$_{52}$ GARCH(1,0) XREG(todas)
- SARIMA (1,0,0)(1,1,0)$_{52}$ GARCH(1,1) XREG(todas)
- SARIMA (1,0,0)(0,1,0)$_{52}$ GARCH(1,0) XREG(significativas)
- SARIMA (1,0,0)(1,1,0)$_{52}$ GARCH(1,0) XREG(significativas)

- SARIMA (1,0,0)(1,1,0)$_{52}$ GARCH(1,1) XREG(significativas)

Segundo o critério MAPE, dos oito modelos analisados, a alternativa SARIMA (1,0,0)(0,1,0)$_{52}$ GARCH(1,1) XREG(significativas) foi a que mostrou o menor MAPE, de 19,54% no conjunto de dados treino, o que pode ser visto na Tabela resumo 37.

Tabela 37 – Resumo dos resultados do modelo 5 para série 2 – Blusa de Mulher

Modelos	MAPE Treino
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,1) XREG(significativas)	**19,54%**
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,0) XREG(todas)	19,86%
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,1) XREG(todas)	19,84%
SARIMA(1,0,0)(1,1,0)$_{52}$ GARCH(1,0) XREG(todas)	20,01%
SARIMA(1,0,0)(1,1,0)$_{52}$ GARCH(1,1) XREG(todas)	20,00%
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,0) XREG(significativas)	19,69%
SARIMA(1,0,0)(1,1,0)$_{52}$ GARCH(1,0) XREG(significativas)	19,93%
SARIMA(1,0,0)(1,1,0)$_{52}$ GARCH(1,1) XREG(significativas)	19,79%

Fonte: elaboração própria.

É fundamental ressaltar que houve uma piora significativa do MAPE do modelo adotado para o conjunto de dados teste, que saiu de 19,54% no conjunto de dados treino para 54,78% para o conjunto de dados teste. Além disso, pode-se afirmar que a inclusão das variáveis exógenas não superou o modelo SARIMA GARCH em termos de acuracidade, indicando que a inclusão das variáveis exógenas não contribuiu para a construção de um modelo mais preciso. Por fim, é possível comparar os dados reais contra os dados previstos para o modelo selecionado na Figura 67.

Modelo SARIMAX GARCH - QTDMENINOCAMISETA	MAPE Treino	MAPE Teste
SARIMA(1,0,0)(0,1,0)$_{52}$ GARCH(1,1)	19,54%	54,78%

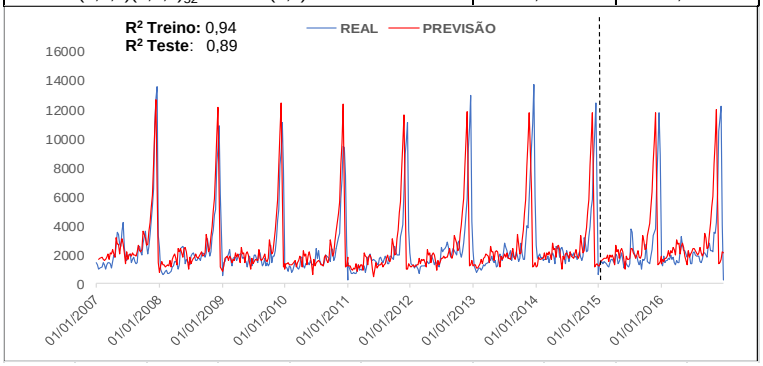


Figura 67 – Gráfico com resultados da melhor alternativa método 5 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

Para execução do método SARIMA GARCH para a série de blusa de menina, o procedimento foi incluir as variáveis exógenas na sua totalidade e somente as significativas nas duas melhores alternativas de modelos verificadas no quarto método. Desta forma, o objetivo foi realizar uma extensão do último método, através do aumento de *inputs*, visando alcançar uma maior assertividade da previsão.

Sendo assim, a partir dos quatro modelos do quarto método, foi inserida as duas variações consideradas até o momento para as outras séries, as quais são: (1) com todas as variáveis exógenas, e (2) somente as variáveis exógenas significativas, o que totalizou oito alternativas de modelos, as quais são:

- SARIMA (1,0,1)$(0,1,0)_{52}$ GARCH(1,0) XREG(significativas)
- SARIMA (2,0,3)$(0,1,0)_{52}$ GARCH(1,1) XREG(todas)
- SARIMA (2,0,3)$(0,1,0)_{52}$ GARCH(1,0) XREG(todas)
- SARIMA (1,0,1)$(0,1,0)_{52}$ GARCH(1,0) XREG(todas)
- SARIMA (1,0,1)$(0,1,0)_{52}$ GARCH(1,1) XREG(todas)
- SARIMA (1,0,1)$(10,1,0)_{52}$ GARCH(1,1) XREG(significativas)
- SARIMA (2,0,3)$(0,1,0)_{52}$ GARCH(1,0) XREG(significativas)
- SARIMA(2,0,3)$(0,1,0)_{52}$ GARCH(1,1) XREG(significativas)

A escolha do modelo mais apropriado se deu pelo *ranking* crescente do MAPE das alternativas no conjunto de dados de treino, onde foi constatado que o modelo com menor MAPE foi o SARIMA (1,0,1)(0,1,0)52 GARCH(1,0) XREG(significativas), com o valor de 19,34%.

Tabela 38– Resumo dos resultados do modelo 5 para série 4 – Blusa de Menina

Modelos	MAPE Treino
SARIMA(1,0,1)$(0,1,0)_{52}$ GARCH(1,0) XREG(significativas)	**19,34%**
SARIMA(2,0,3)$(0,1,0)_{52}$ GARCH(1,1) XREG(todas)	19,63%
SARIMA(2,0,3)$(0,1,0)_{52}$ GARCH(1,0) XREG(todas)	19,66%
SARIMA(1,0,1)$(0,1,0)_{52}$ GARCH(1,0) XREG(todas)	19,49%
SARIMA(1,0,1)$(0,1,0)_{52}$ GARCH(1,1) XREG(todas)	19,45%
SARIMA(1,0,1)$(10,1,0)_{52}$ GARCH(1,1) XREG(significativas)	19,34%
SARIMA(2,0,3)$(0,1,0)_{52}$ GARCH(1,0) XREG(significativas)	19,70%
SARIMA(2,0,3)$(0,1,0)_{52}$ GARCH(1,1) XREG(significativas)	19,60%

Fonte: elaboração própria.

Assim como ocorreu com as demais séries, no caso da série temporal de blusa de menina, houve uma piora significativa, em termos de acuracidade, da aplicação do

modelo selecionado no conjunto de treino para o conjunto de teste, indo de 19,34% para 70,89%. Além disso, foi possível verificar que as variáveis exógenas não ajudaram a gerar previsões mais precisas, evidenciando um MAPE pior que o do método SARIMA GARCH.

De todas as séries analisadas e para todos os métodos estatísticos clássicos estudados, a previsão encontrada para a série de blusa de menina, através do método SARIMA GARCH com variáveis exógenas, foi a que mostrou o pior resultado de MAPE.

Na Figura 68, é evidenciado o gráfico com os dados reais e com os dados previstos pelo modelo adotado para o método em questão para a série de blusa de menina.

Modelo SARIMAX GARCH - QTDMENINABLUSA	MAPE Treino	MAPE Teste
SARIMA(1,0,1)(0,1,0)$_{52}$ GARCH(1,0)	19,34%	70,89%

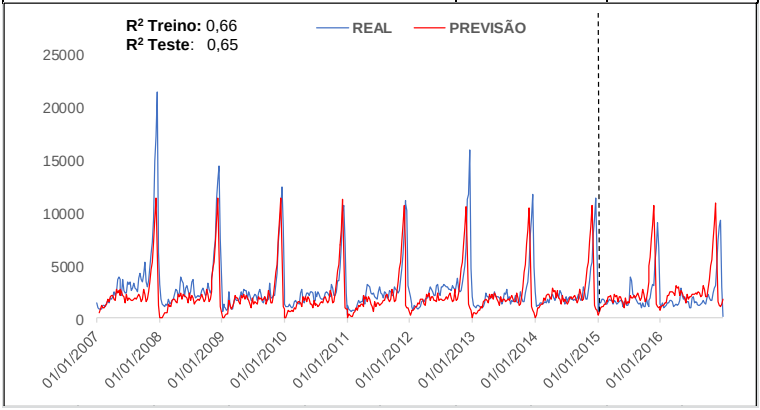


Figura 68 – Gráfico com resultados da melhor alternativa método 5 para a série 4 – Blusa de Menina
Fonte: elaboração própria.

5.6. Método 6 – Atual

O método atual faz parte do grupo de métodos customizados e reflete a forma como a empresa estudada realiza suas previsões para as séries em questão. Como já foi explicado no segundo capítulo deste trabalho, a empresa faz as suas previsões tendo como *baseline* o ano anterior, ajustado por um fator que contempla a expectativa de crescimento da economia e seu reflexo para a companhia

O método atual somente tem uma alternativa de modelo e faz parte do conjunto de métodos que são comparados com o intuito de se elencar o método mais preciso.

Deve-se ressaltar que conseguir apresentar métodos com acuracidade superior à técnica utilizada atualmente pela empresa será um grande resultado alcançado por este trabalho, já que permitirá, de forma prática e objetiva, que a companhia melhore o desempenho da sua área comercial. A seguir, está especificado o MAPE de treino e de teste para cada série, tendo como base a forma de previsão do Lojão do Brás.

Série 1 – Camiseta de Homem

Para a série de camiseta de homem, considerando o mesmo conjunto de dados de treino e de teste utilizado para os demais métodos, obteve-se um MAPE de treino de 31,13% e de teste de 30,01%. Pela primeira vez, foi possível alcançar resultados superiores nos dados de teste. Abaixo, na Tabela 39, temos o quadro resumo dos resultados e também o gráfico, na Figura 69, que compara os dados reais contra as previsões obtidas neste método.

Tabela 39 – Resumo dos resultados do modelo 6 para série 1 – Camiseta de Homem

Modelos	MAPE Treino
Atual	**31,13%**

Fonte: elaboração própria.

Pode-se verificar que a acuracidade do modelo do conjunto de teste foi um pouco superior ao do conjunto de treino, o que indica uma maior precisão dos dados previstos para esta parte da série na fase de treino.

Modelo Atual - QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
Atual	31,13%	30,01%

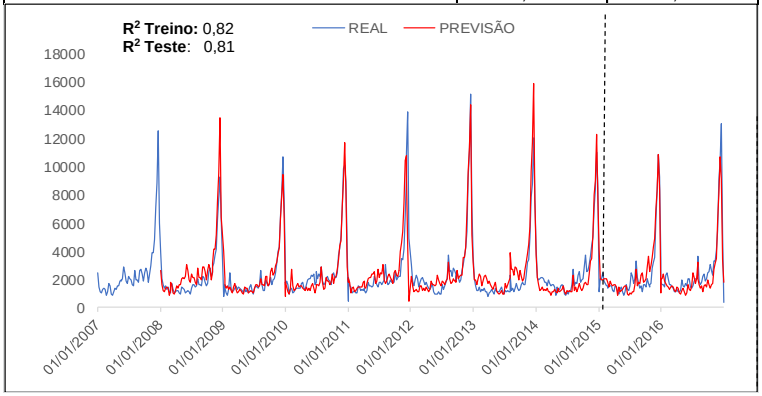


Figura 69 – Gráfico com resultados da melhor alternativa método 6 para a série 1– Camiseta de Homem

Fonte: elaboração própria.

Série 2 – Blusa de Mulher

No caso da série de blusa de mulher, a aplicação do método atual, o qual só tem uma alternativa, gerou um MAPE no conjunto de dados treino de 35,32% e, no conjunto de teste, de 26,75%. Mais uma vez, o teste superou a acuracidade do treino.

Tabela 40 – Resumo dos resultados do modelo 6 para série 2 – Blusa de Mulher

Modelos	MAPE Treino
Atual	**35,32%**

Fonte: elaboração própria.

Na Figura 70, é possível comparar os dados previstos com base no método atual para a série de blusa de mulher contra seus dados reais.

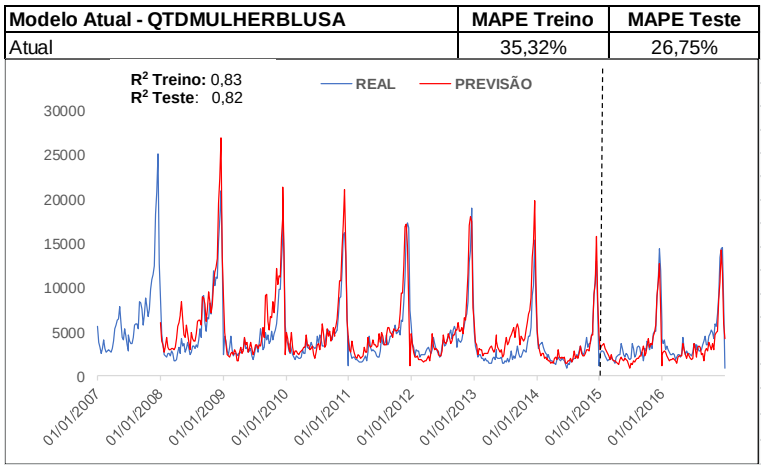


Figura 70 – Gráfico com resultados da melhor alternativa método 6 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

Considerando o mesmo conjunto de dados utilizados até o momento, para o conjunto de dados de treino, a aplicação do método atual, que gera somente uma alternativa de modelo, atingiu um MAPE de 32,13% no conjunto de dados de treino e 29,25% no conjunto de teste.

Tabela 41– Resumo dos resultados do modelo 6 para série 3– Camiseta de Menino

Modelos	MAPE Treino
Atual	**32,13%**

Fonte: elaboração própria.

Na Figura 71, podem-se verificar os dados reais contra os dados previstos para esta metodologia, onde se constatou uma maior precisão nos dados de teste para este método.

Modelo Atual - QTDMENINOCAMISETA	MAPE Treino	MAPE Teste
Atual	32,13%	29,25%

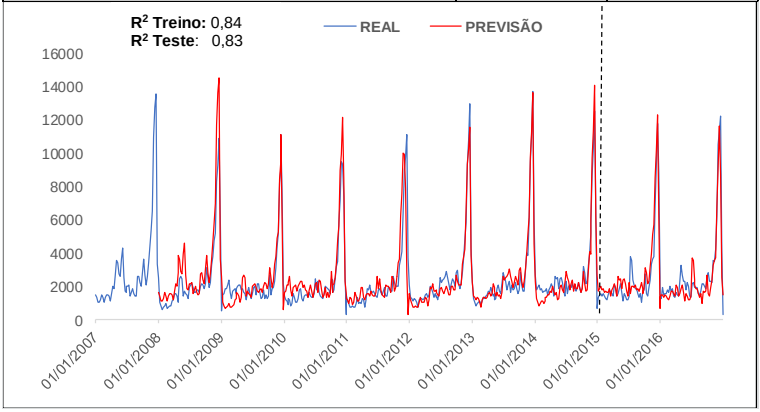


Figura 71 – Gráfico com resultados da melhor alternativa método 6 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

Para a última série, a de blusa de menina, a aplicação do método atual gerou um MAPE no conjunto de treino de 30,13% e, no conjunto de teste, de 35,52%. No método atual, somente para a série de blusa de menina não houve uma melhora no MAPE de teste, o que é na verdade, o padrão mais esperado, diferente do que ocorreu nas séries anteriores.

Tabela 42 – Resumo dos resultados do modelo 6 para série 4 – Blusa de Menina

Modelos	MAPE Treino
Atual	**30,13%**

Fonte: elaboração própria.

Como foi feito até o momento, temos a Figura 72 com os dados reais e os dados previstos, o que permite realizar uma avaliação comparativa.

Modelo Atual - QTDMENINABLUSA	MAPE Treino	MAPE Teste
Atual	30,13%	35,52%

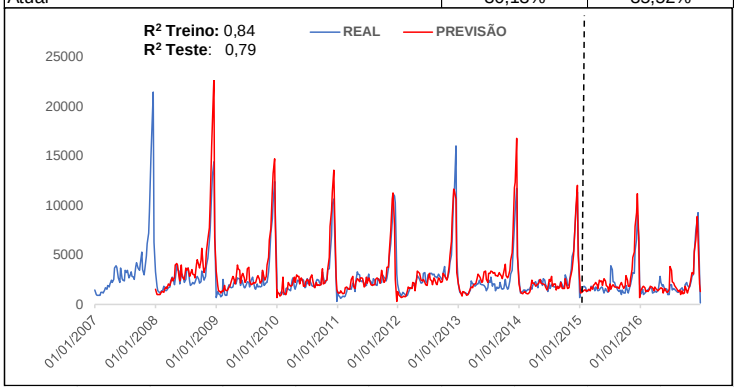


Figura 72 – Gráfico com resultados da melhor alternativa método 6 para a série 4 – Blusa de Menina
Fonte: elaboração própria.

5.7. Método 7 – MLP

A partir do método sete, serão apresentados os resultados obtidos com os modelos avançados. O primeiro método a ser aplicado é a rede neural Perceptron multicamadas com janela fixa (MLP). O algoritmo utilizado neste caso foi a retropropagação. Por usar entradas defasadas, este tipo de rede também pode ser chamada de FTDNN.

Em termos de topologia, onde se especifica a quantidade de camadas intermediárias, a quantidade de neurônios pertencentes a esta camada, os tipos de função de ativação e os tipos de otimizadores, foram testados 126 tipos de redes neurais diferentes para cada série. As variações consideradas foram feitas com base nos seguintes hiperparâmetros:

- **Número de defasagens de entrada:** 52;
- **Número de camadas escondidas**: 1;
- **Número de camadas de saída:** 1;
- **Número de neurônios na camada escondida:** 6,13,26,52,104,156,208;
- **Taxa de aprendizado:** 0.001, 0.01, 0.1;
- **Funções de ativação da camada escondida:** Sigmoide e RELU;
- **Otimizadores:** SGD, ADAM e RMSProp.

Para cada uma das séries serão apresentados os resultados das 10 melhores redes neurais obtidas fase de treinamento.

Série 1 – Camiseta de Homem

Diferentemente dos métodos estatísticos clássicos, nos casos das redes neurais não existe um processo de identificação onde se analisa um conjunto de funções e testes para determinamos um conjunto de modelos adequados. No caso da rede neural, não temos um processo determinístico, mas sim heurístico, onde o objetivo é se aproximar do melhor resultado possível. Neste sentido, é definido um conjunto de possíveis combinações de hiperparâmetros, que são testados e, assim, é possível verificar o modelo que apresentou o resultado mais próximo ao desejado que, no caso deste trabalho, é o modelo que obteve o menor MAPE.

Pode-se observar, por meio da Tabela 43, que a melhor rede para a série de camiseta de homem foi a rede com 156 neurônios na camada intermediária, com taxa de aprendizado η = 0.001, que utiliza a função ReLU na camada escondida e o otimizador Adam, a qual gerou um MAPE de 4,63%. Interessante notar que não foi a rede com o maior número de neurônios na camada escondida que foi a melhor rede. Adicionalmente, percebe-se que todas as 10 melhores redes usaram a função ReLU e, predominantemente, o otimizador Adam. Em nenhuma das 10 redes foi utilizado o otimizador SGD.

Tabela 43 – Resumo dos resultados do modelo 7 para série 1 – Camiseta de Homem

Rede	Qtd Neurônios Camada Intermediária	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**156**	**relu**	**adam**	**0,001**	**4,63%**
Rede 2	156	relu	adam	0,01	4,72%
Rede 3	156	relu	adam	0,1	5,75%
Rede 4	104	relu	adam	0,001	5,81%
Rede 5	26	relu	adam	0,1	5,96%
Rede 6	104	relu	adam	0,01	6,06%
Rede 7	208	relu	adam	0,01	6,69%
Rede 8	104	relu	rmsprop	0,01	6,80%
Rede 9	156	relu	rmsprop	0,1	6,82%
Rede 10	26	relu	adam	0,01	7,48%

Fonte: elaboração própria.

O desempenho da rede anterior no conjunto teste teve uma piora atingindo um MAPE de 15,34%. Mesmo assim, percebe-se que o modelo ficou bem ajustado quando ele é comparado com os dados reais.

Modelo MLP- QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
MLP - n=156,η = 0.001, relu, adam	4,63%	15,34%

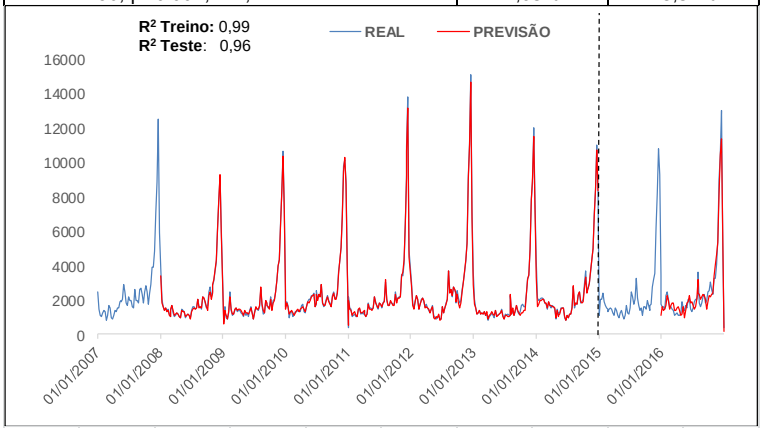


Figura 73 – Gráfico com resultados da melhor alternativa método 7 para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Série 2 – Blusa de Mulher

Para a série de blusa de mulher, foi adotado o mesmo procedimento da série de camiseta de homem para o método MLP, onde não existe um processo de análise determinístico, mas sim um processo heurístico, onde se busca encontrar o modelo que ofereça a melhor aproximação do objetivo estabelecido, tendo em vista um conjunto de possibilidades.

Este conjunto de possibilidades é constituído pela combinação dos diversos hiperparâmetros envolvidos para especificação de uma rede neural multicamada, os quais, para este trabalho, são: quantidade de neurônios da camada escondida, função de ativação da camada escondida, otimizador e taxa de aprendizado. O objetivo buscado é minimizar o MAPE. É importante ressaltar que as variações possíveis de hiperparâmetros são as mesmas para todas as séries estudadas.

A Tabela resumo 44 evidencia as 10 melhores RNAs das 126 possíveis para série de blusa de mulher. A rede que apresentou o menor MAPE no conjunto de dados de treino foi a rede neural com 104 neurônios na camada intermediária, que utilizou a função ReLU, o otimizador Adam e teve uma taxa de aprendizado de 0.001.

Das 10 melhores redes neurais em termos de MAPE, todas utilizaram a função de ativação ReLU e, em sua maioria, o otimizador Adam. E, mais uma vez, não foi a rede com o maior número de neurônios na camada intermediária que apresentou a maior acuracidade nos dados de treino.

Tabela 44 – Resumo dos resultados do modelo 7 para série 2 – Blusa de Mulher

Rede	Qtd Neurônios Camada Intermediária	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**104**	**relu**	**adam**	**0,001**	**4,72%**
Rede 2	26	relu	adam	0,100	5,14%
Rede 3	26	relu	adam	0,010	5,31%
Rede 4	208	relu	adam	0,001	5,42%
Rede 5	104	relu	adam	0,001	5,52%
Rede 6	26	relu	adam	0,001	5,87%
Rede 7	104	relu	rmsprop	0,010	5,99%
Rede 8	156	relu	adam	0,100	6,40%
Rede 9	104	relu	rmsprop	0,100	6,82%
Rede 10	208	relu	rmsprop	0,010	7,22%

Fonte: elaboração própria.

A rede neural escolhida teve um desempenho um pouco pior no conjunto de dados de teste, atingindo um MAPE de 13,28%. Apesar desta evolução negativa do MAPE, mas esperada do conjunto de dados treino para o conjunto de dados teste, o resultado obtido foi bem superior ao visto nos métodos anteriores. Logo, constata-se um bom ajustamento dos dados previstos perante os dados reais, como pode ser observado também na Figura 74.

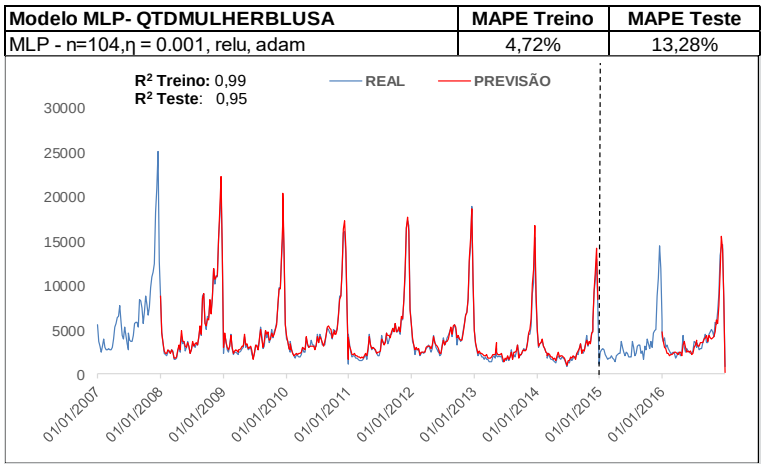


Figura 74 – Gráfico com resultados da melhor alternativa método 7 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

No caso da série de camiseta de menino, foi utilizada a mesma metodologia heurística aplicada nas séries anteriores, onde foram testadas 126 redes neurais, considerando as variações dos seguintes hiperparâmetros: número de neurônios na camada escondida, função de ativação, otimizador e taxa de aprendizado.

Com base no *ranking* do MAPE, das 126 redes neurais feitas, a rede neural com o menor MAPE, de 3,53%, foi a rede com a seguinte topologia: 208 neurônios na camada escondida, que utilizou a função Relu, o otimizador Adam e a taxa de aprendizado menor de 0.001. Neste caso, a melhor rede neural foi a que usou o maior número de neurônios, conforme pode ser observado na Tabela 45 com as 10 melhores redes.

Uma análise do quadro resumo destas 10 redes mais precisas permite verificar que todas utilizaram a função Relu, predominantemente o otimizador Adam e necessitou de um número maior de neurônios na camada escondida, quando é realizada uma comparação com as outras séries.

Tabela 45 – Resumo dos resultados do modelo 7 para série 3 – Camiseta de Menino

Rede	Qtd Neurônios Camada Intermediária	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**208**	**relu**	**adam**	**0,001**	**3,53%**
Rede 2	104	relu	adam	0,010	4,74%
Rede 3	26	relu	adam	0,001	5,51%
Rede 4	26	relu	adam	0,010	5,68%
Rede 5	156	relu	rmsprop	0,010	5,85%
Rede 6	208	relu	adam	0,010	6,13%
Rede 7	104	relu	adam	0,001	6,14%
Rede 8	26	relu	adam	0,100	6,39%
Rede 9	208	relu	rmsprop	0,100	6,42%
Rede 10	156	relu	adam	0,001	6,44%

Fonte: elaboração própria.

A rede neural com menor MAPE no conjunto de dados treino apresentou um MAPE de 13,78% no conjunto de dados teste, valor superior ao encontrado no primeiro período, mas extremamente inferior quando se compara com os resultados obtidos no período de teste dos métodos vistos até o momento. Na Figura 75, observa-se o grau de ajustamento do método MLP para a série de camiseta de menino frente aos dados reais da mesma.

Modelo MLP- QTDMENINOCAMISETA	MAPE Treino	MAPE Teste
MLP - n=208,η = 0.001, relu, adam	3,53%	13,78%

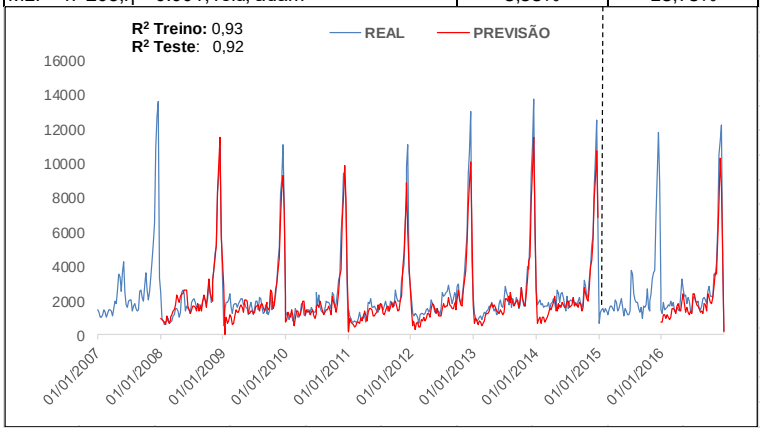


Figura 75 – Gráfico com resultados da melhor alternativa método 7 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

A série de blusa menina teve um comportamento muito similar ao encontrado na série de menino. Com base na mesma metodologia aplicada até o momento e nas mesmas variações de hiperparâmetros, a rede neural que gerou o menor MAPE das 126 redes configuradas foi a rede com 208 neurônios na camada escondida, que utilizou a função Relu, o otimizador Adam e a taxa de aprendizado de 0.001. O MAPE obtido nesta rede foi de 6,26%.

Na Tabela 46 é possível constatar a desempenho das 10 melhores redes neurais, em termos de acuracidade. Todas as 10 redes utilizaram a função Relu e não houve um consenso no otimizador e na taxa de aprendizado. Mais uma vez, verificou-se uma necessidade de uma maior quantidade neurônios na camada intermediária para modelar a série em questão, assim como ocorreu na série de camiseta de menino.

Tabela 46 – Resumo dos resultados do modelo 7 para série 4 – Blusa de Menina

Rede	Qtd Neurônios Camada Intermediária	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**208**	**relu**	**adam**	**0,001**	**6,26%**
Rede 2	156	relu	rmsprop	0,100	6,38%
Rede 3	104	relu	adam	0,010	6,56%
Rede 4	156	relu	adam	0,001	7,26%
Rede 5	156	relu	adam	0,100	7,53%
Rede 6	208	relu	rmsprop	0,001	7,63%
Rede 7	104	relu	adam	0,001	7,93%
Rede 8	208	relu	rmsprop	0,100	8,10%
Rede 9	156	relu	rmsprop	0,010	8,14%
Rede 10	156	relu	rmsprop	0,001	8,38%

Fonte: elaboração própria.

O resultado do MAPE da rede neural escolhida no conjunto de dados teste foi de 15,89%, o qual foi pior do que o valor obtido no conjunto de dados treino, mas muito melhor do que o verificado nos métodos estatísticos clássicos. A seguir, na Figura 76, pode-se constatar a qualidade da previsão do modelo escolhido, por meio da comparação dos reais contra os dados previstos.

Modelo MLP- QTDMENINABLUSA	MAPE Treino	MAPE Teste
MLP - n=208,η = 0.001, relu, adam	6,26%	15,89%

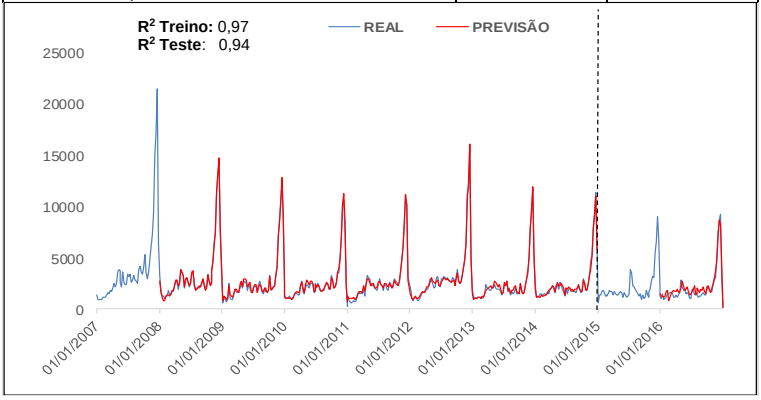


Figura 76 – Gráfico com resultados da melhor alternativa método 7 para a série 4 – Blusa de Menina
Fonte: elaboração própria.

5.8. Método 8 – MLP com exógenas

O oitavo método é uma extensão do sétimo método, pois passa a considerar as variáveis exógenas. Diferentemente do que ocorreu nos métodos clássicos estatísticos, não é necessário especificar as variáveis significativas; os próprios valores dos pesos de cada variável da rede neural expressarão a sua relevância. O algoritmo de treinamento também é o *backpropagation*.

Assim, são feitas as mesmas variações de topologia para construir as redes neurais deste método. São evidenciados 126 modelos para cada série, conforme os hiperparâmetros abaixo e incluindo as variáveis exógenas:

- **Número de defasagens de entrada:** 52;
- **Número de camadas escondidas**: 1;
- **Número de camadas de saída:** 1;
- **Número de neurônios na camada escondida:** 6,13,26,52,104,156,208;

- **Taxa de aprendizado:** 0.001, 0.01, 0.1;
- **Funções de ativação da camada escondida:** Sigmoide e RELU;
- **Otimizadores:** SGD, ADAM e RMSProp.

O próximo passo é fazer a avaliação para cada série.

Série 1 – Camiseta de Homem

A inclusão das variáveis exógenas não melhorou o desempenho da rede neural multicamada com janela fixa de 52 entradas defasadas. Uma das justificativas é que as variáveis exógenas trouxeram alguma não linearidade, que a rede não conseguiu se adequar de uma maneira melhor, quando comparada com o modelo sem tais variáveis. Na Tabela 47, é apresentado o desempenho das 10 melhores redes encontradas. Percebe-se que a rede com melhor acuracidade possui características bem diferentes do modelo MLP. Ela demandou menos neurônios na camada intermediária, somente 26, e utilizou otimizador RMSProp. Deve-se verificar que o otimizador Adam apresentou resultados bem próximos. A taxa de aprendizado e a função de ativação foram as mesmas.

Mais uma vez, todas as 10 melhores redes usaram a função Relu, houve uma divisão de utilização de otimizador entre o Adam e o RMSProp e taxa de aprendizado dominante foi a de 0.001.

Tabela 47 – Resumo dos resultados do modelo 8 para série 1 – Camiseta de Homem

Rede	Qtd Neurônios Camada Intermediária	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**26**	**relu**	**rmsprop**	**0,001**	**33,88%**
Rede 2	26	relu	adam	0,001	33,89%
Rede 3	52	relu	rmsprop	0,001	40,97%
Rede 4	52	relu	adam	0,001	42,30%
Rede 5	104	relu	adam	0,100	44,39%
Rede 6	156	relu	adam	0,001	52,71%
Rede 7	156	relu	adam	0,100	52,75%
Rede 8	156	relu	adam	0,010	52,76%
Rede 9	104	relu	rmsprop	0,001	53,21%
Rede 10	156	relu	rmsprop	0,001	54,73%

Fonte: elaboração própria.

Quando a melhor alternativa foi testada no conjunto teste, verificou-se uma pequena piora, o que fez o MAPE sair de 33,88% para 35,62%. Na Figura 77, temos a comparação entre os dados previstos e os dados reais.

Modelo MLP com exógenas- QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
MLPcom exógenas - n=26,η = 0.001, sigmoid, rmsprop	33,88%	35,62%

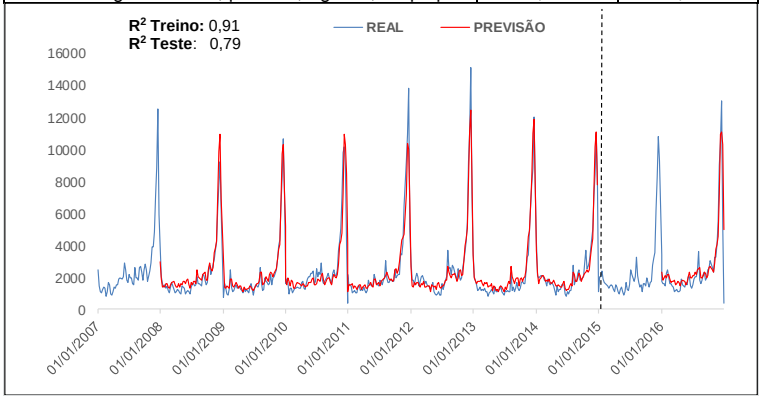


Figura 77 – Gráfico com resultados da melhor alternativa método 8 para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Série 2 – Blusa de Mulher

No caso da série de blusa de mulher, a verificação do método MLP com variáveis exógenas também é uma extensão do sétimo método, onde não se faz necessário realizar a especificação das variáveis significativas, como ocorreu nos métodos estatísticos clássicos.

Assim, é utilizado o mesmo conjunto de hiperparâmetros do sétimo método, juntamente com a inclusão das variáveis exógenas e o uso do algoritmo de *backpropagation.*

A seguir, temos um quadro resumo das 10 melhores redes neurais das 126 construídas. Pode-se constatar que a inclusão das variáveis exógenas piorou significativamente a acuracidade das previsões no conjunto de dados de treino. Neste sentido, pode-se dizer que tais variáveis não ajudaram a realizar uma aproximação mais precisa dos dados reais.

A rede com melhor desempenho foi a rede com 52 neurônios na camada escondida, que utilizou a função Relu, o otimizador RMSProp e a taxa de aprendizado de 0.001 com MAPE de 41,68%. Em relação às demais alternativas, não se pode realizar uma generalização sobre função de ativação, otimizador e número de neurônios. Verificou-se somente que a maior parte das redes utilizou, predominantemente, a taxa de aprendizado de 0.001.

Tabela 48 – Resumo dos resultados do modelo 8 para série 2 – Blusa de Mulher

Rede	Qtd Neurônios Camada Intermediária	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**52**	**relu**	**rmsprop**	**0,001**	**41,68%**
Rede 2	52	relu	adam	0,001	42,29%
Rede 3	26	relu	rmsprop	0,001	44,23%
Rede 4	26	relu	adam	0,001	44,38%
Rede 5	104	relu	rmsprop	0,001	55,70%
Rede 6	156	sigmoid	rmsprop	0,001	58,74%
Rede 7	104	sigmoid	adam	0,001	60,78%
Rede 8	104	sigmoid	adam	0,010	60,78%
Rede 9	104	sigmoid	adam	0,100	60,79%
Rede 10	13	relu	rmsprop	0,001	64,73%

Fonte: elaboração própria.

No conjunto de dados testes, a rede com melhor performance no conjunto de dados treino apresentou um MAPE de 33,21%, superando então o resultado obtido no momento do treinamento, o que não é comum. Na Figura 78, podem-se observar os dados reais e as previsões feitas por meio da rede neural escolhida.

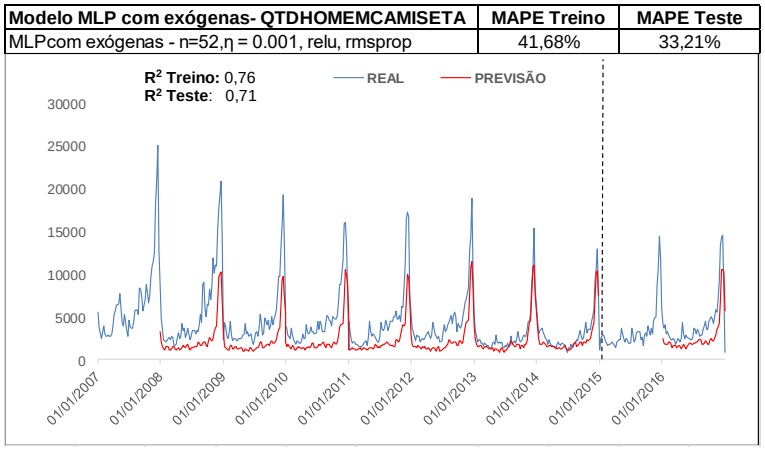


Figura 78 – Gráfico com resultados da melhor alternativa método 8 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

Na série de camiseta de menino foi adotada a mesma abordagem considerada até o momento, que foi o desenvolvimento das 126 redes neurais contemplando o mesmo conjunto de hiperparâmetros com o incremento das variáveis exógenas, as quais não precisam ser selecionadas previamente, pois, por meio dos pesos, a rede neural consegue refletir a relevância de cada uma delas.

A partir da análise da Tabela 49, pode-se afirmar que a rede com melhor performance em termos de precisão, com MAPE de 38,03%, foi a rede com 26 neurônios na camada escondida, que utilizou a função Relu, o otimizador Adam e a taxa de aprendizado de 0.001.

Uma análise das demais redes permite constatar uma predominância da utilização da função sigmoide, o que não ocorreu na MLP sem as exógenas, do otimizador Adam, da quantidade de 104 neurônios na camada escondida e da taxa de aprendizado de 0.001.

Tabela 49 – Resumo dos resultados do modelo 8 para série 3 – Camiseta de Menino

Rede	Qtd Neurônios Camada Intermediária	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**26**	**relu**	**adam**	**0,001**	**38,03%**
Rede 2	52	sigmoid	rmsprop	0,001	41,95%
Rede 3	52	sigmoid	adam	0,001	43,23%
Rede 4	104	sigmoid	adam	0,100	46,93%
Rede 5	104	sigmoid	adam	0,010	47,03%
Rede 6	104	relu	adam	0,001	47,04%
Rede 7	104	sigmoid	rmsprop	0,001	55,70%
Rede 8	104	sigmoid	rmsprop	0,010	55,70%
Rede 9	104	sigmoid	rmsprop	0,100	55,71%
Rede 10	156	sigmoid	adam	0,010	56,90%

Fonte: elaboração própria.

Assim como ocorreu na série de blusa de mulher, a rede escolhida no conjunto de dados treino evidenciou uma precisão um pouco superior no conjunto de dados teste, atingindo o MAPE de 37,15%, o qual ainda é muito pior do que o valor encontrado na rede sem a inclusão das variáveis exógenas para o mesmo conjunto de dados.

Modelo MLP com exógenas- QTDMENINOCAMISETA	MAPE Treino	MAPE Teste
MLPcom exógenas - n=26,η = 0.001, relu, adam	38,03%	37,15%

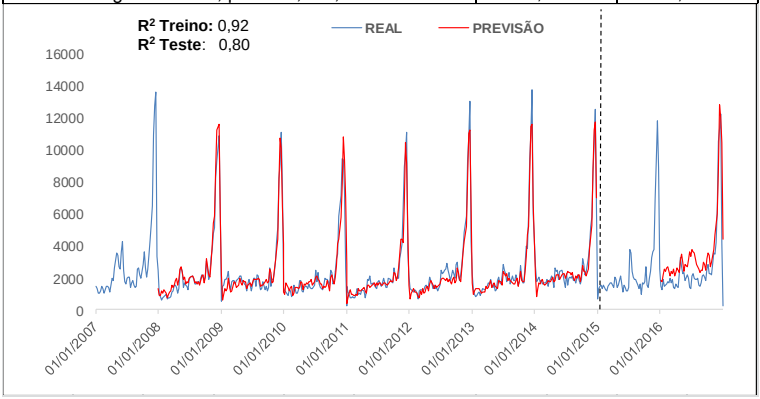


Figura 79 – Gráfico com resultados da melhor alternativa método 8 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

Em relação à série de menina, foi realizada a extensão do sétimo método de previsão com a inclusão das variáveis exógenas, tendo em vista que um dos objetivos deste trabalho é verificar a contribuição que tais variáveis podem oferecer para construir modelos de previsão mais precisos.

Pode-se constatar, mais uma vez, que a inclusão das variáveis exógenas trouxe uma não linearidade que não pode ser tratada pela rede, já que os resultados obtidos no âmbito de precisão foram bem inferiores, quando comparados com o método anterior, que não levou em conta tais *inputs*.

Da mesma forma que foi feito para as demais séries temporais, também foram construídas 126 redes para a série e, abaixo, na Tabela 50, é apresentado um quadro resumo com as 10 melhores alternativas em termos de MAPE. Na série de blusa de menina a rede com maior precisão, mais especificamente com MAPE de 40,85% no treinamento, foi a rede com 52 neurônios na camada escondida, que utilizou a função de ativação Relu, o otimizador RMSProp e a taxa de aprendizado de 0.001.

Das 10 RNAs analisadas, houve uma predominância no uso da função Relu, do otimizador RMSProp e da taxa de aprendizado de 0.001. Depois de verificadas todas as séries, é interessante destacar uma maior frequência do otimizador RMSProp, quando foram introduzidas as variáveis exógenas.

Tabela 50 – Resumo dos resultados do modelo 8 para série 4 – Blusa de Mulher

Rede	Qtd Neurônios Camada Intermediária	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**52**	**relu**	**rmsprop**	**0,001**	**40,85%**
Rede 2	52	relu	adam	0,001	41,44%
Rede 3	26	sigmoid	rmsprop	0,001	43,35%
Rede 4	26	relu	adam	0,001	43,49%
Rede 5	104	relu	rmsprop	0,001	54,59%
Rede 6	104	relu	rmsprop	0,010	54,59%
Rede 7	104	relu	rmsprop	0,100	54,62%
Rede 8	156	relu	rmsprop	0,001	57,56%
Rede 9	104	relu	adam	0,001	59,56%
Rede 10	104	relu	adam	0,010	59,56%

Fonte: elaboração própria.

Em relação ao desempenho no conjunto de dados treino, a rede neural selecionada no treinamento evidenciou um MAPE de 56,53%, um pouco pior do que o encontrado no período de treinamento, o que já é esperado normalmente. Na Figura

80, é apresentado o gráfico para efeito de comparação entre os dados reais e os dados previstos com base no modelo selecionado para esta série.

Modelo MLP com exógenas- QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
MLPcom exógenas - n=52,η = 0.001, relu, rmsprop	40,85%	56,53%

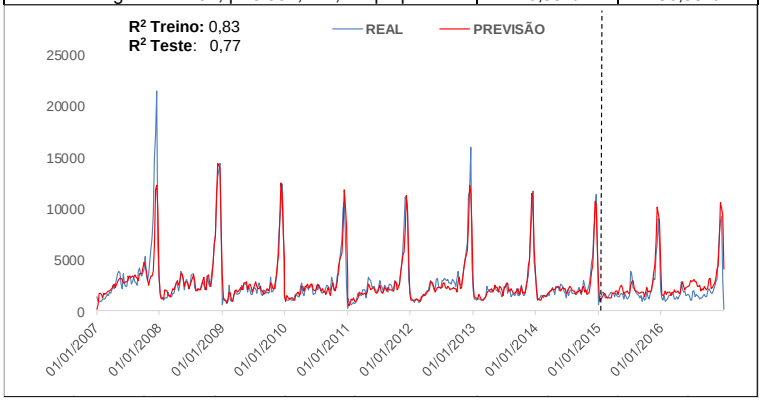


Figura 80 – Gráfico com resultados da melhor alternativa método 8 para a série 4 – Blusa de Menina
Fonte: elaboração própria.

5.9. Método 9 – LSTM

O próximo método a ser detalhado também faz parte dos métodos avançados e tem como grande promessa conseguir modelar a relação de dependência temporal entre as observações da série. Este modelo de rede recorrente é o método mais novo em termos de modelagem. Ao invés de neurônios, utiliza-se blocos de memórias que conseguem, por meio do estado da célula e dos portões de entrada, esquecimento e saída, definir as informações relevantes e compreender o sequenciamento existente em uma série temporal. Adicionalmente, a LSTM é o tipo de rede mais preparada para lidar com problema de explosão ou dissipação de gradiente. Ao mesmo tempo, este tipo de rede neural é mais intensivo computacionalmente. O algoritmo de treinamento é o BPTT, que é uma adaptação do *backpropagation.*

Pode-se afirmar que o LSTM é um refinamento da rede multicamadas Perceptron, não sendo necessário especificar a priori o número de defasagens que devem ser consideradas. Uma diferença fundamental em relação ao método MLP é que a normalização é um pré-requisito para aplicação da rede recorrente LSTM. No caso do MLP, apesar de não ser necessário, é uma boa prática realizar normalização. Desta forma, em todas as redes neurais deste trabalho foi utilizado este procedimento.

Assim como ocorreu no sétimo método e no oitavo método, são feitas 126 variações das redes recorrentes LSTM através das seguintes possibilidades de hiperparâmetros:

- **Número de defasagens de entrada:** 52;
- **Número de camadas escondidas**: 1;
- **Número de camadas de saída:** 1;
- **Blocos de memória:** 6,13,26,52,104,156,208;
- **Taxa de aprendizado:** 0.001, 0.01, 0.1;
- **Funções de ativação da camada escondida:** Sigmoide e RELU;
- **Otimizadores:** SGD, ADAM e RMSProp.

Abaixo é feita a aplicação da rede recorrente LSTM para cada uma das séries.

Série 1 – Camiseta de Homem

Para a série de camiseta de homem, a rede recorrente LSTM apresentou bons resultados para as diversas variações de redes construídas. Por meio da variação da taxa de aprendizado, da quantidade de blocos de memória, da função de ativação e do otimizador, constatou-se que a melhor rede possui a seguinte topologia: 52 blocos de memória, utilizou a função Relu nestes blocos com o otimizador Adam e taxa de aprendizado de 0.001. Com isso, obteve-se um MAPE no treino de 4,87%. Na Tabela 51, pode-se visualizar as 10 melhores redes LSTM encontradas.

Tabela 51 – Resumo dos resultados do modelo 9 para série 1 – Camiseta de Homem

Rede	Qtd Blocos de Memória	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**52**	**relu**	**adam**	**0,001**	**4,87%**
Rede 2	52	relu	adam	0,100	5,18%
Rede 3	156	relu	adam	0,001	5,24%
Rede 4	156	relu	adam	0,100	5,37%
Rede 5	208	relu	adam	0,100	5,38%
Rede 6	52	relu	adam	0,010	5,66%
Rede 7	208	relu	adam	0,001	5,66%
Rede 8	104	relu	adam	0,100	5,70%
Rede 9	156	relu	adam	0,010	5,82%
Rede 10	26	relu	adam	0,100	6,31%

Fonte: elaboração própria.

Todas as 10 melhores redes usaram a função Relu e o otimizador Adam. Além disso, a melhor rede não usou o número maior de blocos de memórias possível, que é de 208 blocos, mas sim apenas 52 blocos de memória.

Considerando a aplicação da melhor rede, vista na Tabela 51, no conjunto de dados de teste, o MAPE obtido foi de 12,98%, um erro baixo frente ao que já foi observado até o momento. No Figura 81, é possível comparar os dados previstos contra os dados reais, onde se constata uma grande aderência.

Modelo LSTM- QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
LSTM - n=52,η = 0.001, relu, adam	4,87%	12,98%

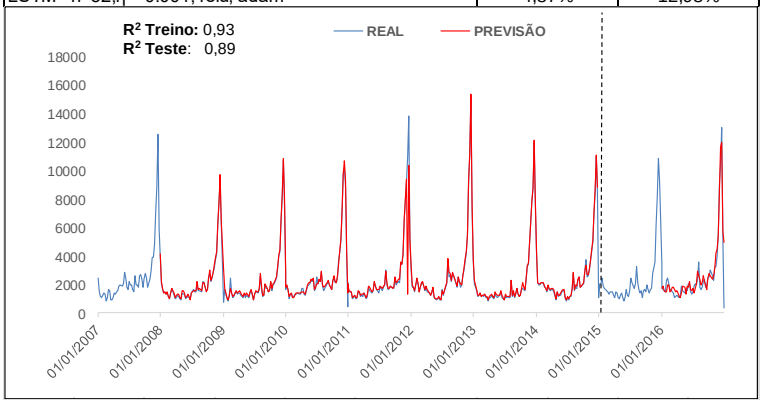


Figura 81 – Gráfico com resultados da melhor alternativa método 9 para a série 1– Camiseta de Homem
Fonte: elaboração própria.

Série 2 – Blusa de Mulher

Para a série de blusa de mulher, o procedimento adotado para construção das redes das RNNs LSTM foi a construção de 126 alternativas de modelos, por meio da variação dos hiperparâmetros estipulados no início deste capítulo, juntamente com a utilização do algoritmo BPTT. Deve-se reforçar que, no caso redes LSTM, é feita a especificação em termos de blocos de memória e não de neurônios na camada escondida, sendo esta estrutura muito mais complexa e preparada para compreender a dependência temporal existente entre os termos da série.

A partir da análise da Tabela 52, verifica-se que a rede LSTM com maior precisão no período de treinamento, atingindo um MAPE de 3,63%, foi a rede com a seguinte especificação: 156 blocos de memória, utilizou a função Relu, o otimizador Adam e teve taxa de aprendizado de 0.01, que foi a rede selecionada.

Em relação às 10 melhores redes, deve-se salientar que todas as redes utilizaram a função Relu e otimizador Adam. Adicionalmente, a maior parte das redes necessitou de um número grade de blocos de memória com, pelo menos, 52 unidades na sua especificação. Em nenhuma das 10 redes foi verificada a utilização do otimizador SGD e da função de ativação sigmoide.

Tabela 52 – Resumo dos resultados do modelo 9 para série 2 – Blusa de Mulher

Rede	Qtd Blocos de Memória	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**156**	**relu**	**adam**	**0,010**	**3,63%**
Rede 2	208	relu	adam	0,010	3,77%
Rede 3	52	relu	adam	0,100	4,37%
Rede 4	104	relu	adam	0,100	4,50%
Rede 5	208	relu	adam	0,001	4,54%
Rede 6	52	relu	adam	0,010	4,60%
Rede 7	104	relu	adam	0,001	4,83%
Rede 8	104	relu	adam	0,010	5,00%
Rede 9	52	relu	adam	0,001	5,23%
Rede 10	156	relu	adam	0,100	5,27%

Fonte: elaboração própria.

A rede escolhida na fase de treinamento obteve uma MAPE de 13,25% na fase de teste, o qual foi um pouco inferior ao desempenho com os dados testes, mas muito superior ao obtido com os métodos estatísticos clássicos e muito próximo do valor obtido com o método MLP. Na Figura 82, podem-se comparar os dados reais contra os dados previstos por meio do modelo selecionado, sendo possível verificar a grande aderência do modelo adotado, existindo uma grande sobreposição.

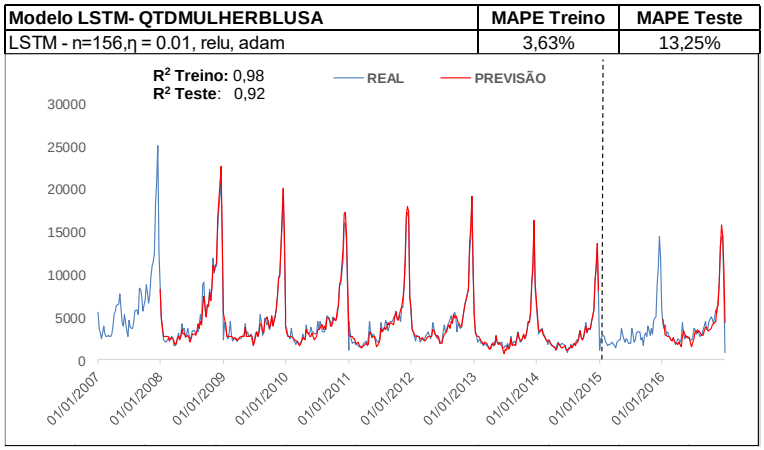


Figura 82 – Gráfico com resultados da melhor alternativa método 9 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

Para a série de camiseta de menino, a aplicação do método de previsão de RNNs LSTM apresentou bons resultados em termos de acuracidade. A metodologia adotada foi a construção das 126 redes neurais através da variação dos seguintes hiperparâmetros: quantidade de blocos de memória na camada escondida, função de ativação, otimizador e taxa de aprendizado.

A RNN LSTM com melhor desempenho MAPE, de 4,40%, na fase de treinamento para esta série temporal foi a rede com a seguinte topologia: 208 blocos de memória na camada escondida, uso da função de ativação Relu, otimizador Adam e taxa de aprendizado de 0.001, conforme pode ser constatado Tabela 53, que contém o resultado das 10 melhores redes para esta série:

Tabela 53 – Resumo dos resultados do modelo 9 para série 3 – Camiseta de Menino

Rede	Qtd Blocos de Memória	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**208**	**relu**	**adam**	**0,001**	**4,40%**
Rede 2	208	relu	adam	0,100	4,98%
Rede 3	156	relu	adam	0,010	5,06%
Rede 4	156	relu	adam	0,001	5,26%
Rede 5	104	relu	adam	0,001	5,32%
Rede 6	26	relu	adam	0,010	5,50%
Rede 7	156	relu	adam	0,100	5,56%
Rede 8	52	relu	adam	0,001	6,06%
Rede 9	156	relu	rmsprop	0,010	6,13%
Rede 10	208	relu	adam	0,010	6,20%

Fonte: elaboração própria.

A partir da análise deste quadro resumo, verifica-se que todas as redes utilizaram a função de ativação Relu e nove redes utilizaram o otimizador Adam. Em termos de blocos de memória na camada escondida, ocorreu uma preponderância de 156 unidades.

A aplicação da rede selecionada no conjunto de dados testes gerou um MAPE de 13,15%, um pouco menos preciso, o que já é esperado, porém muito superior aos resultados obtidos nos métodos clássicos e, ligeiramente, mais preciso que o método MLP. A seguir, na Figura 83, temos um gráfico comparativo dos dados reais contra os dados previstos.

Modelo LSTM- QTDMENINOCAMISETA	MAPE Treino	MAPE Teste
LSTM - n=208,η = 0.001, relu, adam	4,40%	13,15%

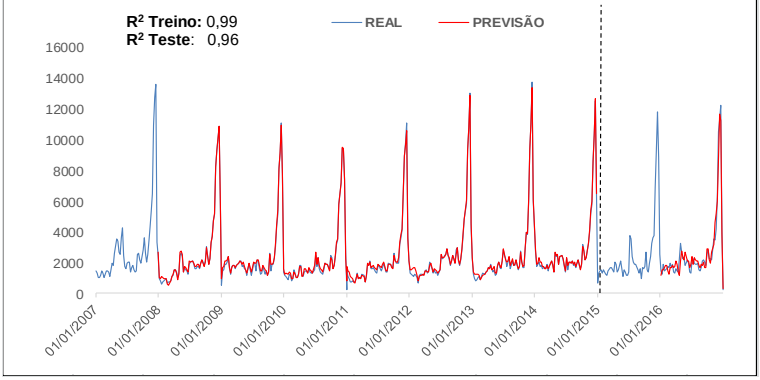


Figura 83 – Gráfico com resultados da melhor alternativa método 9 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

Para a série de blusa de menina, aplicando os mesmos passos utilizados para as demais séries e considerando as 126 redes neurais especificadas, chega-se à conclusão que a RNN LSTM mais precisa para esta série foi a rede neural com MAPE de 4,65% na fase de treinamento. Esta rede possui a seguinte topologia: 156 blocos de memória na camada escondida, utilizou a função de ativação Relu, o otimizador Adam e a taxa de aprendizado de 0.01, o que pode ser observado no resumo da Tabela 54, com o *ranking* das 10 melhores redes para a série em questão.

O exame da Tabela 54 permite constatar que as 10 melhores redes, em sua totalidade, utilizaram a função de ativação Relu e o otimizador Adam. Além disso, para a maior parte das redes houve a necessidade da utilização de uma grande quantidade de blocos de memória e ocorreu um empate na taxa de aprendizado especificada.

Tabela 54 – Resumo dos resultados do modelo 9 para série 4 – Blusa de Menina

Rede	Qtd Blocos de Memória	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**156**	**relu**	**adam**	**0,010**	**4,65%**
Rede 2	104	relu	adam	0,100	4,84%
Rede 3	156	relu	adam	0,001	4,89%
Rede 4	208	relu	adam	0,100	5,16%
Rede 5	208	relu	adam	0,001	5,29%
Rede 6	104	relu	adam	0,001	5,48%
Rede 7	52	relu	adam	0,100	5,62%
Rede 8	104	relu	adam	0,010	5,72%
Rede 9	208	relu	adam	0,010	6,94%
Rede 10	26	relu	adam	0,100	7,23%

Fonte: elaboração própria.

A rede selecionada na fase de treinamento apresentou um MAPE no conjunto de dados de teste de 22,90%, uma precisão bem menor frente ao resultado obtido no conjunto de dados treino e que não superou o resultado obtido no método MLP. Na Figura 84, é possível verificar as previsões obtidas e os dados reais conjuntamente, permitindo a realização de uma análise comparativa.

Modelo LSTM- QTDMENINABLUSA	MAPE Treino	MAPE Teste
LSTM - n=156,η = 0.01, relu, adam	4,65%	22,90%

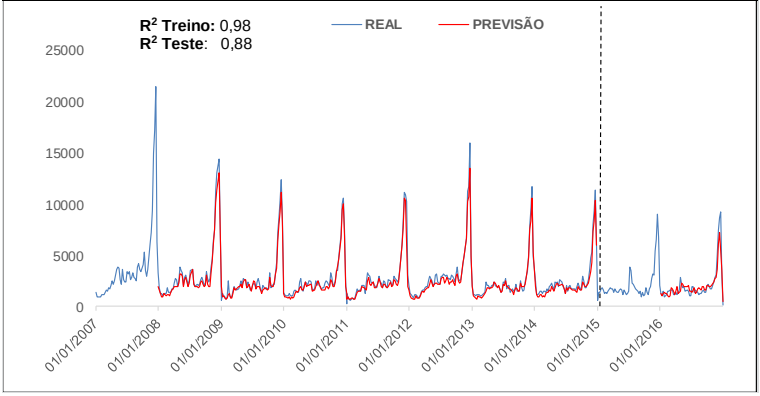


Figura 84 – Gráfico com resultados da melhor alternativa método 9 para a série 4 – Blusa de Menina
Fonte: Elaborado pelo autor

5.10. Método 10 – LSTM com exógenas

Este é o último método de previsão avaliado nesta dissertação. Ele é uma extensão do método anterior com a adição de todas variáveis exógenas, as quais são: TMED, PREC, ICC, TAXA_DESEMPREGO, REND_REAL, MSAD_REAL, CRED_FP, PC_VEST e as 6 *dummies* do calendário varejista.. No caso das redes recorrentes LSTM, assim como foi visto na rede MLP com variáveis externas, não é necessário fazer uma análise de significância destas variáveis, diferentemente do que ocorre nos métodos estatísticos clássicos. Como no LSTM sem exógenas, o algoritmo de treinamento deste método será o BPTT.

As topologias verificadas neste método são as mesmas visualizadas no método do LSTM. Então, os hiperparâmetros utilizados são:

- **Número de defasagens de entrada:** 52;
- **Número de camadas escondidas**: 1;
- **Número de camadas de saída:** 1;

- **Blocos de memória:** 6,13,26,52,104,156,208;
- **Taxa de aprendizado:** 0.001, 0.01, 0.1;
- **Funções de ativação da camada escondida:** Sigmoide e RELU;
- **Otimizadores:** SGD, ADAM e RMSProp.

A seguir, temos a aplicação para cada série temporal do trabalho.

Série 1 – Camiseta de Homem

A inclusão das variáveis exógenas não melhorou o MAPE da rede LSTM, quando é feita a comparação do modelo sem estas variáveis. Tal fato também ocorreu com a rede MLP. Assim, mais uma vez, as variáveis exógenas trouxeram uma não linearidade que a rede não conseguiu interpretar. A melhor rede apresentou a seguinte topologia: 52 blocos de memória, utilização da função de ativação Relu nestes blocos, com o otimizador Adam e taxa de aprendizado de 0.01.

Nas 10 melhores redes, por meio da verificação da Tabela 55, temos uma predominância na utilização de 104 blocos de memória e do otimizador Adam. Em todas as redes foi utilizada a função de ativação Relu para os blocos de memória.

Tabela 55 – Resumo dos resultados do modelo 10 para série 1 – Camiseta de Homem

Rede	Qtd Blocos de Memória	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**52**	**relu**	**adam**	**0,010**	**6,58%**
Rede 2	104	relu	rmsprop	0,100	7,07%
Rede 3	104	relu	rmsprop	0,001	7,48%
Rede 4	104	relu	adam	0,001	7,49%
Rede 5	104	relu	adam	0,010	7,71%
Rede 6	104	relu	adam	0,100	8,15%
Rede 7	104	relu	rmsprop	0,010	8,27%
Rede 8	52	relu	adam	0,100	8,42%
Rede 9	52	relu	adam	0,001	9,70%
Rede 10	26	relu	adam	0,100	9,89%

Fonte: elaboração própria.

A aplicação da melhor rede no conjunto teste gerou um MAPE de 24,52%, sendo assim houve uma piora considerável em relação ao MAPE do conjunto treino. Na Figura 85, pode-se visualizar o gráfico com os dados reais e as previsões obtidas.

Modelo LSTM com exógenas- QTDHOMEMCAMISETA	MAPE Treino	MAPE Teste
LSTM com exógenas- n=52,η = 0.01, relu, adam	6,58%	24,52%

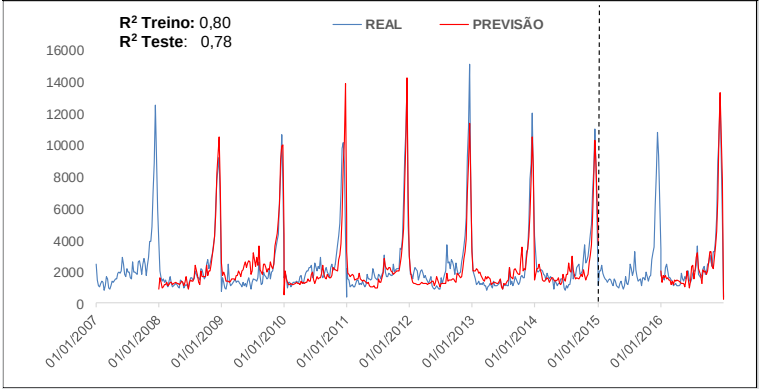


Figura 85 – Gráfico com resultados da melhor alternativa método 10 para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

Série 2 – Blusa de Mulher

Novamente, busca-se verificar se a inclusão de variáveis exógenas pode contribuir de forma positiva para o aumento da precisão do método LSTM. Foi utilizado o mesmo conjunto de hiperparâmetros, o algoritmo BPTT, as mesmas variáveis exógenas e o mesmo conjunto de dados.

Para a série de blusa de mulher, pode-se afirmar, com base na análise da Tabela 56, que a inclusão das variáveis exógenas não contribuiu para aumentar a precisão das previsões no período de treinamento. A topologia da RNN LSTM com o melhor desempenho foi a rede com as seguintes características: 104 blocos de memória na camada escondida, uso da função de ativação Relu, do otimizador RMSProp e de taxa de aprendizado de 0.001, sendo então a rede selecionada e que apresentou um MAPE de 8,05%

A análise da Tabela 56, que evidencia a configuração das dez melhores redes, torna possível constatar que todas as redes utilizaram a função de ativação Relu, em sua maioria o otimizador Adam e a quantidade de 104 blocos de memória na camada escondida.

Deve-se salientar a maior frequência do otimizador RMSProp entre as melhores redes, quando ocorre a inclusão das variáveis exógenas, assim como foi verificado no método MLP com exógenas.

Tabela 56 – Resumo dos resultados do modelo 10 para série 2 – Blusa de Mulher

Rede	Qtd Blocos de Memória	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**104**	**relu**	**rmsprop**	**0,001**	**8,05%**
Rede 2	104	relu	adam	0,100	8,07%
Rede 3	104	relu	adam	0,010	8,45%
Rede 4	52	relu	adam	0,001	8,87%
Rede 5	104	relu	rmsprop	0,100	8,95%
Rede 6	52	relu	adam	0,010	9,78%
Rede 7	104	relu	adam	0,001	10,69%
Rede 8	26	relu	adam	0,010	10,75%
Rede 9	52	relu	rmsprop	0,001	11,16%
Rede 10	26	relu	adam	0,001	11,45%

Fonte: elaboração própria.

A RNN LSTM escolhida na fase de treinamento apresentou um MAPE de 19,76% na fase teste, indicando uma piora da acuracidade do conjunto de dados treino para o conjunto de teste e ela não superou o resultado obtido, com base no mesmo indicador, da RNN LSTM sem a presença das exógenas. Na Figura 86, temos o gráfico comparativo com os dados reais e os dados previstos da RNN selecionada.

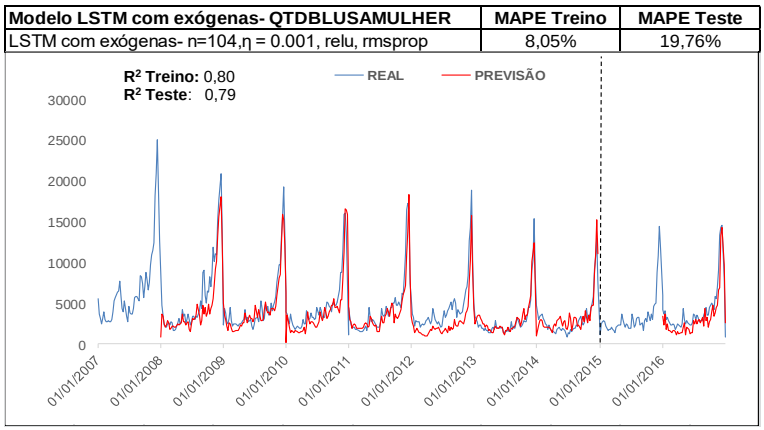


Figura 86 – Gráfico com resultados da melhor alternativa método 10 para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

Série 3 – Camiseta de Menino

Na série de camiseta de menino, visando verificar o impacto da inclusão das variáveis exógenas na precisão das previsões realizadas, pode-se constatar que, das 126 redes RNN LSTM construídas, a rede com melhor desempenho teve um MAPE de 6,23%, pior do que o valor obtido com o método LSTM, o que indica que a inclusão das variáveis exógenas não trouxe ganhos em termos de precisão.

A RNN LSTM com exógenas com melhor MAPE apresentou a seguinte topologia: 156 blocos de memória na camada intermediária, uso da função Relu, do otimizador RMSProp e taxa de aprendizado de 0.001, conforme a Tabela 57, que evidencia os resultados das 10 melhores redes.

A partir da análise da topologia das dez melhores redes para a série em questão, pode-se verificar que todas as redes utilizaram a função de ativação Relu, houve uma divisão equitativa entre o uso do otimizador RMSProp e o uso do otimizador Adam e a maior parte das redes utilizou 104 blocos de memória na camada escondida. Não houve um consenso em termos de taxa de aprendizado.

Tabela 57– Resumo dos resultados do modelo 10 para série 3 – Camiseta de Menino

Rede	Qtd Blocos de Memória	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**156**	**relu**	**rmsprop**	**0,001**	**6,23%**
Rede 2	104	relu	adam	0,010	6,24%
Rede 3	104	relu	adam	0,100	6,58%
Rede 4	104	relu	rmsprop	0,100	7,06%
Rede 5	52	relu	adam	0,010	7,45%
Rede 6	156	relu	adam	0,001	7,68%
Rede 7	104	relu	rmsprop	0,010	8,02%
Rede 8	52	relu	adam	0,001	8,70%
Rede 9	52	relu	rmsprop	0,100	9,10%
Rede 10	52	relu	adam	0,100	9,32%

Fonte: elaboração própria.

A aplicação da RNN LSTM no conjunto de dados teste atingiu um MAPE de 24,99%, bastante inferior ao valor obtido no conjunto de dados treino e ao valor obtido com a LSTM pura, sem as exógenas, que foi de 13,15%. Pode-se observar a aderência das previsões obtidas através da Figura 87, que confronta os dados reais com os previstos.

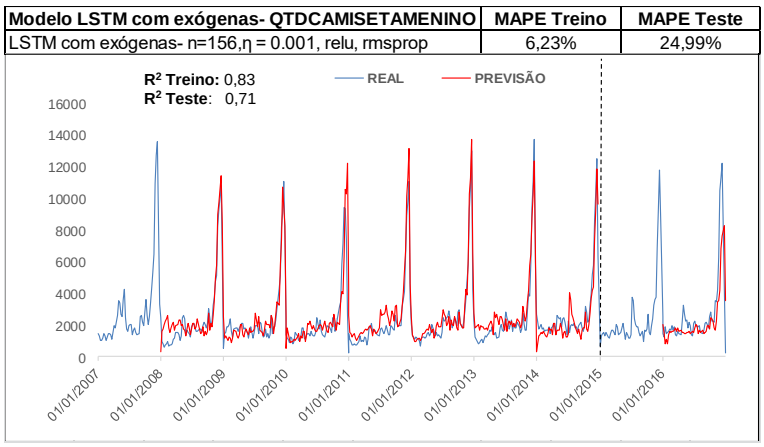


Figura 87– Gráfico com resultados da melhor alternativa método 10 para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

Série 4 – Blusa de Menina

Na série de blusa de menina aplicou-se a mesma metodologia utilizada nas séries anteriores, onde foi realizada uma extensão do nono método LSTM com a inclusão das variáveis exógenas para o mesmo conjunto de 126 redes neurais testadas. Estas redes foram construídas por meio das variações dos seguintes hiperparâmetros: quantidade de blocos de memória, taxa de aprendizado, função de ativação e otimizador.

Assim como ocorreu com todas as outras séries temporais, a inclusão das variáveis exógenas não aumentou a acuracidade de previsão quando comparada com o método LSTM univariado. A melhor rede apresentou um MAPE no período de treinamento de 6,20% e possui a seguinte topologia: 104 blocos de memória na camada escondida, uso da função Relu, otimizador RMSprop e taxa de aprendizado de 0.001, o que é verificado a partir da Tabela 58 com os dados das 10 melhores RNNs LSTM da série de blusa de menina, dentre as 126 estudadas.

Tabela 58 – Resumo dos resultados do modelo 10 para série 4 – Blusa de Menina

Rede	Qtd Blocos de Memória	Função de ativação	Otimizador	Taxa Aprendizado	MAPE Treino
Rede 1	**104**	**relu**	**rmsprop**	**0,001**	**6,20%**
Rede 2	13	relu	rmsprop	0,001	7,37%
Rede 3	208	relu	adam	0,001	7,93%
Rede 4	156	relu	rmsprop	0,100	8,13%
Rede 5	156	relu	rmsprop	0,010	8,20%
Rede 6	208	relu	adam	0,100	8,29%
Rede 7	26	relu	rmsprop	0,001	8,56%
Rede 8	156	relu	adam	0,001	8,78%
Rede 9	156	relu	rmsprop	0,001	8,91%
Rede 10	156	relu	adam	0,010	8,93%

Fonte: elaboração própria.

Uma análise mais detalhada do quadro das 10 melhores RNNs LSTM torna possível afirmar que todas as redes utilizaram a função Relu, houve uma divisão equitativa entre o uso do otimizador Adam e RMSprop, sendo que o último passou a estar muito mais frequente com a inclusão das variáveis exógenas, e houve a necessidade de, pelo menos, 26 blocos memória na camada escondida. A taxa de aprendizado mais evidenciada foi a de 0.001.

A RNN LSTM com exógenas de melhor desempenho na fase de treinamento obteve um MAPE na fase de teste de 27,25%. Isto evidencia uma piora significativa,

mas em certa medida esperada, da fase de treinamento para a fase de teste. Adicionalmente, o resultado obtido foi inferior ao valor encontrado no método LSTM univariado e no método MLP. Na Figura 88, é apresentado um gráfico com os dados reais e os dados previstos por meio da RNN LSTM com exógenas selecionada para a série de blusa de menina.

Modelo LSTM com exógenas- QTBLUSAMENINA	MAPE Treino	MAPE Teste
LSTM com exógenas- n=104,η = 0.001, relu, rmsprop	6,20%	27,25%

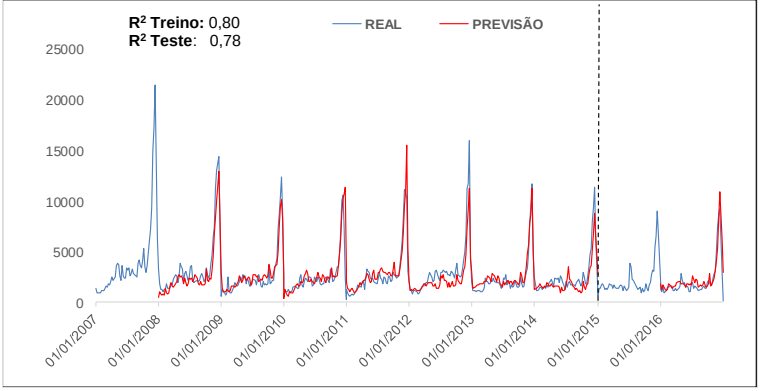


Figura 88 – Gráfico com resultados da melhor alternativa método 10 para a série 4 – Blusa de Menina
Fonte: elaboração própria.

5.11. Comparação dos modelos

Esta parte do trabalho apresenta uma avaliação conjunta de todos os métodos de previsão utilizados. O objetivo é conseguir verificar qual método apresentou a melhor acuracidade em termos de MAPE no conjunto de dados testes. Também será destacado o impacto da inclusão das variáveis exógenas.

A análise será feita para cada uma das séries estudadas, onde serão apresentados de forma conjunta os resultados de acuracidades de cada método para o conjunto de dados treino e dados teste.

5.11.1. Comparação modelos série camiseta de homem

Na análise comparativa para a série de camiseta de homem, constatou-se uma superioridade dos métodos avançados em relação aos métodos estatísticos clássicos

tanto na fase de treino como na fase de teste. Vale ressaltar que somente o método avançado MLP com exógenas atingiu resultados piores que os métodos clássicos.

Já os métodos Naive e o método atual superaram os métodos estatísticos clássicos somente na fase de teste.

Em relação aos métodos avançados, o melhor método foi o MLP na fase de treino, sendo que método LSTM apresentou resultados bem próximos, inclusive superando o MLP na fase de teste.

Cabe ressaltar que as variáveis exógenas tiveram um impacto positivo, contribuindo para a redução do MAPE e, consequentemente, melhorando a acuracidade das previsões somente no caso dos métodos estatísticos clássicos. Já no caso dos métodos avançados, verificou-se que a inclusão das variáveis externas prejudicou, significativamente, o desempenho do MAPE tanto na fase de treino como no teste.

Em relação ao método atual utilizado pela empresa, ele foi superado por praticamente todos os métodos no conjunto de dados treino, com exceção do método MLP com exógenas. E, na fase de teste, ele foi superado pelos seguintes métodos avançados: MLP, LSTM e LSTM com exógenas. Considerando que o melhor método foi o LSTM, obteve-se uma previsão 57% mais precisa que o método vigente com base nos dados de teste.

Conclui-se, então, que para a série de camiseta de homem, por meio da avaliação do valor absoluto do MAPE na fase de teste, que o método mais adequado é o LSTM. Na Figura 89, temos o gráfico comparativo com o resultado de todos os métodos, para a fase de treino e para a fase de teste.

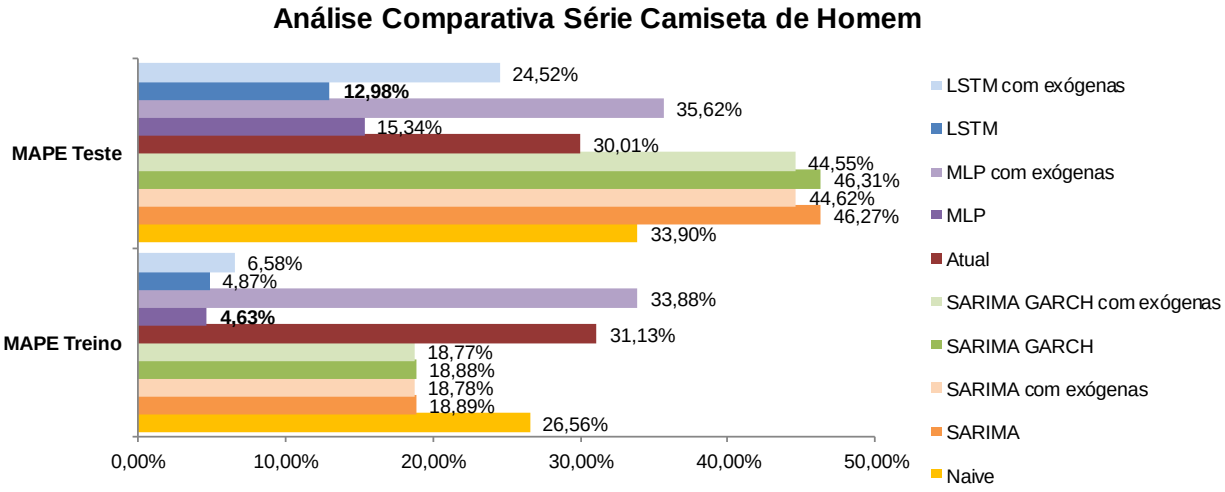


Figura 89 – Gráfico análise comparativa dos métodos de previsão para a série 1 – Camiseta de Homem
Fonte: elaboração própria.

5.11.2. Comparação modelos série blusa de mulher

No caso da série de blusa de mulher, pode-se verificar também que os métodos avançados, com exceção do método MLP com exógenas, conseguiram superar os resultados alcançados pelos métodos estatísticos clássicos e pelo método atual utilizado pela empresa. Importante ressaltar que tal fato ocorreu tanto na fase de treino como na fase de teste.

Um aspecto interessante é que os métodos Naive e o atual conseguiram apresentar resultados melhores que os métodos estatísticos clássicos na fase de teste; o mesmo não ocorreu no período treinamento.

No caso dos métodos avançados, o método que evidenciou o melhor desempenho MAPE na fase de treino foi o método LSTM, seguido de perto pelo método MLP. Tal resultado foi o mesmo encontrado na fase de teste, com a diferença de que os resultados obtidos entre os dois métodos foram muito mais próximos.

Sobre as variáveis exógenas pode-se constatar que, para todos os métodos, elas não contribuíram para gerar previsões mais precisas. Na verdade, é como se a introdução destas variáveis gerasse um ruído que não foi compreendido pelos diversos métodos utilizados.

O método atual utilizado pela empresa foi suplantado pelos métodos Naive, MLP, LSTM e LSTM com exógenas. Em relação ao método com melhor desempenho, o LSTM, houve um incremento de precisão de 50,5% na fase de teste.

Por fim, deve-se reiterar que o método com maior acuracidade para a série de blusa de mulher é o método LSTM, através da verificação do MAPE na fase de teste dos diversos métodos utilizados. Na Figura 90, é apresentado o gráfico com os todos os resultados de todos os métodos, tanto para a fase de treino como para a fase de teste.

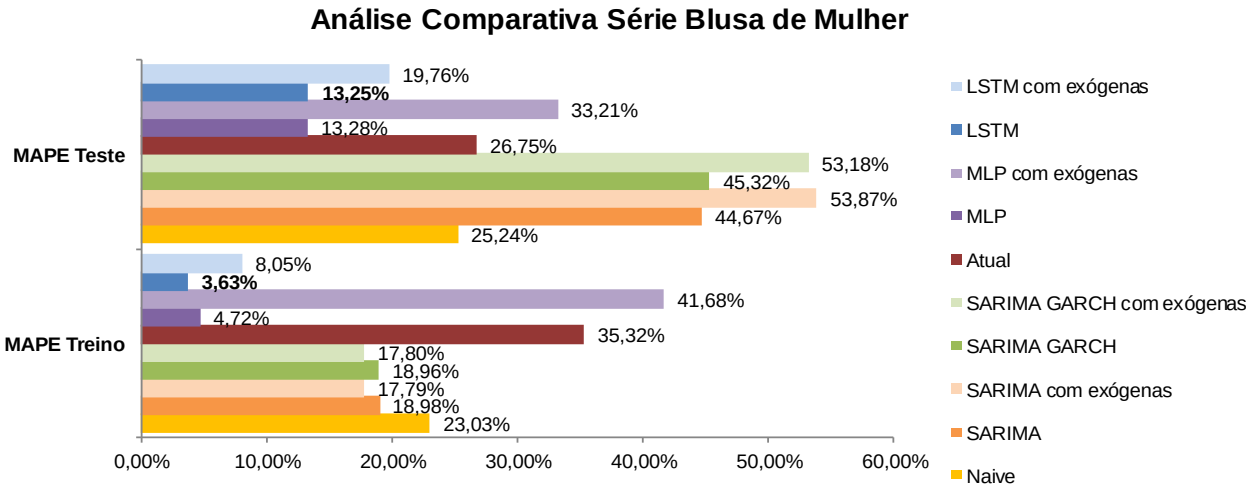


Figura 90 – Gráfico análise comparativa dos métodos de previsão para a série 2 – Blusa de Mulher
Fonte: elaboração própria.

5.11.3. Comparação modelos série camiseta de menino

No caso da série de camiseta de menino, mais uma vez, o desempenho dos métodos avançados foi muito superior aos métodos estatísticos clássicos tanto na fase de teste como na fase de treino, com exceção do método MLP com exógenas. Tal fato também ocorreu com o método Naive e o método atual, que foram superados pelos métodos avançados, porém obtiveram resultados melhores que os modelos estatísticos clássicos na fase de teste.

Uma análise mais específica dos métodos avançados torna possível afirmar que o melhor método foi o MLP na fase de treino e o LSTM na fase de teste. A diferença entre as medidas MAPEs obtidas foi pequena entre os dois métodos na fase de treino, mas a LSTM foi ligeiramente melhor.

As variáveis exógenas só conseguiram apresentar um impacto positivo, em termos de acuracidade, nos métodos estatísticos clássicos e na fase de treino dos mesmos. Para os demais métodos a inclusão destas variáveis não foi capaz de melhorar a precisão das previsões encontradas tanto no período de teste como no período de treino.

O método atual utilizado pela empresa apresentou um resultado na fase de teste melhor do que o resultado na fase treino. No entanto, foi superado por todos os métodos avançados, com exceção do MLP com exógenas. Em relação à rede LSTM na fase de

teste, que foi a rede que teve a maior acuracidade, foi possível aumentar a precisão em 55% frente ao resultado obtido com o método vigente.

Enfim, para a série de camiseta de menino o método mais apropriado, tendo em vista a questão de acuracidade, é o método LSTM, o qual apresentou o menor erro percentual absoluto médio entre todos os métodos escolhidos. Na Figura 91, temos o quadro resumo com os resultados de todos os métodos tanto na fase de treino como na fase de teste.

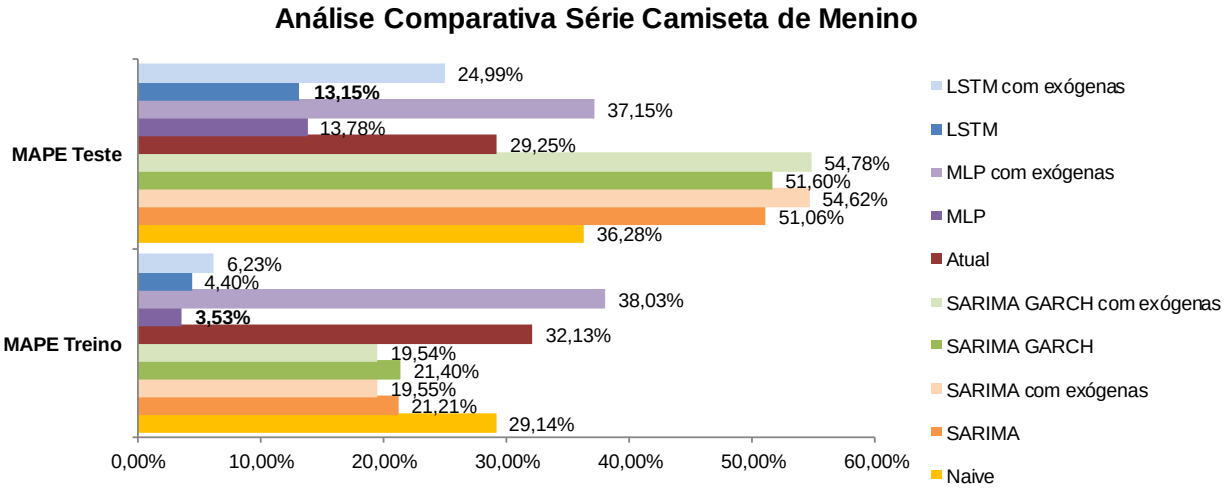


Figura 91 – Gráfico análise comparativa dos métodos de previsão para a série 3 – Camiseta de Menino
Fonte: elaboração própria.

5.11.4. Comparação modelos série blusa de menina

Para a série de blusa de menina, assim como ocorreu com as demais séries temporais estudadas, os métodos avançados superaram os métodos estatísticos clássicos, com a ressalva do método MLP com exógenas, que apresentou resultados inferiores. Tal fato ocorreu tanto no conjunto de dados treino como no conjunto de dados teste.

Em relação aos métodos Naive e ao atual, ambos foram superados pelos métodos avançados ao longo de todo o período de análise.

Uma análise mais específica dos métodos avançados torna possível verificar que o método com maior precisão na fase de treino foi o LSTM. No entanto, na fase de teste o método MLP conseguiu apresentar melhores resultados.

A inclusão das variáveis exógenas somente conseguiu reduzir o erro de previsão para a fase de treino dos métodos estatísticos clássicos. Para os demais métodos é possível afirmar que, tanto na fase de treino como na fase de teste, a inclusão das variáveis exógenas reduziu a precisão das previsões realizadas.

Em relação ao modelo atual, que é o modelo que vigora na área comercial da empresa, na fase de teste ele foi superado, em termos de acuracidade, por todos os modelos avançados. Mais especificamente, o método mais assertivo, o MLP, conseguiu gerar previsões 55% mais precisas que o método atual.

Conclui-se que, para a série de blusa de menina, o método mais adequado é o MLP sem exógenas, considerando os resultados de acuracidade observada para todos os métodos na fase de teste. Na Figura 92, segue o gráfico comparativo com os resultados alcançados pelos diversos métodos de previsão na fase de treinamento e na fase de teste.

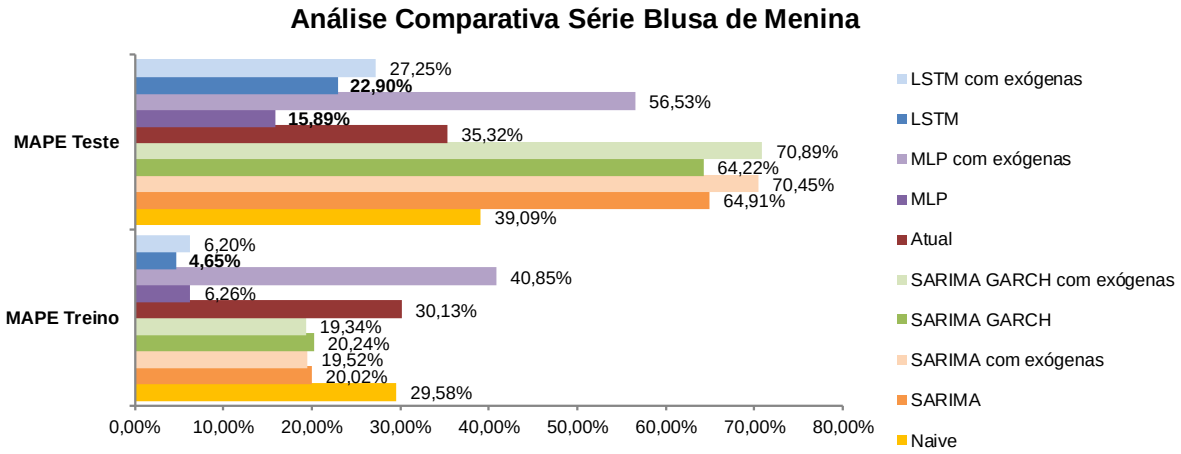


Figura 92 – Gráfico análise comparativa dos métodos de previsão para a série 4 – Blusa de Mulher
Fonte: elaboração própria.

5.11.5. Análise final dos métodos com melhor desempenho

Neste item é apresentada uma análise final contemplando todas as séries temporais estudadas. Desde já, é importante destacar a forte presença de sazonalidade em todas as séries, o que em parte explica a maior precisão dos métodos não lineares na modelagem das mesmas, exatamente por conseguirem captar os fatores que melhor explicam tais variações.

Retomando as particularidades do varejo de moda: forte impacto da sazonalidade e de forma diferente para a previsão das vendas de cada categoria do vestuário, diversas variáveis exógenas gerando distúrbios nas respectivas previsões de vendas, o fato das vendas serem altamente dependente do aspecto de moda, o que faz com a maior parte dos itens tenham que ser renovados constantemente, reduzindo assim a quantidade de dados históricos existentes e a grande granularidade presente em termos de tamanho, cor e vários outros atributos. Todos estes aspectos corroboram para que a previsão de vendas para o setor do varejo de moda moda seja extremamente específica e complexa.

Diante destes pontos, pode-se afirmar que os modelos avançados, representados aqui pelas redes neurais, são mais adequados para o setor estudado, tendo em vista que os dados existentes apresentam elevada flutuação, são incompletos e apresentam grandes distúrbios, sobretudo nas datas comemorativas e devido promoções e ou liquidações realizadas pelos varejistas. Mais especificamente, as seguintes características das redes neurais: adaptabilidade, não linearidade, ser uma função de aproximação univeral e o fato de utilizar uma modelagem não paramétrica, em que não há necessidade de se entender o processo efetivamente, vão ao encontro das especificidades do setor em foco, reforçando, então, a grande aderência na utilização destes métodos para a previsão de vendas.

A análise da Tabela 59, com o resumo dos resultados alcançados em termos de MAPE para cada uma das séries, que é uma medida de erro, onde se busca encontrar o menor valor possível, permite constatar que o método LSTM foi o mais preciso para todas as séries, com exceção da série de blusa de menina, onde o método MLP mostrou ser o mais exato.

Tabela 59 – Resumo dos resultados dos métodos de previsão para a fase de teste

	Resultados Período Teste			
Métodos	**MAPE Camiseta de Homem**	**MAPE Blusa de Mulher**	**MAPE Camiseta de Menino**	**MAPE Blusa de Menina**
1. Naive	33,90%	25,24%	36,28%	39,09%
2. SARIMA	46,27%	44,67%	51,06%	64,91%
3. SARIMA com exógenas	44,62%	53,87%	54,62%	70,45%
4. SARIMA GARCH	46,31%	45,32%	51,60%	64,22%
5. SARIMA GARCH com exógenas	44,55%	53,18%	54,78%	70,89%
6. Atual	30,01%	26,75%	29,25%	35,32%
7. MLP	15,34%	13,28%	13,78%	**15,89%**
8. MLP com exógenas	35,62%	33,21%	37,15%	56,53%
9. LSTM	**12,98%**	**13,25%**	**13,15%**	22,90%
10. LSTM com exógenas	24,52%	19,76%	24,99%	27,25%

Fonte: elaboração própria.

Uma das justificativas para o método MLP ser mais preciso para o caso da série de blusa menina é o fato desta série apresentar a menor quantidade de *outliers* e o fato do impacto sazonal ser menos intenso, quando comparado com o das demais séries.

O maior êxito do método LSTM para três das quatro séries estudadas, tendo como parâmetro o critério de acuracidade, quando comparado ao MLP, pode ser atribuído ao fato de a rede neural recorrente conseguir lidar melhor com a questão da dependência temporal tanto de curto como de longo prazo. A rede neural LSTM consegue aprender o contexto necessário para fazer previsões e trabalha de forma incremental por meio de seus portões de entrada, esquecimento e sáida, adicionando e eliminando as informações necessárias para determinar que a saída prevista seja mais próxima possível da saída real.

Apesar da maior acuracidade dos métodos LSTM e dos métodos MLP, deve-se verificar que os métodos apresentaram erros relativamente altos e com uma grande amplitude: de 12,98% a 70,89%. Os métodos estatísticos clássicos foram os métodos com pior nível de acuracidade.

A aplicação da abordagem GARCH colaborou para solucionar a questão de heterocedasticidade condicional presente em todas as séries, o que está diretamente ligada com a componente de sazonalidade. No entanto, não gerou previsões mais precisas.

Em relação às variáveis exógenas, com exceção da série de camiseta de homem e para os métodos estatísticos clássicos, pode-se afirmar que elas não contribuíram para gerar previsões mais assertivas em nenhum dos métodos estudados na fase de teste. Uma possível justificativa para isto está no fato de as variáveis exógenas terem uma periodicidade mensal, que foi adaptada para semanal, permitindo a sua utilização nos modelos de previsão. Tal procedimento pode ter afetado a forma de interpretação dos modelos dos impactos das variáveis exógenas. Neste sentido, poderia ser interessante realizar um tratamento distinto para adaptar os dados de mensal para semanal e/ou realizar o teste com outras varíaveis exógenas com o intuito de se conseguir modelos mais adequados e precisos para realizar a previsão de vendas.

Vale ressaltar que o método atual foi suplantado por todos os métodos avançados em termos de acuracidade, com exceção do método MLP com exógenas, que evidenciou uma precisão inferior. Em média, o uso dos métodos avançados gerou previsões 54,32% mais precisas que o método de previsão vigente na empresa Lojão do Brás.

Em todos os gráficos com o melhor modelo para cada método e série de produto foi apresentado o valor do coefiente de determinação (R^2), que é uma medida de ajustamento do modelo proposto, tendo em vista os dados coletados. Mais especificamente, esta medida evidencia o quanto da variância do modelo pode ser explicada pelas variáveis contempladas no modelo. Esta medida é muito utilizada na estatística tradicional para verificar o grau de ajustamento de um modelo adotado. No caso das redes neurais, que é foco deste trabalho, geralmente é utilizado o MAPE e o RMSE como métrica para avaliação dos modelos, considerando que o intuito é encontrar o modelo com menor nível de erro. Apesar disto, considerou-se relevante evidenciar os resultados verificados para esta medida estatística.

Assim como ocorreu no caso na medida de acuracidade escolhida para comparação dos métodos deste trabalho, o MAPE, os coeficientes de determinação dos métodos avançados superaram os coeficiente de determinação dos métodos estatísticos clássicos no período de teste. Além disso, a inclusão das variáveis exógenas também pirou o a qualidade do ajuste R^2 para todos os métodos estudados.

Deve-se destacar que considerando o R^2, o MLP foi o método que obteve a melhor quallide de ajustamento. Somente na série de camiseta de menino a rede recorrente LSTM conseguiu superar o resultado da MLP.

Na Tabela 60 é apresentado um quadro resumo dos resultados do coeficiente de determinação para cada categoria de produto e por método no período de teste.

Tabela 60– Resumo do R^2 dos métodos de previsão para a fase de teste

	Resultados Período Teste			
Métodos	**R^2 Camiseta de Homem**	**R^2 Blusa de Mulher**	**R^2 Camiseta de Menino**	**R^2 Blusa de Menina**
1. Naive	0,58	0,69	0,61	0,56
2. SARIMA	0,92	0,71	0,95	0,61
3. SARIMA com exógenas	0,78	0,80	0,89	0,69
4. SARIMA GARCH	0,92	0,71	0,94	0,65
5. SARIMA GARCH com exógenas	0,79	0,59	0,89	0,65
6. Atual	0,81	0,82	0,83	0,79
7. MLP	**0,96**	**0,95**	0,92	**0,94**
8. MLP com exógenas	0,79	0,71	0,80	0,77
9. LSTM	0,89	0,92	**0,96**	0,88
10. LSTM com exógenas	0,78	0,79	0,81	0,78

Fonte: Elaborado pelo autor

Por fim, em termos de métodos de previsão, buscou-se utilizar os métodos mais contemporâneos e apropriados para o objetivo e setor desta dissertação, o que não exclui a possibilidade de uma extensão deste trabalho para aplicação de métodos mais novos, visando confrontar o seu respectivo nível de precisão contra àqueles obtidos nesta dissertação. Mas certamente, o que ficou evidente, foi a maior acuracidade e adequação dos métodos não lineares, exatamente por lidarem com esta característica presente nas séries de quantidade vendidas dos produtos de moda.

6. Considerações Finais

Os objetivos deste trabalho foram: (i) verificar as restrições e limitações existentes para aplicação dos métodos de previsão de vendas para o mesmo setor (ii) identificar os métodos de previsão de vendas empregados no varejo de moda, (iii) aplicar os diversos métodos de previsão para as séries temporais que compõem este trabalho, (iv) comparar a acuracidade obtida com o uso dos métodos avançados em relação aos métodos estatísticos clássicos, (v) comparar os modelos avançados estatísticos clássicos versus o método atual vigente na empresa, e (vi) verificar o impacto da inclusão das variáveis exógenas.

Os dois primeiros objetivos foram verificados na revisão bibliográfica e foi constatado que já existe a aplicação de diversos métodos de previsão de vendas para o varejo de moda, que vão desde os métodos estatísticos clássicos passando pelos métodos avançados, que têm as redes neurais como um dos seus representantes. Em relação às restrições e dificuldades de previsão de vendas no varejo de moda, verificou-se que este setor é afetado por diversos fatores, tais como clima, sazonalidade, preço e promoções praticados pela empresa e pela concorrência, definição do nível de agregação para realização da previsão, ciclo de vida de produtos, que muitas vezes limita o volume de histórico de dados existente para efetivação da previsão, e por mais uma série de variáveis exógenas exploratórias.

Um aspecto extremamente relevante foi a constatação do fato de todas as séries de quantidade vendidas e de preços serem estacionárias, não sendo necessário passar pelo processo de diferenciação. Tal fato é fundamental para aplicação dos métodos estatísticos clássicos, mas não impacta a utilização dos métodos avançados.

Após a construção de cada um dos dez métodos de previsão – Naive, SARIMA, SARIMA com exógenas, SARIMA GARCH, SARIMA GARCH com exógenas, Atual, MLP, MLP com exógenas, LSTM e LSTM com exógenas – foi possível realizar a comparação da acuracidade entre os diversos métodos, através do erro percentual médio absoluto (MAPE). Com base na precisão, pode-se afirmar que os métodos avançados e, mais especificamente, a RNN LSTM apresentou a maior precisão para três das séries estudadas e a quarta série teve uma maior precisão com a rede MLP. Sendo assim, os métodos avançados superaram os métodos estatísticos clássicos.

Perante o modelo atual de previsão de vendas vigente na empresa Lojão do Brás, constatou-se que somente os métodos avançados conseguiram alcançar resultados mais precisos, considerando todos os métodos aplicados neste trabalho. O acréscimo de precisão médio do método LSTM foi de 54,32%, quando comparado com o método corrente. Interessante destacar que os métodos estatísticos clássicos não apresentaram melhor desempenho frente ao atual.

Por fim, cabe ressaltar que as variáveis exógenas não conseguiram contribuir de forma positiva aumentando a precisão dos métodos aplicados nesta dissertação. Pelo contrário, pode-se afirmar que a inclusão das variáveis exógenas trouxe ruídos para as séries, dificultando o processo de modelagem, o que pode ser justificado pela necessidade de adaptação dos dados da maior parte das variáveis exógenas de periodicidade mensal para semanal.

6.1. Considerações sobre os resultados

Os métodos de previsão implementados neste trabalho, sobretudo os métodos avançados, comprovaram que podem trazer grandes ganhos para a previsão das vendas no varejo dada a sua maior acuracidade, o que é determinante para atualidade, tendo em vista a maior concorrência presente no mercado.

O método com maior precisão foi a RNN LSTM, que é o estado de arte na área de redes neurais e é uma das técnicas de aprendizagem profunda. Contudo, este método possui uma alta complexidade e requer um alto esforço computacional. Estes aspectos, inicialmente, podem dificultar a adesão de tal metodologia pelo setor em questão.

No entanto, é preciso constatar que a computação paralela tem crescido de forma significativa nos últimos tempos através das GPUs, tornando mais barata e acessível este tipo de tecnologia. As GPUs, certamente, são potencializadores para o desenvolvimento do aprendizado de máquina e da aprendizagem profunda. Adicionalmente, o volume crescente de dados e a disseminação do conhecimento de *machine learning* tornarão mais fácil a adesão aos métodos mais avançados pelas empresas do setor de varejo de moda.

Como já foi colocado no início deste item, o método com maior nível de acuracidade, considerando as séries estudadas, foi a rede neural recorrente LSTM. Contudo, é importante destacar que as características de cada série contribuíram para determinar níveis de precisão distintos para cada método aplicado. Neste ponto, é importante destacar o aspecto da sazonalidade como um componente difícil de ser modelado e que afetou os níveis de precisão alcançados.

6.2. Considerações sobre as limitações encontradas

Certamente a maior limitação desta dissertação foi a quantidade de dados disponíveis para cada série temporal, sendo que cada uma contou com 524 observações. Para a utilização de redes neurais, quanto maior o volume de dados existente, melhor será o processo de aprendizagem. Então, seria importante ter um uma quantidade maior de observações para aplicação dos métodos avançados.

Outra limitação da pesquisa foi o tempo gasto para treinamento das redes neurais recorrentes LSTM com e sem as variáveis exógenas, sendo maior no segundo caso. Seria possível a construção de um número maior de modelos caso houvesse mais tempo ou fosse feito o investimento em uma GPU mais potente.

Por fim, deve-se considerar a existência de uma grande quantidade de técnicas de previsão de vendas e a possibilidade da utilização de modelos híbridos, que levam em conta na sua implementação a utilização de mais de um modelo, o que dificultou a escolha dos métodos para fazerem parte deste trabalho, sobretudo pelo receio de não ter sido contemplado um método, que fosse realmente mais adequado e preciso para as séries temporais estudadas. Além disso, cabe ressaltar que este campo de estudos é bastante produtivo, sempre apresentando novos métodos e soluções para alcançar resultados superiores.

6.3. Recomendações futuras pesquisas

Em termos de pesquisas futuras, certamente existem muitos trabalhos que podem ser realizados como uma continuação desta dissertação. Tais trabalhos podem ser divididos em três vertentes: (1) tratamento dos dados de entradas; (2) variações e ajustes dos métodos utilizados; e (3) aplicação de outros métodos.

Na vertente de tratamento de dados, pode-se dessazonalizar os dados e eliminar os *outliers* para verificar se as redes neurais apresentarão uma maior convergência e previsões mais precisas.

Além disso, seria interessante testar os mesmos métodos utilizados para um nível de agrupamento menor de produto, mais especificamente, no nível de segmento, onde a empresa busca analisar o estilo de um determinado conjunto de itens. Também seria importante testar outras variáveis exógenas, realizar uma outra forma de adaptação da periodicidade mensal para semanal e implementar outras possíveis defasagens de tais variáveis, com o objetivo de encontrar *inputs* que pudessem efetivamente ajudar na construção de modelos de previsão mais precisos.

Já na vertente de variações e ajustes dos métodos utilizados, mais especificamente para os métodos avançados, poderia se aplicar alguma técnica de regularização, como a de *dropout*, com o objetivo de verificar se as redes neurais terão uma maior generalização e, possivelmente, apresentarem uma maior acuracidade. Adicionalmente, seria interessante refazer a divisão dos dados observados em: treino (70%), validação (20%) e teste (10%). Assim, a inclusão das subamostras de validação poderia contribuir para encontrar modelos mais generalistas e precisos. Outra sugestão seria a construção de RNNs recorrentes LSTM empilhadas, que são as redes LSTM com mais de uma camada intermediária e/ou as redes LSTM Stateful, que tem as suas memórias reiniciadas em determinados momentos, como por exemplo, no início de cada ano.

Em relação à vertente de aplicação de outros métodos, existem muitas possibilidades a serem testadas, tais como: Lógica Fuzzy, Extreme Learning Machine (ELM), que é uma rede neural MLP que não utiliza o *backpropagation* como algoritmo de treinamento, Algoritmo Genético e alguns métodos híbridos, como o MLP com otimização por meio de Algoritmo Genético e o método Neuro-Fuzzy, que utiliza de forma híbrida as técnicas de Rede Neural e Lógica Fuzzy.

Por fim, cabe ressaltar que um grande desafio será a implementação efetiva dos métodos avançados na própria organização estudada, tendo em vista sua complexidade e a falta de processo para integração das previsões geradas com os sistemas da empresa. Tal aspecto, certamente, é um ponto crítico para o Lojão do Brás e para as várias empresas do setor, o que, sem dúvidas, geraria mais um trabalho de cunho exploratório com o intuito de verificar quais abordagens para implementação destas técnicas mais avançadas, de forma a garantir um maior nível de aderência por parte das equipes presentes nas empresas, e permitindo que os ganhos de precisão destes novos métodos sejam capturados em sua totalidade.

Referências Bibliográficas

AIUBE, F. A. L. **Modelos quantitativos em finanças:** com enfoque em commodities.1. ed. Porto Alegre: Bookman, 2013. 472p.

AKSOY, Asli; OZTURK, Nursel; SUCKY, Eric. A decision support system for demand forecasting in the clothing industry. **International Journal of Clothing Science and Technology**, v. 24, n. 4, p. 221-236, 2012.

ALMEIDA, Fernando Carvalho de; PASSARI, Antônio Fabrizio Lima. Previsão de Vendas no Varejo por meio de Redes Neurais. **Revista da Administração**, v. 41, n. 3, p. 257-272, 2006.

ALON, Ilan; QI, Min; SADOWSKI, Robert J. Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. **Journal of Retailing and Consumer Services**, v. 8, n. 3, p. 147-156, 2001.

ARMSTRONG, J. **Sales Forecasting**. 20 jul. 2008. Disponível em SSRN: <http://ssrn.com/abstract=1164602>. Acesso em: 10 nov. 2017.

ARMSTRONG, J. **Principles of Forecasting:** a Handbook for Researchers and Practitioners. Boston: Kluwer Academic Publishers, 2001.

ARMSTRONG, J. Strategic Planning and Forecasting Fundamentals. In: ALBERT, K. **The Strategic Management Handbook**. New York: MacGraw Hill, 1983, p.1-32.

ASIMOVE INSTITUTE. Disponível em: <www.asimovinstitute.org/neural-network-zoo>. Acesso em: 22 dez. 2017.

BERMAN, Berry; EVANS, Joel R. **Retail Management**: a strategic aproach. 7th ed. Upper Saddle River: Prentice Hall, 1998

BOLLERSLEV, Tim. Generalized Auto-Regressive Conditional Heteroskedasticity. **Journal of Econometrics,** Elsevier, vol. 31(3), pages 307-327, 1986.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time Series Analysis**: forecasting and control. 3. ed. Englewood Clifs, NJ. Prentice Hall, 1994.

BRAGA, A. P.; CARVALHO, A. P. L.; LUDERMIR, T. B. **Redes Neurais Artificiais**: teoria e aplicações. 2. ed. Rio de Janeiro: LTC, 2016.

BROWNLEE, Jason. **The Promise of Recurrent Neural Networks for Time Series Forecasting**. 22 maio 2017. Disponível em: <machinelearningmastery.com/promise-recurrent-neural-networks-time-series-forecasting/>. Acesso em: 20 nov. 2017.

BROWNLEE, Jason. **On the Suitability of Long Short-Term Memory Networks for Time Series Forecasting**. 26 maio 2017. Disponível em: <machinelearningmastery.com/suitability-long-short-term-memory-networks-time-series-forecasting/>. Acesso em: 20 nov. 2017.

BUENO, R. L. S. **Econometria de séries temporais.** 2. ed. São Paulo: Cengage Learning, 2008.

CAMELO, Henrique; LUCIO, Paulo; LEAL Junior, João Bosco; CARVALHO, Paulo Cesar Marques de. Previsão de velocidade do vento no nordeste brasileiro através do modelo ARIMAX. **Revista Ciência e Tecnologia**, v. 20, n. 36, jul. 2017.

CATTINI, Orlando. **Em busca da excelência na cadeia têxtil**. 2004. Trabalho apresentado ao 2º. Encontro de Logística Têxtil, São Paulo, 2004. Não publicado.

CELIA, Frank; GARG, Ashish; SZTANDERA, Les; RAHEJA, Amar. Forecasting women's apparel sales using mathematical modeling. **International Journal of Clothing Science and Technology**, v. 15, n. 2, p. 107-125, abr. 2003.

CHOI, Tsan-Ming; HUI, Chi-Leung; YU, Young. **Intelligent fashion forecasting systems**: models and applications. New York: Springer, 2014.

CHU, Ching-Wu; ZHANG, Guoqiang Peter. A comparative study of linear and nonlinear models for aggregate retail sales forecasting. **International Journal of Production Economics**, v. 86, n. 3, p. 217-231, 2003.

CHUNG, J.; GULCEHRE, C., CHO, K; BENGIO, Y. **Empirical evaluation of gated recurrent neural networks on sequence modeling**. In NIPS 2014 Workshop on Deep Learning, December 2014

COPELAND, Michael. Whats the difference between artificial intelligence, machine learning, and deep learning? **NVIDIA**, 29 jul. 2016. Disponível em: <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>. Acesso em: 10 jan. 2018.

DAHL, George E.; SAINATH, Tara N.; HINTON, Geoffrey E. **Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout**. .IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), p.8609-8613. 2013
DATA SCIENCE ACADEMY. **Deep Learning Book**. Disponível em: <deeplearningbook.com.br/>. Acesso em: 15 dez. 2017.

DE OLIVEIRA, Mauri Aparecido; FAVERO, Luiz Paulo Lopes. **Uma breve descrição de algumas técnicas para análise de séries temporais**: séries de Fourier, waveletes, ARIMA, modelos estruturais para séries de tempo e redes neurais. VI Semead: Ensaio de MQI. 2003. Disponível em: <sistema.semead.com.br/6semead/MQI/010MQI%20-Algumas%20Tecnicas%20para%20Anal%20de%20Series%20Temporais.doc>.
Acesso em: 1 jul. 2017.

DIAS, George Paulus Pereira. **Proposta de Processo de Previsão de Vendas para Bens de Consumo**. Disponível em: <www.abepro.org.br/biblioteca/ENEGEP1999_A0687.PDF> Acesso em: 10 mar. 2018.

DORFFNER, Georg. **Neural Networks for Time Series Processing**. Neural Network World. 1996.

Em: Disponível em: <http://img.clubic.com/06675738-photo-machine-learning.jpg>. Acesso em: 10 fev. 2018.

ENGLE, Robert F. Autoregressive conditional heteroskedasticity with estimates of the U.K. inflation. **Econometrica**, v. 50, p. 987-1008, 1982.

ENDERS, Walter. **Applied econometric time series**. 3 ed. Nova York, Wiley, 2009.

FAVA, Vera Lúcia. Análise de Séries de Tempo. In: VASCONCELLOS, Marco Antonio Sandoval; ALVES, Denisard. **Manual de Econometria**. São Paulo: Atlas, 2000. Cap. 9.

FELISONI DE ANGELO, Claudio; ZWICKER, Ronaldo; DIAS FOUTO, Nuno Manoel Martins; LUPPE, Marcos Roberto. Séries temporais e redes neurais: uma análise comparativa de técnicas na previsão de vendas do varejo brasileiro. **Brazilian Business Review** (Portuguese Edition), v. 8, n. 2, p.1-21, 2011.

FERNANDES, L. G. L; PORTUGAL, M. S; NAVAUX, P. O. A. **Previsão de Séries de Tempo**: Redes Neurais Artificiais e Modelos Estruturais. Porto Alegre, 1995.

GERS, Felix A.; SCHMIDHUBER, Jürgen; CUMMINS, Fred. Learning to Forget: Continual Prediction with LSTM. **Neural Computation**, v. 12, n. 10, p. 2451–2471, 2000.

GERS, Felix A.; ECK, Douglas; SCHMIDHUBER, Jürgen. **Applying LSTM to Time Series Predictable Through Time-Window Approaches**. International Institute for Advanced Scientific Studies / Societa Italiana Reti Neuroniche (SIREN). Italian workshop, 12th, 2001, Salerno, Itália.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 4. ed. São Paulo: Atlas, 1994.

GOMES, Pedro A. B., CARVALHO, Eduardo C. de, ARRUDA, Helder M., SOUZA, Paulo de, PASSIN, Gustavo. **Avaliação de Diferentes Arquiteturas de Redes Neurais Recorrentes na Previsão do Nível de Atividade de Abelhas.** XXXVII Congresso da Sociedade Brasileira de Computação, São Paulo, 2017.

GRAVES, Alex. **Generating Sequences With Recurrent Neural Networks.** arXiv:1308.0850v5 [cs.NE] 5. 2014. Department of Computer Science University of Toronto

GUIMARÃES, Roberta Valente. **Uso de Regressão Logística para previsão de fechamento de operações financeiras**: termo de moedas. Escola Politécnica da Universidade de São Paulo. São Paulo, Monografia, 2006.

GUJARATI, Damodar N. **Econometria Básica**. 4. ed. Rio de Janeiro: Editora Campus, 2006.

HAYKIN, Simon S. **Redes neurais**: princípios e práticas. Porto Alegre: Bookman, 2001.

HOCHREITER, S.; SCHMIDHUBER, J. **Long short-term memory**. Neural Computation, v. 9, n. 8, p. 1735-1780, 1997.

HOCHREITER, S.; BENGIO, Y.; FRASCONI, P.; SCHMIDHUBER, Jürgen. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: KREMER, S. C.; KOLEN, J. F. (eds.). **A Field Guide to Dynamical Recurrent Neural Networks**. IEEE Press, 2001.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. **Forecasting**: Principles and Practise. 1. ed. Disponível em: <http://otexts.org/fpp/>2013>2013.

IEMI. **Varejo de moda terá expansão com a classe C**. Disponível em: <http://www.iemi.com.br/varejo-de-moda-tera-expansao-com-a-classe-c/>. Acesso em: 10 abr. 2018.

JUNGSUK, Kim; LANE, Ian. **Accelerating large vocabulary continuous speech recognition on heterogeneous CPU-GPU plataforms**. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),p3291-3295.2014.

KDNUGGETS. Disponível em: <https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html>. Acesso em: 10 jan. 2018.

KINGMA, Diederik P.; BA, Jimmy Lei. **Adam: a Method for Stochastic Optimization**. International Conference on Learning Representations, p. 1-13, 2015.

KOHN, R. Identification Results for Armax Structures. **Econometrica**, v. 47, n. 5, p. 1295-1304, set. 1979.

KOTLER, Philip. **Administração de Marketing**. 10. ed. São Paulo: Prentice Hall, 2000.

LÄNGKVIST, Martin; KARLSSON, Lars; LOUTFI, Amy. A Review of Unsupervised Learning and Deep Learning for Time-Series Modeling. **Pattern Recognition Letters**, v. 42, p. 11-24, 2014.

LEVY, Michael; WEITZ, Barton A. **Administração do Varejo**. 3. ed. São Paulo: Atlas, 2000.

LINDBERG, E; ZACKRISSON, U. **Deciding about the Uncertain: The Use of Forecasts as an Aid to Decison-making**. Scandinavian Journal of Management. v. 7, n. 4, 1991.

MAIA, Alexandre Gori. **Econometria**: conceitos e aplicações. 1. ed. São Paulo: Saint Paul, 2017.

MAKRIDAKIS, S.; WHEELWRIGHT, S. C.; HYNDMAN, R. J. **Forecasting – methods and applications**. 3. ed. Nova York: John Wiley, 1998.

MENEZES JUNIOR, José Maria Pires de. **Redes Neurais dinâmicas parapredição e modelagem não-linear de séries temporais**. Dissertação de Mestrado. UFC. Ceará. 2006.

MONTGOMERY, D.; JOHNSON, L.; GARDINER, J. **Forecasting and Time Series Analysis**. Nova York: McGraw-Hill, 1990.

MORAES, Gustavo Inácio de ; SILVA, Thago Leal Flores da . **Determinantes macroeconômicos das vendas no varejo no período 2004-2011**. Revista de Economia Contemporânea , v. 19, p. 30-48, 2015.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. 2. ed. São Paulo: Edgard Blücher, 2006.
MSV, Janakiram. In The Era of Artificial Intelligence, GPUs Are The New CPUs. **Forbes**, 7 ago. 2017. Disponível em: <www.forbes.com/sites/janakirammsv/2017/08/07/in-the-era-of-artificial-intelligence-gpus-are-the-new-cpus/2/#255681bb4efa>. Acesso em: 5 jan. 2018.

NELSON, M.; HILL, T.; REMUS, W.; O'CONNOR, M. Time series forecasting using neural networks: should the data be deseasonalized first? **Journal of forecasting**, v. 18, n. 5, p. 359-367,1999.

NVIDIA. **Whats-difference-artificial-intelligence-machine-learning-deep-learning-ai.** Disponível em: <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/> Acesso em: 10 de janeiro de 2018.

OLAH, Christopher. Understanding LSTM Networks. **Colah's Blog**. 27 ago. 2015. Disponível em: <http://colah.github.io/posts/2015-08-Understanding-LSTMs>. Acesso em: 3 fev. 2018.

PARENTE, Juracy. **Varejo no Brasil**. São Paulo: Atlas, 2000.

PASQUOTTO, J. L. D. **Previsão de Redes Temporais no Varejo Brasileiro**: uma Investigação Comparativa da Aplicação de Redes Neurais Recorrentes de Elman. 2010. 200f. Dissertação (Mestrado em Administração) USP– Universidade de São Paulo, São Paulo, 2010.

PASSARI, A. **Exploração de Dados Atomizados para Previsão de Vendas no Varejo Utilizando Redes Neurais**. 2003. 150f. Dissertação (Mestrado em Administração). Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo, São Paulo, 2003.

PELLEGRINI, F.R., FOGLIATTO, F.S. **Passos para Implantação de Sistemas de Previsão de Demanda** - Técnicas e Estudo de Caso. Produção, v. 11, n. 1, 2001.

POMMERANZENBAUM, I. R. **Redes Neurais artificiais na predição das principais séries do índice Ibovespa e suas aplicações em sistemas automatizados de negociação**. 2014. 126f. Dissertação (Mestrado em Engenharia Elétrica) UFRJ– Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

PRAHALAD, C. K. **A Riqueza na Base da Pirâmide**. São Paulo: Bookman Companhia, 2009.

PONTI, Moacir A.; COSTA, Gabriel B. Paranhos. **Como funciona o Deep Learning**. Tópicos em gerenciamento de dados e informação. SBC. 1. ed. 2017. Disponível em:< http://conteudo.icmc.usp.br/pessoas/moacir/papers/Ponti_Costa_Como-funciona-o-Deep-Learning_2017.pdf>. Acesso em 15 de janeiro de 2018

R2RT. **Written Memories**: Understanding, Deriving and Extending the LSTM. 26 jul. 2016. Disponível em: <https://r2rt.com/written-memories-understanding-deriving-and-extending-the-lstm.html>. Acesso em: 3 jan. 2018.

RUDER, Sebastian. **An overview of gradient descendent optimization algorithms**. arXiv preprint arXiv:1609.04747. 2016.

SILVA I. N.; SPATTI, D. H.; FLAUZINO R. A. **Redes Neurais Artificiais para engenharia e ciências aplicadas**. 1. ed. São Paulo: Artliber, 2010.

SOCIEDADE BRASILEIRA DE VAREJO E CONSUMO. **O papel do varejo na economia**. Disponível em: <http://www.sbvc.com.br/wpcontent/uploads/2015/04/O-Papel-do-Varejonaeconomia_apresenta%C3%A7%C3%A3ooficial1.pdf>. Acesso em: 20 out. 2017.

THOMASSEY, Sébastien. Sales forecasts in clothing industry: The key success factor of the supply chain management. **International Journal of Production Economics**, v. 128, n. 2, p. 470-483, 2010.

THOMASSEY, Sebastien; HAPPIETTE, Michel; CASTELAIN, Jean Marie. A short and mean-term automatic forecasting system - application to textile logistics. **European Journal of Operational Research**, v. 161, n. 1, p275-p285, 10p, 16 fev. 2005.

THOMASSEY, Sebastien; HAPPIETTE, Michel; CASTELAIN, Jean Marie. An automatic textile sales forecast using fuzzy treatment of explanatory variables. **Journal of Textile and Apparel, Technology and Management**. v. 2, n. IV, p1-15,15p, 2002.

VASCONCELLOS, M. A. S.; ALVES, D. **Manual de econometria.** 1. ed. São Paulo: Atlas, 2000.

VELLASCO, L. F. **Redes Neurais Artificiais**. 2007. ICA – Laboratório de Inteligência Computacional Aplicada, PUC-RJ. Disponível em: <www2.ica.ele.puc-rio.br/Downloads/33/ICA-introdu%C3%A7%C3%A3o%20RNs.pdf>. Acesso em: 1 ago. 2017.

WOOLDRIDGE, Jeffrey M. **Introdução à econometria**: uma abordagem moderna. 4. ed. São Paulo: Cengage Learning, 2015.

WONG, W.K.; GUO, Z.X.. A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. Supply Chain Forecasting Systems, **International Journal of Production Economics**, v. 128, n. 2, p. 614-624, 2010.

WINKLHOFER, H.; DIAMANTOPOULOS, A.; WITT, S. F. Forecasting Practice: a Review of the Empirical Literature and an Agenda for Future Research. **International Journal of Forecasting**, v. 12, p. 193-221, 1996.

XIA, Min; ZHANG, Yingchao; WENG, Liguo; YE, Xiaoling. Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. **Knowledge-Based Systems**, v. 36, p. 253-259, dez. 2012.

ZHANG, G. P.; PATUWO, E.; HU, M. Y. Forecasting with artificial neural networks: the state of the art. **International Journal of Forecasting**, v. 14, p. 35-62, 1998.

APÊNDICES

APÊNDICE 1: CÓDIGO PYTHON REDE MLP

```
# Imports
import numpy
import matplotlib.pyplot as plt
import pandas
import math
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import Dropout
from keras import regularizers
from keras import optimizers
from keras.utils import plot_model
from sklearn.utils import check_array
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error

# Agora podemos definir uma função para criar um novo conjunto de dados como descrito acima
# A função recebe 2 argumentos: o conjunto de dados, que é um array NumPy que queremos converter # em um dataset e look back, que é o número de etapas de tempo anteriores para usar como variáveis
# de entrada para prever. O próximo período de tempo, neste caso, definido como 1.
#Esse padrão criará um conjunto de dados onde X é a quantidade em um determinado momento (t) e Y # é a quantidade no próximo momento (t + 1).

def create_dataset(dataset, look_back = 1):
    dataX, dataY = [], []
    for i in range(len(dataset)-look_back-1):
        a = dataset[i:(i+look_back), 0]
        dataX.append(a)
        dataY.append(dataset[i + look_back, 0])
    return numpy.array(dataX), numpy.array(dataY)# Random seed
numpy.random.seed(7)

# Função MAPE

import numpy as np
def mean_absolute_percentage_error(y_true, y_pred):
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100
def mean_absolute_percentage_errort(y_truet, y_predt):
    return np.mean(np.abs((y_truet - y_predt) / y_truet)) * 100

# Carregando o dataset
dataframe = pandas.read_csv('data/Dados.csv', usecols = [1], engine = 'python', skipfooter = 0)
dataset = dataframe.values
dataset = dataset.astype('float32')

# Normaliza os dados
scaler = MinMaxScaler(feature_range = (0, 1))
dataset = scaler.fit_transform(dataset)

# Split em treino e teste
train_size = int(len(dataset) * 0.8)
test_size = len(dataset) - train_size
train, test = dataset[0:train_size,:], dataset[train_size:len(dataset),:]
print(len(train), len(test))


# Reshape - Perceba o valor da variável look_back
```

```
look_back = 52
trainX, trainY = create_dataset(train, look_back)
testX, testY = create_dataset(test, look_back)


# Modelo MLP

n1 =208
funcaoact= 'relu'
otimizador ='adam'
model = Sequential()
model.add(Dense(n1, input_dim = look_back, activation = funcaoact))
model.add(Dense(1))
sgd = optimizers.SGD(lr=0.001, momentum=0.9, nesterov=True)
adam = optimizers.adam(lr=0.001)
rmsprop = optimizers.RMSprop(lr=0.001)
model.compile(loss = 'mean_squared_error', optimizer = adam)
model.fit(trainX, trainY, epochs=300, batch_size=2, verbose=0)

# Previsões para treino e teste
trainPredict = model.predict(trainX)
testPredict = model.predict(testX)

# Inverte as previsões por conta da normalização
trainPredict = scaler.inverse_transform(trainPredict)
trainY = scaler.inverse_transform([trainY])
testPredict = scaler.inverse_transform(testPredict)
testY = scaler.inverse_transform([testY])

# Performance do Modelo – calculo RMSE e MAPE
trainScore = math.sqrt(mean_squared_error(trainY[0], trainPredict[:,0]))
print('Score em Treino: %.2f RMSE' % (trainScore))
y_true= trainY[0]
y_pred=trainPredict[:,0]
mean_absolute_percentage_error(y_true,y_pred)
print('Score em Treino: %.2f MAPE' % mean_absolute_percentage_error(y_true,y_pred))

# Ajusta os dados de treino para o Plot
trainPredictPlot = numpy.empty_like(dataset)
trainPredictPlot[:, :] = numpy.nan
trainPredictPlot[look_back:len(trainPredict)+look_back, :] = trainPredict

# Ajusta os dados de teste para o Plot
testPredictPlot = numpy.empty_like(dataset)
testPredictPlot[:, :] = numpy.nan
testPredictPlot[len(trainPredict)+(look_back*2)+1:len(dataset)-1, :] = testPredict

# Plot do baseline e previsões
plt.plot(dataset)
plt.plot(trainPredictPlot)
plt.plot(testPredictPlot)
plt.show()
```

APÊNDICE 2: CÓDIGO PYTHON REDE MLP COM EXÓGENAS

```
# Para o plot ficar no notebook ao invés da célula
%matplotlib inline

# Aumento do tamanho da imagem
%pylab inline
pylab.rcParams['figure.figsize'] = (14, 9)

# Imports
import math
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from keras import optimizers
from keras import regularizers
from keras.callbacks import EarlyStopping
from keras.layers import Activation
from keras.layers import Dense
from keras.layers import Dropout
from keras.layers import Input
from keras.models import Sequential
from keras.optimizers import Adam
from math import sqrt
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler

# Carregando o dataset
df = pd.read_csv('data/Dados.csv', engine = 'python', skipfooter = 0)

# Checagem dos dados
df.head(5)

# Checagem do shape
print("Registros na base: {} ".format(df.shape))

# Normaliza os dados
scaler = MinMaxScaler(feature_range = (0, 1))
dataset = scaler.fit_transform(dataset)

# Vamos armazenar a variável dependente no Y
Y = df['Y']

# E retirar ela do conjunto de treinamento
del df['Y']

# E posteriormente usar somente as variáveis independentes para separar os conjuntos de treinamento e testes
X = df

print("Registros na base das variáveis independentes: {} ".format(X.shape))
print("Registros na base das variáveis dependentes: {} ".format(Y.shape))

# Agora vamos transformar os dataframes em arrays do numpy usando a propriedade .values
X = X.values
Y = Y.values

# E em seguida, vamos deixar todos os registros como float32 para melhor cálculo dos tensores
```

```
X = X.astype('float32')

# Aqui normalizamos os dados no range [0,1]
scaler = MinMaxScaler(feature_range = (0, 1))

# E aplicamos o scaler
X = scaler.fit_transform(X)

# Agora vamos dividir a base de treinamento e teste passando o X e o Y como parâmetros e escolhendo 20% como base de teste
X_train, X_test, y_train, y_test = train_test_split(X
                                                    ,Y
                                                    ,test_size=0.2
                                                    ,shuffle=False)
# Vamos deixar o shuffle como falso para ele não embaralhar os dados

print("Registros na base de treinamento (X): {} ".format(X_train.shape))
print("Registros na base de teste (X): {} ".format(X_test.shape))
print("Registros na base de treinamento (Y): {} ".format(y_train.shape))
print("Registros na base de teste (Y): {} ".format(y_test.shape))

#Hiperparâmetros do modelo
batch_size = 2
nb_epoch = 300
n_samples = X_train.shape[0]
n_samples_test = y_train.shape[0]

# Aqui vamos colocar esse callback para se caso o modelo não evoluir depois de 100 épocas ele parar de processar
early_stopping = EarlyStopping(monitor='val_loss'
                               , min_delta=0.001
                               , patience=100)

# Arquitetura 1
model_arch_1 = Sequential()
model_arch_1.add(Dense(66, input_shape=(X_train.shape[1],), kernel_initializer='uniform',
activation='relu'))
model_arch_1.add(Dense(104, kernel_initializer='uniform', activation='relu'))
model_arch_1.add(Dense(1, kernel_initializer='uniform', activation='linear'))
# A saída aqui sempre tem que ser linear
model_arch_1.compile(loss='mean_squared_error', optimizer="adam", metrics=['mse'])
model_arch_1.summary()
# Aqui vamos armazenar algumas informações de histórico do modelo enquanto ele treina
hist_model_arch_1 = model_arch_1.fit(X_train[:n_samples]
                                     , y_train[:n_samples]
                                     , epochs=nb_epoch
                                     , batch_size=batch_size
                                     , validation_data=(X_test[:n_samples_test], y_test[:n_samples_test])
                                     , verbose=1
                                     , callbacks=[early_stopping])

# No final guardamos em um dataframe os valores durante o processamento
final = pd.DataFrame()
final['loss'] = hist_model_arch_1.history['loss']
final['mean_squared_error'] = hist_model_arch_1.history['mean_squared_error']
final['val_loss'] = hist_model_arch_1.history['val_loss']
final['val_mean_squared_error'] = hist_model_arch_1.history['val_mean_squared_error']

# Plots em relação à performance do modelo na base de validação
plt.plot(hist_model_arch_1.history['val_loss'][1:-1])
plt.title('# de Loss na base de validacao')
```

```
plt.ylabel('val loss')
plt.xlabel('epoch')
plt.legend(['ADAM'], loc='upper right')
plt.show()
plt.plot(hist_model_arch_1.history['val_mean_squared_error'][1:-1])
plt.title('# MSE na base de validacao')
plt.ylabel('val_mean_squared_error')
plt.xlabel('epoch')
plt.legend(['ADAM'], loc='upper right')
plt.show()

# Nessa variavel batch, vamos salvar as predições do modelo como um todo
batch = X.shape[0]

# Em seguida fazemos as predições, e armazenamos na variavel y_pred
 y_pred_train = model_arch_1.predict(X_train, batch_size=batch, verbose=2)
 y_pred_test = model_arch_1.predict(X_test, batch_size=batch, verbose=2)

# Inverte as previsões por conta da normalização
y_pred_train = scaler.inverse_transform( y_pred_train)
y_pred_test = scaler.inverse_transform(y_pred_test)

# Função MAPE
def mean_absolute_percentage_error(y_true, y_pred):
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100

# Vamos passar para int para não haver conflito de tipos.
    y_true_train = y_train.astype('int')
    y_pred_train = y_pred_train.astype('int')
    y_true_test = y_test.astype('int')
    y_pred_test = y_pred_test.astype('int')

# Plot das predições com os dados originais
plt.plot(y_pred_test[:-1], label = "prediction",c = "b")
plt.plot(y_true_test[:-1],label = "actual data",c="r")
plt.yticks([])
plt.xticks([])
plt.title("comparison between prediction and actual data")
plt.legend()

pd.DataFrame(y_true_test).head(20)
pd.DataFrame(y_pred_test).head(20)

# Calculo MAPE e RMSE

mape_treino = mean_absolute_percentage_error(y_true_train, y_pred_train)
mape_teste = mean_absolute_percentage_error(y_true_test, y_pred_test)
rmse_treino = sqrt(mean_squared_error(y_true_train, y_pred_train))
rmse_teste = sqrt(mean_squared_error(y_true_test, y_pred_test))
print("%s, %s, %s, %s, %s   " (mape_treino,rmse_treino,mape_teste,rmse_teste))
```

APÊNDICE 3: CÓDIGO PYTHON REDE LSTM

```
# Imports
import math
import numpy
import pandas
import matplotlib.pyplot as plt
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error

# Converte array em matriz
def create_dataset(dataset, look_back=1):
    dataX, dataY = [], []
    for i in range(len(dataset)-look_back-1):
        a = dataset[i:(i+look_back), 0]
        dataX.append(a)
        dataY.append(dataset[i + look_back, 0])
    return numpy.array(dataX), numpy.array(dataY)

# Random seed
numpy.random.seed(7)

# Carrega o dataset
dataframe = pandas.read_csv('data/Dados.csv', usecols = [1], engine = 'python')
dataset = dataframe.values
dataset = dataset.astype('float32')

# Normaliza os dados
scaler = MinMaxScaler(feature_range = (0, 1))
dataset = scaler.fit_transform(dataset)

# Split em treino e teste
train_size = int(len(dataset) * 0.67)
test_size = len(dataset) - train_size
train, test = dataset[0:train_size,:], dataset[train_size:len(dataset),:]

# Reshape em X=t e Y=t+1
look_back = 52
trainX, trainY = create_dataset(train, look_back)
testX, testY = create_dataset(test, look_back)

# Reshape do input para: [samples, time steps, features]
trainX = numpy.reshape(trainX, (trainX.shape[0], 1, trainX.shape[1]))
testX = numpy.reshape(testX, (testX.shape[0], 1, testX.shape[1]))

# Cria o modelo LSTM
model = Sequential()
model.add(LSTM(52, input_shape = (1, look_back)))
model.add(Dense(1))
model.compile(loss = 'mean_squared_error', optimizer = 'adam')
model.fit(trainX, trainY, epochs = 300, batch_size = 2, verbose = 2)

# Previsões
trainPredict = model.predict(trainX)
testPredict = model.predict(testX)
```

```
# Inverte as previsões por conta da normalização
trainPredict = scaler.inverse_transform(trainPredict)
trainY = scaler.inverse_transform([trainY])
testPredict = scaler.inverse_transform(testPredict)
testY = scaler.inverse_transform([testY])

# Calcula o RMSE e o MAPE
rmse_treino = math.sqrt(mean_squared_error(trainY[0], trainPredict[:,0]))
rmse_teste = math.sqrt(mean_squared_error(testY[0], testPredict[:,0]))
y_true= trainY[0]
y_pred=trainPredict[:,0]
y_truet= testY[0]
y_predt=testPredict[:,0]
mape_treino = mean_absolute_percentage_error(y_true,y_pred)
mape_teste = mean_absolute_percentage_errort(y_truet,y_predt)
mape = mean_absolute_percentage_error(y_true,y_pred)
 print("%s, %s, %s, %s " % (mape_treino,rmse_treino,mape_teste,rmse_teste))

# Ajusta as previsões em treino para o plot
trainPredictPlot = numpy.empty_like(dataset)
trainPredictPlot[:, :] = numpy.nan
trainPredictPlot[look_back:len(trainPredict)+look_back, :] = trainPredict

# Ajusta as previsões em teste para o plot
testPredictPlot = numpy.empty_like(dataset)
testPredictPlot[:, :] = numpy.nan
testPredictPlot[len(trainPredict)+(look_back*2)+1:len(dataset)-1, :] = testPredict

# Plot do baseline e previsões
plt.plot(scaler.inverse_transform(dataset))
plt.plot(trainPredictPlot)
plt.plot(testPredictPlot)
plt.show()
```

APÊNDICE 4: CÓDIGO PYTHON REDE LSTM COM EXÓGENAS

```
%matplotlib inline

# Imports
import csv
import math
import matplotlib.pyplot as plt
import numpy as np
import pandas
import pandas as pd
import time
from keras.layers import Dense
from keras.layers import LSTM
from keras.layers.core import Dense, Activation, Dropout
from keras.layers.recurrent import LSTM
from keras.models import Sequential
from keras import losses
from pandas import read_csv
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler

# Random seed
np.random.seed(7)

# Carregando o dataset
df = pandas.read_csv('data/Dados.csv', engine = 'python', skipfooter = 0)

# Conversão da variável data
df['DATA'] = pd.to_datetime(df['DATA'])

# Guarda a data em um objeto
DATA = df['DATA']

# Adiciona o index
df.index.name = 'DATA'

# Mark index
idx = 0
data = df['DATA']

# Deleta a data
del df['DATA']

# Nova coluna para o index
new_col = data

# Insere o index
df.insert(loc=idx, column='DATA', value=new_col)

# Seta o index com a variável data
df.set_index('DATA', inplace=True)

# Check dos dados
df.head(3)


# Aqui ainda temos nosso shape com 15 colunas e 521 linhas
```

```
df.shape

# Aqui nesse values pegamos apenas os valores do dataframe e armazenamos em um objeto.
# Como estamos usando somente o values, o Python vai partir do princípio que esses dados são
# do tipo que o NumPy consegue realizar as multiplicações de matrizes
values = df.values
values.shape

# Aqui vamos fazer o scaling dos dados do objeto values entre o intervalo de [0,1]
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaled = scaler.fit_transform(values)
scaled.shape

# Aqui vamos fazer o scaling dos dados do objeto values entre o intervalo de [0,1]
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaled = scaler.fit_transform(values)
scaled.shape

# Essa função irá realizar a conversão dos dados da time series para um objeto
# com o formato de dados supervisionados.
# Em outras palavras, essa função vai colocar as variáveis categóricas na base de dados
def to_supervised(data,dropNa = True,lag = 1):
    df = pd.DataFrame(data)
    column = []
    column.append(df)
    for i in range(1,lag+1):
        column.append(df.shift(-i))
    df = pd.concat(column,axis=1)
    df.dropna(inplace = True)
    features = data.shape[1]
    df = df.values
    supervised_data = df[:,:features*lag]
    supervised_data = np.column_stack( [supervised_data, df[:,features*lag]])
    return supervised_data

# Funções de erro para verificar a acuracidade do modelo
def func_mean_absolute_percentage_error(y_true, y_pred):
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100

def func_mean_absolute_error(y_true, y_pred):
    return mean_absolute_error(final_pred,actual_pred)

def func_mean_squared_error(y_true, y_pred):
    return mean_squared_error(final_pred,actual_pred)

Os timesteps são o número de lags (defasagens) que serão usadas
# Essa variável será importante para a definição dos dados para o tipo supervisionado (variáveis
de input)
# e vai auxiliar no momento em que tivermos que passar essas informaçõe ao Keras na
definição da LSTM
timeSteps = 52

# Aqui realizamos a conversão para supervisionado usando a quantidade de timesteps para
cada uma das features originais
supervised = to_supervised(scaled,lag=timeSteps)

# Aqui visualizamos como ficou o nosso conjunto de dados
pd.DataFrame(supervised).head()
```

```
# Como podemos ver temos agora 781 features novas (15 colunas * 52 defasagens + 1 variável Y defasada que dada pela função)
supervised.shape

# Aqui vamos pegar a quantidade original de features que ainda está no nosso dataframe df, que no caso é 15
features = df.shape[1]
features

# Quantidade de lags que estamos aplicando (nesse caso como estabelecemos 52, vamos usar o mesmo número para não haver confitos)
lags = timeSteps

train_hours = 419

# Aqui agora vamos dividir o nosso X (variáveis independentes) e o nosso Y (variável dependente)
X = supervised[:,:features*timeSteps]
y = supervised[:,features*timeSteps]

# Checando agora os nossos conjuntos de dados
print (X.shape, y.shape)

# Aqui vamos dividir os nossos conjuntos de treinamento e teste
x_train = X[:train_hours,:]
x_test = X[train_hours:,:]
y_train = y[:train_hours]
y_test = y[train_hours:]
print (x_train.shape,x_test.shape,y_train.shape,y_test.shape)

# Conversão dos dados para aplicar na LSTM. dimensões = (amostras, timesteps, features)
x_train = x_train.reshape(x_train.shape[0], timeSteps, features)
x_test = x_test.reshape(x_test.shape[0], timeSteps, features)
print (x_train.shape,x_test.shape)

# Hiperparâmetros do modelo
dimensoes = x_train.shape[2]
steps_predictions = timeSteps
n1=13
otimizador = 'adam'
nr_epochs = 300
qty_batch_size = 2

# Aqui definimos um modelo com 50 unidades, passando como input 52 timesteps e 15 features
model = Sequential()
model.add(LSTM(n1, input_shape = (timeSteps,x_train.shape[2])))
model.add(Dense(1))

# Sumário do modelo
model.summary()

# Aqui compilaremos o modelo usando o Mean Absolute Error como função de perda e o otimizador ADAM
model.compile(loss =  'mean_squared_error',optimizer = otimizador)

# Nessa parte atribuiremos o treinamento no objeto history em que terá os dados do treinamento da propriedade .fit
history =  model.fit( x_train
                     ,y_train
                     ,validation_data = (x_test, y_test)
                     ,epochs = nr_epochs
```

```
                ,batch_size = qty_batch_size
                ,verbose = 2
                ,shuffle = False)

# Vamos colocar na escala principal que usamos para realizar a normalização dos dados
y_pred = model.predict(x_train)
x_train = x_train.reshape(x_train.shape[0],x_train.shape[2]*x_train.shape[1])

# Concatenação das predições com os valores da base de teste
inv_new = np.concatenate((y_pred, x_train[:,-14:]) , axis =1)

# Transformação da variável inv_new
inv_new = scaler.inverse_transform(inv_new)

# Resultset final com as predições
final_predtrain = inv_new[:,0]

# Resultset com a variável de teste
y_train = y_train.reshape(len(y_train), 1)

# Concatenação dos dados de teste Y com os valores X da base de teste
inv_new = np.concatenate((y_train, x_train[:,-14:]) ,axis = 1)

# Transformação da variável inv_new
inv_new = scaler.inverse_transform(inv_new)

# Resultset final com as informações da base de teste
actual_predtrain = inv_new[:,0]

# Resultset final com as predições
final_predtrain[:-1]
y_true=actual_predtrain
y_pred=final_predtrain

mape_treino = func_mean_absolute_percentage_errort(y_true,y_pred)
rmse_treino = math.sqrt((mean_squared_error(y_true,y_pred))

# Vamos colocar na escala principal que usamos para realizar a normalização dos dados
y_pred = model.predict(x_test)
x_test = x_test.reshape(x_test.shape[0],x_test.shape[2]*x_test.shape[1])

# Concatenação das predições com os valores da base de teste
inv_new = np.concatenate((y_pred, x_test[:,-14:]) , axis =1)

# Transformação da variável inv_new
inv_new = scaler.inverse_transform(inv_new)

# Resultset final com as predições
final_pred = inv_new[:,0]

# Resultset com a variável de teste
y_test = y_test.reshape(len(y_test), 1)


# Concatenação dos dados de teste Y com os valores X da base de teste
inv_new = np.concatenate((y_test, x_test[:,-14:]) ,axis = 1)

# Transformação da variável inv_new
inv_new = scaler.inverse_transform(inv_new)

# Resultset final com as informações da base de teste
```

```
actual_pred = inv_new[:,0]

# Resultset final com as predições
final_pred[:-1]
y_truet =actual_pred
y_predt =final_pred

mape_teste = func_mean_absolute_percentage_errort(y_truet,y_predt))
rmse_teste = math.sqrt((mean_squared_error(y_truet,y_predt))))
print("%s, %s, %s, %s " % (mape_treino,rmse_treino,mape_teste,rmse_teste))
```