UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIA E CIÊNCIAS AMBIENTAIS
CURSO DE ENGENHARIA DE PRODUÇÃO

JONATHAN SANTOS SOARES E SILVA

MODELO DE PREVISÃO DE BOLSAS DE SANGUE BASEADO EM APRENDIZADO DE MÁQUINA

MOSSORÓ
2020

JONATHAN SANTOS SOARES E SILVA

MODELO DE PREVISÃO DE BOLSAS DE SANGUE BASEADO EM APRENDIZADO DE MÁQUINA

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Engenharia de Produção.

Orientador: Prof. Dr. Breno Barros Telles Do Carmo

MOSSORÓ

2020

S586m Silva, Jonathan Santos Soares e.
Modelo de previsão de bolsas de sangue baseado em aprendizado de máquina / Jonathan Santos Soares e Silva. - 2020.
36 f. : il.

Orientador: Breno Barros Telles do Carmo.
Coorientadora: Miriam Karla Rocha.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Engenharia de Produção, 2020.

1. ARIMA. 2. Séries temporais. 3. Aprendizagem de máquina. 4. Previsão de bolsas de sangue. I. Carmo, Breno Barros Telles do, orient. II. Rocha, Miriam Karla, co-orient. III. Título.

JONATHAN SANTOS SOARES E SILVA

MODELO DE PREVISÃO DE BOLSAS DE SANGUE BASEADO EM APRENDIZADO DE MÁQUINA

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Engenharia de Produção.

APROVADO EM: 11/12/2020

BANCA EXAMINADORA

Prof. Dr. Breno Barros Telles do Carmo
Presidente

Profa. Dra. Miriam Karla Rocha
Membro

Profa. Ms. Gabrielli Yamashita
Membro

THOMAS EDSON ESPINDOLA GONCALO:05354414440
Assinado de forma digital por THOMAS EDSON ESPINDOLA GONCALO:05354414440
Dados: 2020.12.14 07:17:25 -03'00'

Prof. Dr. Thomas Edson Espíndola Gonçalo
Membro

AGRADECIMENTOS

Agradeço a Deus pelo discernimento a mim retratado e sustentação.

Agradeço a meus pais Joelma Batista e Josivan Santos durante todo esse tempo de acolhimento, sustento e por acreditar em mim.

Agradeço a minha companheira Bruna Fernanda pelo apoio em momento difíceis e direcionamentos.

Agradeço ao Prof. Dr. Thomas Edson pelo seu belo trabalho profissional.

Agradeço ao meu orientador Prof. Dr. Breno Barros pelo seu belo papel como professor e mentor do meu trabalho e a Prof. Dra. Miriam Karla e Me. Gabrielli Yamashita pela ajuda neste trabalho.

Agradeço aos meus colegas de graduação Sabrina Almeida, Pedro Lucas, Lucas Chagas e Luanna Castro pelos momentos de estudo e descontração.

A logística é o caminho mais curto entre dois pontos.

Ronaldo Paschoaloni

RESUMO

A previsão demanda de bolsas de sangue tem o propósito de contribuir para uma gestão mais eficiente dos hemocentros, direcionando campanhas de coleta mais efetivas e, consequentemente, enfrentando os problemas de rupturas de estoques que ameaçam a qualidade do serviço prestado. O problema, então, está na eficiência na gestão de bolsas de sangue, uma vez que não se tem um direcionamento ou previsão demanda desses itens em hemocentros da região. O método proposto está organizado em cinco fases: (i) extração dos conjuntos de dados através de bancos de dados relacionais, (ii) organização do banco de dados, (iii) verificação da consistência da base de dados, (iv) aplicação do método ARIMA para previsão de bolsas de sangue e (v) análise do desempenho do modelo de previsão. Foram obtidos os conjuntos de dados com visualização de suas respectivas séries temporais. Posteriormente os dados foram tratados para aplicação do método ARIMA em seguida, onde foram identificados dados discrepantes nos tipos sanguíneos AB+, AB-, B+, B- e A-. Estes tipos sanguíneos foram tratados por meio dos testes de normalidade de Shapiro-Wilk e os valores discrepantes foram substituídos pela média. De acordo com a representação de autocorrelações, autocorrelações parciais e testes de estacionariedade de Adfuller, foi feito a configuração dos parâmetros usados pelo modelo de previsão. Também foi implementado um algoritmo automático (auto.ARIMA) de detecção dos melhores parâmetros a serem usados. Após a separação dos conjuntos de dados em treino e teste, foi verificado o desempenho do modelo para os tipos sanguíneos, e dois deles (A+ e O+) apresentaram menores erros proporcionais uma vez que estes possuem seus valores de bolsas de sangue mais elevados. O modelo também apresentou indicativos de previsão de acordo com a demanda real, sejam elas negativas ou positivas. O tipo sanguíneo AB- obteve maior erro, tornando seu uso em casos práticos não indicado. Foi constatado também que os conjuntos de dados possuem grande aleatoriedade nos dados, podendo ser interessante a análise multivariada de séries temporais que abranjam períodos festivos, cirurgias eletivas, dentre outros.

Palavras-Chave: ARIMA. Séries temporais. Aprendizagem de máquina. Previsão de bolsas de sangue.

ABSTRACT

The demand forecast for blood bags has the purpose of contributing to a more efficient management of blood centers, directing more effective collection campaigns and, consequently, facing the problems of stockouts that threaten the quality of the service provided. The problem, then, lies in the efficiency in the management of blood bags, since there is no targeting or forecast demand for these items in blood centers in the region. The proposed method is organized in five phases: (i) extraction of data sets through relational databases, (ii) organization of the database, (iii) verification of the consistency of the database, (iv) application of the method ARIMA for forecasting blood bags and (v) analyzing the performance of the forecasting model. Data sets were obtained with visualization of their respective time series. Subsequently, the data were treated for application of the ARIMA method afterwards, where discrepant data were identified in blood types AB +, AB-, B +, B- and A-. These blood types were treated using Shapiro-Wilk's normality tests and outliers were replaced by the mean. According to Adfuller's representation of autocorrelations, partial autocorrelations and stationarity tests, the parameters used by the forecast model were configured. An automatic algorithm (auto.ARIMA) was also implemented to detect the best parameters to be used. After separating the data sets in training and testing, the performance of the model for blood types was verified, and two of them (A + and O +) showed lower proportional errors since they have their highest blood bag values. The model also presented forecast indicators according to real demand, whether negative or positive. The blood type AB- obtained a greater error, making its use in practical cases not indicated. It was also found that the data sets have great randomness in the data, and the multivariate analysis of time series that cover festive periods, elective surgeries, among others, may be interesting.

Keywords: ARIMA. Time series. Machine Learning. Blood bags forecast.

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

Os países mais ricos são os que mais arrecadam e fazem análises laboratoriais para a obtenção de bolsas de sangue, sendo responsáveis por 42% de todo o sangue coletado em nível mundial, abrigando, entretanto, somente 16% população (ORGANIZAÇÃO PAN-AMERICANA DE SAÚDE, 2020).

Segundo Araújo (2020), o número de doações de sangue adequado para um país varia entre 3% e 5% da população. Vale salientar que em períodos festivos, como carnaval, festas juninas e de final de ano, a demanda por hemocomponentes é maior, segundo Haubert (2020), sendo necessário, portanto, uma quantidade mais elevada de doadores ativos. A demanda por hemocomponentes para tratamentos hospitalares pode ser avaliada de uma forma analítica, levando, possivelmente, a conclusões sobre sua previsibilidade, de acordo com Gurgel (2014).

No Brasil, os hemocentros são organizações que administram bolsas de sangue adquiridas. Atualmente, não se tem um sistema integrado entre população e hemocentro, tornando a doação de sangue uma atitude espontânea de cada cidadão, segundo Marade (2019). Alguns estudos, como o de Carmo (2019), aponta para uma comunicação mais eficiente entre as partes, tornando a arrecadação mais alinhada com as necessidades dessas unidades hospitalares e a população.

De acordo com Jacobsen (2018), na gestão em hemocentros, a demanda e as doações incertas das bolsas de sangue e seu respectivo índice de alta perecibilidade torna esse processo complexo. Ainda de acordo com o mesmo autor, para contornar essa situação, podem ser utilizados métodos estatísticos ou de simulação para identificação de um direcionamento de períodos com maior possibilidade de demanda.

De acordo com Silva (2020), os métodos de análise como *bootstrap* podem ser utilizados para realizar a previsão de demanda de produtos. Entretanto, essa estratégia pode gerar superdimensionamento dos níveis de estoque. Por outro lado, o uso de métodos autorregressivos como o ARIMA é indicado para esses casos, por ser capaz de tratar dados não-estacionários e por meio de médias móveis.

A previsibilidade de hemocomponentes ou bolsas de sangue podem ser estudadas dentro de uma perspectiva dos algoritmos de aprendizagem de máquina. Os métodos SARIMA são indicados para trabalho com dados que apresentam comportamento sazonal, enquanto que os métodos *Facebook Prophet* e ARIMA são mais recomendados para séries temporais com dados estacionários ou não-

estacionários (dados em que média, variância e autocorrelação não variam ao longo do tempo), uma vez que permite o controle de um parâmetro dos dados para torná-los estacionários (AHMAR, 2018).

Ainda de acordo com Ahmar (2018), o método ARIMA é capaz de estabelecer previsões mais próximas da realidade, dado que esse modelo é um subconjunto da regressão linear e trabalha com dados anteriores para realizar previsões, tendo como finalidade analisar uma série temporal para uma previsibilidade mais detalhada e precisa dos dados.

A partir de tal análise quantitativa, é possível evitar rupturas de estoques destes itens ou períodos de estoques em níveis abaixo do recomendado. A falta desses itens acarreta problemas na saúde pública e privada, uma vez que tratamento oncológicos, cirurgias eletivas e tratamento de pacientes advindos de acidentes de trânsito e trabalho dependem dos hemocomponentes, inviabilizando os tratamentos, de acordo com Haubert (2020).

Além disso, a disponibilidade desses materiais em períodos em que a demanda é baixa poderia ser ajustada conforme a necessidade. Bolsas de sangue se dividem em hemocomponentes e estes possuem um tempo de validade em estoque limitado, de acordo com Jacobsen (2018). Assim, o uso de modelos capazes de prever as quantidades ideais de bolsas de sangue a serem coletadas em determinados períodos pode facilitar o planejamento de campanhas de arrecadação e ajudar a controlar os níveis de estoque desses itens.

Assim, este trabalho tem por objetivo propor um modelo capaz de prever a demanda de bolsas de sangue para os tipos sanguíneos A+, A-, B+, B-, AB+, AB-, O+ e O-. Para tanto, uma base de dados de consumo de bolsas de sangue foi trabalhada por meio de um modelo de aprendizagem de máquina. Foram determinados os coeficientes (*p, d e q)* para cada tipo sanguíneo na aplicação do ARIMA. Em seguida, foram feitas as previsões para o ano de 2017 e janeiro de 2018, e comparadas com os dados reais. Posteriormente, foram feitas as visualizações de previsão em comparação com os dados reais, análises das possíveis causas de erros nas previsões e desempenho do modelo de previsão.

2. REFERENCIAL TEÓRICO

A revisão de literatura está organizada da seguinte forma: método ARIMA; aplicações do método ARIMA e processos utilizados para implementação do método. Ainda, do ponto de vista teórico, não foram encontradas pesquisas para previsão de demanda de diversos tipos sanguíneos, sendo assim, esta pesquisa contribuirá para a academia e escopo da saúde.

2.1 MÉTODO ARIMA

O ARIMA é um método de previsão em séries temporais que utiliza três parâmetros (p, d e q), que são a autorregressão (AR) para o parâmetro p, média móvel (MA) para o parâmetro q e possui ainda um termo (d) de diferenciação (I) para trabalhar com dados estacionários ou não-estacionários, segundo Benvenuto (2020). Com isso, de acordo com Kahn (2020), o método ARIMA apresenta grande habilidade em lidar com mudanças de tempos relacionados, ou seja, períodos de tempo que tem relação um com o outro.

O ARIMA tem a forma da Equação 1 como processo básico que gera a série temporal de previsão, de acordo com Ogneva (2018):

$$y_t = \Theta_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \quad (1)$$

Onde:

y_t: valor atual no período t;

ε_t: erro aletório no período t;

$\varphi_i(1, 2, \ldots, p)$: parâmetro autorregressivo do modelo;

$\theta_j(1, 2, \ldots, q)$: parâmetro de média móvel do modelo.

O ARIMA possui ainda um código de execução que pode contribuir para previsões mais acuradas: o auto.ARIMA. Este cria modelos de previsão automaticamente, sem a necessidade de análises de autocorrelação, autorrelação parcial e testes de estacionariedade como o de Adfuller, de acordo com Yermal (2017). Também é possível criar sistemas integrados, que ao receberem os dados, automaticamente já fazem previsões para um período subsequente, tornando uma gestão de estoque, por exemplo, mais eficiente e dinâmica.

Algoritmos de previsão com o uso do ARIMA ou auto.ARIMA possuem grandes aplicações no campo acadêmico e profissional, uma vez que proporcionam maior eficiência na gestão de seus respectivos negócios, como na indústria e saúde. Algumas aplicações podem ser conferidas nos tópicos subsequentes.

2.1.1 Aplicações da ferramenta ARIMA

Para Alghamdi (2019), o modelo de aprendizagem de máquina com o uso do ARIMA e auto.ARIMA apresentaram grandes resultados para a previsão de períodos de grande tráfego em rodovias, assegurando uma melhor alocação de recursos como segurança rodoviária para esses locais e possibilitando uma assistência mais pontual e eficaz.

Siami-namini (2018) apresenta o método ARIMA utilizado em dados para previsibilidade financeira, mas indicando que um grande conjunto de dados para treino não necessariamente proporciona um bom desempenho para previsão, resultando ainda em um comportamento aleatório dos dados na previsão.

Jain (2017) utiliza o ARIMA para previsão do clima. De acordo com o autor, o clima é difícil de prever devido sua grande variabilidade e dependência de outros fatores ou variáveis que o influenciam. No seu trabalho, foram utilizadas métricas de desempenho como o MAE (erro médio absolute), RMSE (raiz quadrada média) e o MAPE (erro médio absoluto percentual), que torna perceptível o quão os dados previstos diferem dos dados reais.

O método ARIMA pode ser implementado também para projetos futuros de investimentos governamentais, como no estudo realizado por Sardokie (2020), em que se tem uma previsibilidade do aumento do consumo de energia elétrica para os anos posteriores no país de Gana. A previsão indica um crescimento no consumo de energia elétrica, apontando a importância de investimentos governamentais para a área, evitando possíveis faltas de energia no país.

De acordo com o trabalho de Unhapipat (2018), o ARIMA foi aplicado em um conjunto de dados de visitas turísticas na região de Bumthang, em Butão. O modelo foi capaz de prever os dados com 91% de acurácia, sugerindo um uso prático conveniente. Este modelo pode preparar o setor hoteleiro, administrando melhor seus funcionários e recursos, bem como direcionar investimentos que podem tornar a região ainda mais atrativa.

Uma área de grande interesse nas indústrias é a de planejamento, programação e controle da produção, que envolve custos logísticos, de pessoal e compras de matérias-primas, de acordo com Ramos (2015). Ainda de acordo com o mesmo autor, o ARIMA ajuda significativamente uma indústria nesse sentido, uma vez que pode contribuir para a previsão do consumo de seus produtos nos períodos subsequentes.

É possível também a aplicação de modelos de previsão com o ARIMA ou auto.ARIMA para a disseminação de doenças, como trabalhado por Awan (2020), que fez o uso do auto.ARIMA para detector quais os melhores parâmetros *p*, *d* e *q* que melhor preveem diagnósticos positivos do COVID-19 em países europeus.

Pode-se, então, aplicar o mesmo método para previsão de bolsas de sangue, por meio de configurações e algoritmos que podem ser implementados nos conjuntos de dados, como no trabalho de Jacobsen (2018) que utilizou simulação na gestão de bolsas de sangue.

Drackley (2012) realizou uma previsão de demanda de bolsas de sangue baseada em três variáveis: idade, sexo e demandas de bolsas de sangue ao longo do tempo. Obteve-se como resultados que a demanda iria gerar rupturas de estoque na região de pesquisa, atraindo atenções de secretarias da saúde para o problema.

2.2 ADFULLER, SHAPIRO-WILK, *AIC* E MÉTRICAS DE DESEMPENHO

Para uma melhor implementação do método ARIMA, utiliza-se de alguns processos para otimização do conjunto de dados e análises de desempenho do modelo. Esses processos podem ser trabalhados pelo teste de estacionariedade de Adfuller, teste de normalidade de Shapiro-Wilk, Métrica *AIC* (Critério de Informação de Akaike) e métricas de desempenho, que serão descritas a seguir.

2.2.1 Teste de estacionariedade de Adfuller

Alguns testes podem ser implementados juntamente a análise de séries temporais para um resultado mais eficaz. Um desses testes é o teste de estacionariedade de Adfuller. Este teste verifica se a série temporal obedece a uma estacionariedade, que acontece quando os dados variam regularmente em torno de uma média. O teste tem uma hipótese nula (H0), que se não for rejeitada, significa que a série temporal não é estacionário e possui alguma estrutura que depende do tempo. Caso a hipótese nula (H0) venha a ser rejeitada, tem-se a hipótese alternativa

(H1), e indica que os dados são estacionários e não dependem do tempo, de acordo com Brownlee (2020). Ainda de acordo com o mesmo autor, para esta análise, utiliza-se a obtenção de p-valores para hipótese mais adequada.

No algoritmo de estudo de estacionariedade de Adfuller (com 95% de confiança, por exemplo) quando o p-valor de um conjunto de dados for maior que 0,05, sugere-se não rejeitar a hipótese nula (H0), ou seja, os dados não são estacionários. Por outro lado, quando o p-valor assumir valor menor ou igual a 0,05, rejeitamos a hipótese nula (H0) e adota-se a hipótese alternativa (H1), ou seja, os dados são estacionários. Esta análise é importante pois um dos parâmetros a serem utilizados no ARIMA (*d*), deve conter esta informação de estacionariedade do conjunto de dados a ser trabalhado.

2.2.2 Teste de normalidade de Shapiro-Wilk

Outro teste que pode ser usado para o tratamento de dados, é o teste de Shapiro-Wilk, em que, semelhante ao teste de escionariedade, tem-se uma análise de p-valores. Este teste verifica se o conjunto de dados trabalhado possui distribuição normal, de acordo com Oppong (2016). Caso o conjunto de dados venha a apresentar uma distribuição normal, dados discrepantes ou faltantes podem ser substituídos pela média, a depender do algoritmo de previsão a ser utilizado, pois algumas ferramentas de previsão são sensíveis a dados discrepantes e/ou não permitem o trabalho com dados faltantes.

2.2.3 Métrica *AIC*

Para o Auto.ARIMA, utiliza-se a métrica de Critério de Informação de Akaike (*AIC*). Esta métrica, que de acordo com Velasco (2019), obtém cada conjunto de parâmetro *p, d* e *q*. Cada conjunto, gera um desempenho *AIC*, dessa forma, pode indicar qual o melhor conjunto de parâmetros que pode fornecer uma previsibilidade mais eficaz.

2.2.4 Métricas de desempenho

Após uma melhor otimização do modelo de previsão, tem-se uma necessidade de uma análise mais detalhada e mensurável entre o que é previsto e o que é realizado. Para problemas desse tipo (previsão de dados), utiliza-se algumas métricas de desempenho como *RMSE, MAE* e *MAPE.* Estas métricas são as mais utilizadas

para a análise de erros entre dados de treino e teste e abordam os erros de diferentes ângulos, pois tem-se o erro absoluto e um erro percentual, de acordo com Ngo (2019).

As métricas *RMSE* e *MAE* baseiam-se em valores absolutos e podem ser definidas pelas Equações 2 e 3. Já o *MAPE* fornece um erro percentual médio e pode ser verificado na Equação 4.

$$MAE = \frac{\sum_{n=1}^{N} |\hat{r}_n - r_n|}{N} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (\hat{r}_n - r_n)^2}{N}} \quad (3)$$

$$MAPE = \frac{\sum \frac{|\hat{r}_n - r_n|}{\hat{r}_n}}{N} \quad (4)$$

Onde:

$\hat{r}_n$: valor previsto;

r_n: valor real ou valor de teste;

N: quantidade de pares de valores previsto e real.

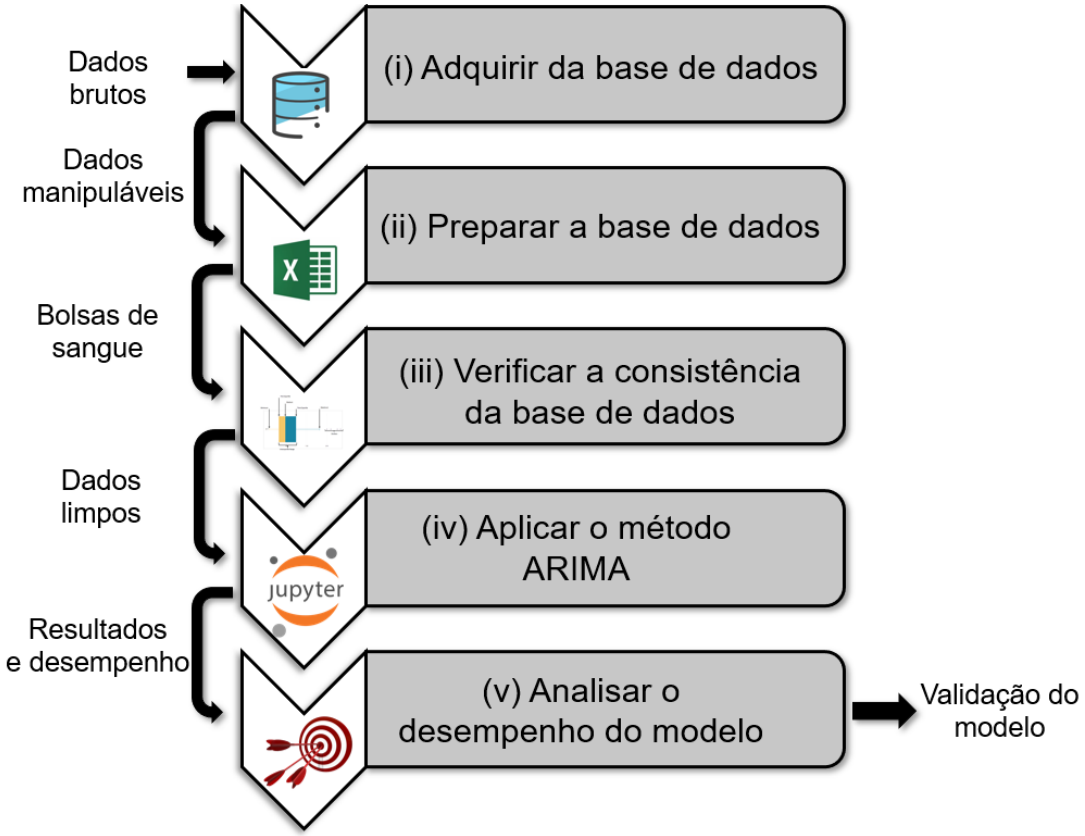
Pela Equação 2 (*MAE*), pode-se perceber que a equação faz a soma de todos os erros, sejam eles negativos ou positivos, o que mostra se os erros são predominantemente negativos ou positivos, de acordo com Wang (2018). Ainda de acordo com o mesmo autor, a Equação 3 (*RMSE*), torna os erros negativos em positivos (através do quadrado do valor, independente se o valor é negativo ou positivo), faz a soma de todos os erros e depois retira o quadrado aplicado anteriormente através da raiz quadrada. Este último mostra a magnitude dos erros, independentes se eles são negativos ou positivos.

O *MAPE* fornece uma visualização mais relativa entre os erros, e percebe-se, em termos percentuais, o quanto o previsto diferiu em relação ao de fato realizado.

3. MÉTODO DE PESQUISA

Com o objetivo de prever as bolsas de sangue dos tipos sanguíneos a partir de uma base de dados de um hemocentro, foram aplicados os métodos de previsão de demanda baseados em aprendizagem de máquina. A Figura 1 apresenta as etapas previstas para a realização da pesquisa.

Figura 1 - Representação do processo do método de pesquisa



Fonte: autoria própria, 2020.

A primeira etapa (extração de conjuntos de dados através de bancos de dados relacionais) visou coletar os dados de demanda, extraídos do software Hemovida (programa utilizado pelo Hemocentro regional estudado). A série histórica disponível no referido software apresentou dados de demanda entre os anos de 2008 e 2018 de bolsas de sangue de um hemocentro regional.

A segunda fase (organização do banco de dados) visou preparar e organizar a base de dados, estabelecendo a variável alvo (bolsas de sangue totais), ou seja, a variável que se deseja prever. Uma vez que a base de dados apresentou a demanda por hemocomponentes, para transformar esse dado na demanda por bolsas de

sangue totais (produto colhido junto ao doador), foi considerada a relação proposta por Carmo et al. (2019), sendo possível definir a variável alvo “bolsas de sangue” através da Equação 5.

$$\text{bolsas de sangue} = \text{máx}\,(\text{CP};\ \text{CH} + \text{CHPL};\ \text{PF}) \quad (5)$$

Onde:
CP: concentrado de plaquetas;
CH: concentrado de hemácias;
CHPL: concentrado de hemácias pobre em leucócitos;
PF: plasma fresco.

Na terceira fase, a consistência dos dados foi verificada, de forma a identificar a existência de valores discrepantes e faltantes. Também foi realizado o teste de *Shapiro-Wilk*, identificando se a base de dados segue ou não uma distribuição normal por meio das análises de p-valores (OPPONG, 2016), para um possível tratamento dos dados. Caso a base de dados apresente dados faltantes ou discrepantes, eles seriam substituídos pela média do conjunto de dados, se a base de dados for similar a uma distribuição normal (DO NASCIMENTO, 2012), caso contrário, não seria feita essa substituição. Devido o ARIMA ser sensível a dados discrepantes, esses dados podem indicar um maior erro nas métricas de desempenho que virão a ser utilizadas. Vale ressaltar que não existe uma regra para este procedimento, mas é conveniente para o bom ajuste de séries temporais, pois o algoritmo ARIMA é sensível a dados discrepantes, de acordo com Silva (2019).

Na quarta fase, o modelo ARIMA foi aplicado, sendo implementados os parâmetros *p*, *d* e *q* através das autocorrelações entre os dados, teste de estacionariedade de Adfuller e autocorrelação parcial. Os parâmetros *p, d* e *q* também foram mensurados de maneira exaustiva, para uma melhor adequação para o modelo de previsão. Isto foi realizado de acordo um ranking realizado pelo próprio algoritmo, onde o *AIC* é levado em conta (quanto menor, mais indicado é a configuração *p, d* e *q*).

Na quinta fase, foi feita uma análise dos erros obtidos, por meio de uma separação da base de dados entre treino (anos de 2008 a 2016) e teste (2017 a janeiro de 2018). Para a obtenção destes erros, foram escolhidas a raiz quadrada do erro

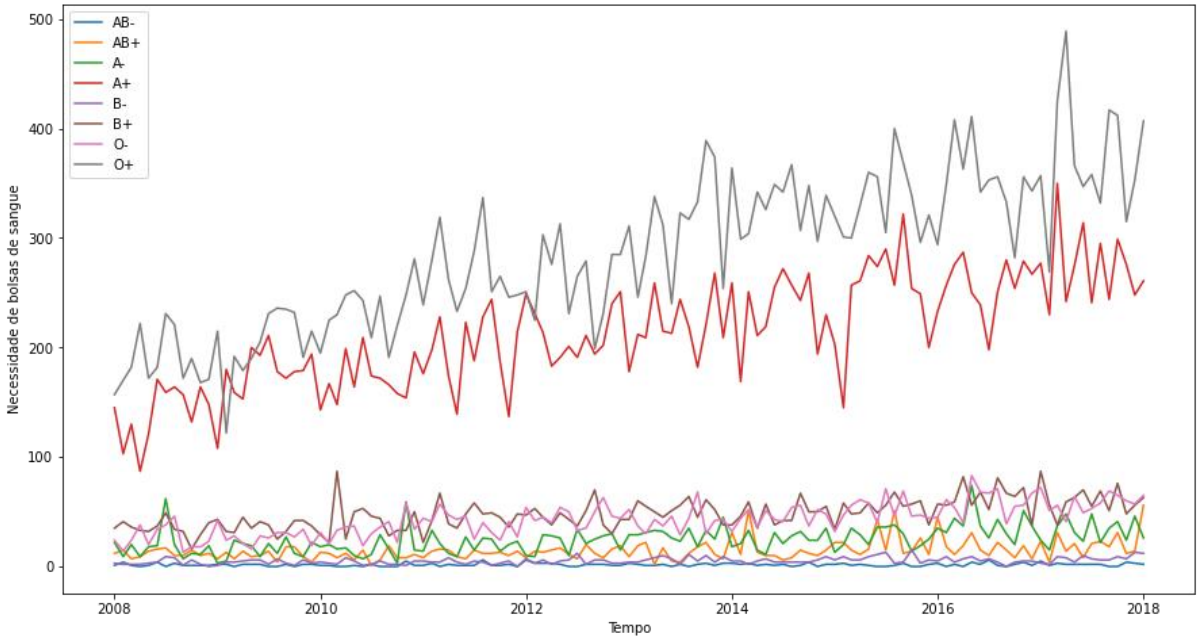
médio (*RMSE*), o erro médio absoluto (*MAE*) e o erro médio percentual absoluto (*MAPE*). Desta forma, pode-se, então, verificar a precisão da previsão e visualizar o desenvolvimento dos dados reais ao longo do tempo juntamente com os dados previstos.

4. ANÁLISE DOS RESULTADOS

Os dados foram coletados por meio da extração do software Hemovida, que apresenta uma série contínua de dez anos, entre 2008 e 2018. Esses dados eram os mais recentes e consistentes que constavam no sistema. A extração, então, foi feita por meio de softwares de banco de dados relacionais.

Após a extração dos dados, eles foram agregados a partir da Equação 5 para obtenção da variável alvo “bolsas de sangue”. A Figura 2 apresenta os dados de demanda dos hemocomponentes agregados em bolsas de sangue para cada tipo sanguíneo, totalizando 8 conjuntos de dados (A+, A-, B+, B-, AB+, AB-, O+, O-).

Figura 2 - Representação do conjunto de dados para os tipos sanguíneos (A+, A-, B+, B-, AB+, AB-, O+, O-)



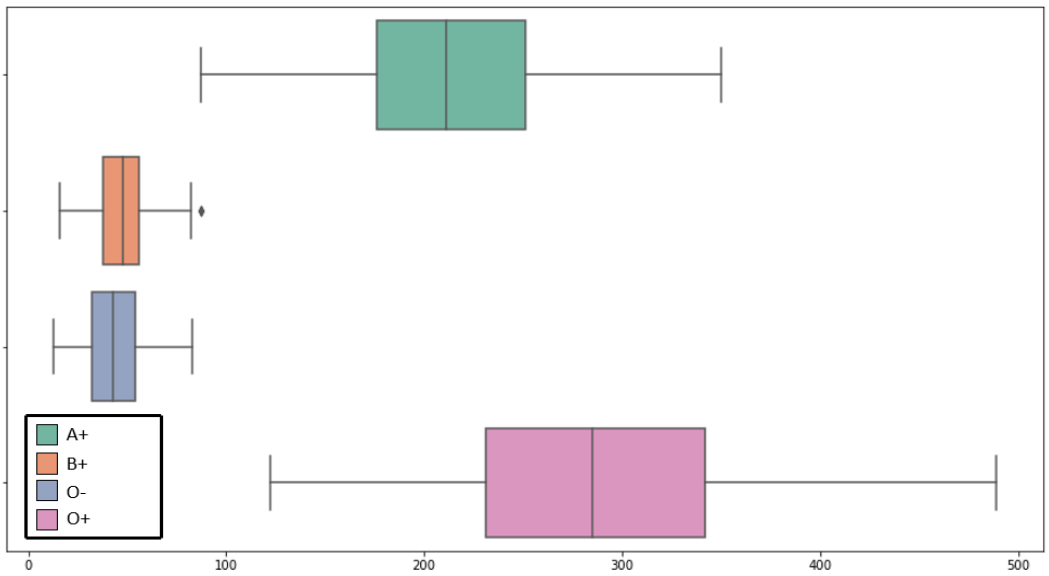
Fonte: autoria própria, 2020.

Por meio do gráfico apresentado na Figura 2, constata-se que os tipos sanguíneos não obedecem a uma sazonalidade e a demanda apresenta tendência em alguns casos (A+ e O+), justificando a não utilização de um modelo SARIMA. Neste caso, o ARIMA é mais indicado, mesmo para os conjuntos de dados A+ e O+, que apresentam comportamento tendencial.

Para todos os tipos sanguíneos não foram constatados dados faltantes, dado que a agregação dos dados, realizada por meio da Equação 5, considera o maior valor dentre os hemocomponentes para estabelecer a necessidade de bolsas de sangue.

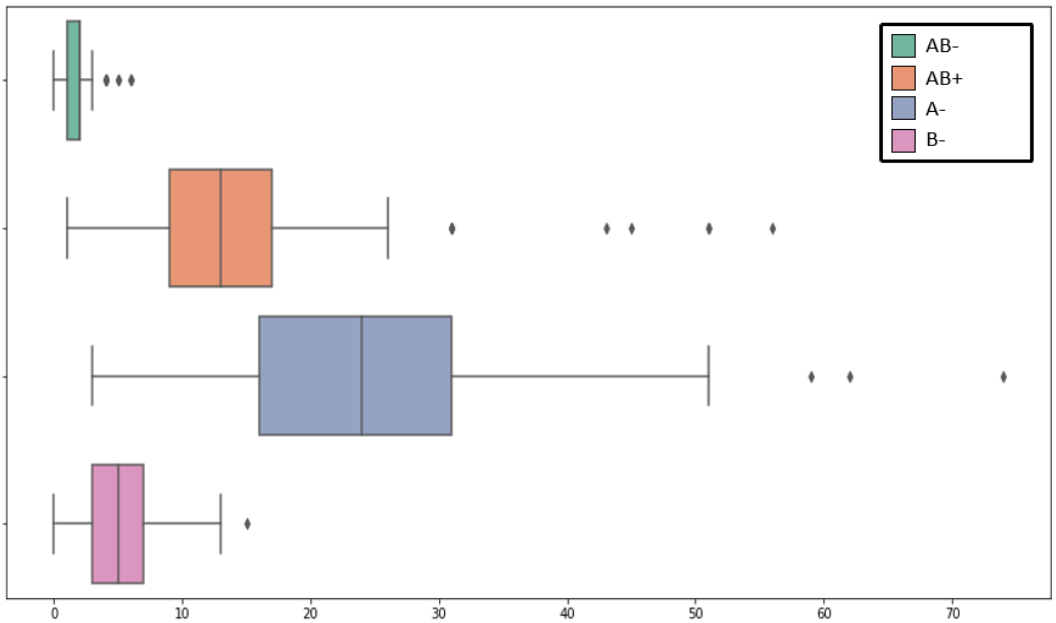
Os potenciais dados discrepantes foram identificados por meio de gráficos tipo *boxplot*, conforme apresentado nas Figuras 3 e 4.

Figura 3 - *Boxplots* para os tipos sanguíneos A+, B+, O- e O+



Fonte: autoria própria, 2020.

Figura 4 - *Boxplots* para os tipos sanguíneos AB-, AB+, A- e B-



Fonte: autoria própria, 2020.

Foram identificados dados discrepantes para os tipos sanguíneos B+, AB-, AB+, A- e B-. Considerando que o modelo ARIMA não permite trabalhar com dados faltantes (em caso de exclusão dos dados discrepantes) e que os dados discrepantes geram muita incerteza no modelo, eles poderiam ser substituídos pela média, de acordo com Do Nascimento (2012), caso apresentassem comportamento similar a uma distribuição normal.

Para identificação deste comportamento, foi utilizado o teste de distribuição normal de *Shapiro-Wilk* com 95% de confiança. O resultado do teste para os conjuntos de dados pode ser conferido na Tabela 1.

Tabela 1 – p-valores para o teste de distribuição normal de Shapiro-Wilk

Tipo sanguíneo	p-valor	Distribuição normal
AB-	0,0000	Não
AB+	0,0000	Não
A-	0,0002	Não
A+	0,7652	Sim
B-	0,0002	Não
B+	0,1089	Sim
O-	0,3834	Sim
O+	0,1935	Sim

Fonte: autoria própria, 2020.

O único tipo sanguíneo que apresenta um conjunto de dados similar a uma distribuição normal e dados discrepantes foi B+. Dessa forma, seus dados discrepantes foram substituídos pela média. Para os outros tipos sanguíneos que apresentaram dados discrepantes, mas não demonstraram um conjunto de dados similar a uma distribuição normal (AB-, AB+, A- e B-), foi aplicado o método de previsão a esses conjuntos sem a substituição dos dados discrepantes pela média. Isso pode indicar um erro mais elevado em relação aos outros conjuntos de dados nas métricas de desempenho utilizadas.

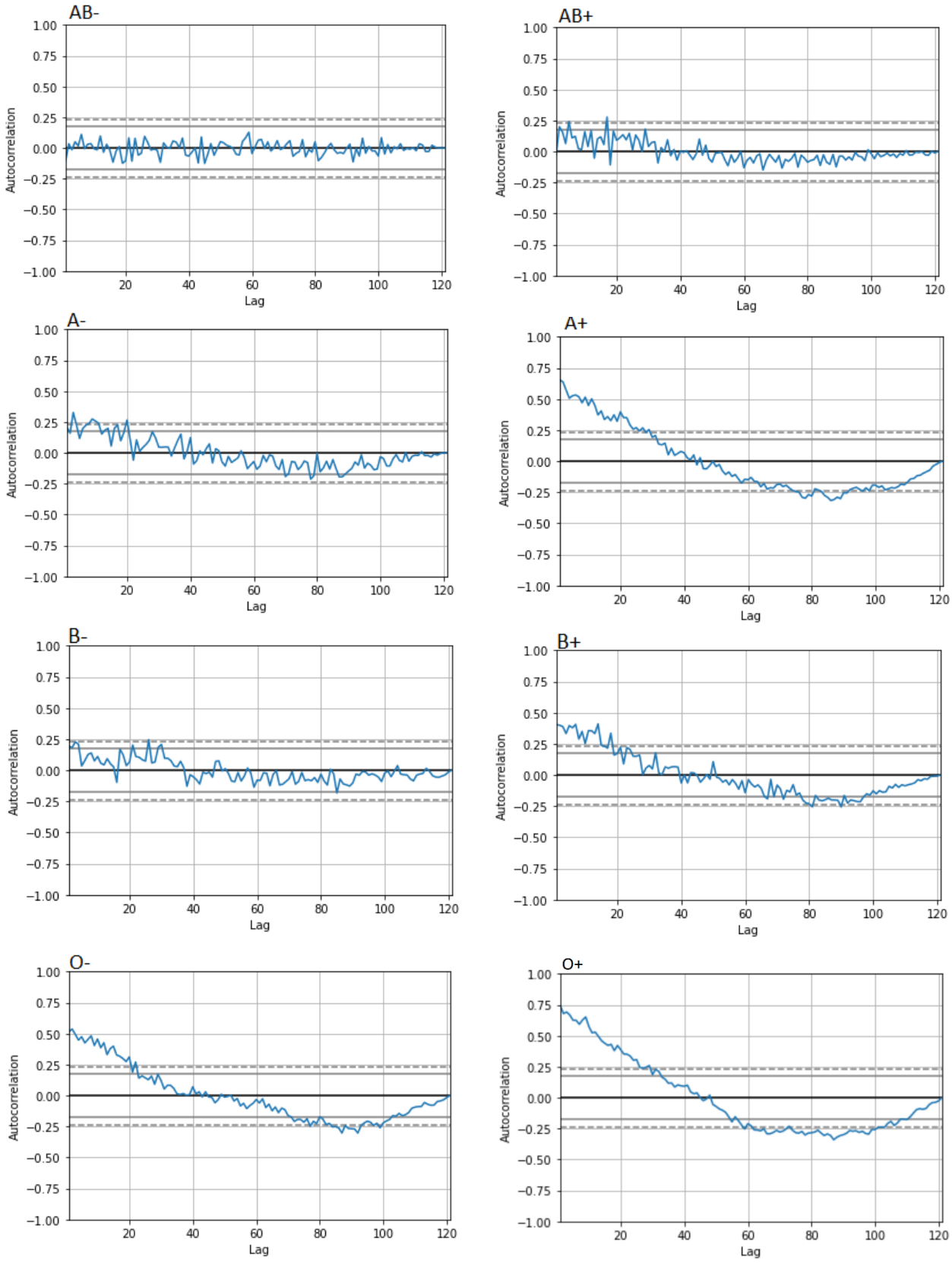
A ocorrência de dados discrepantes nos conjuntos de dados indicados pode ser explicada pelo volume reduzido de demanda destas bolsas de sangue (AB-, AB+, A-, B-, B+), pois, de acordo com Santa Casa de Misericórdia (2020), percebe-se que esses tipos sanguíneos possuem baixa representatividade na população. Desta

forma, quaisquer valores um pouco acima dos valores costumeiros já se tornam discrepantes.

Após a aquisição da base de dados, organização e verificação de consistência, foram obtidos os parâmetros do método ARIMA por meio do diagrama de autocorrelação, que estabeleceu o parâmetro *p*, conforme ilustrado na Figura 5.

Nesta representação, o intervalo para os valores de *p* será todo aquele *lag* que se encontre acima da zona de insignificância (linha tracejada). O intervalo de valores que *p* pode adotar para cada tipo sanguíneo podem ser conferidos na Tabela 2.

Figura 5 - Diagramas de autocorrelação para os tipos sanguíneos A+, A-, B+, B-, AB+, AB-, O+ e O-



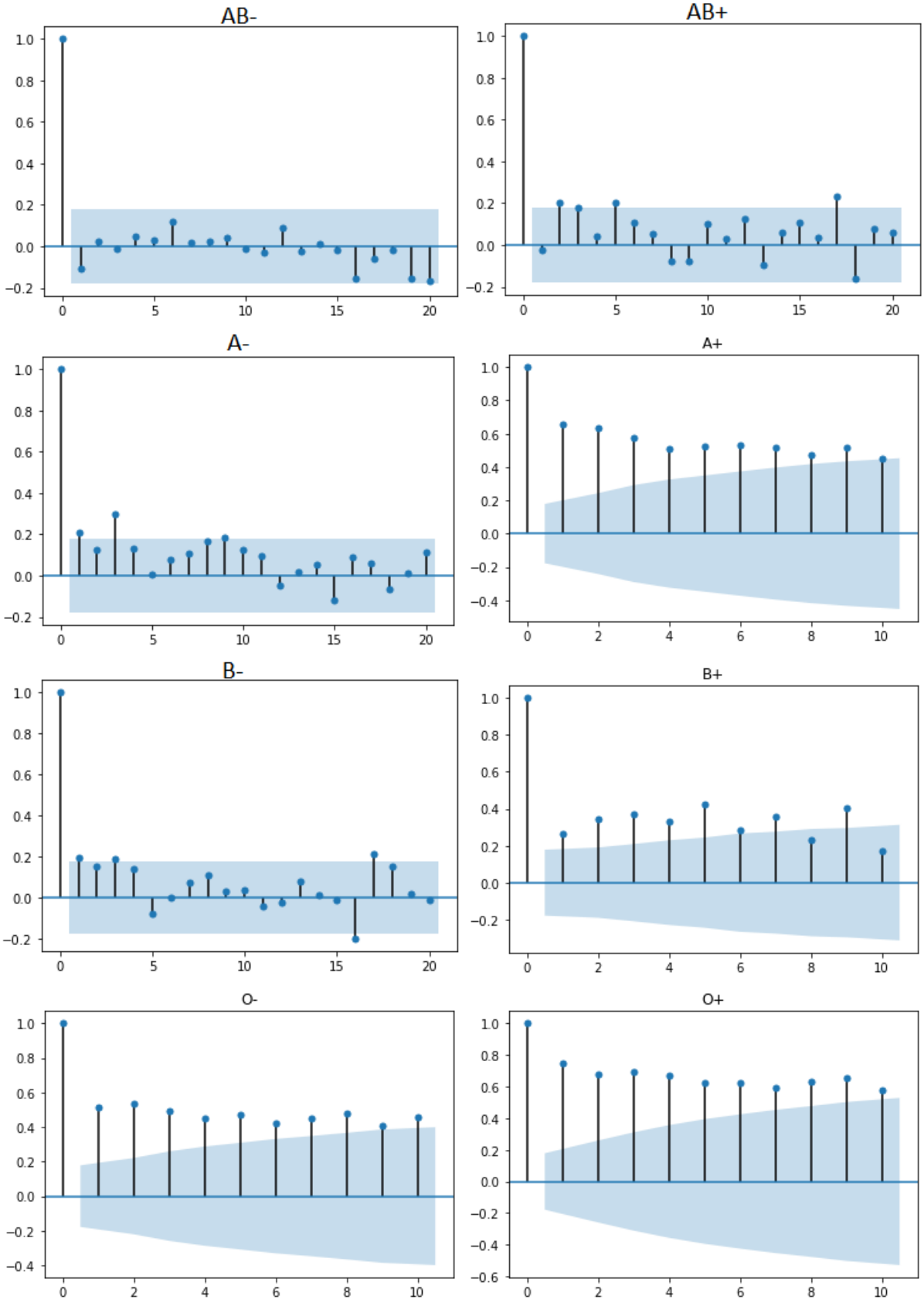
Fonte: autoria própria, 2020.

Para o parâmetro *d* (estacionariedade), foi utilizado o teste de Adfuller de estacionariedade, que pode ser usado para identificar se o conjunto de dados de cada tipo sanguíneo são estacionários, de acordo com Yuan (2020).

O teste trabalha com p-valor, adotando uma confiança de 95%, então, se o p-valor do teste for menor que 5%, indica que, de acordo com Kang (2020), que o conjunto de dados é estacionário. Sendo assim, o parâmetro *d* a ser utilizado é 0, caso o conjunto de dados seja indicado como não-estacionário, o parâmetro adotará seu valor como 1. O intervalo de valores que o parâmetro *d* pode adotar podem ser conferidos na Tabela 2, para cada tipo sanguíneo.

O parâmetro *q* é obtido por meio de uma representação de autocorrelação parcial, conforme ilustrado na Figura 6. O valor de *q* pode ser obtido analogamente ao parâmetro *p* (o intervalo a ser usado é todo aquele *lag* que se encontra além da zona de insignificância). Os intervalos de valores de *q* para cada tipo sanguíneo podem ser conferidos na Tabela 2.

Figura 5 - Autocorrelação parcial (parâmetro *q*) para cada típico sanguíneo



Fonte: autoria própria, 2020.

Assim, os possíveis parâmetros *p*, *d* e *q* são apresentados para cada tipo sanguíneo na Tabela 2.

Tabela 2 - Parâmetros utilizados para os tipos sanguíneos

Tipo sanguíneo	*p*	*d*	*q*
AB-	0	0	1
AB+	0	0	1 a 2
A-	0	1	1 a 4
A+	0 a 25	1	1 a 11
B-	0	0	1 a 4
B+	0 a 8	1	1 a 9
O-	0 a 20	1	1 a 10
O+	0 a 25	1	1 a 10

Fonte: autoria própria, 2020.

Também foi proposto um algoritmo para detectar quais os melhores parâmetros *p, d* e *q*. Esse algoritmo percorre toda a lista da Tabela 2 de parâmetros *p, d* e *q,* pois são todos os possíveis valores para se adotar para cada parâmetro, com diferentes combinações. Cada linha percorrida gera um desempenho da métrica *AIC* – Akaike Information Criterion, que mensura a qualidade de um modelo estatístico, de acordo com Cavanaugh (2019). Ao término do algoritmo, o conjunto de parâmetros é escolhido de acordo com a métrica *AIC*. Esse conjunto de parâmetros escolhido tende a ser o melhor a ser configurado para o modelo de previsão.

Para o auto.ARIMA, obteve-se como resultado os dados da Tabela 3, em que se tem os valores de *AIC* seguidos de seus parâmetros *p*, *d* e *q* mais indicados.

Tabela 3 – Parâmetros utilizados de acordo com mensuração *AIC*

Tipo sanguíneo	*AIC*	*p*	*d*	*q*
A+	1078,32	0	1	1
O+	1112,29	0	1	1
A-	853,94	0	1	1
O-	846,52	0	1	1
B+	860,45	0	1	2
AB+	807,74	0	1	1
B-	561,41	0	1	1
AB-	417,64	1	1	1

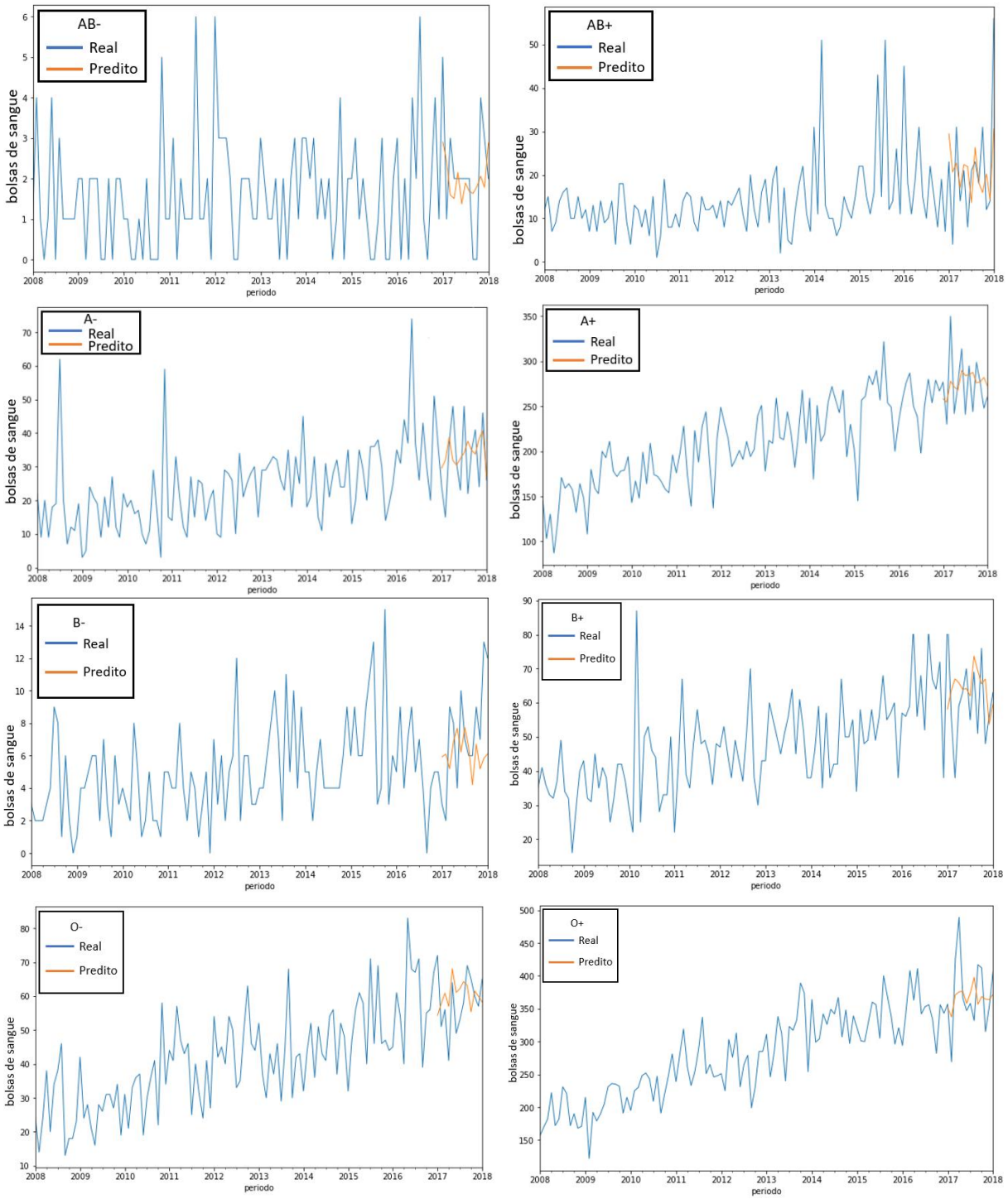
Fonte: autoria própria, 2020.

Optou-se pela ferramenta auto.ARIMA para a escolha do conjunto de parâmetros *p*, *d* e *q* para cada tipo sanguíneo, uma vez que o algoritmo percorreu todos os parâmetros possíveis da Tabela 2. As análises de autocorrelação, estacionariedade e autocorrelação parcial serviram, então, para se dimensionar quais os valores máximos que *p, d* e *q* poderiam ser percorridos pelo auto.ARIMA. Isso é interessante pois o auto.ARIMA demanda muito processamento computacional e ter um limite a ser percorrido por esses parâmetros ajuda nesse quesito. Como o auto.ARIMA automaticamente gera um resultado mais conciso e preciso do conjunto de configurações *p*, *d* e *q,* isto torna o modelo de previsão mais interessante para o problema em questão em relação a acurácia.

Com os parâmetros obtidos, foi feito a separação entre os dados de treino e teste. Para treino, o modelo trabalhou com os dados de 2008 a 2016. Para teste, optou-se por trabalhar do início de 2017 a janeiro de 2018.

Com a implementação do modelo de previsão do ARIMA, foi realizada a previsão para 2017 a janeiro de 2018. A representação gráfica dos dados reais e preditos para os tipos sanguíneos pode ser verificada na Figura 7. Os dados reais estão em azul e os preditos estão representados em laranja.

Figura 6 - Representação dos dados reais de demanda de bolsas de sangue (azul) e previsão de bolsas de sangue (laranja)



Fonte: autoria própria, 2020.

Nos tipos sanguíneos com uma certa tendência de crescimento, como acontece em A+ e O+, vide Figura 7, o algoritmo consegue acompanhar essas tendências, mantendo-se dentro dos intervalos de dados de bolsas de sangue. Pode-se verificar

também que o modelo de previsão segue as demandas de bolsas de sangue em cada período (mês), sejam crescentes ou decrescentes, apesar de não ser exato.

Para os tipos sanguíneos com valores de bolsas de sangue menores, como AB-, visualmente apresenta grandes erros, indicando uma má previsão dos valores, e, consequentemente, não se indica o uso desse tipo de previsão para casos práticos.

Após as configurações dos parâmetros e aplicação do modelo ARIMA para todos os tipos sanguíneos e visualização entre os dados previstos e reais, foram verificados seus desempenhos com métricas com o *RMSE, MAE e MAPE*, de acordo com a Tabela 4.

Tabela 4 - Métricas de desempenho dos dados de teste

Tipo sanguíneo	*RMSE*	*MAE*	*MAPE*
A+	32,038	26,325	9,68%
O+	51,631	41,846	11,20%
A-	10,379	8,304	31,58%
O-	9,221	7,965	14,30%
B+	14,450	11,111	20,24%
AB+	11,077	8,425	68,28%
B-	3,591	3,051	53,37%
AB-	1,275	1,085	0,00%

Fonte: autoria própria, 2020.

Pela Tabela 4, constata-se que os erros absolutos (*RMSE* e *MAE*) tendem a acompanhar a magnitude dos valores de bolsas de sangue (por exemplo: a magnitude das bolsas de sangue de A+ estão na casa das centenas e os erros absolutos são maiores em comparação ao tipo O-, que está na casa das dezenas, e verifica-se com erros absolutos menores). Já para os erros percentuais (*MAPE*) acontece o inverso, os erros percentuais menores foram obtidos para os tipos sanguíneos com magnitudes mais elevadas.

O conjunto de dado AB- (com pequeno escopo de valores), obteve um *MAPE* igual a 0,00%. Esse resultado se justifica pelos dados possuírem valores pequenos e baixa variância. Assim, quando é realizada a previsão, ela também apresenta valores pequenos e, quando é feito a comparação entre previsto e real, temos dados muito semelhantes, o que ocasiona um erro percentual ínfimo.

Para o modelo de previsão em questão, o ARIMA se demonstrou indicado para o problema em questão, uma vez que consegue lidar com dados estacionários ou não estacionários e consegue demonstrar acompanhamento da demanda, seja ela crescente ou decrescente.

5. CONCLUSÃO

A modelagem de previsão de bolsas de sangue para todos os tipos sanguíneos foi feita baseadas em autocorrelações, autocorrelações parciais e testes de estacionariedade, atingindo previsões dentro do escopo dos conjuntos de dados, ou seja, satisfatoriamente dentro os valores habituais, devido justamente sua grande vantagem em relação aos demais modelos de previsão, como flexibilidade para previsão, testes de adequação do modelo e previsões obtidas diretamente do modelo.

Tem-se que, apesar da adaptação a dados estacionários ou não-estacionários, a aleatoriedade contribui muito para os erros obtidos nas previsões, algo que os modelos de previsões não conseguem trabalhar devido às incertezas que podem ser associadas às séries temporais, fatores externos, por exemplo, como outras variáveis que pode contribuir para essa variabilidade.

Ficou constatado, então, que o ARIMA é mais indicado para conjuntos de dados com escopo de bolsas de sangue mais elevado (A+ e O+), pois obteve-se os menores índices na métrica *MAPE*. É evidenciado uma previsão para os tipos sanguíneos restantes (A-, AB+, B-, B+ e O-) com desempenho regular a vista de um modelo de previsão, pois, de acordo com a Figura 7, a previsão acompanha o desenvolvimento das demandas, mas pela Tabela 4, a previsão atinge patamares mais elevados (em relação aos tipos sanguíneos A+ e O+) na métrica *MAPE*. Já para o tipo sanguíneo AB-, o modelo não atinge satisfatoriamente uma previsão precisa, não indicando uma aplicação prática.

Para o ponto de visto de um gestor de um hemocentro, a previsão é capaz de tornar mais eficiente uma gestão de bolsas de sangue, uma vez que o modelo de previsão demonstrou sugestões de crescimento ou decrescimento nas demandas de bolsas de sangue, apesar de apresentar erros nas métricas de desempenho utilizadas.

De modo a tornar a previsão mais precisa, seria interessante uma análise com a presença de mais variáveis, ou seja, além da própria série temporal, a presença de dados que correlatam crescimento ou decrescimento das demandas de bolsas de sangue, tornando a análise em um modelo de previsão multivariado.

Também seria interessante a análise de outros métodos de previsão disponíveis na literatura e, da mesma forma, analisar a acurácia dos mesmos.

REFERÊNCIAS

AHMAR, Ansari Saleh et al. Modeling data containing outliers using ARIMA additive outlier (ARIMA-AO). arXiv preprint arXiv:1803.00257, 2018.

ALGHAMDI, Taghreed et al. Forecasting traffic congestion using ARIMA modeling. In: 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019. p. 1227-1232.

ARAÚJO, Ericka. Número de doadores de sangue no Brasil está abaixo da meta do OMS. Disponível em: <https://brasilcampinas.com.br/numero-de-doadores-de-sangue-no-brasil-esta-abaixo-da-meta-do-oms.html>. Acesso em: 25 out. 2020.

AWAN, Tahir Mumtaz; ASLAM, Faheem. Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. **Journal of Public Health Research**, v. 9, n. 3, 2020.

BENVENUTO, Domenico et al. Application of the ARIMA model on the COVID-2019 epidemic dataset. **Data in brief**, p. 105340, 2020.

BROWNLEE, Jason. **How to Check if Time Series Data is Stationary with Python**. Disponível em: <https://machinelearningmastery.com/time-series-data-stationary-python/>. Acesso em: 12 dez. 2020.

Carmo BBT, Lacerda DF, Queiroz PPG, Oliveira VESF, Lira lLB, Silva AAS (2019) Doar – Sistema para gestão de hemocomponentes. Mossoró (RN): Departamento de Engenharias e Ciências Ambientais. Relatório Técnico. Patrocinado pela Universidade Federal Rural do Semi Árido.

CAVANAUGH, Joseph E.; NEATH, Andrew A. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 11, n. 3, p. e1460, 2019.

DO NASCIMENTO, R. M. et al. Algoritmo de detecção e correção de outliers para previsão de carga. 2012.

GURGEL, Julia Lorena Marques; DO CARMO, Breno Barros Telles. Dimensionamento do estoque de derivados de sangue em um hemocentro do Brasil baseado em um modelo de gestão de estoques e previsão de demanda. **Revista Produção Online**, v. 14, n. 1, p. 264-293, 2014.

HAUBERT, Ana Caroline. **Demanda para doação de sangue é maior no inverno.** Disponível em: <https://diariodamanha.com/noticias/demanda-para-doacao-de-sangue-e-maior-no-inverno/>. Acesso em: 13 dez. 2012.

JACOBSEN, Victor. Gestão de Estoques de Bolsas de Sangue em um Hospital Público: uma Abordagem Usando Simulação. 218. 108 f. Tese (Engenharia, área Civil, habilitação Produção Civil.) – Engenharia de Produção Civil, Santa Catarina, 2018.

JAIN, Garima; MALLICK, Bhawna. A study of time series models ARIMA and ETS. Available at SSRN 2898968, 2017.

KANG, Eugine. Time Series: Check Stationarity. Disponível em: <https://medium.com/@kangeugine/time-series-check-stationarity-1bee9085da05>. Acesso em: 25 out. 2020.

KHAIR, Ummul et al. Forecasting error calculation with mean absolute deviation and mean absolute percentage error. In: **Journal of Physics: Conference Series**. IOP Publishing, 2017. p. 012002.

KHAN, Md Munir Hayet; MUHAMMAD, Nur Shazwani; EL-SHAFIE, Ahmed. Wavelet based hybrid ANN-ARIMA models for meteorological drought forecasting. **Journal of Hydrology**, v. 590, p. 125380, 2020.

MARADE, Chinmay et al. Forecasting Blood Donor Response Using Predictive Modelling Approach. 2019.

NGO, Ngoc-Tri. Early predicting cooling loads for energy-efficient design in office buildings by machine learning. **Energy and Buildings**, v. 182, p. 264-273, 2019.

OGNEVA, D. S.; GOLEMBIOVSKII, D. Yu. Option Pricing with Arima-Garch Models of Underlying Asset Returns. **Computational Mathematics and Modeling**, v. 29, n. 4, p. 461-473, 2018.

OPPONG, Felix Boakye; AGBEDRA, Senyo Yao. Assessing univariate and multivariate normality. a guide for non-statisticians. **Math. Theory Modeling**, v. 6, n. 2, p. 26-33, 2016.

ORGANIZAÇÃO PAN-AMERICANA DE SAÚDE. Representação da OPAS no Brasil. Brasília, 2020. Disponível em: <https://www.paho.org/bra/index.php?option=com_content&view=article&id=6196:doacoes-de-sangue-sao-essenciais-durante-pandemia-de-covid-19-afirma-opas&Itemid=875>. Acesso em: 24 out. 2020.

RAMOS, Patrícia; SANTOS, Nicolau; REBELO, Rui. Performance of state space and ARIMA models for consumer retail sales forecasting. **Robotics and computer-integrated manufacturing**, v. 34, p. 151-163, 2015.

SANTA CASA DE MISERICÓRDIA. Histórico. São Paulo. Disponível em: <https://www.santacasasp.org.br/portal/site/doe-sangue/pub/4587/web-site-hemocentro-sangue-tipo-sanguineo>. Acesso em: 31 out. 2020.

SARKODIE, Samuel Asumadu. Estimating Ghana's electricity consumption by 2030: An ARIMA forecast. **Energy Sources, Part B: Economics, Planning, and Policy**, v. 12, n. 10, p. 936-944, 2017.

SATO, Renato Cesar. Gerenciamento de doenças utilizando séries temporais com o modelo ARIMA. **Einstein (São Paulo)**, v. 11, n. 1, p. 128-131, 2013.

SIAMI-NAMINI, Sima; NAMIN, Akbar Siami. Forecasting economics and financial time series: ARIMA vs. LSTM. arXiv preprint arXiv:1803.06386, 2018.

SILVA, Amim Alleff de Souza. Adaptação do Modelo de Gestão de Estoques em Reposição Periódica por Meio do Bootstrap para Dados Não Paramétricos. 2020. 53 f. Tese (Graduação em Engenharia da Produção) – Universidade Federal Rural do Semi-Árido, Mossoró, 2020.
SILVA, Raimundo Matheus Pereira da. Predição de fluxo de tráfego de automóveis em praças de pedágio. 2019.

UNHAPIPAT, Suntaree et al. ARIMA model to forecast international tourist visit in Bumthang, Bhutan. In: **Journal of Physics Conference Series**. 2018. p. 012023.

VELASCO, Julián A.; GONZÁLEZ-SALAZAR, Constantino. Akaike information criterion should not be a “test” of geographical prediction accuracy in ecological niche modelling. **Ecological Informatics**, v. 51, p. 25-32, 2019.

WANG, Weijie; LU, Yanmin. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. In: **IOP Conference Series: Materials Science and Engineering**. 2018. p. 012049.

YERMAL, Lakshmi; BALASUBRAMANIAN, P. Application of auto arima model for forecasting returns on minute wise amalgamated data in nse. In: 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2017. p. 1-5.

YUAN, Chen Zhi; SAN, Wong Weng; LEONG, Teo Wing. Determining Optimal Lag Time Selection Function with Novel Machine Learning Strategies for Better Agricultural Commodity Prices Forecasting in Malaysia. In: Proceedings of the 2020 2nd International Conference on Information Technology and Computer Communications. 2020. p. 37-42.