UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PAULO VICTOR RIOS PINTO

PREVISÃO DA CAPACIDADE DE PROCESSAMENTO EM COMPUTADORES PESSOAIS UTILIZANDO ARIMA E ASSINATURA COMPORTAMENTAL

DISSERTAÇÃO

PONTA GROSSA
2021

PAULO VICTOR RIOS PINTO

PREVISÃO DA CAPACIDADE DE PROCESSAMENTO EM COMPUTADORES PESSOAIS UTILIZANDO ARIMA E ASSINATURA COMPORTAMENTAL

Processing Capacity Forecast in Personal Computers Using Arima and Behavioral Signature

Dissertação apresentada como requisito parcial à obtenção do título de Mestre em Ciência da Comutação do programa de Pós-Graduação em Ciência da Computação da Universidade Tecnológica Federal do Paraná – Campus Ponta Grossa.

Área de Concentração: Sistemas e Métodos de Computação

Orientador: Prof. Dr. Lourival Aparecido de Góis

PONTA GROSSA

2021

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Ponta Grossa

PAULO VICTOR RIOS PINTO

PREVISÃO DA CAPACIDADE DE PROCESSAMENTO EM COMPUTADORES PESSOAIS UTILIZANDO ARIMA E ASSINATURA COMPORTAMENTAL

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Ciência Da Computação da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Sistemas E Métodos De Computação.

Data de aprovação: 01 de Setembro de 2021

Prof Lourival Aparecido De Gois, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Augusto Foronda, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Rodolfo Miranda De Barros, Doutorado - Universidade Estadual de Londrina (Uel)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 13/09/2021.

Dedico este trabalho à minha família que me incentivou e me apoiou, e ao Pai Celestial que proporcionou os meios para a sua realização.

AGRADECIMENTOS

Existem muitos que contribuíram para essa grande realização em minha vida, tanto diretamente quanto indiretamente. Desde a minha família que me motivou e me deu apoio, ao Pai Celestial que me deu saúde e discernimento, até aos colegas com as caronas e ajudas durante a trajetória.

Agradeço por aprender que sou um filho de Deus e tenho a capacidade de aprender e progredir através do conhecimento.

Agradeço pelo exemplo de minha mãe Elvira, ao ensinar o poder transformador da educação na vida das pessoas.

Agradeço profundamente a minha esposa Geruza e meu filho Amon, pelo sacrifício de tempo e atenção cedidos e por ser um dos principais motivadores desse feito.

Agradeço também ao meu Orientador Prof. Dr. Lourival Aparecido de Gois, por acreditar em mim desde o processo de seleção dos alunos e por me guiar com sua sapiência neste mundo de descobertas.

Não poderia esquecer de mencionar o Vinicius Schultz, que além de colega de curso foi um companheiro e amigo durante esse tempo, no qual conseguimos alcançar nossos objetivos e ajudar um ao outro.

Agradeço a cada um que contribuiu para a realização deste trabalho dentro de sua esfera de atuação.

RESUMO

Delinear o comportamento de um recurso importante para um ambiente distribuído pode acarretar uma interação mais assertiva entre o gestor e os fornecedores e consumidores do mesmo, tornando seu consumo mais racional. A finalidade de uma Assinatura Comportamental é traçar o comportamento padrão do uso de um ou de um conjunto de recursos necessários para se estabelecer a capacidade que um equipamento possui de servir o ambiente distribuído a que pertence. Este trabalho apresenta uma proposta que permite a criação de uma Assinatura Comportamental através do modelo *ARIMA*, tendo como base o consumo da *CPU* em computadores pessoais, possibilitando desta forma, uma sistemática que permite a previsão deste consumo através da atualização contínua desta Assinatura. Os métodos elaborados foram avaliados através do Critério de Informação de Akaike (*AIC*), em comparação com um gerador automático de modelos *ARIMA*. Os dados obtidos e suas análises demonstraram resultados superiores aos métodos gerados pelo modelo automático, apresentando um nível de assertividade maior nessa proposta de previsibilidade e são tratados em detalhes nesta dissertação de mestrado. Outro motivador deste estudo é que as Assinaturas Comportamentais obtidas possuem um tamanho físico que viabilizará a transferência para o módulo gestor com baixo overhead de transmissão.

Palavras-chave: Séries Temporais. Previsibilidade. Recursos Computacionais. Assinatura Comportamental. *ARIMA*

ABSTRACT

Outlining the behavior of an important resource for a distributed environment can lead to a more assertive interaction between the manager and its suppliers and consumers, making its consumption more rational. The purpose of a Behavioral Signature is to trace the standard behavior of the use of one or a set of resources necessary to establish the capacity of a piece of equipment to serve the distributed environment to which it belongs. This work presents a proposal that allows the creation of a Behavioral Signature through the ARIMA model, based on the CPU consumption in personal computers, thus enabling a system that allows the prediction of this consumption through the continuous updating of this Signature. The developed methods were evaluated using the Akaike Information Criterion (AIC), in comparison with an automatic ARIMA model generator. The data obtained and their analysis showed superior results to the methods generated by the automatic model, presenting a higher level of assertiveness in this proposal of predictability and are treated in detail in this master's dissertation. Another motivator of this study is that the Behavioral Signatures obtained have a physical size that will enable the transfer to the manager module with low transmission overhead.

Keywords: Time series. Predictability. Computational Resources. Behavioral Signature. *ARIMA*

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

ACF	*Autocorrelation Function*
ADF	*Augmented Dickey-Fuller*
AE	*Absolute Error*
AIC	*Akaike Information Criterion*
AR	*Autoregressive Models*
ARIMA	*Autoregressive Integrated Moving Average*
ARMA	*Autoregressive Moving Average*
CPU	*Central Processing Unit*
DES	*Double Exponential Smoothing*
EMA	*Exponential Moving Average*
GPU	*Graphics Processing Unit*
IoT	*Internet of Things*
LSTM	*Long Short Term Memory*
MA	*Moving Average*
MAD	*Mean Absolute Deviation*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MMER	*Magnitude of Error Divided by Estimate*
MSE	*Mean Squared Error*
NMSE	*Normalized Mean Squared Error*
PACF	*Partial Autocorrelation Function*
PC	*Personal Computer*
R^2	*R Squared (Coefficient of Determination)*
RMSE	*Root Mean Squared Error*
SGR	Sistema de Gerenciamento de Recursos

SARIMA *Seasonal Autoregressive Integrated Moving Average*

ROC *Receiver Operation Characteristic*

SUMÁRIO

1 INTRODUÇÃO

Também conhecida como Assinatura Digital (HERNANDES, 2016), no que tange a recursos computacionais a Assinatura Comportamental é uma abordagem bem recente na área de Métodos Computacionais, essa ferramenta é utilizada para traçar o comportamento padrão de um componente frequentemente mensurado. Este processo torna possível a disponibilização visual e clara das características referentes a um equipamento computacional (SENGER; GOIS, 2018).

Os indicadores que apresentam o nível de uso dos recursos computacionais de uma máquina expressam as características relacionadas ao seu estado em determinado instante. Diante disso a interpretação destas informações proporciona um maior controle sobre a capacidade de utilização dos mesmos para diversas finalidades.

Através dos métodos de previsão baseados em séries temporais torna-se viável, por exemplo, realizar o dimensionamento de cargas de trabalho para provisionamento de recursos na nuvem e consequentemente diminuir o atraso da inicialização deste ambiente (OUHAME; HADI, 2019; MARINHO *et al.*, 2018; BENIFA; DHARMA, 2018; DUC *et al.*, 2019). Isso é possível em razão do conhecimento prévio obtido através da observação dos recursos, possibilitando a previsão da quantidade de recursos que devem ser definidos na parametrização do ambiente.

Na literatura existem diversos métodos baseados em séries temporais que são capazes de prever o comportamento de um recurso previamente observado. De modo a exemplificar esta variedade Chatfield (2005) apresenta uma classe de métodos de previsão que se baseiam em equações simples de atualização para calcular o comportamento futuro da série, conhecida como Amortecimento Exponencial. Porém estes métodos se aplicam a cenários específicos e normalmente são utilizados para introduzir os principais conceitos relacionados à análise de séries temporais.

Os métodos Autorregressivos são apropriados ao trabalhar com séries temporais que geralmente não são estacionárias e de tamanho curto, inclusive quando o objetivo da análise da série busca a previsão de valores futuros (SHUWAI; STOFFER, 2011). Box e Jenkins fizeram com que o modelo *ARIMA* se destacasse neste grupo de métodos devido a alguns aspectos, onde um deles está relacionado

à identificação de um modelo apropriado para cada série temporal de forma iterativa. Bisgaard (2011) diz que Box e Jenkins reconheceram a subjetividade no processo de seleção do modelo, realizando os ajustes baseados em um estudo das propriedades existentes na série temporal em questão.

Devido ao problema relacionado à falta de conhecimento da entidade que gerencia os recursos existentes em um ambiente distribuído, o presente trabalho propõe o uso do modelo *ARIMA* a fim de prover uma Assinatura Comportamental para descrever o recurso observado. Deste modo a entidade gerenciadora terá um maior conhecimento sobre a disponibilidade dos recursos, tornando a interação entre eles mais efetiva.

1.1 OBJETIVO PRINCIPAL

Prever a carga de processamento em Computadores Pessoais por meio da Assinatura Comportamental baseada no modelo *ARIMA*.

1.2 OBJETIVOS ESPECÍFICOS

- Criar o módulo responsável por obter as séries temporais relacionadas ao processamento dos computadores pessoais pertencentes aos voluntários da pesquisa;
- Gerar a Assinatura Comportamental através do modelo *ARIMA*, baseando-se nos dados de processamento dos computadores pessoais observados;
- Validar o modelo *ARIMA* utilizado na geração da Assinatura Comportamental, por meio do método *AIC (Akaike Information Criterion)*.
- Realizar a previsão dos valores reais de um dia específico, a partir dos valores existentes na Assinatura Comportamental previamente obtida para o respectivo dia da semana;
- Realizar uma análise comparativa através das métricas *MAE* e *RMSE*, que avaliam a diferença entre a observação prevista e a observação real.

1.3 JUSTIFICATIVA

Diante de um cenário baseado em processos de planejamento e tomada de decisão, a capacidade de previsão de eventos ou comportamentos futuros se torna uma contribuição crítica neste ambiente (MONTGOMERY, 2008). Por este motivo a abordagem de obtenção da disponibilidade de um determinado recurso através da verificação frequente de sua utilização em intervalos de tempo pré-determinados, apesar de efetiva apresenta alguns inconvenientes como, custo computacional, overhead de transmissão e confiabilidade nos índices levantados.

Valliyammai e Thamarai (2015) dizem que um dos propósitos de prever o comportamento de recursos em um ambiente distribuído visa a utilização eficiente do recurso no ambiente auxiliando no processo de alocação e minimizando o custo computacional. Ademais, a previsão e controle estão inclusos nos objetivos da análise de Séries Temporais citados por Brockwell (2002) e Chatfield (2005), os quais permitem conhecer os valores futuros da série e controlar alguns aspectos dentro do sistema em que o recurso está inserido, tais como a qualidade ou assertividade de um sistema físico ou econômico.

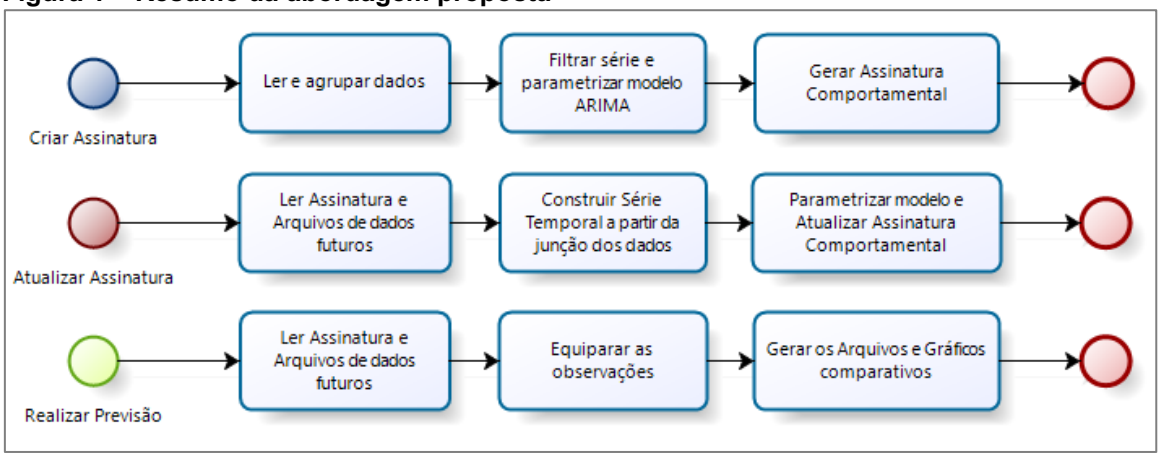
Baseado nessas premissas, estes estudos são relevantes para o desenvolvimento de uma abordagem que permita a análise de uma série temporal e com isto, definir sua tendência visando à obtenção de um comportamento sistematizado no tempo, de modo que seja possível a transferência destes dados com baixo overhead para o SGR (Sistema de Gerenciamento de Recursos), permitindo a este uma previsão de baixo erro e confiável para seus módulos de alocação.

1.4 METODOLOGIA

Este trabalho tem como objetivo criar um componente responsável pela previsibilidade de recursos computacionais em computadores pessoais utilizando a Assinatura Comportamental gerada a partir do modelo *ARIMA*. Para tanto, foi realizada uma revisão sistemática a respeito os métodos e técnicas voltados à

previsibilidade, o que permitiu identificar um baixo número de pesquisas relacionadas a este tema.

Figura 1 – Resumo da abordagem proposta



Fonte: Autoria Própria (2021)

Como mostra a Figura 1, a abordagem proposta foi concebida para possibilitar a criação da Assinatura Comportamental referente ao recurso observado, atualizar a Assinatura gerada anteriormente e prever o comportamento do recurso por meio da que foi gerada nos processos anteriores. Deste modo torna-se necessário validar o nível de assertividade do processo de previsão através da comparação da Assinatura e dos valores reais do recurso os quais não se tornaram conhecidos pela mesma durante o processo de sua construção.

Os dados obtidos a partir das observações do recurso foram agrupados de duas maneiras distintas: a primeira consiste em separar os dados por dia da semana (segunda-feira, terça-feira, etc.) devido à possibilidade de haver uma grande similaridade no comportamento da entidade observada. E a segunda, separar os dados em dois grupos, finais de semana e demais dias. Mais detalhes a respeito do conjunto de dados são apresentados na seção 5.

1.5 ORGANIZAÇÃO DO TRABALHO

Este trabalho é composto por quatro capítulos. O capítulo 1 com a introdução, os objetivos, a justificativa e a metodologia do trabalho. No capítulo 2 são apresentados os conceitos referentes às séries temporais, computadores pessoais e Assinatura Comportamental. No capítulo 3 é desenvolvido o

mapeamento sistemático para que sejam identificados os trabalhos referentes ao tema de previsibilidade de recursos através de séries temporais. E finalmente, o capítulo 4 apresenta a abordagem utilizada neste trabalho.

2 REFERENCIAL TEÓRICO

A fim de fundamentar a construção do conhecimento, serão expostos os principais conceitos sobre as áreas que foram estudadas ao decorrer da presente pesquisa.

2.1 SÉRIES TEMPORAIS

O uso de séries temporais para análise de dados, estudo de comportamentos e padrões, tendo em vista a previsibilidade de um recurso computacional, teve origem na área de econometria, conhecida também como o estudo do comportamento financeiro. Esta abordagem foi expandindo gradativamente sua aplicação em diversas áreas, adquirindo uma importância essencial na literatura (WOOLDRIDGE, 2007).

Uma série temporal pode ser definida como um conjunto de dados organizados em uma sequência de observações quantitativas sobre um sistema ou processo, os quais foram obtidos ao longo do tempo e em intervalos de tempo iguais entre si (PAL; PRAKASH, 2017). Podendo ser representada através da notação apresentada na equação 2.1, a qual indica o instante da observação *S* em sua ordem cronológica *t*:

$$S_t = s_1, s_2, s_3, \dots s_t \tag{2.1}$$

Diante de uma realidade onde a criação, manipulação, troca e armazenagem de dados são tão frequentes, o que ou quais motivos justificam a obtenção de uma quantidade considerável de dados através do monitoramento de um sistema ou processo qualquer? A motivação provém dos resultados obtidos ao realizar uma análise cuidadosa de um conjunto de dados caracterizado como uma série temporal, na qual se pode atingir ao menos um dos seguintes propósitos (CHATFIELD, 2005):

1. Entender e interpretar quais as forças subjacentes que produzem o estado observado de um sistema ou processo no decorrer do tempo;
2. Prever o estado futuro do sistema ou processo através das características identificadas no processo analítico.

Existem diversos métodos criados com o fim de modelar e prever o comportamento de recursos com eficiência. Porém a natureza variável dos dados referente às observações deste recurso faz com que um novo método proposto não seja válido para todas as aplicações avaliadas.

A vasta utilização de séries temporais na busca pela resolução de problemas de pesquisa atraiu a atenção de vários pesquisadores nas últimas décadas, apresentando um alto grau de relevância e contemporaneidade relacionadas ao problema de pesquisa desenvolvido neste trabalho.

Pandith e Babu (2018) indicam que para atingir seu principal objetivo, o qual é desenvolver um modelo que descreva o comportamento dos dados coletados e consequentemente prever valores futuros, os modelos de previsão baseados em séries temporais utilizam-se essencialmente do passado. Quanto mais obsoleta for a origem deste histórico contínuo, as previsões realizadas por meio destes dados tendem a ser mais assertivas.

Devido a esta capacidade de "conhecer" o futuro, a gama de aplicações é extremamente vasta, podendo auxiliar não somente na área da Ciência da Computação, mas nas áreas de Engenharia Elétrica (DU *et al.*, 2018) e Climatologia (VOYANT *et al.*, 2017) por exemplo. Porém ao concentrarmos nesta área, observamos que o uso da presciência pode auxiliar em problemas de pesquisa relacionados a redes de computadores (YOO; SIM, 2016; HUA *et al., 2018*), *cloud computing* (MARINHO *et al.*, 2018; BENIFA; DHARMA, 2018; DUC *et al.*, 2019), *grid computing* (VALLIYAMMAI; THAMARAI, 2015), *clusters* (SENGER; GOIS, 2017; 2018), sistemas veiculares (ZHANG *et al.*, 2016), dentre outros.

A análise de Séries Temporais faz uso de diversos conceitos inerentes ao processo de análise dentre os quais alguns deles serão citados de forma recorrente durante a construção deste trabalho, portanto uma breve contextualização será realizada a seguir.

2.1.1 Estacionariedade

Na teoria da estimativa estatística existe uma suposição muito importante onde, para que a estatística da amostra seja confiável, a população não passa por mudanças sistêmicas ou fundamentais sobre os indivíduos da amostra ou sobre o

tempo durante o qual os dados foram coletados. Essa suposição garante que as estatísticas da amostra não sejam alteradas e serão mantidas para entidades que estão fora da amostra usada para sua estimativa (PAL; PRAKASH, 2017).

Isso também se aplica na análise de séries temporais, onde a média, variância e autocorrelação estimadas possam ser usadas como base para ocorrências futuras. Tal suposição é denominada como estacionariedade, a qual exige que as estruturas internas da série não sejam alteradas ao longo do tempo.

Bisgaard (2011) afirma que a suposição de que uma série temporal gerada através de dados reais seja estacionária é algo bastante irreal, porém para que o processo analítico tenha sucesso estes dados devem possuir uma confiabilidade estatística que reside na estacionariedade. Uma série temporal S_t é considerada estacionária quando a mesma possui as seguintes características:

1. Média constante: $E(S_t) = \mu(t)$, para todo t = 1, 2, 3,...;
2. Variância constante: $Var(S_t) = \sigma_\varepsilon^2$;
3. Autocorrelação sem variação em relação ao tempo: $E(S_t, S_{t+h}) = g(h)$.

Caso uma série temporal seja definida como não estacionária, para realizar o processo de análise dos dados será necessário realizar o processo de diferenciação deste conjunto de dados, o que tornará a série estacionária e passível de prevê-la.

2.1.2 Diferenciação

Como a manipulação de uma série estacionária é algo raro de ser encontrado (PAL; PRAKASH, 2017; BISGAARD, 2011), um processo realizado de forma recorrente junto a estes dados é a diferenciação, o qual consiste em remover os sinais de tendências e reduzir a variância, permitindo tratar os dados como de uma série estacionária.

Bisgaard (2011) identifica dois tipos de não estacionariedade, a principal causa, como tratar o conjunto de dados e o nome do processo realizado, apresentados no Quadro 1.

Quadro 1 – Tipos de não Estacionariedade

Tipo	Causa	Resolução	Processo

1	Mudança de Nível; e Variabilidade Homogênea	Primeira Diferença $\nabla s_t = s_t - s_{t-1}$	Não Estacionário de Primeira Ordem
2	Alteração do Nível e Inclinação; e Variabilidade Homogênea	Segunda Diferença $\nabla^2 s_t = \nabla(s_t - s_{t-1})$ $= (s_t - s_{t-1}) - (s_{t-1} - s_{t-2})$ $= s_t - 2s_{t-1} + s_{t-2}$	Não Estacionário de Segunda Ordem

Fonte: Autoria Própria (2021)

Sendo s_t uma observação no tempo t em determinada série temporal, a sua diferenciação se dá na diferença entre o seu valor no tempo t e o seu valor no tempo $t-1$. Caso seja necessário diferenciar os dados d vezes, supondo que $w_t = \nabla s_t = s_t - s_{t-1}$ pode ser utilizada a seguinte equação $w_t = \nabla^d s_t$.

Durante este processo existem duas preocupações as quais são denominadas subdiferenciação e superdiferenciação. A primeira ocorre ao diferenciar poucas vezes uma série temporal de modo a não torná-la estacionária, e a segunda refere-se a diferenciar uma série já estacionária.

Como ambos os casos podem acarretar em problemas durante o processo de análise, o teste *ADF* verifica a existência de raízes unitárias na regressão automática, o que indicaria não estacionariedade e consequentemente a necessidade de diferenciação (BISGAARD, 2011).

2.1.3 *Augumented Dickey-Fuller*

Primeiramente, para identificar a existência de propriedades que caracterizam a série como não estacionária é necessário realizar testes estatísticos conhecidos como testes de raiz unitária. O teste pode ser realizado através do seguinte modelo:

$$Y_t = \rho Y_{t-1} + e_t \tag{2.2}$$

Sendo u_t o termo de erro aleatório conhecido como ruído branco, caso possua média zero, variância constante e ausência de autocorrelação. A existência da raiz unitária é dada quando o coeficiente $\rho = 1$.

De modo a expor as hipóteses de teste, a equação 2.2 pode ser reescrita da seguinte maneira:

$$\Delta Y_t = (\rho - 1)Y_{t-1} + e_t \\ = \gamma Y_{t-1} + e_t \tag{2.3}$$

Onde neste caso Δ representa o operador de primeira diferença e $\gamma = \rho - 1$. As hipóteses que serão testadas são denominadas hipótese nula (H_0) e a hipótese alternativa (H_1) apresentadas na equação 2.4, as quais podem ser representadas de duas maneiras, ao aplicar a substituição demonstrada na equação 2.3 em relação ao coeficiente γ (CAZAROTTO, 2006).

$$H_0: \rho = 1 \leftrightarrow H_0: \gamma = 0 \\ H_1: \rho < 1 \leftrightarrow H_0: \gamma < 0 \tag{2.4}$$

A hipótese nula se refere à presença da raiz unitária ou não estacionariedade na série temporal, enquanto a hipótese alterativa sugere estacionariedade da série.

Por fim, a equação 2.5 representa o teste *ADF*, o qual provê uma capacidade maior de testes devido ao componente Autorregressivo incluído através das várias ocorrências Δy_{t-p} (ALVES, 2008).

$$\Delta Y_t = \gamma Y_{t-1} + \Delta Y_{t-1} + \Delta Y_{t-2} + \cdots + \Delta Y_{t-p} + e_t \tag{2.5}$$

2.1.4 Autocorrelação

É definida como sendo a dependência serial linear entre determinada observação, s_t e s_{t+h} ao longo do tempo, fornecendo o grau de relação entre os valores de uma série temporal em momentos diferentes. Sendo assim este conceito possui um papel fundamental na compreensão dos valores passados para obter informações necessárias na previsão dos valores futuros (BROCKWELL, 2002). Caso o resultado seja positivo isso indica que as observações possuem correlação, sendo negativo possuem correlação inversa, e zero não possuem correlação linear.

A autocorrelação pode ser expressa por meio da equação 2.6, sendo uma função do intervalo de tempo h independente do índice de tempo real t. Tal definição permite que a autocorrelação seja uma propriedade independente do tempo, podendo ser utilizada para inferir o comportamento futuro da série temporal (PAL; PRAKASH, 2017).

$$E(s_t, s_{t+h}) = g(h) \tag{2.6}$$

Enders (2015) diz que todas as correlações indiretas estão presentes na função de autocorrelação de qualquer processo Autorregressivo, em contrapartida a autocorrelação parcial entre s_t e s_{t+h} elimina os efeitos dos valores no intervalo entre s_{t+1} e s_{t+h-1}.

2.1.5 Modelo ARIMA

Conhecido como um dos principais e mais populares modelos baseados em séries temporais, para compreender o conceito que envolve este modelo é preciso definir os modelos que compõem o *ARIMA*. Em seguida serão explorados os modelos Autorregressivos (AR) e de Médias Móveis (MA).

Swamynathan (2017) define o modelo Autorregressivo (AR) como uma involução da série temporal em relação a si mesma, onde a combinação linear de observações passadas é usada na previsão de observações futuras. O modelo está baseado na importância dos valores anteriores no processo de previsão, também conhecido como observações defasadas ou atrasadas, onde o valor atual pode ser apresentado como uma função baseada em valores passados devido à correlação existente entre as observações da série analisada.

O modelo Autorregressivo é definido como $AR(p)$, onde p refere-se à sua ordem, os modelos de primeira ordem $AR(1)$, segunda ordem $AR(2)$ e p-ésima ordem $AR(p)$ são definidos respectivamente como:

$$S_t = \emptyset_1 \, \varepsilon_{t-1} + \varepsilon_t \tag{2.7}$$

$$S_t = \emptyset_1 \varepsilon_{t-1} + \emptyset_2 \varepsilon_{t-2} + \varepsilon_t \tag{2.8}$$

$$S_t = \emptyset_1 \varepsilon_{t-1} + \emptyset_2 \varepsilon_{t-2} + \cdots + \emptyset_p \varepsilon_{t-p} + \varepsilon_t \tag{2.9}$$

Nas equações 2.7, 2.8 e 2.9, $\emptyset_i$ é o coeficiente do modelo, ε_t representa um erro no tempo t, e p está relacionado à ordem do modelo. Como o modelo AR utiliza a relação de dependência existente entre uma observação e um intervalo de observações no passado, este modelo exige como parâmetro o valor de p o qual representa o número de observações atrasadas que devem ser consideradas (PAL; PRAKASH, 2017).

No modelo de Médias Móveis (MA) utiliza-se os erros de previsões passadas ou defasamentos dos erros aleatórios de previsão, no qual a dependência entre determinada observação e um erro residual são obtidos através da aplicação do modelo a um conjunto de observações atrasadas q, ajudando no ajuste de ocorrências imprevistas de modo a não resultar em grandes discrepâncias (SWAMYNATHAN, 2017).

$$S_t = \alpha - \theta_1 \varepsilon_{t-1} + \varepsilon_t \quad (2.10)$$

$$S_t = \alpha - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} + \varepsilon_t \quad (2.11)$$

$$S_t = \alpha - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2.12)$$

Nas equações 2.10, 2.11 e 2.12 são expostos consecutivamente, os modelos de Médias Móveis de primeira ordem $MA(1)$, segunda ordem $MA(2)$ e q-ésima ordem $MA(q)$, onde ε_t representa um erro no tempo t, α é a intercepção média (PAL; PRAKASH, 2017).

Além dos modelos AR e MA, o modelo ARIMA possui um componente de integração responsável pela aplicação do processo de diferenciação d vezes, o qual irá "estacionarizar" a série temporal. Este componente é o primeiro a ser aplicado no conjunto de dados, tornando a série previsível, em seguida aplica-se o modelo ARMA.

O processo de diferenciação é detalhado na seção 2.1.2, para representar este processo será alterada a incógnita $\emptyset$ por φ na equação que define o modelo ARIMA (BISGAARD, 2011).

$$S_t = \sum_{i=1}^{p+d} \varphi_i s_{t-i} + \varepsilon_t \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (2.13)$$

2.1.6 Parâmetros do Modelo

No processo de previsibilidade é necessário definir corretamente os parâmetros de entrada do modelo $ARIMA(p, d, q)$, os quais variam de acordo com a série em análise. Os parâmetros são necessários para que cada um dos componentes apresentados anteriormente trabalhe em conjunto, de modo a alcançar um modelo bem definido capaz de delinear o comportamento da entidade previamente observada.

O teste *ADF* permite avaliar a estacionariedade da série, como citado na seção 2.1.3, onde o resultado do teste com valor de $p > 0.05$ significa que os dados não são estacionários. Isso leva à diferenciação da série até que a mesma se torne estacionária, a partir disso é obtido o número necessário de diferenciações o qual definirá o valor do parâmetro de Integração (d) do modelo *ARIMA*.

Como a autocorrelação parcial mensura a correlação entre duas observações sem considerar a influência dos valores intermediários, a sua utilização é indicada para definir o valor do parâmetro Autorregressivo (p) (PAL; PRAKASH, 2017; CHATIFIELD, 2005). Pal e Prakash afirma que, a função de autocorrelação define bem a correlação serial de erro e é, portanto, usada para detectar o parâmetro de Médias Móveis (q).

Geralmente o valor dos parâmetros p e q são menores ou iguais a 3, sendo necessária ao menos uma quantidade $n = 50$ observações para análise e o intervalo ($lags$) das funções de autocorrelação e autocorrelação parcial em torno de $n/4$. Wei (2006) apresenta as características teóricas do ACF e PACF para processos estacionários a fim de auxiliar o processo de definição destes parâmetros, apresentadas através do Quadro 2.

Quadro 2 – Padrões Teóricas do ACF e PACF para Processos Estacionários

Processo	ACF	PACF
AR(p)	Declive exponencial ou onda senoidal amortecida	Corte após $lag\ p$
MA(q)	Corte após $lag\ q$	Declive exponencial ou onda senoidal amortecida
ARMA(p, q)	Queda após o atraso ($q - p$)	Queda após o atraso ($p - q$)

Fonte: Adaptado de (Wei, 2006).

Segundo Wei (2006) existe uma necessidade de seguir este modelo na definição dos parâmetros, pois em muitos casos a variação amostral e a correlação entre os resultados das funções *ACF* e *PACF* muitas vezes disfarçam os padrões teóricos.

2.1.7 Métricas de Acurácia

Existem diversas métricas responsáveis pela verificação da diferença entre o valor real e o valor previsto, resultando no erro obtido sob a respectiva unidade de medida, a seguir serão apresentados algumas destas ferramentas.

O MAPE é definido pela equação 2.14, sendo R_t o valor real e P_t o valor previsto. Embora simples esta métrica penaliza os erros negativos, tendenciado aos métodos de previsão com valores mais baixos (TOFALIS, 2015).

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{R_t - P_t}{R_t} \right| \tag{2.14}$$

O MAE é definido pela equação 2.14, sendo R_t o valor real e P_t o valor previsto. Esta métrica é bem difundida para medição do erro de previsão em séries temporais, a qual usa a mesma escala dos dados inseridos no processo (HYNDMAN; KOEHLER, 2006).

$$MAE = \frac{\sum_{t=1}^{n} |R_t - P_t|}{n} \tag{2.15}$$

O RMSE é definido pela equação 2.14, sendo R_t o valor real e P_t o valor previsto. Como esta métrica eleva o erro ao quadrado, ela é muito útil para identificação de picos (*outliers*). Vale ressaltar que para todas as métricas apresentadas, quanto menor o valor melhor será o resultado já que ele representa o erro de previsão.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (R_t - P_t)^2} \tag{2.15}$$

2.2 ASSINATURA COMPORTAMENTAL

Para definir o comportamento padrão de uma entidade específica e viabilizar a sua previsão, será estabelecido um perfil de utilização através do seu monitoramento. Um método que permite a criação deste tipo de perfil é denominado como Assinatura Comportamental, o qual determina o desempenho de um recurso computacional de qualquer natureza dentro de um contexto para assim demonstrar de forma embasada o seu funcionamento regular (SENGER; GOIS, 2018).

O processo de construção da Assinatura Comportamental se dá através da obtenção constante dos dados relacionados ao recurso desejado em intervalos pré-estabelecidos arbitrariamente, organizados de acordo com o instante em que a observação foi realizada, assim caracterizando uma série temporal. Posteriormente estes dados serão gravados em uma base de dados, os quais serão normalizados e utilizados como entrada em um algoritmo responsável pela definição da Assinatura (SENGER; GOIS, 2017).

Como citado previamente, a utilização de Assinatura Comportamental possui um âmbito consideravelmente amplo em relação a sua aplicabilidade, dentre os quais se podem citar dois cenários: o primeiro está relacionado ao comportamento de uma rede de computadores possibilitando a identificação de anomalias durante o monitoramento das redes, auxiliando a descoberta de contextos críticos tanto para sobrecarga quanto para queda de desempenho (ZACARON, 2012; ADANIYA, 2012; ZACARON, 2013; PENA, 2014; HERNANDES, 2016); e o segundo refere-se ao desempenho de computadores pessoais permitindo definir a capacidade computacional existente nos clusters formados por computadores pessoais através de diversos recursos (SENGER; GOIS, 2017; 2018).

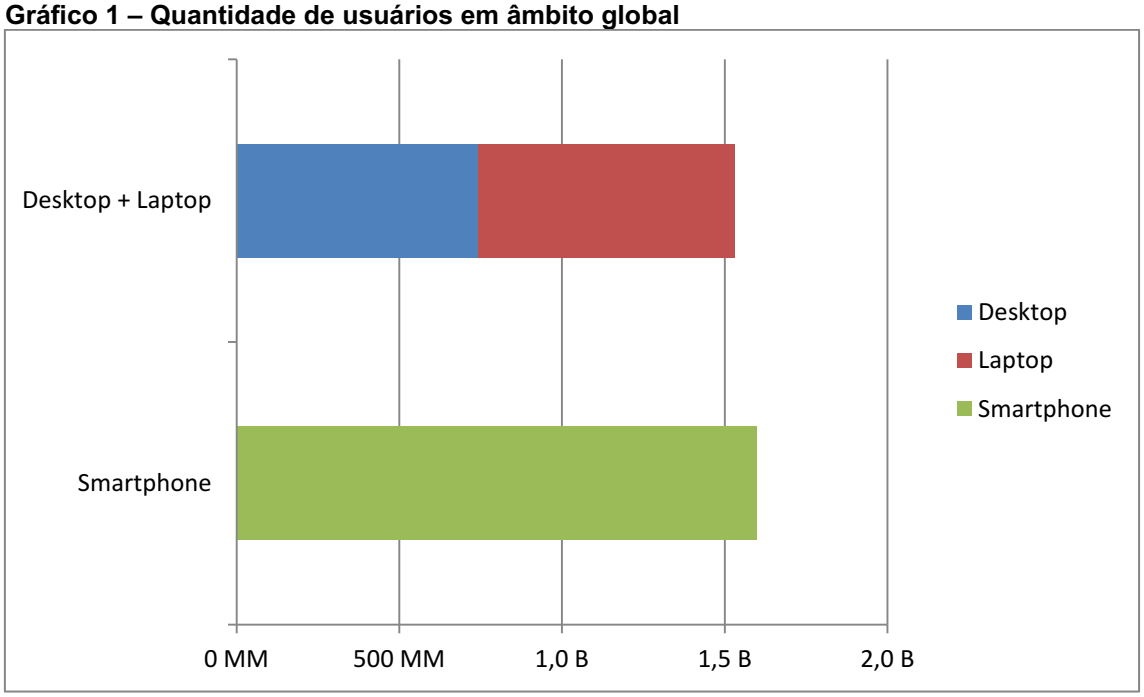
No presente trabalho será utilizada a previsibilidade para definir o comportamento padrão de um recurso computacional específico, de modo a possibilitar uma melhor utilização do mesmo. O componente a ser analisado será a *CPU*, a qual é responsável por realizar os cálculos aritméticos, lógicos e pelo processamento de entrada e saída de dados.

2.3 COMPUTADORES PESSOAIS

O computador pessoal foi o primeiro equipamento com poder computacional de alto desempenho para uso individual a se tornar acessível à comunidade em geral. Atualmente o *Smartphone* é o equipamento com maior adoção na história do mundo moderno, o número de equipamentos móveis supera de forma contundente a quantidade total de televisores ou computadores pessoais em contexto mundial (Meeker, 2014).

Através do gráfico 1 observamos que a quantidade de *PCs* (*Personal Computers*) foi mantida na última década com cerca de um bilhão e meio de equipamentos, considerando tanto os modelos Desktop quanto os Laptops. Meeker (2014) cita estes aparelhos separadamente, o que torna a diferença mais acentuada entre os *Smartphones* (1,6 bilhões), os *PCs Desktop* (743 milhões) e os *PCs Laptop* (789 milhões).

Gráfico 1 – Quantidade de usuários em âmbito global



Fonte: Adaptado de (Meeker, 2014).

De acordo com a IDC (2019) apesar dos problemas envolvendo países com grande poder tecnológico e de produção, como a China e os EUA, o mercado de PCs (inclusive *Laptops*) continua ativo, no terceiro trimestre de 2019 foram registradas 70,4 milhões de unidades embarcadas em âmbito mundial. Neste

período destaca-se o aumento de remessas na América Latina devido ao segmento empresarial, e os setores bancário, varejista e manufatureiro.

A importância de definir os *PCs* como ferramenta de estudo reside principalmente na exploração de seu tempo de processamento ocioso com o propósito de direcionar o seu uso. Um case deste tipo de abordagem se dá com o projeto da IBM denominado *World Community Grid*, que foi iniciado em 2004. Este projeto permite que qualquer usuário de PC, *Smartphone* ou *Tablet* doe o poder computacional ocioso para que seja utilizado em pesquisas que envolvem a saúde, pobreza e sustentabilidade (IBM, 2020).

O projeto da IBM é formando por mais de 650 mil voluntários espalhados em dezenas de países que doaram o poder de processamento ocioso de milhões de computadores físicos individuais, criando um *grid computing* que até 2011 tinha possibilitado mais de 400 mil anos em tempo de execução (IBM, 2011).

O Grupo *Pande* do Departamento de Química da Universidade de Stanford administra o projeto *Folding@Home,* existente desde 2000 que usa a computação distribuída para simular a dinâmica de proteínas utilizando o tempo de processamento ocioso tanto da *CPU* quanto da *GPU* oriundos de computadores pessoais (FAH, 2020).

3 ESTADO DA ARTE

Tendo em vista o levantamento de informações sobre os trabalhos realizados na área de pesquisa em questão, faz-se necessário a execução de um processo denominado revisão sistemática. Este é iniciado através da busca em bases de pesquisa que mantém tais documentos, no decorrer desta atividade pode-se realizar a filtragem de modo a identificar os trabalhos mais relevantes para uso no desenvolvimento da pesquisa.

O presente capítulo está organizado em nove seções principais. A seção 3.1 apresenta o método no qual a revisão sistemática se baseia. A seção 3.2 mostra as questões de pesquisa que serão respondidas. A seção 3.3 traz as bases de busca utilizadas no processo. A seção 3.4 define o termo utilizado nas buscas e apresenta os resultados das buscas realizadas nas respectivas bases de dados. A seção 3.5 utiliza os filtros para refinar os resultados com maior critério. A seção 3.6 busca solucionar as questões de pesquisa levantadas na seção 3.2 através do estudo realizado sobre os trabalhos resultantes da seção 3.5. A seção 3.7 utiliza a revisão realizada para descrever aspectos relevantes identificados durante a revisão. E a seção 3.8 aborda as considerações finais do capítulo.

3.1 MÉTODO DE REVISÃO SISTEMÁTICA

Fundamentada sobre a busca de evidências através da definição, análise e resolução de questões da pesquisa, a metodologia proposta por Kitchenham e Charters (2007) permitirá a definição do escopo e o nível de abrangência no qual a pesquisa se limitará. Deste modo serão obtidos como resultado os trabalhos mais substanciais e assertivos no que se diz respeito ao tema de pesquisa.

Os critérios de inclusão e exclusão definidos antes do processo de filtragem é um aspecto de extrema importância nesta metodologia, pois a sua definição melhora consideravelmente o refinamento da busca, identificando os trabalhos que não possuem tanta relevância apesar de estarem relacionados como tema.

3.2 QUESTÕES DE PESQUISA

O principal objetivo da pesquisa foi desenvolver um levantamento dos métodos baseados em séries temporais mais utilizados na previsão da utilização de recursos computacionais. Portanto, é necessário estabelecer as questões que nortearão as buscas e oferecerão os dados necessários para a apresentação dos resultados obtidos.

Foram definidas as seguintes perguntas:

1. P1: Qual o número de citações de cada trabalho?
2. P2: Quais foram os métodos baseados em séries temporais utilizados?
3. P3: Quais métricas de avaliação foram utilizadas?
4. P4: Que recurso se buscava prever o comportamento?
5. P5: Qual a quantidade anual de publicações sobre o tema?
6. P6: Em qual área foi aplicada a previsão?

3.3 ESCOLHA DAS BASES DE PESQUISA

A obtenção dos dados iniciou-se por meio de cinco bases de dados diferentes, utilizando a mesma frase no período entre 2014 e 2019, porém cada um destes repositórios tem especificidades no processo de busca, o Quadro 3 indica os detalhes das configurações de busca que foram aplicadas.

É válido ressaltar que as configurações de busca variam para cada repositório devido às opções disponíveis em cada um deles, no entanto as configurações realizadas tinham como propósito abranger o maior número de trabalhos publicados nas respectivas fontes.

Quadro 3 – Definição dos Repositórios de Pesquisa

Repositório	Endereço Eletrônico	Configurações de Busca
Google Scholar	https://scholar.google.com.br	“Title”
ACM Digital Library	https://dl.acm.org	“Title” Opção “The ACM Guide for Computer Literature”
Science Direct	https://www.sciencedirect.com	“Abstract, Title and Keywords”
Scopus	https://www.scopus.com	“Article Title, Abstract, Keywords”
IEEE Xplore	https://ieeexplore.ieee.org	“All Metadata”

Fonte: Autoria Própria (2021)

3.4 DEFINIÇÃO E EXECUÇÃO DAS BUSCAS

Os repositórios definidos para a realização das buscas possuem em grande parte publicações em inglês, portanto este idioma será utilizado na formulação do termo de busca. Apropriando-se das palavras-chave que permeiam a pesquisa e utilizando o operador lógico *AND*, foi definida a seguinte *string* de busca: *Time Series AND Predictability AND Computational Resources*.

Os trabalhos obtidos nesta etapa, por meio da *string* de busca previamente definida, foram agrupados pelo seu respectivo repositório conforme apresentado na Tabela 1.

Tabela 1 – Total Inicial de Publicações

Google Scholar	ACM	Science Direct	Scopus	IEEE	Total
1008	5	180	0	45	1238

Fonte: Autoria Própria (2021)

Através da Tabela 1 observa-se que o repositório *Scopus* não retornou publicações realizadas com ao menos uma das palavras utilizadas na *string* de busca, por outro lado a base do *Google Scholar* encontrou um maior número de trabalhos relacionados. Considerando o fato de que o mesmo trabalho possa ter sido identificado em mais de uma busca, assim como pode não possuir um nível de relação mínimo junto ao tema de pesquisa em questão, é fundamental a aplicação de filtros sobre os resultados atuais.

3.5 ESTRATÉGIA DE FILTRAGEM

Com a finalidade de identificar os trabalhos de maior expressividade em meio aos resultados já obtidos, realiza-se o processo de filtragem dividido em três etapas.

Inicialmente foi realizada uma verificação ingênua com o seguinte critério:

1. Exclusão de trabalhos duplicados, encontrados tanto dentro do respectivo repositório quanto em outro.

Durante a primeira etapa foram identificadas 82 publicações duplicadas e 257 citações nos resultados do Google Scholar, resultando num total de 669 publicações. Em seguida uma nova varredura foi realizada a fim de identificar prioritariamente a relação com o tema de pesquisa explorado, contendo os critérios abaixo:

1. Inclusão de trabalhos que apresentassem em seu título, palavras-chave e resumo, relação com os principais métodos de previsão através de séries temporais;
2. Classificação dos trabalhos validados através do critério anterior em três grupos para a próxima análise, os quais possuem muita relação com o tema, razoável relação com o tema e pouca ou nenhuma relação;
3. Exclusão de trabalhos com pouca ou nenhuma relação com o tema.

Tabela 2 – Segunda Etapa de Filtragem

Google Scholar	ACM	Science Direct	Scopus	IEEE	Total
Grupo 1 – Muito relacionado					
10	2	4	0	1	17
Grupo 2 – Razoavelmente relacionado					
20	2	5	0	10	37
Grupo 3 – Pouco ou não relacionado					
639	1	171	0	34	845

Fonte: Autoria Própria (2021)

Como descrito na Tabela 2, após a segunda etapa 17 publicações foram incluídas ao Grupo 1, as quais possuem maior relação com o tema de pesquisa, por meio da análise do título, palavras-chave e resumo. No Grupo 2 foram incluídos 37 trabalhos realizados com um nível de relação aceitável. E no Grupo 3 foram removidas 838 publicações sem relação.

Por fim, foram realizadas leituras prévias dos trabalhos, em busca de um maior detalhe sobre o nível de similaridade ou divergência dos métodos e ferramentas utilizadas, realizando a exclusão dos resultados com pouca ou nenhuma relação com o tema de pesquisa os quais são apresentados na Tabela 3.

Tabela 3 – Terceira Etapa de Filtragem

Google Scholar	ACM	Science Direct	Scopus	IEEE	Total
21	2	4	0	10	37

Fonte: Autoria Própria (2021)

Depois da etapa mais exaustiva de leitura e revisão das publicações foi obtido o total de 9 trabalhos para uso na resolução das questões definidas previamente e durante todo o desenvolvimento da pesquisa, conforme apresentado na Tabela 4.

Tabela 4 – Total Final de Publicações

Google Scholar	ACM	Science Direct	Scopus	IEEE	Total
4	1	3	0	1	9

Fonte: Autoria Própria (2021)

A seleção final dos trabalhos após a filtragem está presente no Quadro 4, onde para cada trabalho foi definido um código identificador (S_n, sendo que S faz menção ao documento e *n* refere-se ao número sequencial do mesmo), que será utilizado ao longo desta pesquisa como referência para os respectivos documentos.

Quadro 4 – Trabalhos Finais Selecionados **(continua)**

Id.	Autor	Título	Ano de Publicação
S1	Zhang, F.; De Grande, R.; Boukerche, A.	*Accuracy Analysis of Short-term Traffic Flow Prediction Models for Vehicular Clouds*	2016
S2	Benifa, J. V. B.; Dharma, D.	*HAS: Hybrid auto-scaler for resource scaling in cloud environment*	2018
S3	Du, P.; Wang, J.; Yang, W.; Niu, T.	*Multi-step ahead forecasting in electrical power system using a hybrid forecasting system*	2018
S4	Marinho, C. S. S.; Moreira, L. O.; Coutinho, E. F.; Costa Filho, J. S.; Sousa, F. R. C.; Machado, J. C.;	*LABAREDA: A Predictive and Elastic Load Balancing Service for Cloud-Replicated Databases*	2018

Quadro 4 – Trabalhos Finais Selecionados **(conclusão)**

S5	Voyant, C.; Motte, F.; Fouilloy, A.; Notton, G.; Paoli, C.; Nivet, M.	*Forecasting method for global radiation time series without training phase: Comparison with other well-known prediction methodologies*	2017
S6	Duc, T. L.; Leiva, R. G.; Casari, P.; Östberg, P.	*Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey*	2019
S7	Farias, R.	*Time series forecasting based on classification of dynamic patterns*	2015
S8	Yoo, W.; Sim, A.	*Time-Series Forecast Modeling on High-Bandwidth Network Measurements*	2016
S9	Hua, Y.; Zhao, Z.; Liu, Z.; Chen, X.; Li, R.; Zhang, H.	*Traffic Prediction Based on Random Connectivity in Deep Learning with Long Short-Term Memory*	2018

Fonte: Autoria Própria (2021)

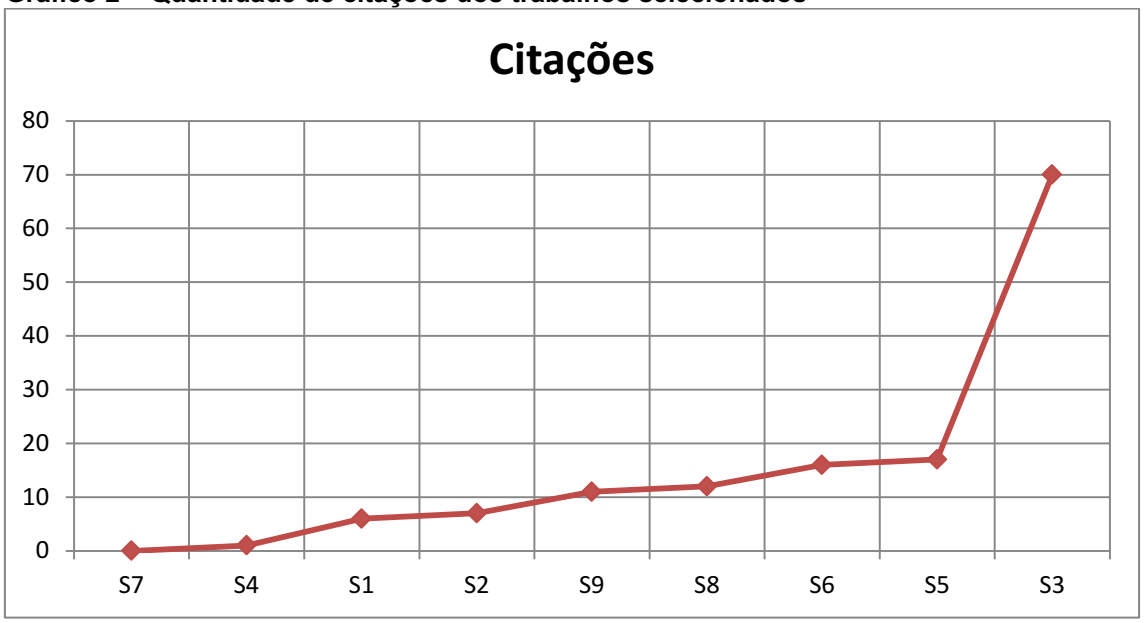
3.6 RESOLUÇÃO DAS QUESTÕES

Nesta seção serão apresentadas as respostas obtidas através do mapeamento sistemático, para as questões estabelecidas no início do processo. Para tal, foram utilizadas fichas de leitura onde foram registrados os pontos chaves identificados durante o estudo de cada trabalho selecionado a fim de documentar o desenvolvimento do presente documento.

Pergunta 1: Qual o número de citações de cada trabalho?

O Gráfico 2 indica que o trabalho S3 possui o maior número de citações, em seguida 4 trabalhos com citação acima de 10. Dos 9 trabalhos apenas 1 não possuem citações até o presente momento, isso representa cerca de 11% dos artigos estudados, sendo que apenas um foi publicado em 2019.

Gráfico 2 – Quantidade de citações dos trabalhos selecionados



Fonte: Autoria Própria (2021)

Pergunta 2: Quais foram os métodos baseados em séries temporais utilizados?

Tabela 5 – Métodos baseados em Séries Temporais

Método de previsão	Total de Trabalhos
ARIMA	2
AR	2
SARIMA	2
ARMA	1
EMA	1
DES	1

Fonte: Autoria Própria (2021)

Apesar da diversidade de métodos identificados, os quatro primeiros métodos são variações do método Autorregressivo e fundamentados na regressão linear, concentrando cerca de 77% do resultado apresentado na Tabela 5. Segundo Marinho *et al.* (2018) e Hua *et al.* (2018), o modelo *ARIMA* é um dos modelos mais populares dentro da classe de métodos de regressão linear e um dos mais utilizados na comunidade acadêmica, apresentando no trabalho de Du *et al.* (2018) um bom desempenho em modelos lineares. Este modelo procura tornar uma série não estacionária em estacionária através dos parâmetros de defasagem e diferença.

Este modelo normalmente é usado off-line, ou seja, sem a necessidade de atender um requisito específico de tempo.

Já o modelo *SARIMA* obteve resultados expressivos assim como outro método baseado em redes neurais que foi utilizado em um experimento, tendo como problema de pesquisa a previsão do fluxo de tráfego em três segmentos de rodovia na Inglaterra (ZANG; DE GRANDE; BOUKERCHE, 2016).

É válido citar também que outros métodos foram utilizados nos trabalhos relacionados no Quadro 4, os quais não são baseados em modelos estatísticos de regressão. Dentre os métodos que não são fundamentados no processo de regressão linear, os mais recorrentes foram os métodos híbridos (S2, S3, S7), métodos baseados em aprendizado supervisionado (S1) e os baseados em redes neurais (S8, S9).

Pergunta 3: Quais métricas de avaliação foram utilizadas?

Tabela 6 – Métricas utilizadas para validação

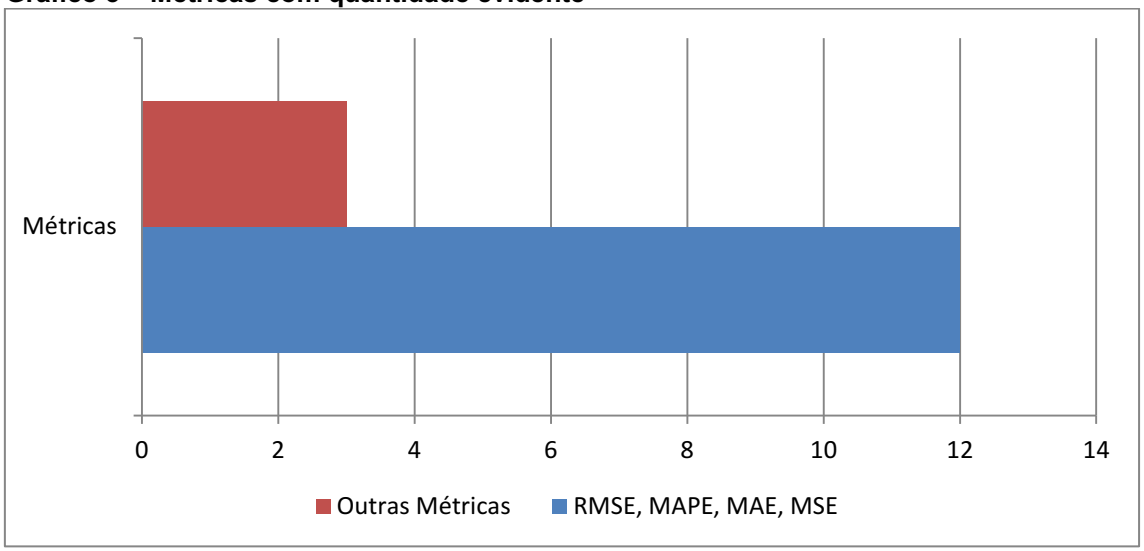
Métrica de validação	Total de Trabalhos
RMSE	5
MAPE	3
MAE	2
MSE	2
MAD	1
R^2	1
AE	1

Fonte: Autoria Própria (2021)

Por intermédio da Tabela 6 nota-se certa variedade de métricas utilizadas nos trabalhos selecionados. As mais utilizadas foram *RMSE* (S2, S4, S5, S6, S7), *MAPE* (S1, S3, S4), *MAE* (S3, S8) e *MSE* (S1, S9).

O Gráfico 3 evidencia o uso das quatro métricas citadas anteriormente, a qual representa quatro vezes mais em relação às outras métricas utilizadas nos trabalhos encontrados.

Gráfico 3 – Métricas com quantidade evidente



Fonte: Autoria Própria (2021)

Pergunta 4: Que recurso se buscava prever o comportamento?

Tabela 7 – Recursos estudados para previsão

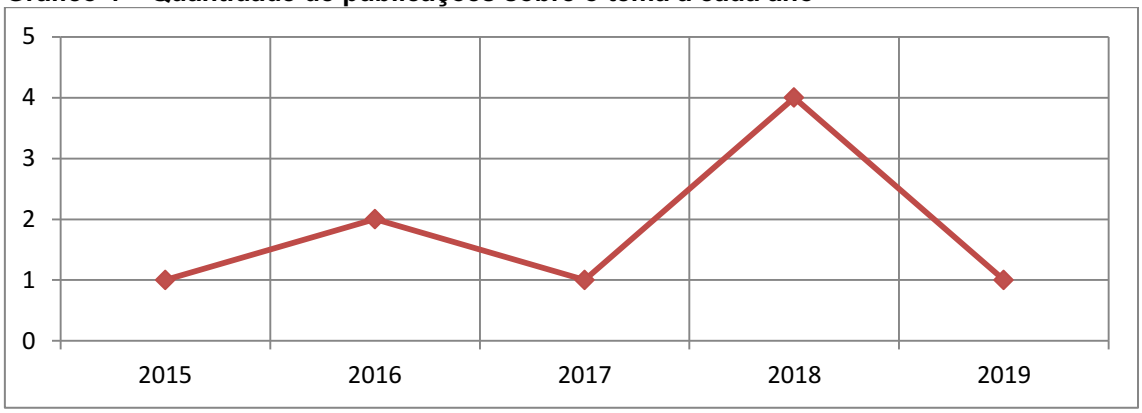
Recurso	Total de Trabalhos
Alocação de recursos na nuvem	3
Tráfego de dados	1
Tráfego de dados	1
Tráfego em Rodovias	1
Sistema de Energia	1
Demanda de água	1
Radiação Global	1

Fonte: Autoria Própria (2021)

De acordo com a Tabela 7 constata-se um esforço considerável despendido através de pesquisas relacionadas a recursos da área de Redes de Computadores, especificamente sobre recursos alocados na nuvem (*Cloud Computing*). A justificativa para que 44% dos recursos analisados estejam concentrados na grande área de Redes de Computadores se dá devido a crescente demanda de serviços remotos, tanto no âmbito profissional junto às empresas que contratam esse tipo de ativo quanto em relação ao consumidor final através de dispositivos móveis e inteligentes, como *Smartphones*, *SmartTV*, entre outros.

Pergunta 5: Qual a quantidade de publicações realizadas sobre o tema a cada ano?

Gráfico 4 – Quantidade de publicações sobre o tema a cada ano



Fonte: Autoria Própria (2021)

Existe uma quantidade crescente de pesquisas sendo desenvolvidas a respeito do tema, com um pico em 2018, superando a média apresentada no Gráfico 4 indicando uma tendência de crescimento no número de materiais publicados, mostrando que o problema tratado é relevante e atual.

Pergunta 6: Em qual área foi aplicada a previsão?

Tabela 8 – Áreas de aplicação da previsão

Área	Total de Trabalhos
Cloud Computing	3
Redes de Computadores	1
Telecomunicação	1
Clima	1
Biologia	1
Sistemas Embarcados	1
Engenharia Elétrica	1

Fonte: Autoria Própria (2021)

Novamente, constata-se a importância da área de Redes de Computadores, inclusive a subárea de *Cloud Computing*, porém vale ressaltar a diversidade de pesquisas que aderem ao uso de séries temporais e os métodos de previsão como ferramentas. Na Tabela 8 foi feita esta separação, pois o artigo S9 é bem específico

ao tratar de uma rede de computadores, porém sem especificar se a mesma é local ou remota.

3.7 TRABALHOS RELACIONADOS

A partir da revisão apresentada e após um estudo ainda mais aprofundado dos trabalhos selecionados, serão analisados os aspectos relacionados à efetividade no uso do método *ARIMA*, a periodicidade dos dados utilizados e as ressalvas apontadas pelos autores.

Segundo Zhang (2016), dentre os métodos estatísticos de previsão o modelo *ARIMA* é aplicável à maioria das séries temporais da vida real, produzindo resultados bem precisos. Com a ressalva sobre o processo de definição dos parâmetros para o modelo *ARIMA*, o qual é complexo e requer uma grande quantidade de dados históricos. O autor utiliza uma variação do modelo *ARIMA* Sazonal ou *SARIMA*, o qual apresentou o melhor resultado em cenários de curto prazo (30 dias). Apesar de não ser o objetivo da pesquisa, os algoritmos foram avaliados em um cenário de longo prazo (200 dias), onde o modelo *SARIMA* obteve o pior resultado embora havendo pouca diferença para os demais.

Outro trabalho que traz resultados positivos na utilização do *SARIMA* é o de Farias (2015), que considera o modelo flexível, se adaptando a maioria dos cenários, de forma simples e eficaz. O *SARIMA* se encaixou bem em sua proposta de previsão da demanda de água em janela de curto prazo (diária).

Benifa (2018) utiliza em seus experimentos outras variações do modelo *ARIMA*, obtendo resultados efetivos através do componente *AR* em comparação com o *ARMA* e com outros modelos baseados em Aprendizagem de Máquina (*Machine Learning*).

Marinho et al. (2018) apresenta bons resultados do modelo *ARIMA* em janelas de 3 a 8 minutos, afirmando que janelas curtas ajudam a ajustar um ambiente altamente dinâmico. Este contexto evidenciou a necessidade de atualizar constantemente os parâmetros do modelo, o que elevou o custo computacional da solução.

Hua *et al.* (2018) expõe uma limitação do *ARIMA* em sua pesquisa, onde o nível de variação presente nas séries temporais é muito alto, sendo assim o modelo se concentrou nos valores médios obtidos no processo de regressão linear. O autor

não menciona a existência de algum processo responsável pela atualização dos parâmetros do modelo a fim de superar o problema da dinamicidade, como mencionado por Marinho *et al.* (2018). Por fim, é proposta uma abordagem mais eficiente em alguns aspectos, porém muito custosa.

No trabalho de previsão de energia elétrica, Du *et al.* (2018) utiliza apenas métodos baseados em Inteligência Artificial e descarta a utilização de métodos estatísticos como o modelo *ARIMA*, justificando que tais algoritmos não são indicados em cenários de janelas de curto prazo (intervalo de minutos ou horas) e manipulando séries temporais não lineares.

Assim como Zhang (2016), Voyant *et al.* (2017) reafirma o desafio de definir os parâmetros do modelo Autoregressivo, e sugere a sua utilização com dados históricos entre 200 e 600 dias. O autor utiliza o *Kalman Filter* para realizar previsões de janelas curtas, com baixo custo computacional e alta assertividade. O mesmo não menciona aspectos a atualização dos parâmetros do modelo AR, cita apenas a atualização do *Kalman Filter*.

Duc *et al.* (2019) diz que dentre as abordagens avaliadas, incluindo tanto métodos estatísticos quanto de Aprendizagem de Máquina, não há um que se destaque de maneira evidente. Apesar da precisão significativa obtida pelas redes neurais em muitos casos, existem três requisitos críticos nesta abordagem que são: uma quantidade extremamente grande de dados para treinamento, tempo de computação extenso e uso de hardware altamente especializado.

A proposta apresentada por Yoo (2016) mostra consistência em lidar com mudanças repentinas, e o mesmo segue a abordagem de atualização regular realizada por Marinho *et al.* (2018). Como os dados utilizados neste trabalho possuem o componente sazonal, o autor utiliza um modelo composto por *STL* (*Seasonal and Trend decomposition using Loess*) e *ARIMA* que mostrou ser mais assertivo do que o *SARIMA*.

É importante enfatizar que a grande diversidade dos modelos Autorregressivos apresentados é natural devido às variações existentes na estrutura das séries temporais, as quais são analisadas durante a fase de treinamento com o propósito de definir os parâmetros do modelo *ARIMA*. Devido a sua capacidade de adaptação, quando um dos componentes tem a ordem ou parâmetro zerado o modelo é modificado. Por exemplo, “AR(1)I(0)MA(0)” mostra que o modelo *ARIMA*

variou para o modelo AR devido aos parâmetros definidos, fazendo com que fossem ‘desativados’ os componentes de Integração e Média Móvel.

Através do trabalho elaborado por Voyant *et al.* (2017) observamos que o modelo *ARIMA* é mais indicado para ser usado como método de previsão off-line, não exigindo um retorno rápido e síncrono, diferente do objetivo do seu trabalho. Porém o trabalho de Marinho *et al.* (2018), mostra que é possível utilizar o *ARIMA* em contextos de previsão manipulando séries temporais curtas, dependendo de aspectos como o tempo de processamento da previsão.

Quadro 5 – Detalhamento dos experimentos realizados nas pesquisas

Autor	ARIMA foi Efetivo	Tamanho das Séries	Ressalvas	Atualização do ARIMA
Zhang (2016)	Sim	30 dias ou 4 semanas	Parametrização Complexa.	-
Benifa (2018)	Sim	39 dias ou 5 semanas	-	-
Du *et al.* (2018)	-	15 dias	Não indicado para uso em séries não lineares.	-
Marinho *et al.* (2018)	Sim	< 10 minutos	Atualização frequente, alto custo computacional.	Sim
Voyant *et al.* (2017)	Não	< 10 horas ou por minuto	Parametrização Complexa.	Não
Farias (2015)	Sim	Diária	-	Não
Yoo (2016)	Sim	60 dias ou 8 semanas	-	Sim
Hua *et al.* (2018)	Não	76 dias ou 10 semanas	Baixa capacidade de adaptação em contexto com alta variação	Não

Fonte: Autoria Própria (2021)

Outro aspecto importante é a evidência da utilização de um componente responsável pela atualização dos parâmetros do modelo *ARIMA*, principalmente em cenários onde ocorre uma grande variação nos dados que compõem as séries temporais, exigindo que seja explorada a capacidade de adaptação do modelo em questão. De acordo com o Quadro 5, os trabalhos onde o modelo *ARIMA* não foi efetivo, não exploraram a possibilidade de atualizar o modelo de acordo com a necessidade ou após a variação da série temporal em questão.

Já que o processo de atualização dos parâmetros do modelo gera um custo computacional relativamente alto e constante, deve ser considerada a possibilidade de estabelecer um nível crítico da assertividade da previsão, de modo a reduzir a quantidade de atualização dos parâmetros e o custo computacional. Ou utilizar um método híbrido, onde parte seja responsável pela previsão online e a outra parte responsável pela previsão off-line.

Apesar da grande contribuição de cada um dos trabalhos apresentados até o momento, não houve menção à análise do comportamento dos computadores pessoais e sua subutilização e ociosidade dentro de alguns cenários.

3.8 CONSIDERAÇÕES

Através do mapeamento sistêmico foi possível examinar diversos pontos referentes ao tema de previsão de recursos através de séries temporais. Inicialmente uma grande quantidade de publicações foi encontrada, exigindo grande esforço durante o processo de filtragem que reduziu consideravelmente o número de trabalhos selecionados para um estudo mais detalhado. As informações resultantes ao fim desta revisão proporcionaram o auxilio necessário para estabelecer o fundamento no qual o presente trabalho foi desenvolvido.

4 PROPOSTA DE ABORDAGEM DE PREVISÃO

Como proposta de projeto de pesquisa desenvolvido utilizou o modelo *ARIMA* na criação do componente de geração da Assinatura Comportamental, visando representar o comportamental padrão observado e utilizado para previsão. Para isso foi necessário compreender as séries temporais utilizando ferramentas estatísticas de análise como *ACF*, *PACF* e *Augumented Dickey-Fuller* para definir os parâmetros do modelo.

Em busca de uma assertividade mais efetiva, foi desenvolvido um módulo responsável pela atualização da Assinatura Comportamental, o qual ajustará a Assinatura de acordo com a observação mais recente.

As métricas de avaliação utilizadas na verificação da eficiência da previsão foram *MAE* e *RMSE*, pois utilizam a mesma unidade do valor que se busca prever. Posteriormente foi feita uma análise comparativa com os resultados obtidos para verificar o nível de assertividade do modelo criado.

A seguir serão detalhados os processos de geração, atualização e previsão.

4.1 PARAMETRIZAÇÃO E MODELAGEM

Para estabelecer o valor do parâmetro p associado ao número de máximo de defasagens aplicadas ao modelo Autorregressivo, foi utilizada a função de autocorrelação parcial a fim de identificar o número de atrasos significativos a partir das primeiras correlações até o momento em que o seguinte não tenha um nível de significância considerável.

Para definir o parâmetro d do modelo *ARIMA*, relacionado ao grau de diferenciação necessária para tornar a série estacionária, foi aplicado o teste *ADF* para avaliar a estacionariedade da série diferenciando-a até o ponto em que os critérios que definem tal característica fossem atendidos.

Por fim, a definição do parâmetro q está associada ao modelo de Médias Móveis. Para estabelecer o seu valor foi utilizada a função de autocorrelação, identificando o número de atrasos significativos, sendo que este valor foi responsável por indicar a quantidade de defasagens relacionadas ao erro de previsão do modelo.

O modelo utilizado no processo foi configurado com os parâmetros definidos anteriormente. Em seguida, um segundo modelo foi criado de forma automática, através de uma biblioteca denominada "*pyramid-arima*", a fim de se realizar a validação do modelo inicial através do Critério de Informação de Akaike (*AIC*). O *AIC* baseia-se na existência de um modelo "real" que é desconhecido, porém permite descrever os dados. O seu cálculo é descrito como $AIC = -2L + 2K$, onde L é o log de verossimilhança máxima e K é o número de parâmetros do modelo. Tal método consiste em estimar o erro de previsão do respectivo modelo, buscando estabelecer o seu nível de qualidade, aquele que apresentar o menor valor será considerado o que possui melhor ajuste.

4.2 GERAÇÃO DA ASSINATURA COMPORTAMENTAL

Tendo em vista a automatização do processo de criação, previsão e atualização das Assinaturas Comportamentais necessárias no desenvolvimento deste trabalho, foi realizado o desenvolvimento de um sistema responsável pela sistematização destas operações. O sistema foi construído na linguagem Python, a qual oferece diversas bibliotecas que permitem a análise e manipulação de séries temporais e possui uma sintaxe de fácil compreensão.

Inicialmente será apresentada a estrutura responsável pela geração da Assinatura Comportamental, tendo como pré-requisito os arquivos gerados pelo sistema de obtenção do processamento citado no tópico 1.2, o qual define a estrutura do arquivo de dados como apresentado na Figura 2, contendo a data, o horário e o percentual de uso da *CPU*. Estes dados são obtidos a cada segundo, gravados em um arquivo texto e enviados para um servidor ao atingirem o tamanho de 4 Megabytes, devido à limitação no tamanho do arquivo para o tráfego na rede.

A partir de um conjunto de dados pertencentes ao mesmo computador pessoal e formatados de acordo com a estrutura apresentada anteriormente, o processo de geração da Assinatura Comportamental pode ser iniciado, conforme representado na Figura 3.

Figura 2 – Estrutura do arquivo de dados

```
18/09/2019 08:42:41 | 0,00
18/09/2019 08:42:42 | 8,67
18/09/2019 08:42:43 | 1,68
18/09/2019 08:42:44 | 1,19
18/09/2019 08:42:45 | 14,40
18/09/2019 08:42:46 | 4,55
18/09/2019 08:42:47 | 8,64
18/09/2019 08:42:48 | 5,07
18/09/2019 08:42:49 | 7,06
18/09/2019 08:42:50 | 8,99
18/09/2019 08:42:51 | 32,61
18/09/2019 08:42:52 | 0,00
18/09/2019 08:42:53 | 7,23
```

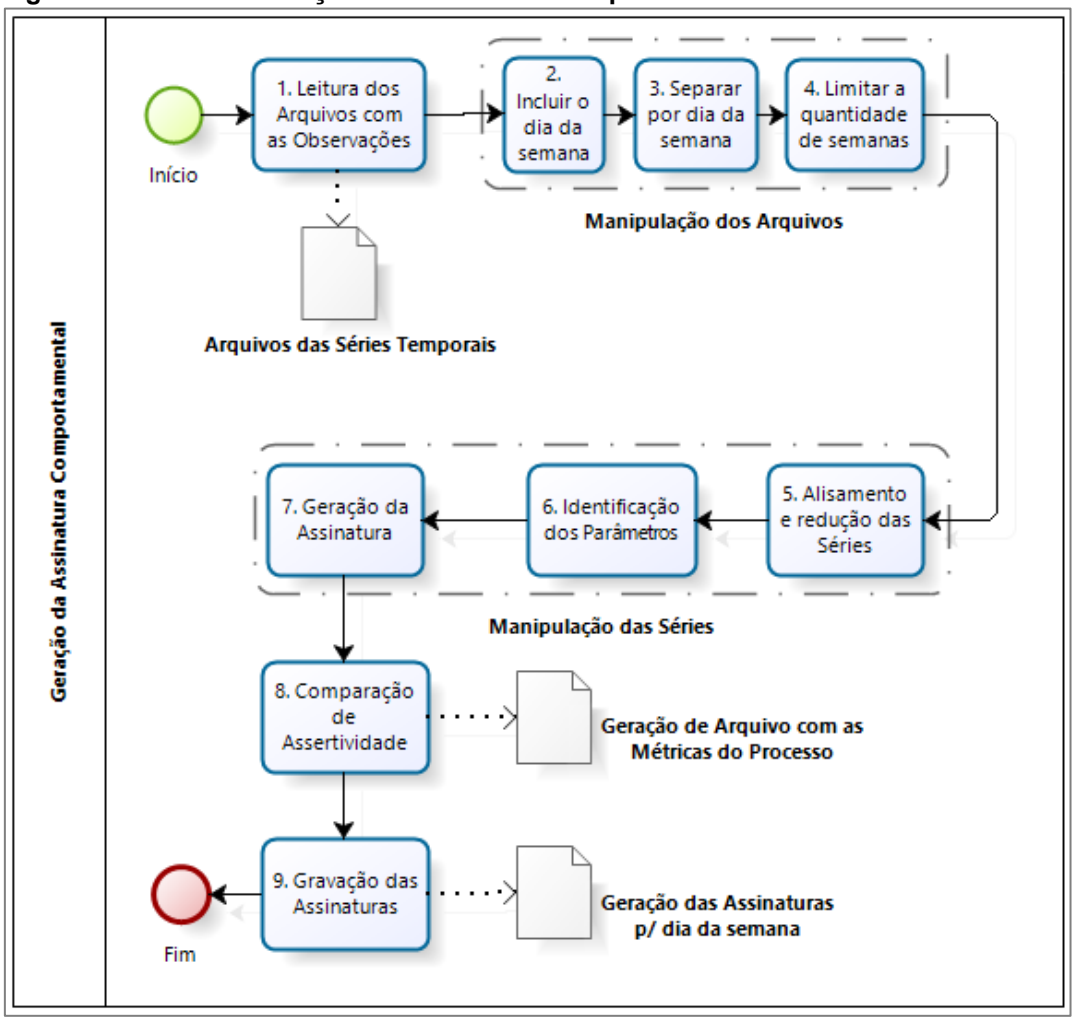
Fonte: Autoria Própria (2021)

O processo de geração da Assinatura é iniciado através da leitura dos n arquivos obtidos a partir da observação de um *PC*, o qual possui apenas as informações de data, hora e índice de uso da *CPU*. Nos passos 2 e 3 ocorre o processo de manipulação e organização dos dados recebidos, onde é realizada a inclusão da referência do dia da semana para cada data registrada permitindo a divisão das observações por dia da semana, visto que o comportamento de uso do recurso pode variar dependendo do contexto no qual está inserido.

O passo 4 limita a quantidade de semanas de cada conjunto de dados, visto que existe uma relação a ser tratada entre o nível de assertividade e a quantidade de dados analisados, aumentando ou diminuindo o desempenho do processo de geração. O passo 5 inicia o processo de manipulação e análise das séries, onde se faz o alisamento por meio de médias móveis com uma janela de tamanho 100, segundo Senger e Gois (2018) este valor permite a realização de uma análise com alta precisão apresentando as variações na Assinatura de forma mais confiável. Ao alterar o tamanho da janela, tendo uma redução de 86.400 para 864 observações, será reduzida de maneira expressiva a quantidade de dados a serem manipulados assim como o tempo de processamento.

No passo 6 são mensurados os parâmetros p, d, q do modelo *ARIMA*, fazendo uso respectivamente da função de autocorrelação parcial, do teste *ADF* e da função de autocorrelação. Assinatura é gerada no passo 7 por meio do *ARIMA*, e o nível de assertividade é confrontado com o modelo automatizado para validar o modelo recém-criado.

Figura 3 – Fluxo de Geração da Assinatura Comportamental



Fonte: Autoria Própria (2021)

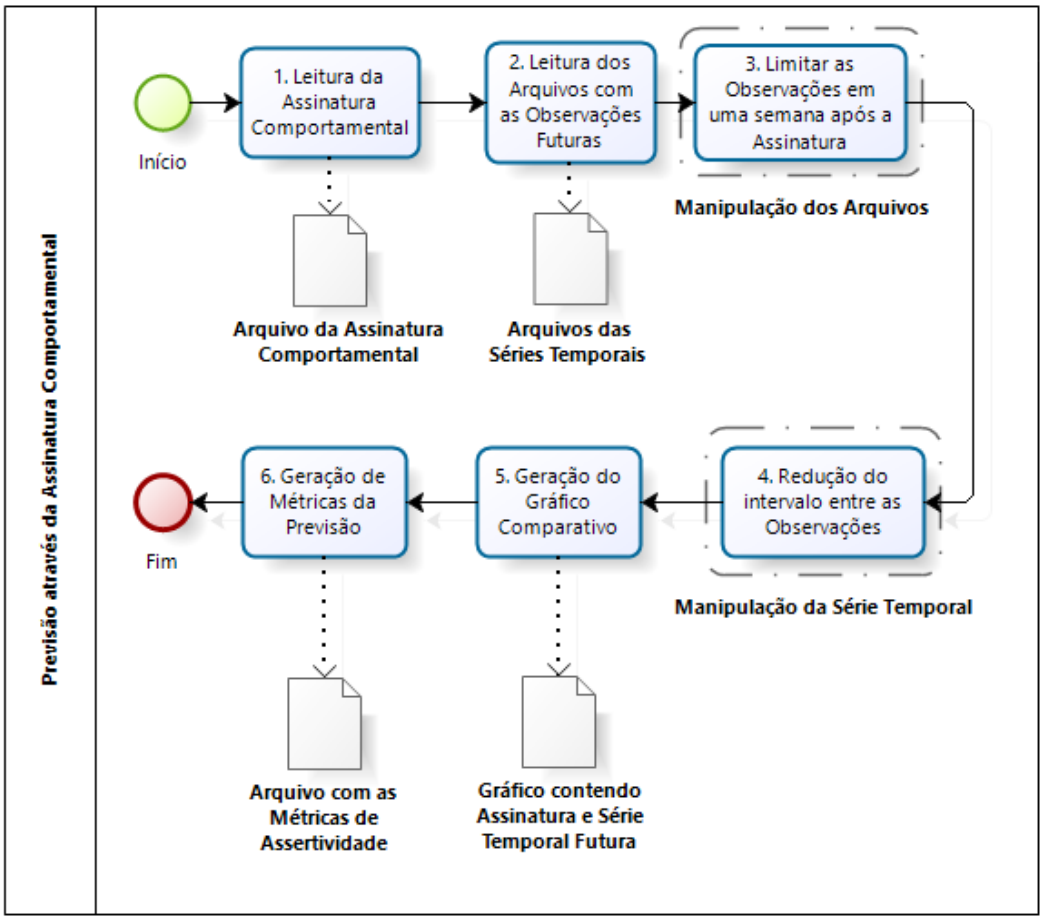
Por fim a Assinatura é gravada em um arquivo *pickle*, o qual serializa os dados reduzindo seu tamanho final e tornando o processo de gravação e posterior leitura mais rápidos.

4.3 PREVISÃO ATRAVÉS DA ASSINATURA COMPORTAMENTAL

A operação subsequente se baseia no processo de previsão, descrito na Figura 4, o qual consiste no cerne do presente trabalho de pesquisa. Possuindo uma Assinatura Comportamental relacionada ao nível de processamento de determinado computador pessoal, é possível realizar sua previsão futura, visto que a previsão foi restringida para a semana seguinte. Supondo que a Assinatura portada seja relacionada ao dia 07/01/2021 (terça-feira), o qual foi o dia mais recente utilizado na construção da Assinatura, espera-se que tal informação seja comparada com os

dados obtidos no dia 14/01/2021 sendo a informação imediata na qual a Assinatura Comportamental foi concebida com o propósito de prever sem conhecê-la de antemão.

Figura 4 – Fluxo de Previsão através da Assinatura Comportamental



Fonte: Autoria Própria (2021)

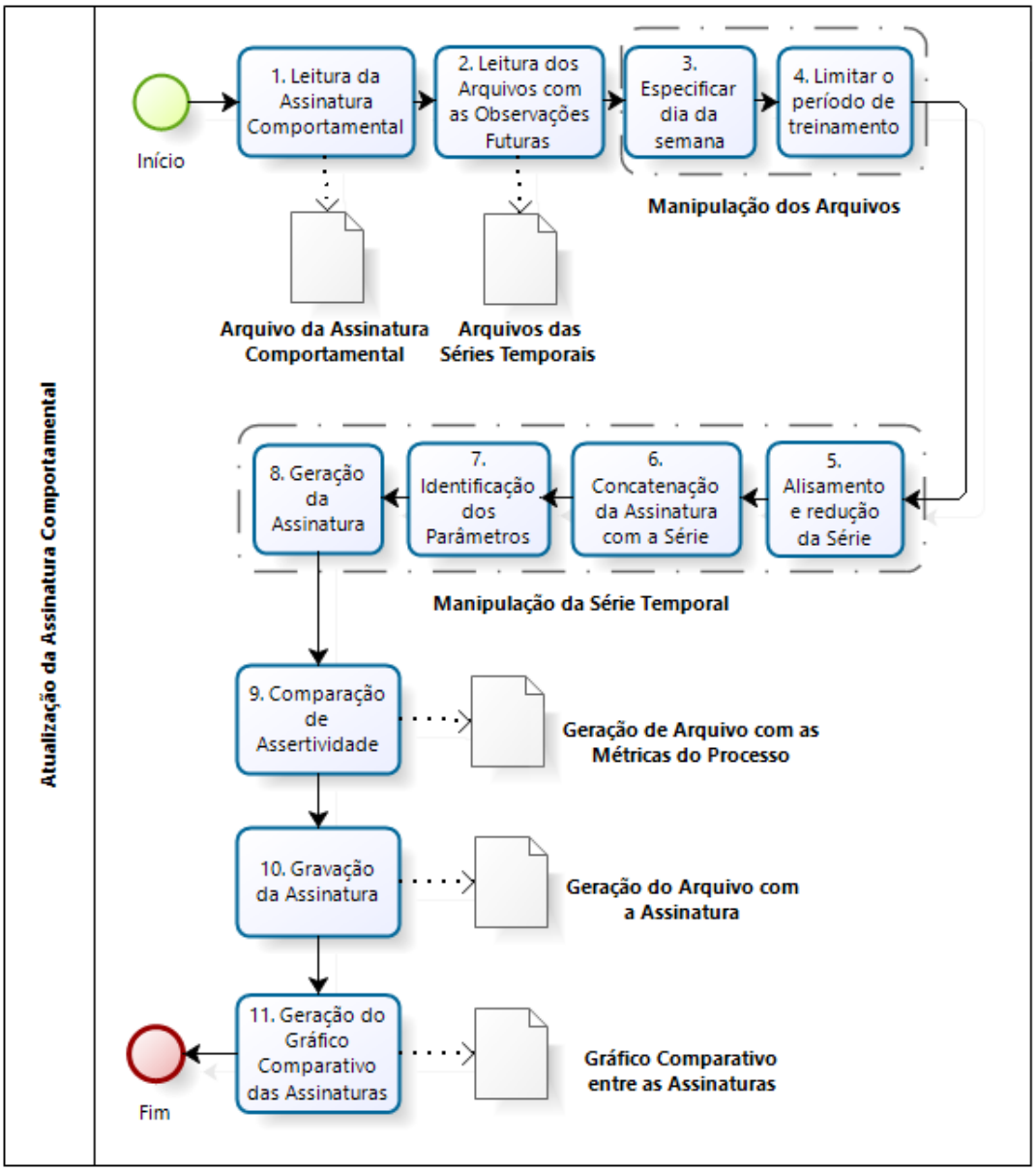
O processo de previsão é iniciado com a leitura do arquivo de Assinatura no formato *pickle* (passo 1), e das séries temporais gerados posteriormente (passo 2), onde tais séries são limitadas somente nos dados referentes à semana posterior à Assinatura (passo 3). No passo 4 a série temporal lida é reduzida, devido à alteração do intervalo de segundos para minutos.

O passo 5 consiste na geração de um gráfico comparativo entre a Assinatura e os valores reais da série temporal, permitindo a visualização entre a previsão e os dados verdadeiros. E por fim são aplicadas as métricas *MAE* e *RMSE*, a fim de mensurar o nível de assertividade da previsão.

4.4 ATUALIZAÇÃO DA ASSINATURA COMPORTAMENTAL

Com o auxílio da Figura 5, será ilustrado o processo de atualização da Assinatura Comportamental, o qual é extremamente necessário para manter um nível satisfatório de confiança em relação ao uso da mesma, visto que a entidade em análise precisa ser constantemente monitorada e acompanhada, fazendo com que a Assinatura siga as variações concernentes ao comportamento da entidade observada, tendo como foco, o nível de uso da *CPU*.

Figura 5 – Fluxo de Atualização da Assinatura Comportamental



Fonte: Autoria Própria (2021)

De maneira similar ao processo de geração da Assinatura, a atualização inicia com a leitura tanto da Assinatura gerada previamente (passo 1), quanto das séries que possuem as observações mais recentes do mesmo computador (passo 2). Os passos 3 e 4 são responsáveis por filtrar as séries temporais de acordo com o dia da respectiva Assinatura e restringir o limite de semanas para treinamento, no passo 5 as séries são submetidas ao processo de alisamento e redução definindo um valor médio dentro do intervalo de 60 observações e diminuindo o tempo de execução do processo.

No passo 6 ocorre a concatenação da Assinatura com os dados recentes, formando uma nova série temporal, a qual é analisada no passo 7 com o objetivo de definir os parâmetros necessários na construção de um novo modelo a partir do *ARIMA*. Por fim a nova Assinatura é gerada, compara-se o grau de assertividade com o modelo automático da biblioteca "*pyramid-arima*" (passo 9), a nova Assinatura é gravada em um arquivo no formato *pickle* (passo 10), e um gráfico de sobreposição apresenta as Assinaturas antes e depois da atualização (passo 11).

4.5 CONSIDERAÇÕES

Este capítulo apresentou uma abordagem para previsão de um recurso utilizando Assinatura Comportamental, assim como a geração e atualização da mesma.

Dentre os pontos relevantes contidos na proposta têm-se: a comparação do processo de geração com uma biblioteca automatizada para criação de modelos *ARIMA*; avaliação da assertividade tanto do modelo *ARIMA* através do parâmetro *AIC*, quanto do resultado da previsão com os valores reais através das métricas mencionadas na seção 1.2 e o tamanho reduzido da Assinatura Comportamental.

A abordagem proposta possui os fluxos de Geração da Assinatura Comportamental, Previsão a partir da Assinatura gerada e Atualização da Assinatura, os quais compõem uma aplicação que centraliza tais processos e a avaliação deles é apresentada no próximo capítulo.

5 APLICAÇÃO DA ABORDAGEM PROPOSTA E RESULTADOS OBTIDOS

Durante a análise do mapeamento sistemático dos trabalhos relacionados, observou-se que um dos testes de avaliação do modelo *ARIMA* resultante após a parametrização é realizado pelo *AIC*, apresentado na seção 4.1. Este método procura estabelecer o nível de qualidade do modelo resultante, onde aquele que apresentar o menor valor será considerado o que possui melhor ajuste.

Os outros meios de avaliação são as métricas que comparam o valor real do recurso, não incluído no conjunto de dados conhecidos pelo modelo que realizará a previsão, com o valor sugerido, resultante do processo de previsão a partir das inferências realizadas sobre os valores anteriores. Dentro desse universo de métricas, foram utilizadas as métricas citadas na seção 1.2, *MAE* e *RMSE*, detalhadas a seguir.

A avaliação da abordagem proposta foi realizada conforme as duas formas citadas anteriormente. A seção 5.1 descreve o processo de criação das ferramentas que compõem a proposta, telas principais de interação com o usuário e define as tecnologias usadas na sua construção. A seção 5.2 relata os resultados da avaliação e a eficácia da abordagem; bem como as limitações identificadas. E a seção 5.3 descreve as considerações finais deste capítulo.

5.1 CONSTRUÇÃO DA ABORDAGEM PROPOSTA

O passo inicial para a construção da abordagem se deu com a criação do Módulo de Observação responsável pela obtenção de dados do computador no qual o programa tenha sido executado. Esta ferramenta captura a cada segundo o percentual de utilização do processador, grava em um arquivo de texto, de acordo com a estrutura apresentada na Figura 2, seção 4.2.

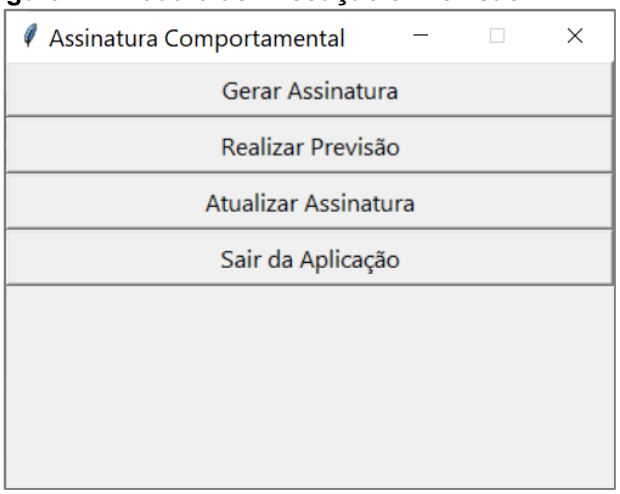
Ao atingir o tamanho de 4 MB o arquivo é enviado a um servidor que centraliza estes dados para uso posterior, os arquivos são rotulados para facilitar a identificação da origem dos mesmos. Após a primeira execução o programa é iniciado automaticamente quando liga o computador, ele opera em segundo plano e pode ser visualizado nos ícones da bandeja do sistema, como mostra a Figura 6.

Figura 6 – Módulo de Observação

Fonte: Autoria Própria (2021)

O Módulo de Execução e Previsão que manipula os arquivos de texto possui os processos de geração da Assinatura, realização de previsão a partir de uma Assinatura e atualização da Assinatura, os quais são descritos na seção 4 e exibidos na Figura 7. Um dos maiores desafios da construção deste módulo foi a parametrização correta do modelo *ARIMA*, principalmente em relação à interpretação conceitual e sua codificação lógica.

Figura 7 – Módulo de Execução e Previsão



Fonte: Autoria Própria (2021)

O Quadro 6 apresenta de forma sintetizada os parâmetros de entrada, que são requeridos para a realização de cada processo, assim como os dados de saída resultantes que viabilizaram a análise e as conclusões relativas à assertividade do módulo criado.

Quadro 6 – Síntese dos Processos do Módulo de Execução e Previsão

Processo	Entrada	Saída
Gerar Assinatura	1 - Arquivos de texto contendo as séries temporais; 2 - A quantidade de semanas para o treino	1 – Um arquivo no formato pickle contendo a Assinatura Comportamental 2 – Um arquivo de texto contendo a comparação entre o modelo gerado pelo módulo e o modelo criado pela biblioteca de geração automática de modelos ARIMA
Realizar Previsão	1 - Arquivo no formato pickle contendo a Assinatura Comportamental; 2 - Arquivos de texto contendo as séries temporais posteriores à última data que compõe a Assinatura Comportamental	1 – Um gráfico comparativo entre a Assinatura Comportamental e a Série Temporal que buscava prever; 2 – Um arquivo de texto contendo as métricas de erro da previsão;
Atualizar Assinatura	1 - Arquivo no formato pickle contendo a Assinatura Comportamental; 2 - Arquivos de texto contendo as séries temporais posteriores à última data que compõe a Assinatura Comportamental; 3 – A quantidade de semanas para serem consideradas na atualização	1 – Um gráfico comparativo entre a Assinatura antiga e a Assinatura atualizada; 2 – Um arquivo no formato pickle contendo a Assinatura Comportamental atualizada; 3 – Um arquivo de texto contendo a comparação entre o modelo gerado pelo módulo e o modelo construído através da biblioteca de criação de modelos ARIMA

Fonte: Autoria Própria (2021)

5.1.1 Tecnologias Utilizadas

O Módulo de Observação foi desenvolvido utilizando a linguagem de programação *C#*, que permitiu a criação de três Threads responsáveis por obter os dados de processamento, enviar o arquivo para o servidor e responder as interações do usuário no programa.

No Módulo de Execução e Previsão a construção ocorreu através da linguagem de programação *Python*, utilizando as seguintes bibliotecas:

1. *Matplotlib*: responsável pela visualização de dados através de gráficos (MATPLOTLIB, 2021 - https://matplotlib.org/);
2. *Numpy*: fornecendo uma variedade de rotinas rápidas para manipulação de dados em objetos multidimensionais (NUMPY, 2021 - https://numpy.org/);
3. *Pandas*: que provê ferramentas para manipulação de dados, séries temporais, leitura e gravação de arquivos (PANDAS, 2021 - https://pandas.pydata.org/);
4. *Pyramid-arima*: também conhecida como pmdaria, essa biblioteca oferece recursos para análise de série temporal, inclusive automatizando a

parametrização e geração de modelos ARIMA (PYRAMID-ARIMA, 2021 - https://pypi.org/project/pmdarima/);

5. *Scikit-learn*: responsável pelas ferramentas de avaliação dos modelos gerados (SCIKIT-LEARN, 2021 - https://scikit-learn.org/);
6. *Statsmodels*: fornece modelos estatísticos como o ARIMA, testes de hipótese como o ADF e análise estatística de dados como as funções de autocorrelação ACF e PACF (STATSMODELS, 2021 - https://www.statsmodels.org/);
7. *Tkinter*: para prover a interface da aplicação (TKINTER, 2021 - https://docs.python.org/3/library/tkinter.html)

5.1.2 Detalhes de Implementação

A construção do modelo *ARIMA* adotado na abordagem em questão se dá através dos seguintes passos: inicialmente, os dados reunidos são suavizados através de Média Móvel utilizando uma janela de tamanho 100 (SENGER; GOIS, 2018); em seguida, busca os parâmetros para a criação de um modelo ARIMA ideal.

No processo de definição dos parâmetros, o termo *'d'* está relacionado à estacionariedade, diferenciando a série temporal até que o resultado do teste *ADF* seja menor ou igual ao valor 0.05, significando que o teste possui 95% de intervalo de confiança (PAL; PRAKASH, 2017). Em seguida determina-se a ordem do termo autorregressivo *'p'*, identificando as correlações significativas dentro das primeiras 50 observações através da *PACF*. Analogamente ao processo que indica o termo 'p', o estabelecimento da ordem do termo de média móvel 'q' se dá variando apenas a utilização do *ACF* (WEI, 2006).

Após a definição dos parâmetros exigidos para a criação do modelo *ARIMA*, de acordo com as características da série temporal em análise, a Assinatura Comportamental é resultado da sua execução. Logo em seguida é executada a função *auto-arima* presente na biblioteca *"pyramid-arima"*, a qual utiliza a estratégia de força bruta, buscando o modelo ideal através da variação dos parâmetros *(p, d, q)* em um intervalo previamente definindo.

Deste modo, definimos que o valor inicial de cada parâmetro submetido para uso do algoritmo *auto-arima* possui o valor zero, e os valores finais são os identificados pelo modelo em questão, limitando assim o espaço de busca e o tempo

de processamento. Além disso, foi indicada a aplicação do teste *ADF* para verificar o ponto relacionado à estacionaridade, o qual foi utilizado na construção do modelo manual.

Na seção seguinte serão contrapostos os resultados obtidos e realizada uma exploração a partir dos dados decorrentes da execução dos algoritmos.

5.2 AVALIAÇÃO DA ABORDAGEM PROPOSTA

Com o intuito de validar a abordagem previamente apresentada o processo de avaliação será organizado em duas fases: na parte inicial validará o modelo *ARIMA* adotado para a criação da Assinatura Comportamental comparando-o com o modelo gerado através da biblioteca automatizada "*pyramid-arima*" utilizando a métrica *AIC*; na etapa final será avaliada a assertividade do processo comparando a Assinatura Comportamental com a série temporal que se pretendia prever.

As análises realizadas a seguir utilizaram dados de um computador pessoal utilizado especialmente em horário comercial, utilizado para construção de software com as seguintes especificações:

- Processador Intel Core i7-8565U;
- Número de núcleos 4;
- Nº de threads 8;
- Frequência baseada em processador 1.80 GHz;
- Frequência turbo Max 4.60 GHz;
- Cache 8 MB Intel® Smart Cache.

5.2.1 Modelo *ARIMA*

Com o propósito de validar os modelos criados para cada série temporal específica, reunimos os dados de monitoramento obtidos no período de 22/04/2020 à 08/09/2020, depois geramos os modelos *ARIMA* a partir dos dados da primeira semana coletada, segunda semana, seguindo até a oitava semana. A seguir serão apresentados os resultados dos modelos comparando o modelo gerado manualmente e o modelo proveniente da biblioteca "*pyramid-arima*" na Tabela 9, utilizando dados de uma a oito semanas para treinamento do modelo de acordo com os trabalhos efetivos relacionados no Quadro 4.

Tabela 9 – Comparação do modelo ARIMA para Assinatura e "*pyramid-arima*"

(continua)

Semanas de Treinamento	Dia da Semana	Modelo	*AIC*	Tempo de Execução (segundos)
1	Segunda-feira	Assinatura	1864.003	0.2150
		"*pyramid-arima*"	1875.302	0.4149
	Terça-feira	Assinatura	1547.894	0.1996
		"*pyramid-arima*"	1591.843	0.3587
	Quarta-feira	Assinatura	1403.636	0.2103
		"*pyramid-arima*"	1410.505	0.4457
	Quinta-feira	Assinatura	1617.144	0.2062
		"*pyramid-arima*"	1627.487	0.4698
	Sexta-feira	Assinatura	1856.280	0.2128
		"*pyramid-arima*"	2071.304	0.4284
2	Segunda-feira	Assinatura	18722.975	1.9408
		"*pyramid-arima*"	18733.933	4.5004
	Terça-feira	Assinatura	17719.374	1.5358
		"*pyramid-arima*"	17763.516	4.1481
	Quarta-feira	Assinatura	28262.740	3.1540
		"*pyramid-arima*"	28270.166	3.4422
	Quinta-feira	Assinatura	17015.989	1.7308
		"*pyramid-arima*"	17026.187	4.0746
	Sexta-feira	Assinatura	1856.280	0.2390
		"*pyramid-arima*"	2071.304	0.4400
3	Segunda-feira	Assinatura	18722.975	2.2646
		"*pyramid-arima*"	18733.933	4.3485
	Terça-feira	Assinatura	30575.089	4.2200
		"*pyramid-arima*"	30615.815	6.8433
	Quarta-feira	Assinatura	48356.300	4.8402
		"*pyramid-arima*"	48363.223	7.4868
	Quinta-feira	Assinatura	32319.518	4.2203
		"*pyramid-arima*"	32328.172	8.3382
	Sexta-feira	Assinatura	37799.867	4.6937
		"*pyramid-arima*"	38219.644	14.9732
4	Segunda-feira	Assinatura	56793.196	8.3124
		"*pyramid-arima*"	56809.774	13.1415
	Terça-feira	Assinatura	40864.116	7.9181
		"*pyramid-arima*"	40914.702	11.2632
	Quarta-feira	Assinatura	72937.060	9.7168

Tabela 9 – Comparação do modelo ARIMA para Assinatura e "*pyramid-arima*"

(continuação)

Semanas de Treinamento	Dia da Semana	Modelo	*AIC*	Tempo de Execução (segundos)
		"*pyramid-arima*"	72944.554	10.6947
	Quinta-feira	Assinatura	47960.428	8.1809
		"*pyramid-arima*"	47969.182	12.9036
	Sexta-feira	Assinatura	54944.862	8.3850
		"*pyramid-arima*"	55402.910	14.1170
5	Segunda-feira	Assinatura	78166.813	12.2599
		"*pyramid-arima*"	78177.157	15.8956
	Terça-feira	Assinatura	58976.404	12.6559
		"*pyramid-arima*"	59027.684	18.8549
	Quarta-feira	Assinatura	92310.120	12.9631
		"*pyramid-arima*"	92317.615	16.2024
	Quinta-feira	Assinatura	61117.487	12.6696
		"*pyramid-arima*"	61125.617	15.0589
	Sexta-feira	Assinatura	70977.759	13.9034
		"*pyramid-arima*"	71458.109	19.7100
6	Segunda-feira	Assinatura	103206.134	17.5044
		"*pyramid-arima*"	103213.609	20.2087
	Terça-feira	Assinatura	76513.571	19.4047
		"*pyramid-arima*"	76554.761	24.2140
	Quarta-feira	Assinatura	109188.805	22.4883
		"*pyramid-arima*"	109196.281	24.0691
	Quinta-feira	Assinatura	82268.500	17.2424
		"*pyramid-arima*"	82274.451	21.9557
	Sexta-feira	Assinatura	83677.415	19.3804
		"*pyramid-arima*"	84235.868	21.2230
7	Segunda-feira	Assinatura	121444.394	22.2004
		"*pyramid-arima*"	121452.601	19.4045
	Terça-feira	Assinatura	108986.998	23.6791
		"*pyramid-arima*"	109032.154	32.4744
	Quarta-feira	Assinatura	129485.030	27.9460
		"*pyramid-arima*"	129492.455	25.6322
	Quinta-feira	Assinatura	96720.600	23.4021
		"*pyramid-arima*"	96726.488	20.5063
	Sexta-feira	Assinatura	98080.042	20.5711
		"*pyramid-arima*"	98669.306	27.5307
8	Segunda-feira	Assinatura	138085.771	24.9962

Tabela 9 – Comparação do modelo ARIMA para Assinatura e "*pyramid-arima*"

(conclusão)

Semanas de Treinamento	Dia da Semana	Modelo	*AIC*	Tempo de Execução (segundos)
		"*pyramid-arima*"	138094.641	26.1719
	Terça-feira	Assinatura	124181.231	27.3821
		"*pyramid-arima*"	124227.417	26.2111
	Quarta-feira	Assinatura	147333.981	26.5964
		"*pyramid-arima*"	147341.589	24.0820
	Quinta-feira	Assinatura	109549.630	24.3554
		"*pyramid-arima*"	109555.639	24.0166
	Sexta-feira	Assinatura	98080.042	20.5696
		"*pyramid-arima*"	98669.306	26.3927

Fonte: Autoria Própria (2021)

A diferença média do *AIC* é de 102, onde o modelo gerado para uso da Assinatura possui um índice menor que o modelo automático. Apesar da pouca diferença na grande maioria das ocorrências, na sexta-feira os valores cresceram a cada semana indo de 215 a 589 de diferença mesmo com os parâmetros iguais.

Em relação ao tempo de processamento, existe uma diferença média de 2 segundos a mais entre tempo do processo gerado por meio da abordagem apresentada e o tempo do processo automático. Entretanto, a partir da semana 7 o tempo de processamento do processo automático passa a ter um desempenho melhor do que o processo realizado no módulo de Execução e Previsão, apesar de que a semana 7 concentra os piores índices de processamento.

Observa-se que o custo de processamento muda por completo de um para dois dígitos da semana 4 para a semana 5, sendo um indício de contraindicação do uso de semanas maiores do que 4 para treinamento. Do mesmo modo os valores do AIC começam a ficar consideravelmente altos, tendo seu valor médio variando da faixa de 5 mil para 7 mil, mostrando que o nível de qualidade dos modelos está a quem do modelo "real".

5.2.2 Geração da Assinatura Comportamental

Para verificar a assertividade das Assinaturas criadas na Tabela 9, elas serão submetidas ao módulo de Execução e Previsão com o intuito de contrapor os valores das observações inferidas e as observações reais. Os dados serão

confrontados com os valores da semana subsequente à última semana que constitui a respectiva Assinatura Comportamental, e logo adiante serão mostrados os valores originados pela diferença entre os dados na Tabela 10.

Tabela 10 – Avaliação da Previsão através da métrica RMSE e MAE

(continua)

Semanas de Treinamento	Dia da Semana	RMSE	MAE
1	Segunda-feira	37.84	25.5
	Terça-feira	36.14	26.07
	Quarta-feira	44.88	35.33
	Quinta-feira	22.67	12.93
	Sexta-feira	-	-
2	Segunda-feira	-	-
	Terça-feira	23.16	15.04
	Quarta-feira	28.4	18.73
	Quinta-feira	23.15	10.97
	Sexta-feira	-	-
3	Segunda-feira	-	-
	Terça-feira	18.72	9.65
	Quarta-feira	11.07	6.4
	Quinta-feira	27.64	15.07
	Sexta-feira	10.68	7.33
4	Segunda-feira	26.59	16.57
	Terça-feira	29.32	18.92
	Quarta-feira	29.38	19.94
	Quinta-feira	24.5	13.77
	Sexta-feira	24.31	13.24
5	Segunda-feira	16.66	10.0
	Terça-feira	25.02	15.87
	Quarta-feira	20.54	13.17
	Quinta-feira	52.04	44.88
	Sexta-feira	17.43	10.26
6	Segunda-feira	23.36	14.13
	Terça-feira	26.81	17.41
	Quarta-feira	29.89	16.68
	Quinta-feira	20.62	11.56
	Sexta-feira	28.23	20.22
7	Segunda-feira	15.56	8.85

Tabela 10 – Avaliação da Previsão através da métrica RMSE e MAE

(conclusão)

Semanas de Treinamento	Dia da Semana	RMSE	MAE
	Terça-feira	28.56	18.53
	Quarta-feira	30.83	18.55
	Quinta-feira	23.43	12.67
	Sexta-feira	-	-
8	Segunda-feira	34.54	21.84
	Terça-feira	32.26	21.86
	Quarta-feira	15.68	9.81
	Quinta-feira	35.0	25.73
	Sexta-feira	-	-

Fonte: Autoria Própria (2021)

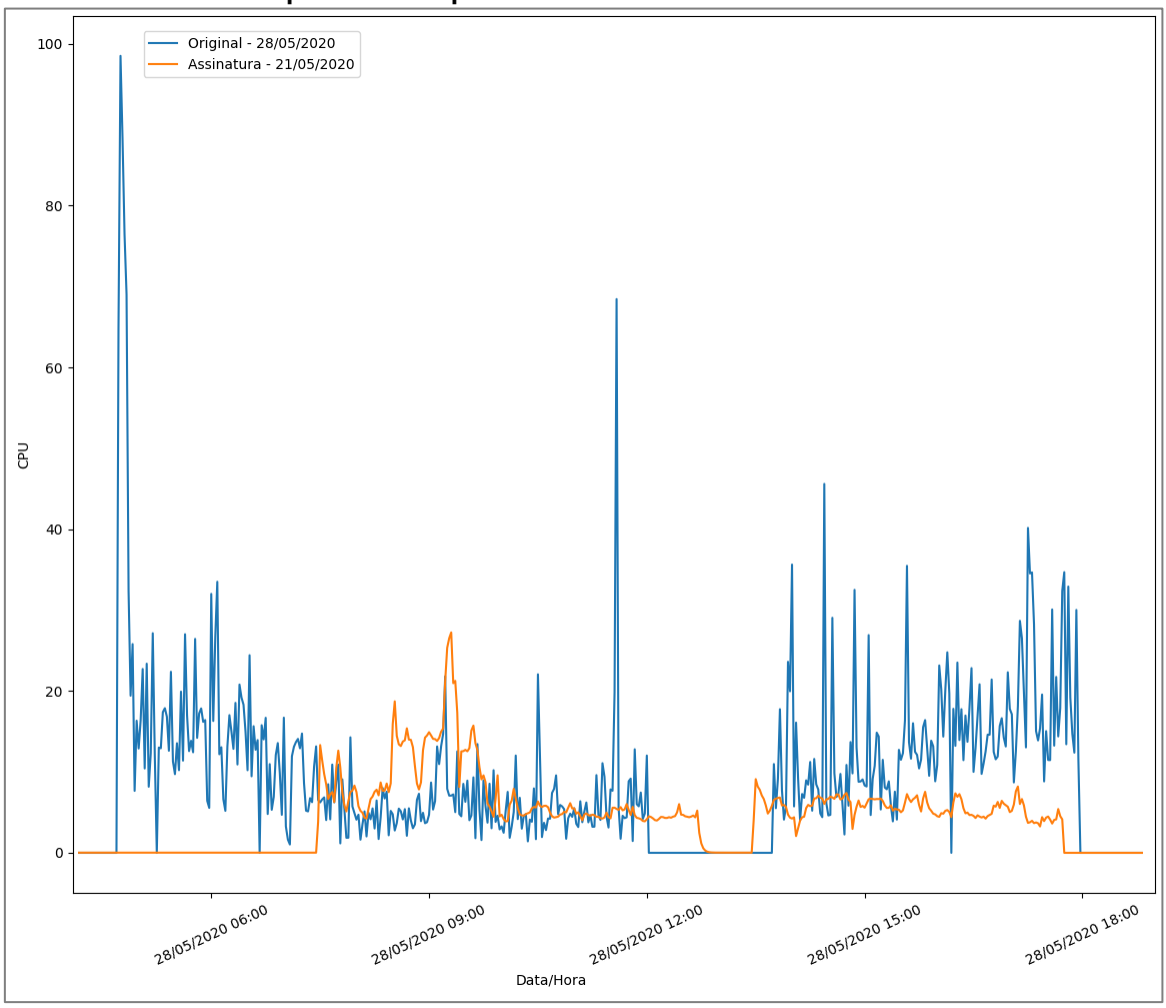
A ausência de previsão na sexta-feira das duas primeiras semanas tem como causa a ausência de dados no dia 01 de maio (feriado), onde na primeira ocorrência não havia dados para prever na semana seguinte à data da Assinatura gerada com base na semana anterior ao feriado. Na segunda ocorrência, como o algoritmo sempre busca prever os dados da semana subsequente à data mais recente que compõe a Assinatura, e não houve atualização da mesma em decorrência da ausência de dados, a mesma tentativa foi realizada.

O mesmo ocorre na segunda-feira da semana 2 e 3, assim como nas semanas 7 e 8 no dia de sexta-feira, sendo consequência da não utilização do equipamento e levando à ausência de dados nos dias 11 de maio e 12 de junho de 2020.

Em relação aos dados, percebe-se que os piores resultados se concentram na primeira e na última semana, os motivos seriam respectivamente a pouca quantidade de dados para treinamento e uma variação de abrupta de comportamento fora do padrão encontrado nos dados.

Os erros mais latentes ocorreram na quinta-feira da semana 5 e na quarta-feira da semana 6, identificados pelo *RMSE* e pela diferença entre o *RMSE* e o *MAE*. O primeiro caso se deu devido a dois *outliers* (dados discrepantes) e a uma mudança de comportamento de 4 horas na série temporal prevista, como mostra o Gráfico 5.

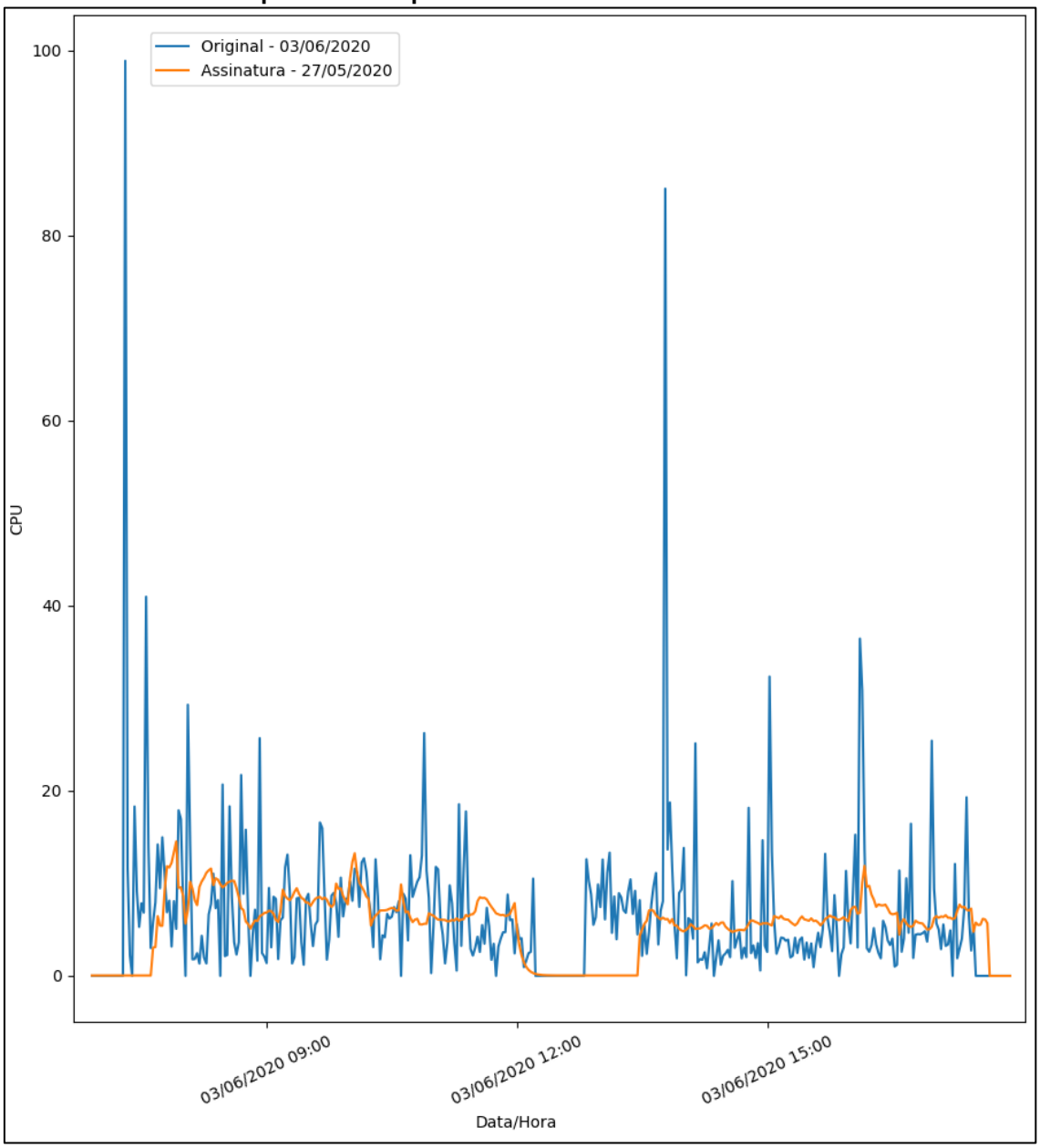
Gráfico 5 – Gráfico de previsão de quinta-feira da semana 5



Fonte: Autoria Própria (2021)

O segundo evento foi resultante de dois *outliers* na série prevista, vistos no Gráfico 5, apesar do baixo índice que representa a média absoluta de erros, de acordo com a Gráfico 6.

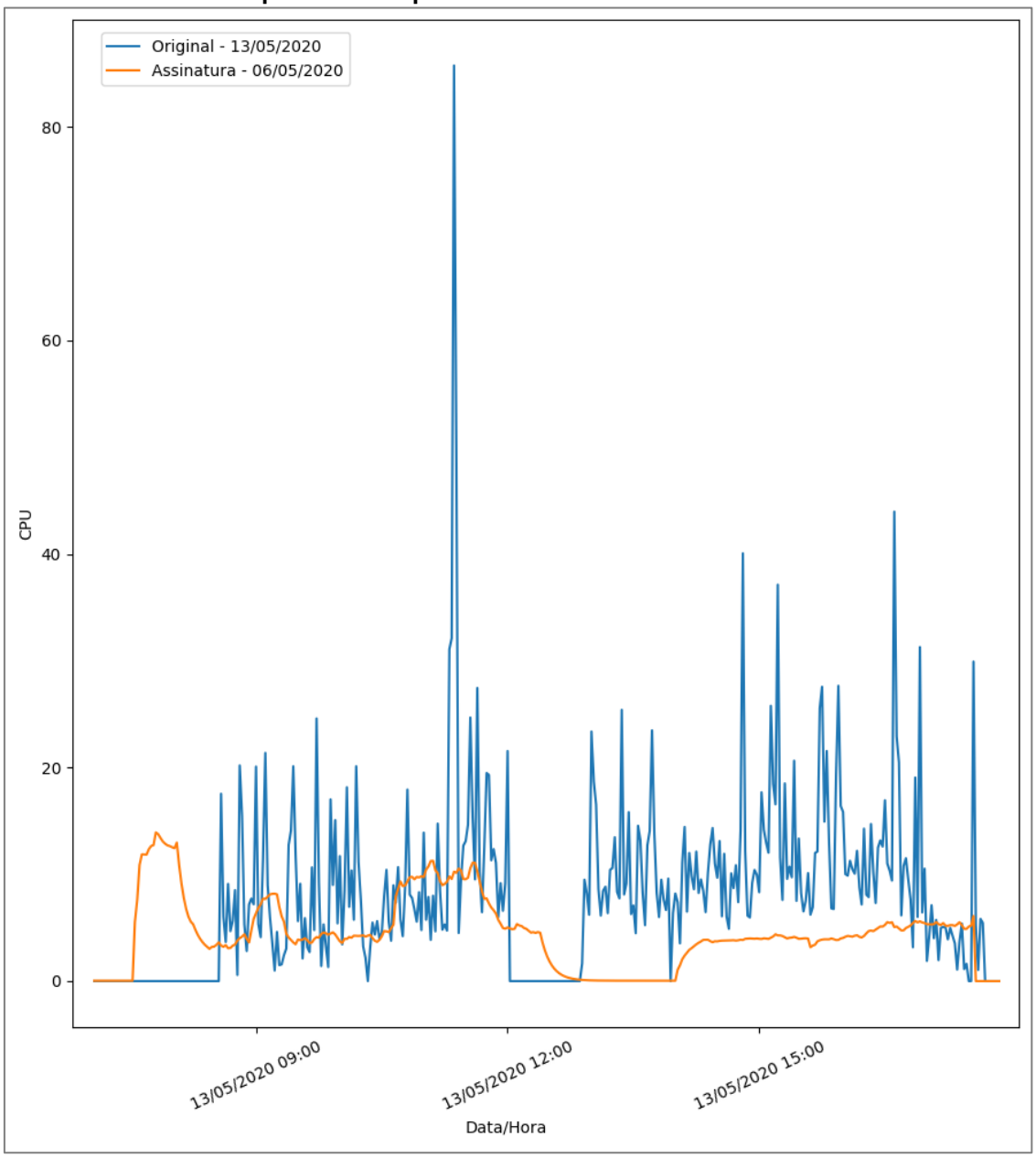
Gráfico 6 – Gráfico de previsão de quarta-feira da semana 6



Fonte: Autoria Própria (2021)

A semana 3 acumula os melhores resultados, tendo a menor diferença entre o *RMSE* e o *MAE*, revelando uma maior assertividade nas previsões. Dentre esse conjunto de dados destacam-se os dias de quarta-feira e sexta-feira. Através do Gráfico 7, observa-se que na quarta-feira houve apenas um *outlier* e sem grande variação no comportamento entre a Assinatura Comportamental e a série temporal prevista.

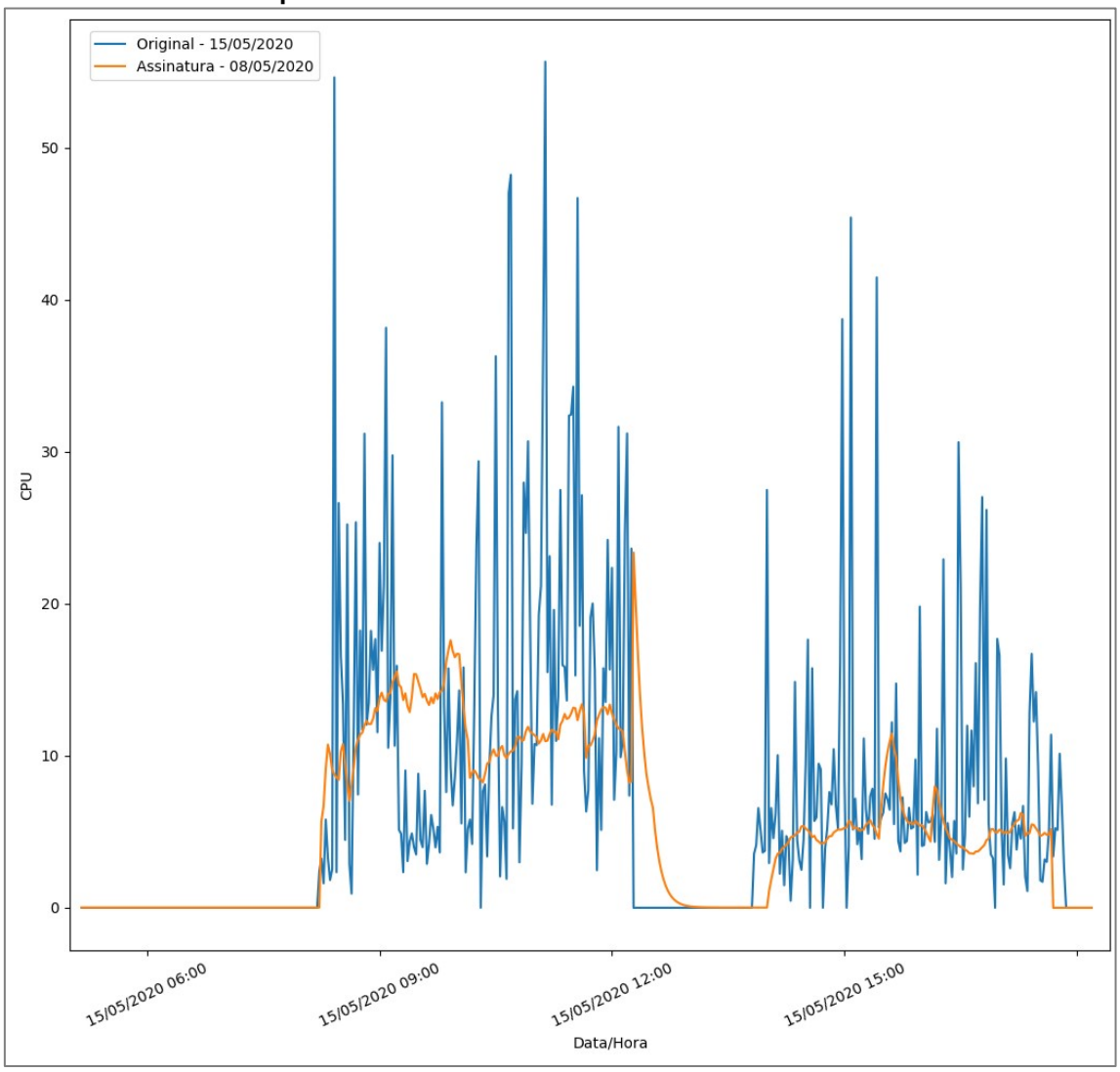
Gráfico 7 – Gráfico de previsão de quarta-feira da semana 3



Fonte: Autoria Própria (2021)

Com auxílio do Gráfico 8, percebe-se que na sexta-feira a Assinatura Comportamental se manteve na média da série temporal prevista e sem a ocorrência de *outliers*.

Gráfico 8 – Gráfico de previsão de sexta-feira da semana 3



Fonte: Autoria Própria (2021)

Observa-se que os *outliers* encontrados nas séries temporais que representam os valores reais a serem previstos causam um grande impacto negativo no resultado das métricas de avaliação. Outro aspecto que aumenta o índice de erro é a mudança abrupta de comportamento, mas uma semana de treinamento é o suficiente para reajustar a Assinatura Comportamental, como é o caso da ocorrência da quinta-feira da semana 5, que gerou o pior índice e na semana seguinte já se adaptou bem à variação.

5.2.3 Atualização da Assinatura Comportamental

Com o intuito de validar o processo de atualização da Assinatura, cruzamos os dados da Assinatura gerada com 8 semanas de treinamento, os dados da

Assinatura atualizada com 1 semana de treinamento a partir da Assinatura da semana 7 e os dados da Assinatura resultante do processo de atualização com 3 semanas de treinamento a partir da Assinatura da semana 5.

Ao utilizar o processo de atualização da Assinatura, disponível no Módulo de Execução e Previsão, para gerar a Assinatura Comportamental da semana 8 identificamos alguns pontos que devem ser considerados.

Tabela 11 – Comparação do modelo ARIMA para Assinatura e "*pyramid-arima*" no processo de atualização

(continua)

Semanas de Treinamento	Dia da Semana	Modelo	*AIC*	Tempo de Execução (segundos)
8	Segunda-feira	Assinatura	138085.771	24.9962
		"*pyramid-arima*"	138094.641	26.1719
	Terça-feira	Assinatura	124181.231	27.3821
		"*pyramid-arima*"	124227.417	26.2111
	Quarta-feira	Assinatura	147333.981	26.5964
		"*pyramid-arima*"	147341.589	24.0820
	Quinta-feira	Assinatura	109549.630	24.3554
		"*pyramid-arima*"	109555.639	24.0166
	Sexta-feira	Assinatura	98080.042	20.5696
		"*pyramid-arima*"	98669.306	26.3927
8 - Atualizada com 1 semana	Segunda-feira	Assinatura	5105.494	0.4259
		"*pyramid-arima*"	2472.425	0.6244
	Terça-feira	Assinatura	4960.261	0.3482
		"*pyramid-arima*"	1986.336	0.4896
	Quarta-feira	Assinatura	4552.771	0.3293
		"*pyramid-arima*"	2063.819	0.4458
	Quinta-feira	Assinatura	4375.422	0.4910
		"*pyramid-arima*"	1901.073	0.4143
	Sexta-feira	-	-	-
		-	-	-
8 - Atualizada com 3 semanas	Segunda-feira	Assinatura	48331.718	4.1430
		"*pyramid-arima*"	45332.398	6.5296
	Terça-feira	Assinatura	48996.396	4.0389
		"*pyramid-arima*"	45996.614	8.9752
	Quarta-feira	Assinatura	41757.100	4.5523
		"*pyramid-arima*"	39492.962	7.8657

Tabela 11 – Comparação do modelo ARIMA para Assinatura e "*pyramid-arima*" no processo de atualização

(conclusão)

Semanas de Treinamento	Dia da Semana	Modelo	*AIC*	Tempo de Execução (segundos)
	Quinta-feira	Assinatura	39893.612	4.0738
		"*pyramid-arima*"	37919.408	7.1080
	Sexta-feira	Assinatura	18953.578	1.9555
		"*pyramid-arima*"	16840.702	4.1010

Fonte: Autoria Própria (2021)

A Tabela 11 mostra que os dados resultantes do processo de atualização onde a Assinatura Comportamental da semana 8 foi gerada a partir da atualização da Assinatura da semana 7 com mais 1 semana de dados para treinamento do modelo, obtiveram índices *AIC* piores, em relação ao processo automático, do que os dados gerados com 8 semanas de treinamento. Por outro lado, a Assinatura atualizada a partir da atualização da Assinatura da semana 5 com mais 3 semanas de dados para treinamento do modelo teve um resultado melhor, mas não tão melhor quanto o do processo automático.

A despeito das métricas de previsibilidade, com auxílio da Tabela 12 fica evidente que os valores do *RMSE* obtidos na atualização utilizando 3 semanas são na sua maioria iguais ou melhores do que os resultados alcançados na geração com 8 semanas de treinamento. Isso mostra que quando houver a necessidade de atualização da Assinatura Comportamental, isso seja realizado com 3 semanas para treinamento do modelo.

Tabela 12 – Avaliação da Atualização através da métrica RMSE e MAE

(continua)

Semanas de Treinamento	Dia da Semana	RMSE	MAE
8	Segunda-feira	1192.78	34.54
	Terça-feira	1040.63	32.26
	Quarta-feira	245.80	15.68
	Quinta-feira	1224.84	35.00
	Sexta-feira	-	-
8 - Atualizada com 1 semana	Segunda-feira	1281.25	35.79
	Terça-feira	864.14	29.4
	Quarta-feira	278.53	16.69
	Quinta-feira	1569.42	39.62

Tabela 12 – Avaliação da Atualização através da métrica RMSE e MAE

(conclusão)

Semanas de Treinamento	Dia da Semana	RMSE	MAE
	Sexta-feira	-	-
8 - Atualizada com 3 semanas	Segunda-feira	1192.84	34.54
	Terça-feira	1039.52	32.24
	Quarta-feira	249.62	15.80
	Quinta-feira	1225.05	35.00
	Sexta-feira	-	-

Fonte: Autoria Própria (2021)

Essa constatação pode ser vista no trabalho de Yoo (2016), o qual afirma que a suposição de estacionariedade nos dados se mantém até 8 semanas no conjunto de dados de treinamento, isso levou o autor à formulação da hipótese de que atrasar as atualizações do modelo em pelo menos uma semana não degradaria o erro de previsão, ao invés de atualizar e reajustar o modelo sempre que novos dados de medição estiverem disponíveis.

5.3 CONSIDERAÇÕES

De acordo com a métrica *AIC* para avaliação de modelos *ARIMA*, a abordagem proposta possui um processo de geração de modelo similar ao processo oriundo da biblioteca automática *"pyramid-arima"*, tendo como destaque uma média de tempo de processamento melhor.

Quanto à quantidade de semanas para composição de uma Assinatura Comportamental, o melhor valor seria de 3 a 4 semanas tomando como base o tempo de processamento para criação do modelo e os resultados das métricas de previsibilidade. Esse resultado corrobora com a afirmação de Zhang (2016) que sugere 4 semanas de treinamento para o modelo e Benifa (2018) que utiliza 5 semanas.

E a atualização da Assinatura pode ser realizada considerando a Assinatura gerada há três semanas antes, utilizando os dados de 3 semanas posteriores para a atualização da Assinatura Comportamental seguinte, porém a assertividade da assinatura resultante ainda dependerá dos dados utilizados no processo.

6 CONCLUSÃO

Por se tratar de um tema pouco explorado em relação ao termo de Assinatura Comportamental, encontramos poucos trabalhos na revisão sistemática. A maioria dos trabalhos encontrados apresentam os algoritmos de previsão através de séries temporais aplicados como componentes de um sistema mais complexo, como esse trabalho representa o início do trabalho de pesquisa é natural haver um detalhamento maior no componente responsável pela previsibilidade de uma entidade.

A partir das informações adquiridas através da revisão, esta dissertação concebeu uma abordagem capaz de criar uma Assinatura Comportamental utilizando o modelo *ARIMA* com o intuito de prever o nível de processamento de Computadores Pessoais. A disposição interna da abordagem é definida com dois módulos que realizam a captação das informações de um computador pessoal através do seu monitoramento (Módulo de Observação) e que concentram os processos de previsão através de uma Assinatura, geração e atualização da Assinatura Comportamental (Módulo de Execução e Previsão).

A abordagem foi avaliada mediante a criação dos módulos descritos previamente, validação do modelo *ARIMA* gerado para compor a Assinatura Comportamental por meio do método *AIC* e uma análise da previsão através das métricas *MAE* e *RMSE* para mensurar o nível de assertividade.

Após a realização dos testes comprovou que a proposta inicial da abordagem foi atendida, principalmente pelo fato de ter uma Assinatura Comportamental que condiz com as séries temporais utilizadas na sua composição.

Dentre as contribuições desta pesquisa, destacam-se: a revisão sistemática referente ao uso de Séries Temporais na previsão de Recursos Computacionais; o Módulo de Observação que tem o propósito de fornecer os dados brutos e o Módulo de Execução e Previsão que centraliza as operações principais relacionadas à Assinatura Comportamental.

6.1 DESAFIOS E RESTRIÇÕES

Uma das limitações que a abordagem apresenta está no módulo de Observação, este processo pode ser melhorado através da mudança de arquitetura na obtenção dos dados de um computador pessoal. Atualmente os dados são centralizados em um arquivo e enviados para um servidor, isso poderia ser transformado para uma simples requisição HTTP com o identificador da máquina, o nome do recurso monitorado e seu valor.

Isso facilitaria o tratamento de erros para envio, teria uma maior segurança ao reunir esses dados em um banco de dados e ofereceria maior disponibilidade. Podendo melhorar até o desempenho do módulo de Execução e Previsão, que consequentemente deveria ser adaptado para ler os dados de um banco de dados e não de arquivos.

6.2 TRABALHOS FUTUROS

Visto que o presente trabalho tem vários pontos de extensão, como sugestão de trabalhos futuros tem-se:

- Criar um módulo que consuma as Assinaturas Comportamentais geradas e que seja capaz de tomar decisão de roteamento de tarefas para Computadores que tenham baixo uso de recursos;
- Aplicar outro algoritmo ao processo de criação de Assinatura Comportamental, a fim de comparar a sua assertividade e identificar os pontos positivos e negativos;
- Realizar mais avaliações com outras métricas de assertividade de previsão, com um conjunto maior de dados, e com cenários mais variados.

Além dos pontos de continuidade previstos na abordagem, há também limitações que podem ser resolvidas futuramente:

- Apresentar uma solução para o módulo de Observação, o qual em caso de falhas pode comprometer o processo a obtenção dos dados de monitoramento;
- Avaliar o tempo de execução dos módulos individualmente visando encontrar pontos de melhoria em relação ao desempenho.

REFERÊNCIAS

ADANIYA, M. H. A. C. **Detecção de Anomalias utilizando Firefly Harmonic Clustering Algorithm e Assinatura Digital de Segmente de Rede**. 2012. 83 f. Dissertação (Mestrado) – Ciência da Computação, Universidade Estadual de Londrina. Londrina, 2012.

ALVES, F. A. **Comparação de Testes de Raiz Unitária e Cointegração em Modelos de Longa Dependência**. 2008. 58 f. Dissertação (Mestrado) – Estatística, Universidade Federal de Minas Gerais. Belo Horizonte, 2008.

BENIFA, J. V. B.; DHARMA, D. HAS: Hybrid auto-scaler for resource scaling in cloud environment. **Journal of Parallel and Distributed Computing**, v. 120, p. 1-15, out. 2018.

BISGAARD, S. **Time Series Analysis and Forecasting by Example**. New Jersey: John Wiley & Sons, 2011.

BROCKWELL, P. **Introduction to Time Series and Forecasting**. 2ed. New York: Springer, 2002.

CAZAROTTO, S. **Teste de Raiz Unitária em Modelo Painel:** Uma aplicação a Teoria da Paridade Real de Juros na América Latina. 2006. 71 f. Dissertação (Mestrado) – Economia, Universidade Federal de Santa Catarina. Florianópolis, 2006.

CHATFIELD, C. **The Analysis of Time Series**: An Introduction. 6ed. Boca Raton: CRC Press, 2005.

DU, P.; *et al.* Multi-step ahead forecasting in electrical power system using a hybrid forecasting system. **Renewable Energy**, v. 122, p. 533-550, jul. 2018.

DUC, T. L.; *et al.* Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey. **ACM Computing Surveys,** v. 52, n.5, p. 1-39, set. 2019.

ENDERS, W. **Applied Econometric Time Series**. 4ed. New Jersey: John Wiley & Sons, 2015.

FAH, 2020. Disponível em: <https://foldingathome.org/about/>. Acesso em: 14 mai. 2021.

Farias, R. L. **Time series forecasting based on classification of dynamic patterns**. 140 f. Thesis (PhD) – Computer Science and Engineering, IMT School for Advanced Studies Lucca. Lucca, 2015.

HERNANDES, P. R. G. **Detecção de Anomalias com Assinatura Digital utilizando Algoritmo Genético e Análise de Fluxos IP**. 109 f. Dissertação (Mestrado) – Ciência da Computação, Universidade Estadual de Londrina. Londrina, 2016.

HUA, Y.; *et al.* Traffic Prediction Based on Random Connectivity in Deep Learning with Long Short-Term Memory. In: Vehicular Technology Conference (VTC-Fall). 88., 2018, Chicago. **Proceedings...** Chicago: Illinois, 2018. p. 1-6.

HYNDMAN, R.; KOEHLER, A. Another look at measures of forecast accuracy. . **International Journal of Forecasting,** v. 22, p. 679-688, dez. 2006.

IBM, 2020. Disponível em: <https://www.idc.com/getdoc.jsp?containerId=prUS45584619>. Acesso em: 14 mai. 2021.

IBM, 2011. Disponível em: <https://www.ibm.com/ibm/history/ibm100/us/en/icons/worldgrid/>. Acesso em: 14 mai. 2021.

IDC, 2019. Disponível em: <https://www.idc.com/getdoc.jsp?containerId=prUS45584619>. Acesso em: 14 mai. 2021.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. [S.l.], 2007.

MARINHO, C. S. S.; *et al.* LABAREDA: A Predictive and Elastic Load Balancing Service for Cloud-Replicated Databases. **Journal of Information and Data Management,** v. 9, n.1, p. 199-208, jun. 2018.

MEEKER, M. Internet Trends, 2014. Disponível em: <https://www.slideshare.net/kleinerperkins/internet-trends-2014-05-28-14-pdf/8-8Global_Users_of_TVs_vs >. Acesso em: 10 fev. 2021.

MONTGOMERY, D. C. **Introduction to Time Series Analysis and Forecasting**. New Jersey: John Wiley and Sons, 2008.

OUHAME, S.; HADI, Y. Multivariate Workload Prediction Using Vector Autoregressive and Stacked LSTM Models. In: New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society. 19., 2019, New York. **Proceedings...** New York: Association for Computing Machinery, 2019. p. 1-7.

PAL, A.; PRAKASH, P. **Pratical Time Series Analysis**: Master Time Series Data Processing, Visualization, and Modeling using Python. UK: Packt, 2017.

PANDITH, T. S. N.; BABU, C. N. Development of pso based hybrid lssvm model for time series prediction. In: International Conference for Convergence in Technology (I2CT). 2., 2017, Mumbai. **Proceedings...** Mumbai: [S.I.], 2017. p. 376-381.

PENA, E. H. M. **Um Sistema para Detecção de Anomalias que utiliza Assinatura Digital de Segmente de Rede, ARIMA Adaptativo e Lógica Paraconsistente**. 131 f. Dissertação (Mestrado) – Ciência da Computação, Universidade Estadual de Londrina. Londrina, 2014.

SENGER, W.; GOIS, L. A. Assinatura Comportamental e Detecção de Anomalias Utilizando K-means. In: Workshop de Pesquisas em Computação dos Campos Gerais (WPCCG). 2., 2017, Ponta Grossa. **Anais...** Ponta Grossa: UTFPR, 2017. p. 61-64.

____. **Arquitetura para Clusterização de Recursos Baseado em seu Poder Computacional utilizando Algoritmo Hierárquico e Assinatura Comportamental**. 80 f. Dissertação (Mestrado) – Ciência da Computação, Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2018.

SHUMWAY, R. H; STOFFER, D. S. **Time Series Analysis and Its Applications**: With R Examples. 3ed. New York: Springer, 2011.

SWAMYNATHAN, M. **Mastering Machine Learning with Python in Six Steps**: A Practical Implementation Guide to Predictive Data Analytics Using Python. Bangalore: Apress, 2017.

TOFALIS, C. A better measure of relative prediction accuracy for model selection and model estimation. **Journal of the Operational Research Society,** v. 66, p. 1352-1362, ago. 2015.

VALLIYAMMAI, C.; THAMARAI, S. S. Grid resource selection based on network performance prediction. In: International Conference on Advanced Computing (ICoAC). 70., 2015, Chennai. **Proceedings...** Chenai: [S.I.], 2015. p. 1-6.

VOYANT, C.; *et al.* Forecasting method for global radiation time series without training phase: Comparison with other well-known prediction methodologies. **Energy,** v. 120, p. 199-208, fev. 2017.

WEI, W. W. S. **Time Series Analysis**: Univariate and Multivariate Methods. 2ed. New York: Pearson Education, 2006.

WOOLDRIDGE, J. M. **Introdução à Econometria**: Uma Abordagem Moderna. São Paulo: Thomson Learnig, 2007.

YOO, W.; SIM, A. Time-Series Forecast Modeling on High-Bandwidth Network Measurements. **Journal of Grid Computing,** v. 14, p. 463–476, jun. 2016.

ZACARON, A. M.; *et al.* Assinatura Digital de Segmento de Rede Utilizando Analise de Fluxos e Clusterização K-means. In: SEMISH - Seminário Integrado de Software e Hardware. 39., 2012, Curitiba. **Anais do XXXII Congresso da Sociedade Brasileira de Computação (CSBC)**. Curitiba: [S.I], 2012.

ZACARON, A. M. **Assinatura Digital de Segmente de Redes utilizando Análise de Fluxos e Clusterização K-Means**. 109 f. Dissertação (Mestrado) – Ciência da Computação, Universidade Estadual de Londrina. Londrina, 2013.

ZHANG, F.; DE GRANDE, R.; BOUKERCHE, A. Accuracy Analysis of Short-term Traffic Flow Prediction Models for Vehicular Clouds. In: ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor and Ubiquitous Networks. 13., 2016, New York. **Proceedings...** New York: Association for Computing Machinery, 2016. p. 19-26.