UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE CIÊNCIAS SOCIAIS APLICADAS

DEPARTAMENTO DE CIÊNCIAS CONTÁBEIS E ATUARIAIS

GRADUAÇÃO EM CIÊNCIAS ATUARIAIS

ESTUDO E PREVISÃO DO CRIME CIBERNÉTICO

LAIS MARIA MUNIZ

Brasil
2021

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE CIÊNCIAS SOCIAIS APLICADAS

DEPARTAMENTO DE CIÊNCIAS CONTÁBEIS E ATUARIAIS

GRADUAÇÃO EM CIÊNCIAS ATUARIAIS

ESTUDO E PREVISÃO DO CRIME CIBERNÉTICO

LAIS MARIA MUNIZ

Monografia apresentada ao Departamento de Ciências Contábeis e Atuariais da Universidade Federal de Pernambuco, como requisito necessário para a obtenção do grau de Bacharel em Ciências Atuariais.

Orientador (a): Renata Gomes Alcoforado

Brasil
2021

Dedico este trabalho a todos aqueles que me ajudaram ao longo desta caminhada.

AGRADECIMENTOS

Primeiramente, quero agradecer a Deus, por sempre me ajudar e me dar forças para superar os obstáculos e alcançar meus objetivos.

Agradeço a minha familia, por todo encorajamento e inspiração, através de gestos e palavras.

Deixo um agradecimento especial a minha orientadora, Renata Alcoforado, por todo suporte, paciência e confiança. Sem seus ensinamentos e seu incentivo eu não teria conseguido.

Por fim, deixo um agradecimento eterno a todos que direta ou indiretamente me ajudaram e acreditaram em mim durante esse processo.

RESUMO

Quando se está conectado a internet se está exposto a diversos riscos, um deles é o crime cibernético. Esta grande ameaça da atualidade vem causando prejuízo para organizações e consumidores, principalmente porque a maioria dos indivíduos utiliza o armazenamento em formato eletrônico. O aumento da frequência dos ataques fez com que a proteção das informações se tornasse mais difícil e, desta forma, torna-se evidente a relevância de estudos acerca deste tema que tem sido tão abordado pela população em geral. O presente trabalho teve como objetivo realizar a modelagem e previsão dos ataques cibernéticos ocorridos no estado Norte Americano da Califórnia no período entre 2005 e 2019. Para isso forão utilizados 3 diferentes modelos de séries temporais: o modelo autoregressivo integrado de médias móveis (ARIMA), o de heterocedasticidade condicional autoregressiva generalizada (GARCH) e o alisamento exponencial. Através do software estatístico R, foram realizadas a modelagem dos dados e também a previsão para cada um dos modelos. Como resultado, foi possível observar que o modelo ARIMA apresentou uma previsão com valores que acompanhavam as curvas mas diferiam muito dos dados originais em nivelamento, gerando erros médios elevados; o modelo GARCH exibiu um intervalo de confiança para as previsões em conformidade com a realidade dos dados; já o alisamento exponencial gerou valores que se distanciaram dos dados reais tanto em nivelamento quanto no seguimento da volatilidade.

Palavras-chave: Crime cibernético, Séries temporais, ARIMA, GARCH, Alisamento Exponencial

ABSTRACT

When you are connected to the internet you are exposed to several risks, one of them is cyber crime. This major threat today is causing harm to organizations and consumers, mainly because most individuals use storage in electronic format. The increase in the frequency of attacks has made the protection of information more difficult, thus, cybersecurity practices and legislation aimed at this topic have been gaining more and more attention. The increase in the frequency of attacks made the protection of information more difficult and, thus, the relevance of studies on this topic that has been so discussed by the general population becomes evident. The present work aimed at modeling and predicting the cyber attacks that have occurred in the US state of California between 2005 and 2019. For this, 3 different time series models were used: the integrated autoregressive model of moving averages (ARIMA), generalized autoregressive conditional heteroskedasticity (GARCH) and exponential smoothing. Using the statistical software R, data modeling and forecasting were performed for each of the models. As a result, it was possible to observe that the ARIMA model presented a forecast with values that followed the curves but differed from the original leveling data, generating high average errors; the GARCH model exhibited a confidence interval for the predictions in accordance with the reality of the data; exponential smoothing, on the other hand, generated values that distanced themselves from real data both in leveling and following volatility.

Keywords: Cyber crimes, Time series, ARIMA, GARCH, Exponential smoothing

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

O ser humano, desde sua origem, se desenvolve constantemente e também busca transformar o ambiente ao seu redor. A isto se dá o nome de avanço tecnológico, quando se utilizam técnicas e aprendizados para aperfeiçoar e facilitar a execução de alguma atividade. Um dos avanços com maior alcance na modernidade é a internet.

A internet é um recurso utilizado atualmente por $59,5\%$ da população, segundo o relatório da Organização das Nações Unidas (ONU) do início de 2021. Esta ferramenta está presente em quase todas as atividades que existem na época atual e funciona como facilitador para muitas delas. A tecnologia fez com que tudo se tornasse conectado mas, isto acabou gerando, além das facilidades, alguns prejuízos, tal como o crime cibernético.

Esse tipo de crime ocorre desde muito tempo atrás porém, ao longo dos tempos e com a evolução tecnológica, a quantidade de crimes desse formato aumentou consideravelmente, afetando os mais diversos tipos de organizações ao redor do mundo (PETERS; SHEVCHENKO; COHEN, 2018). O relatório da Check Point Research (CPR) expõe alguns resultados sobre esse aumento, em média, na quantidade de ataques de hackers por semana, no mundo. Em 2021, segundo este mesmo relatório, o crescimento médio foi de 40% em relação aos números de 2020.

Pela importância supracitada, torna-se evidente a necessidade de estudos acerca deste tema que tem sido tão abordado pela população em geral. A análise de um conjunto de dados é um método considerável para aprofundar a compreensão do comportamento de dada circunstância e, também, possibilita uma melhor visualização das características de sua evolução, como por exemplo a frequência ao longo do tempo e intensidade dos ataques (PENG *et al.*, 2018).

Segundo o relatório da empresa de segurança, Netscout, o país que mais sofre ataques cibernéticos, atualmente, é os Estados Unidos, com 1,33 milhão de ataques ocorridos entre o 1º e o 3º trimestre de 2021. A Califórnia é o estado Norte Americano que mais sofreu ataques nos últimos anos, segundo a base de dados Privacy Rights Clearinghouse.

O objetivo do trabalho é analisar o comportamento das violações de dados ocorridas no estado da Califórnia, no período entre 2005 e 2019, utilizando modelos de séries temporais. Será realizada, também, uma previsão, para que os alvos desses ataques possam ter a possibilidade de alocar recursos de defesa adicionais quando necessário.

Embora a importância da previsão seja óbvia, só recentemente, quando foi proposto o uso de métodos inovadores, essa questão foi investigada. Esse pouco progresso na previsão de ataques pode ser uma consequência da falta de dados reais e modelos preditivos que sejam fáceis de se utilizar (PENG *et al.*, 2018).

A metodologia consiste em estudar, a partir de modelos de séries temporais, os dados observados e, em seguida, realizar uma previsão para os ataques futuros. Para isto, foram escolhidos três modelos: o modelo autoregressivo integrado de médias móveis (ARIMA), o processo de heterocedasticidade condicional autoregressiva generalizada (GARCH) e o alisamento exponencial. Os dados foram analisados por cada método e, por fim, foi observado o resultado para cada um deles. O trabalho será realizado utilizando o software estatístico R, que funciona como um ambiente livre para computação estatística e gráficos. Nele foram executados os processos de análise e previsão necessários para cada método.

A estrutura do trabalho está organizada da seguinte forma: o Capítulo 2 contém o referencial teórico; a descrição da metodologia é feita no Capítulo 3, no Capítulo 4 expomos e discutimos os resultados de cada método utilizado na pesquisa; por fim, no Capítulo 5, são feitas as considerações finais da pesquisa.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

2.1 Avanços Tecnológicos

Segundo sua etimologia, a tecnologia pode ser interpretada como o conhecimento técnico e científico e suas aplicações a um campo particular, ou seja, o uso de técnicas e do conhecimento adquirido para transformar o trabalho com a arte, encontrar a resolução de um problema ou a executar melhor uma tarefa específica. Pode-se dizer que é algo que acontece há mais tempo do que se pode imaginar, pois pode-se perceber o desenvolvimento tecnológico, por exemplo, desde os tempos primitivos com a descoberta do fogo e nas ferramentas básicas criadas a partir de elementos rústicos bem como nas inovações que têm surgido na atualidade. Nota-se que o avanço é contínuo e incalculável (CHATFIELD, 2021).

É possível observar que a tecnologia se desenvolveu imensamente desde a criação do primeiro computador (Electronic Numerical Integrator and Computer - ENIAC), em 1943. A evolução dos computadores e a difusão da internet possibilitou vários aspectos da modernidade como, por exemplo, a comunicação, que se desenvolveu e passou a funcionar de maneira extremamente prática. As pessoas estão totalmente conectadas mesmo vivendo em lugares muito distantes umas das outras, pois o armazenamento e o compartilhamento de dados é muito mais simples e acessível (FELIPE; LINS, 2013).

Os avanços alcançados pela tecnologia atingiram um ponto no qual existem diversas possibilidades de armazenar, utilizar e manipular os dados. A informação que era dispersa tornou-se organizada possibilitando melhores análises e estudos. As mensagens instantâneas agilizam processos que no passado poderiam durar anos e há também a utilização de processos digitalizados, que contribui para a durabilidade da informação contida no documento. A partir disso, várias ações podem ser realizadas com maior eficiência e muitos riscos podem ser minimizados (FINKELSTEIN; FINKELSTEIN, 2020).

Porém, quanto mais conectados, mais exposto se está. De acordo com Talesh (2018), os

dados fornecidos através da comunicação eletrônica acabam se tornando uma oportunidade para pessoas mal intencionadas. Tais indivíduos acabam realizando ações para benefício pessoal e malefício do outro. O ataque cibernético é um exemplo dessas ações.

2.2 Riscos Cibernéticos

O crime cibernético é um tipo de risco que tem sido um grande motivador de preocupações no mundo todo, pois as mais variadas organizações têm sofrido ataques dessa espécie como, por exemplo, agências do governo, universidades, setores financeiros e, em geral, todas as indústrias, como os serviços de emergência e saúde (PETERS; SHEVCHENKO; COHEN, 2018). Este risco pode ser causado por diferentes tipos de ataques e causar, consequentemente, diferentes problemas (SHIM, 2012).

Bonner (2012) afirmou que esse tipo de risco está entre as novas grandes ameaças enfrentadas por empresas e consumidores e envolve a ação de transgressores da lei que procuram efetivamente atacar os detentores de dados e adquirir informações pessoais através do uso de equipamentos eletrônicos, computadores, tecnologia da informação e realidade virtual. À medida que as pessoas dependem cada vez mais da comunicação eletrônica e também as organizações coletam e mantêm mais informações sobre seus consumidores, a oportunidade para os malfeitores causarem problemas para as organizações e o público aumenta exponencialmente (TALESH, 2018).

Torna-se, então, uma fonte de ameaças em constante mudança, algumas delas bastante sofisticadas e que podem ser ajustadas quase que instantaneamente para conter possíveis defesas. Técnicas e ferramentas utilizadas pelos praticantes destes ataques podem ser atualizadas rapidamente e também compartilhadas. A implantação é rápida, amplamente dispersa e adaptável (BONNER, 2012).

Os danos cibernéticos podem ser de diversos tipos: econômico, psicológico, reputacional, social, etc (AGRAFIOTIS *et al.*, 2018). Um exemplo desta categoria de risco foi a violação de dados ocorrida na Sony, em 2011. Tal exposição resultou num custo de dezenas de bilhões de dólares, incluindo políticas de proteção contra roubo de identidade para usuários, exames e investigação forense eletrônica. Há também as responsabilidades enfrentadas em ações judiciais coletivas que acusam a Sony de negligência e violação de privacidade. Este caso demonstrou o impacto devastador que os ataques cibernéticos podem ter nas empresas. Sem proteção adequada contra riscos cibernéticos, o risco de perdas desastrosas é altíssimo (BONNER, 2012).

Gerenciar a segurança cibernética da melhor forma tornou-se algo crucial, pois as informações do consumidor, financeiras e de saúde são cada vez mais armazenadas em formato eletrônico. Hackers, malware, vírus, software de rastreamento, escutas telefônicas, espionagem, ligações automáticas e solicitações podem levar ao roubo de identidade e comprometimento de

informações pessoais (TALESH, 2018).

2.3 Cibersegurança

Lessig (1998) afirmou que estávamos entrando em uma era na qual a privacidade, em qualquer sentido desse termo, seria fundamentalmente alterada, uma era em que a extensão do monitorado e o alcance do que é pesquisável é muito maior do que qualquer outra que conhecemos. E é exatamente neste cenário que vive-se atualmente, onde as informações pessoais são de fácil acesso para qualquer um que queira utilizá-las. Tudo está se tornando conectado à internet, o que significa que tudo está se tornando hackeável (MANWORREN; LETWAT; DAILY, 2016).

Proteger essas informações é algo que tem sido cada vez mais difícil para as organizações, pois a frequência de ataques cibernéticos tem aumentado de forma contínua (COHEN *et al.*, 2019), por outro lado, a defesa dessas informações tem se tornado, cada vez mais, uma prioridade (PANDA *et al.*, 2021) e este assunto tem recebido bastante atenção ultimamente.

Segundo o governo dos Estados Unidos, a segurança cibernética é um dos desafios econômicos e de segurança nacional mais sérios enfrentado, como nação. Esta ameaça está presente no mundo todo e em números crescentes. Muitas empresas relatam milhares de ataques todos os meses, variando de triviais a extremamente graves. Vários bilhões de conjuntos de dados são violados anualmente. Todos os anos, os hackers produzem cerca de 120 milhões de novas variantes de malware (POPPENSIEKER; RIEMENSCHNITTER, 2018).

O Instituto Nacional de Padrões e Tecnologia (NIST) já reconheceu que é quase impossível eliminar esse risco totalmente, restando a opção de implantar controles técnicos e não técnicos para diminuí-lo (HERATH; HERATH, 2011). Os serviços de gerenciamento de risco cibernético não apenas reduzem o risco mas também constroem ativamente o significado de conformidade (TALESH, 2018). Uma forma de fortalecer a segurança contra esses ataques é a partir do uso de seguro cibernético. Muitas empresas e organizações ainda não possuem essa prática que auxilia na sua defesa e prevenção contra o risco cibernético. Políticas que encorajem a adoção dessa modalidade de seguro são muito úteis para expandir essa forma de proteção (KSHETRI, 2020).

Alguns vêem o seguro cibernético como um método mais barato e mais simples para a proteção contra o risco cibernético. Preferem transferir todo o risco para a seguradora e deixam de investir num sistema próprio de segurança (KSHETRI, 2020). O foco principal das seguradoras cibernéticas é evitar a violação de dados, cumprindo todas as leis de privacidade. Para isso, são fornecidas orientações para que as organizações se preparem de forma adequada para possíveis ameaças. No Regulamento Geral de Proteção de Dados Pessoais Europeu n° 679, (General Data Protection Regulation - GDPR), por exemplo, é enfatizado que deve-se considerar o enquadramento legal na modelagem dos sinistros, por exemplo, pois a cobertura pode incluir

multas e penalidades (ZELLER; SCHERER, 2021).

Quando um ataque acontece várias ações devem ser executadas como, por exemplo, identificar a fonte e a causa da violação de dados, restaurar os processos de rede que podem ter sido danificados como resultado da violação, etc. Por isso, deve existir um especialista em segurança da informação cibernética para lidar com esses problemas (TALESH, 2018).

Os governantes, por sua influência e intervenção, podem executar um papel valioso no incentivo de práticas que diminuam o risco cibernético entre as empresas e organizações. Estimular o crescimento do mercado que atua na mitigação do risco cibernético é uma forma de promover o aumento da eficiência das novas políticas e ações acerca da segurança nas redes (BONNER, 2012).

2.4 Políticas e Legislação

Nos últimos 5 anos têm sido criadas ou aperfeiçoadas diversas leis de privacidade, regulamentos e diretrizes no setor de segurança de dados, isto porque, na mesma medida em que as ocorrências de violações de segurança aumentam, cresce também o número de pessoas que buscam soluções legais para se prevenir contra possíveis danos (TALESH, 2018). Tais danos podem ter efeitos que ultrapassam os dados e software da empresa, fazendo com que a mesma incorra em responsabilidade. Ou seja, ela será obrigada a pagar uma indenização às partes que sofrerem algum tipo de dano (GORDON; LOEB; SOHAIL, 2003).

Apesar de todas essas medidas que vêm sendo tomadas, as ameaças à segurança cibernética não estão prestes a desaparecer. Os ataques continuam acontecendo e as empresas precisam estar preparadas para proteger seus clientes, suas informações, sua reputação e seus resultados financeiros (MANWORREN; LETWAT; DAILY, 2016).

Na Holanda, as partes responsáveis pela segurança forneceram informações, a partir de campanhas, sobre as vulnerabilidades cibernéticas e possíveis medidas a serem tomadas para que a segurança fosse aumentada e possíveis perdas evitadas. Outros países seguiram este mesmo procedimento como forma de minimizar o impacto dos riscos cibernéticos (KSHETRI, 2020).

A preocupação em relação aos direitos à privacidade vêm aumentando conforme o passar dos anos em âmbito mundial. Muitos países incorporaram este tópico às suas leis no último século, ou passaram a abordar tal assunto de maneira distinta. A Organização das Nações Unidas (ONU) é um exemplo de organização que atua com foco na proteção dos direitos do indivíduo, expondo o lado negativo da modernidade e suas tecnologias (FINKELSTEIN; FINKELSTEIN, 2020).

O Marco Civil da Internet, Lei Nº 12.965/14, também possui artigos que visam a proteção à confidencialidade e inviolabilidade da vida privada digital e os fluxos de tráfego da internet. Garantindo também que a guarda e disponibilização de registros de conexão e de acesso à aplicações a internet resguardem a intimidade, honra e imagem de seus usuários. Os artigos 1º e

8º do Marco Civíl da Internet estabelecem:

> Art. 1º Esta Lei estabelece princípios, garantias, direitos e deveres para o uso da internet no Brasil e determina as diretrizes para atuação da União, dos Estados, do Distrito Federal e dos Municípios em relação à matéria.
>
> Art. 8º A garantia do direito à privacidade e à liberdade de expressão nas comunicações é condição para o pleno exercício do direito de acesso à internet (Marco Civil da Internet, 2014).

Contudo, ainda que essa preocupação exista há anos, esteja disposta na legislação e tenha sido divulgada por outros meios, é perceptível que a atenção dada a este assunto intensificou consideravelmente nos últimos anos, tanto pela população, no geral, quanto pelo legislador (FINKELSTEIN; FINKELSTEIN, 2020).

O Regulamento Geral de Proteção de Dados Pessoais Europeu n° 679, é um exemplo disso. Ele entrou em vigor no ano de 2018 e trouxe consigo uma nova percepção sobre a proteção de dados pessoais. Seu objetivo foi reforçar e unificar a proteção de dados pessoais na União Europeia por meio de uma adaptação dos princípios à sociedade da informação, que cada vez mais realiza coleta e tratamento de dados pessoais, físicos ou digitais, por meio da internet, ou não. O atigo 1º do RGPD estabelece:

> Artigo 1º
>
> Objeto e objetivos
>
> 1. O presente regulamento estabelece as regras relativas à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados.
>
> 2. O presente regulamento defende os direitos e as liberdades fundamentais das pessoas singulares, nomeadamente o seu direito à proteção dos dados pessoais.
>
> 3. A livre circulação de dados pessoais no interior da União não é restringida nem proibida por motivos relacionados com a proteção das pessoas singulares no que respeita ao tratamento de dados pessoais. (Regulamento Geral sobre a Proteção de Dados, 2016)

Um outro exemplo é a lei brasileira n° 13.709/2018 ou Lei Geral de Proteção de Dados (LGPD), que teve como texto base a GDPR e entrou em vigor no ano de 2020. O principal objetivo dessa nova legislação é proteger o cidadão no que diz respeito ao armazenamento, por parte de empresas públicas e privadas, do fluxo de suas informações no meio digital, assegurando sua privacidade e liberdade. Medidas como a criação da GDPR e da LGPD são necessárias ao desenvolvimento social saudável tanto dentro do mundo virtual quanto fora dele.

CAPÍTULO 3

PROCEDIMENTOS METODOLÓGICOS

3.1 Base de dados

A base de dados escolhida para a modelagem foi a Privacy Rights Clearinghouse, que traz em sua composição as informações de relatos das violações de dados ocorridas entre os anos de 2005 e 2019, nos Estados Unidos. Tais violações podem ser de qualquer tipo como, por exemplo, ataques de hackers, vazamentos não propositais por algum funcionário, documentos físicos roubados ou perdidos, etc. Dentre as informações disponíveis na base estão a data em que o acidente foi divulgado, a quantidade de dados e o local violado, o tipo de violação e a descrição do incidente (PCR).

3.1.1 Processamento dos dados

A base contém milhares de informações sobre todos os estado dos Estados Unidos e, por isso, foi selecionada uma amostra para ser utilizada no trabalho. Analisando a base, foram escolhidos os incidente que ocorreram na Califórnia, pois é o local onde a maioria das violações ocorreram e que teve a maior quantidade de dados violados. Estes dados foram separados também pelo tipo de ataque, sendo selecionados apenas os ataques classificados como "HACK", que são definidos como "hackeado por uma parte externa ou infectado por malware".

No período observado, 470 ataques ocorreram. Desses dados, 90% foram utilizados para a modelagem, enquanto os 10% restantes foram reservados para se utilizar na comparação com a previsão.

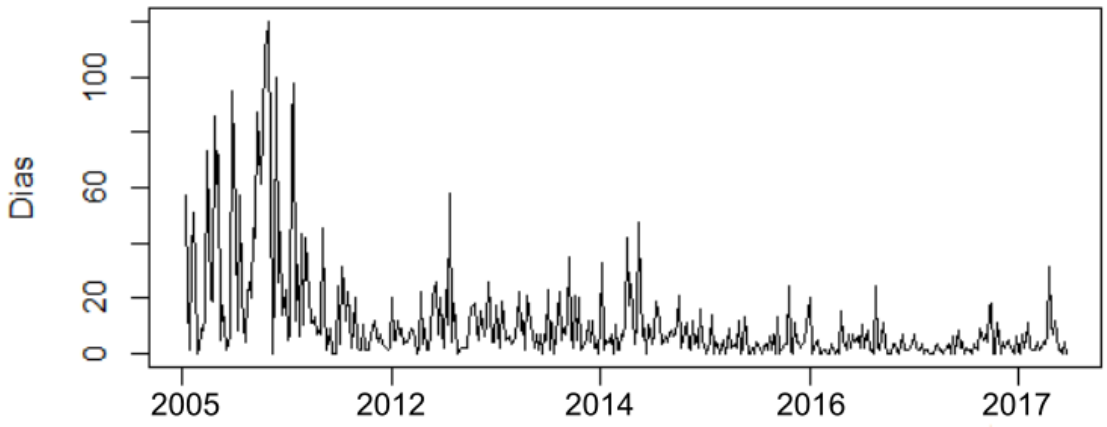
A base está organizada na seguinte ordem: data em que o ataque se fez público; total de dados violados; companhia que sofreu o ataque; cidade em que ocorreu e o tipo da organização atacada. Na última coluna, os tipos de organizações estão representados por siglas e a Tabela 3.1.1 indica a legenda para cada uma delas:

Tabela 3.1.1 – Tipos de organização

BSF	Negócios (serviços financeiros e de seguros)
BSR	Negócios (varejo / comerciante, incluindo varejo online
BSO	Negócios (outros)
EDU	Instituições educacionais
MED	Cuidados de saúde, prestadores de serviços médicos e serviços de seguro médico
GOV	Governo e Militar
NGO	Organizações sem fins lucrativos

A modelagem foi feita a partir do intervalo de tempo entre os ataques ocorridos no período, e este intervalo foi conseguido calculando a diferença entre as datas em que os ataques ocorreram. O resultado pode ser observado na Figura 3.1.1 . O eixo de x apresenta a passagem de tempo a cada 100 ataques ocorridos, então os 100 primeiros ataques ocorreram entre os anos de 2005 e 2012, seguindo desta forma para o restante.

Figura 3.1.1 – Intervalo de dias entre os ataques cibernéticos



A seguir, serão apresentados os três modelos de séries temporais que serão utilizados na modelagem e previsão do trabalho.

3.2 Modelo ARIMA

O modelo autorregressivo integrado de médias móveis (ARIMA) compõe uma forma bastante utilizada na análise de modelos paramétricos. Ele é composto pela junção do modelo autoregressivo (AR) e do modelo de médias móveis (MA) e é um processo frequentemente discutido em estudos de previsão (BALTAR, 2009).

O modelo AR é utilizado quando existe autocorrelação entre os dados observados. O modelo MA é utilizado quando há autocorrelação entre os resíduos. Há também o modelo autoregressivo de médias móveis (ARMA), que é utilizado quando há autocorrelação entre as observações e entre os resíduos. Quando os dados utilizados no modelo ARMA são uma

diferença dos dados originais então essa série é uma integral do modelo e segue um modelo autorregressivo integrado de médias móveis (JEKLIN, 2016).

Segundo Morettin e Toloi (1987), Um processo autorregressivo de ordem p, (AR (p)), pode ser definido da seguinte forma:

$$Y_t = \sum_{p=1}^{P} \phi_p Y_{t-p} + \varepsilon_t$$

em que, Y_t são os dados sobre os quais o ARIMA será aplicado, ϕ_p são os coeficientes autoregressivos e ε_t é o ruído branco.

Enquanto um processo de médias móveis de ordem q, (MA (q)), é definido como segue abaixo:

$$Y_t = \varepsilon_t - \sum_{q=1}^{Q} \theta_q \varepsilon_{t-q}$$

em que, θ_q são os coeficientes de média móvel.

Dessa maneira, combinando os esquemas acima, os processos autoregressivo de médias móveis de ordem p e q, (ARMA(p,q)) seguem o seguinte modelo:

$$Y_t = \sum_{p=1}^{P} \phi_p Y_{t-p} + \varepsilon_t - \sum_{q=1}^{Q} \theta_q \varepsilon_{t-q}.$$

Por sua vez, o processo autorregressivo integrado de médias móveis de ordem p, d e q, (ARIMA(p,d,q)), seguirá o seguinte processo:

$$\triangle^d Y_t = \sum_{p=1}^{P} \phi_p Y_{t-p} + \varepsilon_t - \sum_{q=1}^{Q} \theta_q \varepsilon_{t-q}.$$

Em que o parâmetro integrado (d) representa o número de transformações que ocorrem para tornar a série estacionária, ou seja, $\triangle^d Y_t$ é a série original Y_t diferenciada "d"vezes.

O modelo ARIMA é construído conforme o seguinte ciclo:
I. Especificação do modelo
II. Identificação do modelo
III. Estimação dos parâmetros do modelo
IV. Verificação ou diagnóstico do modelo ajustado

Caso o modelo não seja adequado, o ciclo deverá ser repetido, retornando para a identificação do modelo, que pode ser considerada a fase crítica do processo, já que vários modelos podem ser identificados para uma mesma série.

3.3 Modelo GARCH

O processo de heterocedasticidade condicional autorregressiva generalizada (GARCH), foi introduzido por Bollerslev, em 1986, e se baseia nos processos de heterocedasticidade condicional autorregressiva (ARCH) que, apesar de fornecerem muitas contribuições, apresentam alguns problemas para a modelagem da volatilidade. O GARCH, no entanto, melhora a especificação original adicionando variância condicional defasada, que atua como um termo de suavização. Um modelo GARCH pode descrever a volatilidade utilizando menos parâmetros do que um ARCH (MORETTIN; TOLOI, 1987).

Um processo de heterocedasticidade condicional autorregressiva generalizada de ordem p e q, (GARCH(p,q)), é definido como:

$$X_t = \sqrt{h_t}\varepsilon_t$$
$$h_t = \alpha_0 + \sum_{p=1}^{P} \alpha_p X_{t-p}^2 + \sum_{q=1}^{Q} \beta_q h_{t-q}$$

$$\varepsilon_t \text{ é i.i.d.}$$
$$\alpha_0 > 0$$
$$\alpha_p >= 0$$
$$\beta_q >= 0$$
$$\sum_{i=1}^{z}(\alpha_i + \beta_i) < 1, z = max(p, q)$$

em que, X_t é a série que será utilizada e h_t representa a estimativa do desvio padrão condicional, ou seja, a volatilidade.

O termo alfa, indica quanto a última observação observado tem de influência sobre a variância condicional hoje, já o termo beta, indica quanto a volatilidade do período anterior deve influenciar a volatilidade hoje.

De acordo com Morettin e Toloi (1987), geralmente, é difícil identificar a ordem de um modelo GARCH. É recomendado que sejam utilizados modelos de ordem baixa ((1,1), (1,2), ...) para depois se escolher um com base em alguns critérios como, por exemplo, AIC ou BIC, valores da assimetria e curtose, da log-verossimilhança ou de alguma função de perda. Quanto aos estimadores dos parâmetros, ele são definidos a partir do método de máxima verossimilhança condicional.

A previsão de volatilidade neste modelo pode ser calculada de maneira similar a do modelo ARMA. Num modelo GARCH(1,1), por exemplo, a equação ficaria no seguinte forma:

$$\hat{h}_t(1) = \alpha_0 + \alpha_1 X_t^2 + \beta_1 h_t$$
$$\hat{h}_t(l) = \alpha_0 + \alpha_1 X_t^2(l-1) + \beta_1 h_t(l-1), l >= 1$$

3.4 Alisamento Exponencial

Modelos de alisamento ou suavização exponencial são populares devido a sua simplicidade, à eficiência computacional e ao seu nível de precisão. Esta classe de modelos pode se adequar a diferentes séries temporais e, por isso, existem métodos específicos para cada situação (a série pode ser localmente constante ou pode apresentar tendência ou sazonalidade) (MORETTIN; TOLOI, 1987).

Os modelos utilizados no trabalho serão a suavização exponencial simples (SES) e a suavização exponencial de Holt (SEH). A SES é aplicado em séries localmente constantes, enquanto a SEH é para as série que apresentam tendência (MORETTIN; TOLOI, 1987).

A SES é uma média ponderada que estabelece maiores pesos às observações mais recentes e pode ser definida matematicamente como:

$$Y_t = \alpha Y_t + (1-\alpha)Y_{t-1}$$

$$Y_0 = Y_1$$

$$t = 1, ..., N$$

$$0 < \alpha < 1$$

em que Yt é definido como o valor exponencialmente suavizado e α é o fator de suavização.

A previsão para a SES é dada pelo último valor suavizado exponencialmente:

$$\hat{Y}_t(h) = \overline{Y}_t, \forall h > 0$$

$$\hat{Y}_t(h) = \alpha Y_t + (1-\alpha)\hat{Y}_{t-1}(h+1)$$

Como pode-se notar, só serão utilizadas para a previsão a observação mais recente, a previsão imediatamente anterior e o valor da constante α, que pode ser determinado a partir de alguns critérios como, por exemplo, pelo tipo de autocorrelação entre os dados e custo de previsão ou, mais objetivamente, escolhendo o valor que fornecer a "melhor previsão"das observações já obtidas. Esta etapa do processo, porém, acaba por ser sua maior desvantagem, já que encontrar o valor mais apropriado para a constante de suavização é uma tarefa muito difícil (MORETTIN; TOLOI, 1987).

Enquando no SES somente o nível da série é modelado, a SEH apresenta, como diferença, a modelagem também da tendência a partir de uma nova constante de suavização.

O nível e a tendência serão estimados, no instante *t*, por:

$$\overline{Y}_t = \alpha Y_t + (1-\alpha)(\overline{Y}_{t-1} + \hat{T}_{t-1}); t = 2, ..., N; 0 < \alpha < 1$$

$$\hat{T}_t = \beta(\overline{Y}_t - \overline{Y}_{t-1}) + (1-\beta)\hat{T}_{t-1}; t = 2, ..., N; 0 < \beta < 1$$

Sendo α e β as constantes de suavização do modelo.

A previsão para a SEH pode ser obtida adicionando-se ao valor de $\overline{Y}_t$ a tendência da série multiplicada pelo número de passos a frente que ser quer prever. Ou seja:

$$\hat{Y}_t(h) = \overline{Y}_t + h\hat{T}_t, \forall h > 0$$

CAPÍTULO 4

RESULTADOS E DISCUSSÃO

4.1 Estatísticas descritivas

Através de análises, foi identificado que o tipo de organização que mais sofreu ataques e também que teve a maior quantidade de dados violados durante o intervalo de tempo observado foi a de negócios (BSO), que engloba serviços como github, snapchat, apple, etc. Os valores encontrados podem ser observados na Tabela 4.1.1.

Tabela 4.1.1 – Estatística dos dados

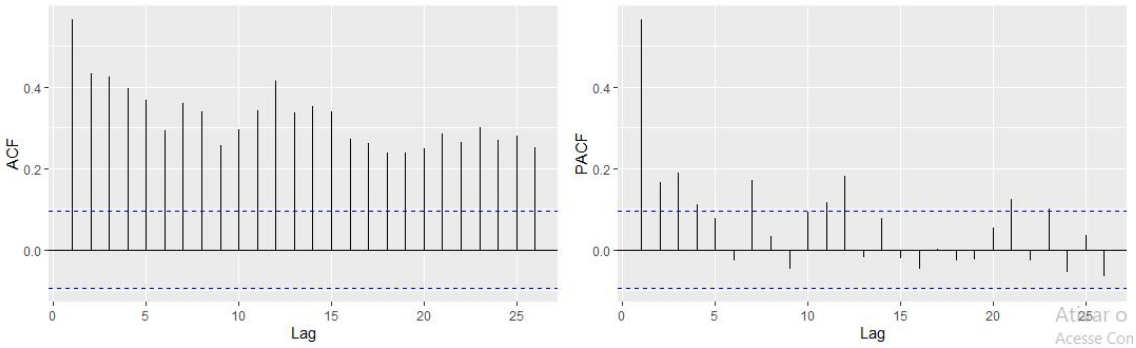
Tipo de Organização	Quantidade de Ataques	Quantidade de Dados
BSO	183	5.008.360.951
MED	92	10.740.903
BSR	68	221.423.600
BSF	52	15.063.796
EDU	51	41.148.562
GOV	19	50.647
NGO	5	4.115

Observando o gráfico do intervalo de tempo entre os ataques também foi possível perceber que nos anos iniciais havia uma lacuna de dias maior do que a que se observa nos anos finais, indicando que os ataques se tornaram mais frequentes do que costumavam ser.

4.2 Modelo ARIMA

Através das funções acf e pacf, disponíveis no pacote "stats"do R, os gráficos de autocorrelação (ACF) e autocorrelação parcial (PACF) da série foram conseguidos. Eles indicaram a necessidade da diferenciação, como mostra a figura 4.2.1 e isso foi confirmado com o resultado obtido na função auto.arima, disponível no pacote "forecast". Essa função retorna, dada uma série de dados, parâmetros eficientes para a modelagem da mesma.

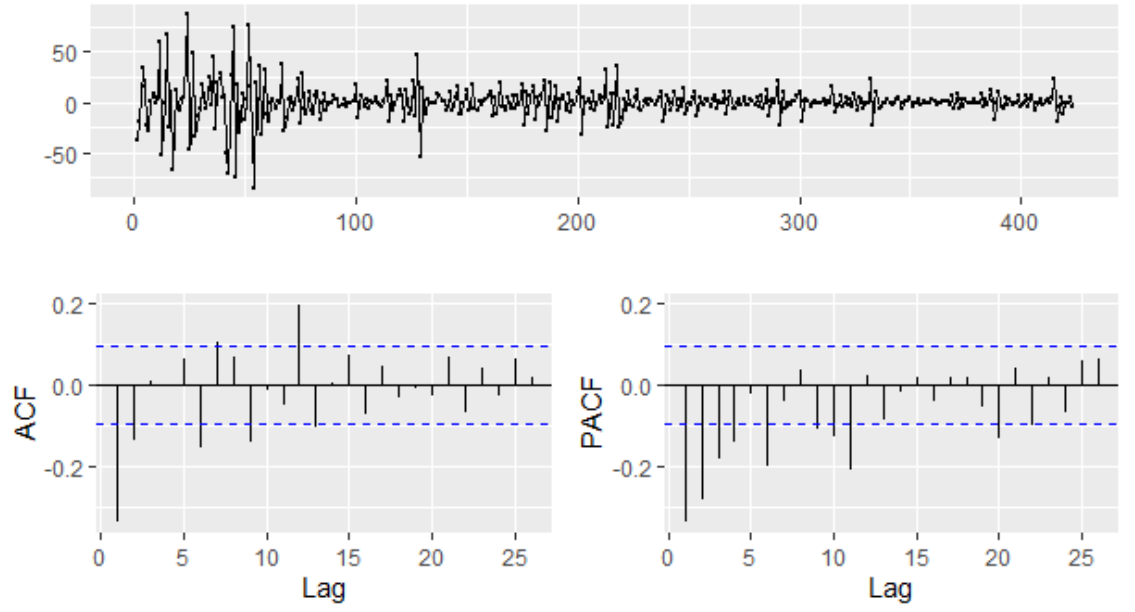
Figura 4.2.1 – Gráficos de Autocorrelação e Autocorrelação Parcial



A partir da função auto.arima, foram encontrados dois resultados, o modelo ARIMA(1,1,2) e o ARIMA(1,1,1), indicando que os mesmos seriam eficientes na modelagem e que os dados precisariam ser diferenciados uma vez em ambos os casos. O processo de diferenciação consiste na subtração entre os dados subsequentes, ou seja, a diferença entre os dados originais tornaram-se os novos dados.

O teste de Dickey-Fuller foi realizado para os dados diferenciados e constatou-se que a série possuía a propriedade da estacionariedade, o que facilita a análise para os modelos. Esta verificação indica uma série estacionária quando o p-valor é menor do que 0,05, e a série estudada apresentou o p-valor = 0,01. Os dados e os gráficos de correlação e correlação parcial ficaram no formato indicado pelo gráfico 4.2.2 após esta transformação:

Figura 4.2.2 – Intervalo de dias entre os ataques cibernéticos - 1º Diferença



Considerando os gráficos de ACF dos resíduos dos modelos nas figuras 4.2.3 e 4.2.4, onde quase todos os valores de "atrasos"estão contidos dentro dos limites e os resultados obtidos

no teste Ljung-Box (p-valor = 0,9234 e p-valor = 0,9514), é possível observar que os dois modelos testados estão ajustados pois ambos possuem resíduos que se comportam como ruído branco .

Figura 4.2.3 – Resíduo - ARIMA (1,1,2)

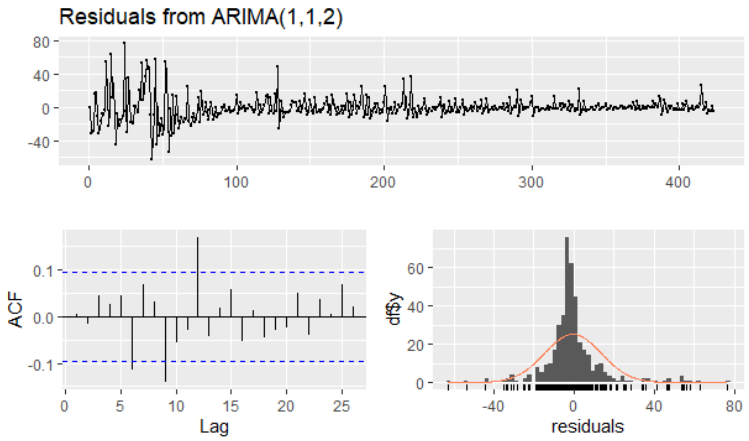
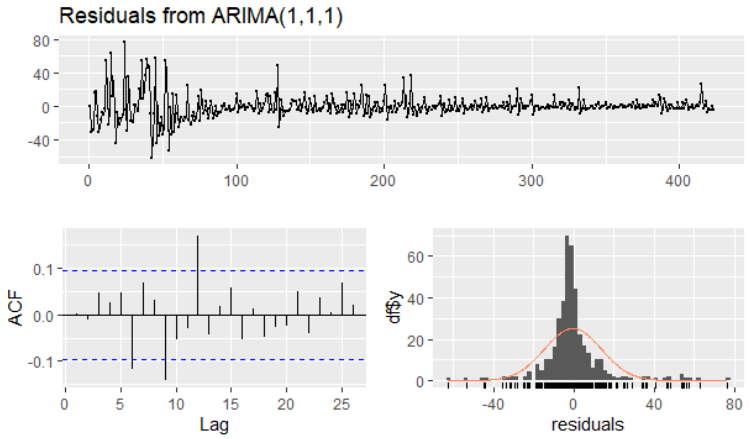


Figura 4.2.4 – Resíduo - ARIMA (1,1,1)

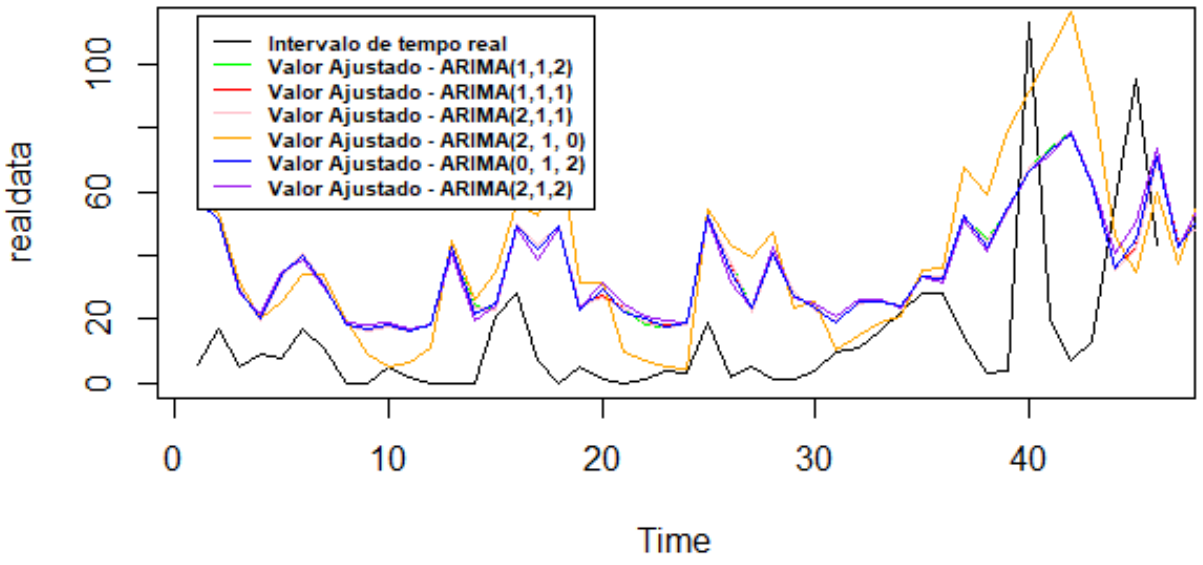


Além dos modelos retornados pela função auto.arima, outros também foram analisados. Foram eles: ARIMA(0,1,1), ARIMA(1,1,0), ARIMA(2,1,1), ARIMA(2,1,0), ARIMA(0,1,2), ARIMA(2,1,2). A partir do teste Ljung-Box, foi identificado que os dois primeiros não estavam ajustados para a previsão, então o processo foi continuado apenas para os outros modelos.

A Figura 4.2.5 apresenta a comparação entre os valores ajustados dos modelos ARIMA com os dados reais observados. Os resultados dos modelos ARIMA(1,1,2), ARIMA(1,1,1), ARIMA(2,1,1), ARIMA(2,1,0), ARIMA(0,1,2) e ARIMA(2,1,2) foram bastante semelhantes, o que acabou dificultando a visualização das retas, que quase não apresentam diferenças entre si. Já o modelo ARIMA(2,1,0) se distinguiu das restantes, traçando um caminho mais distinto dos dados reais.

Nota-se que, em todos os casos, o valor ajustado dos modelos apresenta bastante desconformidade em relação aos dados observados. Ao calcular o erro quadrático médio (EQM)

Figura 4.2.5 – Dados reais x Valor ajustado



das previsões, foi possível observar valores bastante elevados, principalmente para o modelo ARIMA(2,1,0), como é apresentado na Tabela 4.2.2.

Tabela 4.2.2 – Erro quadrático médio

	EQM
ARIMA(1,1,2)	492,3689
ARIMA(1,1,1)	491,9916
ARIMA(2,1,1)	492,2757
ARIMA(2,1,0)	1175,822
ARIMA(0,1,2)	488,1422
ARIMA(2,1,2)	491,4423

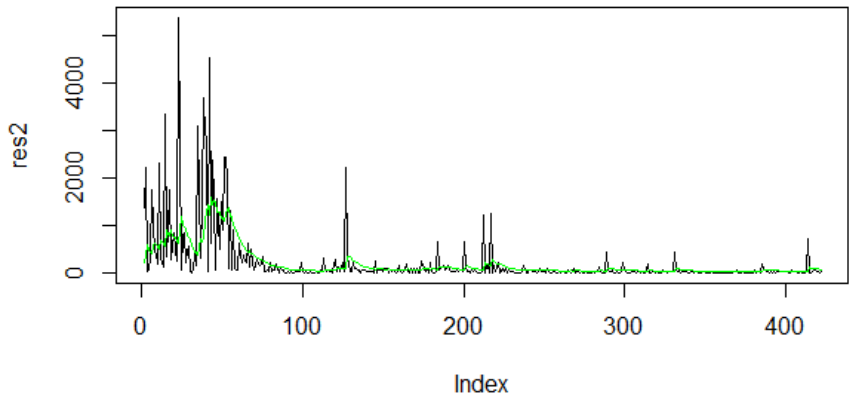
4.3 Modelo GARCH

Para conseguir iniciar a modelagem, alguns pacotes são necessários no R como, por exemplo, o "rugarch"que é utilizado na especificação e estimativa do modelo GARCH e através do qual as principais funções relacionadas a este modelo puderam ser acessadas.

Os dados, inicialmente, foram especificados e ajustados através das funções "ugarchspec"e "ugarchfit", respectivamente. A primeira é utilizada para especificar o tipo de modelo garch que será aplicado no modelo e a segunda para encontrar os melhores parâmetros para a estimação.

Através destas funções, é possível identificar diversos aspectos do modelo como, por exemplo, os resíduos, os coeficientes, a volatilidade, ect. Os resultados encontrados foram utilizados para criar o gráfico 4.3.1:

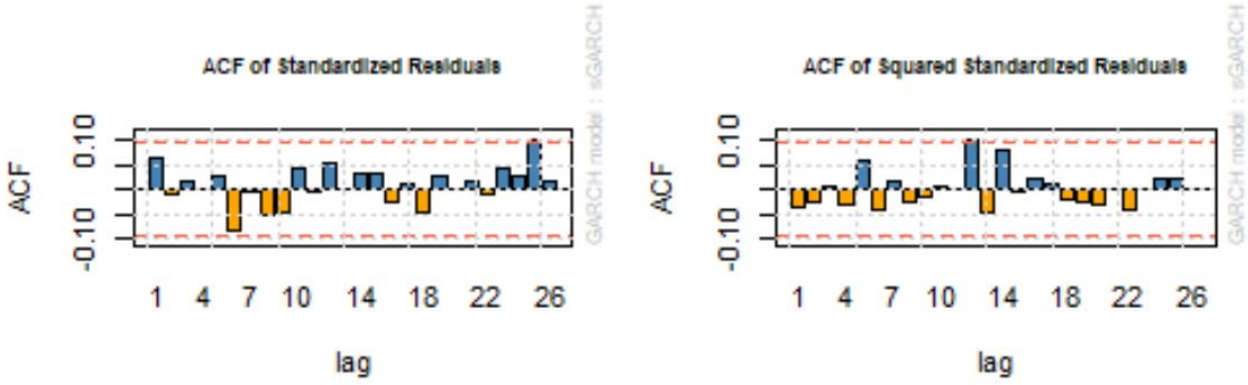
Figura 4.3.1 – Resíduo e Variância - GARCH(1,1)



Na Figura 4.3.1, é ilustrado o gráfico dos resíduos e da variância condicional dos dados, sendo o primeiro representado pela cor preta e o segundo pela cor verde. Através do estudo dessa volatilidade é que será possível realizar a estimação e as previsões sobre a variabilidade ao longo do tempo.

Observando o gráfico de correlação dos resíduos obtido na estimação do GARCH, apresentado na Figura 4.3.2 , nota-se que todas as barras estão dentro do intervalo, então é possível dizer que o modelo estimado está ajustado.

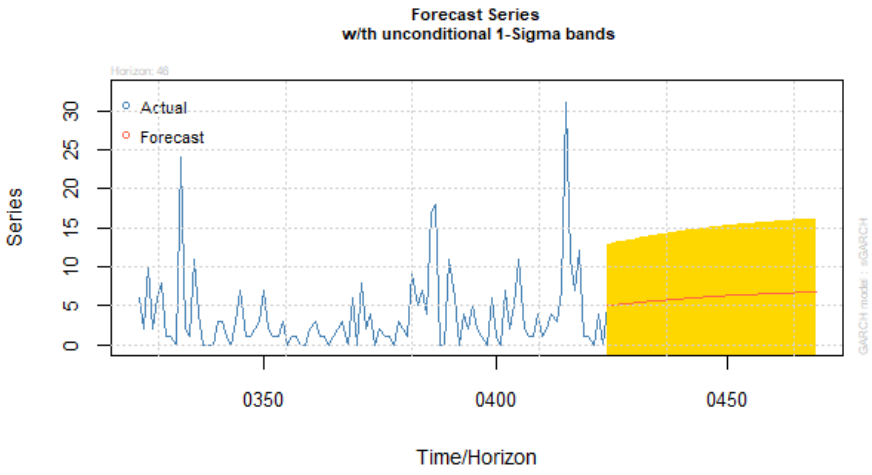
Figura 4.3.2 – ACF - erro e erro quadrático



A previsão, realizada através da função "ugarchforecast", resultou na Figura 4.3.3. Nela, o traçado azul são os dados observados, a reta vermelha representa os valores da previsão exibida pelo modelo e a área em amarelo é o intervalo estimado para a variabilidade dos dados. Foi

observado que quase todos os valores dos dados reais estão dentro deste intervalo fornecido, o que indica que a previsão está em conformidade com a realidade. Em relação ao erro médio quadrado calculado, o resultado foi igual a 567,8893, um valor ainda maior do que os que foram encontrados na modelagem do ARIMA.

Figura 4.3.3 – Previsão - GARCH(1,1)



4.4 Alisamento exponencial

Para a série estudada, foi testado o alisamento exponencial simples (SES) e também o de Holt, considerando a tendência para a modelagem e previsão dos dados. Na Tabela 4.4.3, estão os resultados encontrados para os dois modelos:

Tabela 4.4.3 – Valor inicial, parâmetros de suavização e valor p do teste de Ljung-Box para o modelo

Modelo	a_0	b_0	α	β	Ljung-Box (p)
SES	2,868087	-	0,2974452	-	0,0008133
Holt	0,6265246	-0,8234837	0,595843	0,1619757	0,3488

De acordo com o resultado do teste Ljung-Box, o modelo de alisamento exponencial simples não é adequado para previsão pois não possui resíduos descorrelacionados ao nível de significância de 5%.

A Figura 4.4.1 apresenta os dado reais e ajustados, em preto e vermelho, respectivamente, e o intervalo fornecido pelo modelo, traçado em azul. Nota-se que este intervalo se distancia demais tanto dos dados observados quanto dos valores gerados pelo próprio modelo.

Pode-se observar, na Figura 4.4.2, os dados observados e o valor ajustado do modelo de holt numa comparação mais próxima. É notável que o resultado não acompanha a variabilidade e se distancia bastante dos dados reais.

Figura 4.4.1 – Alisamento Exponencial Holt-Winters com Tendência - Ampliada

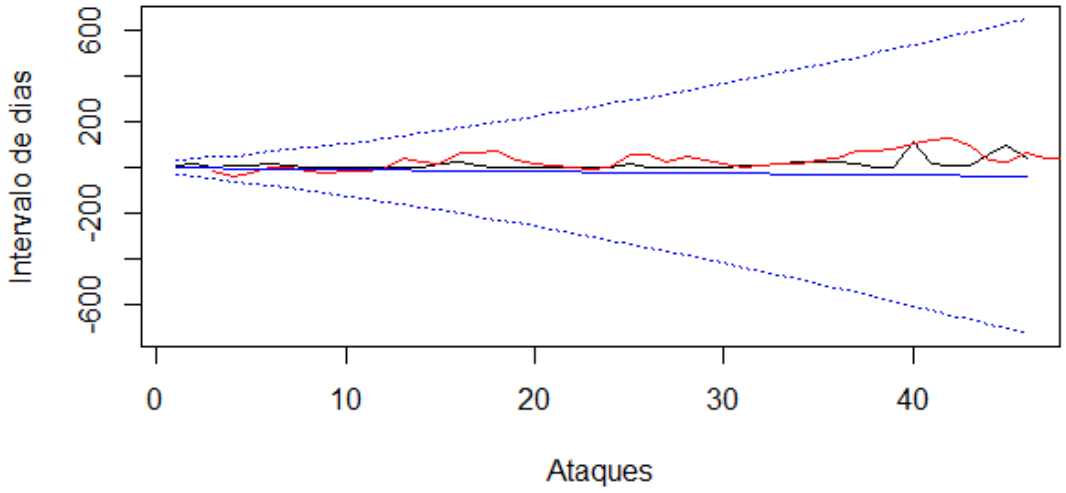
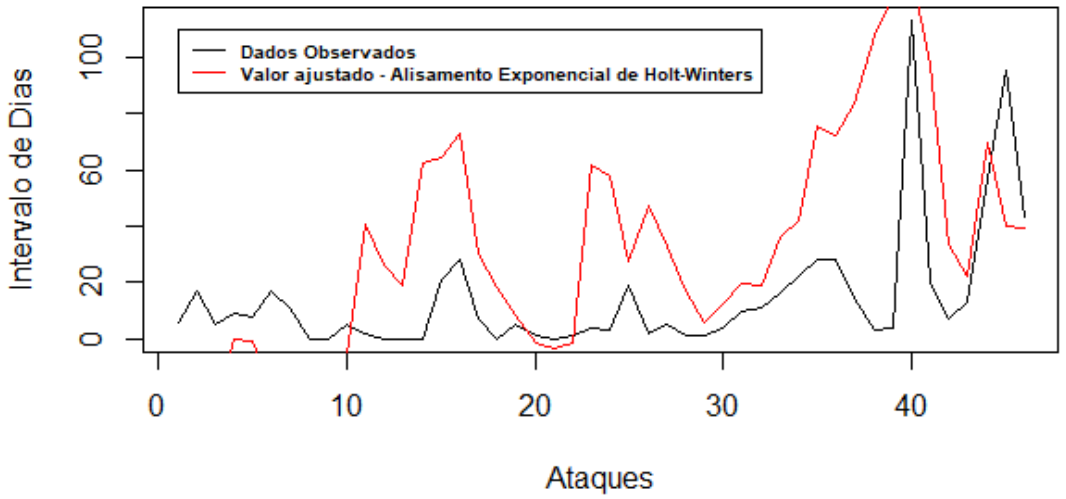


Figura 4.4.2 – Alisamento Exponencial Holt-Winters com Tendência



Ao calcular o erro quadrático médio, o valor encontrado foi bastante elevado, como era esperado, dado os gráficos de previsão. O resultado foi igual a 1958,1.

CAPÍTULO 5

CONSIDERAÇÕES FINAIS E SUGESTÕES PARA TRABALHOS FUTUROS

O presente trabalho teve como objetivo principal analisar os dados de ataques cibernéticos da base de dados do estado da Califórnia e realizar a modelagem e previsão de acontecimentos futuros, seguindo os modelos de séries temporais ARIMA, GARCH e alisamento exponencial.

Analisando a base escolhida foi observado que o tipo de organização mais tem sido alvo de ataques é o de negócios, identificado na base de dados como BSO (Business - others), pois não inclui os negócios financeiros, de seguros ou comércio. Diversas são as companias inseridas nesta categoria como, por exemplo, as de redes sociais, hóteis, serviços de vários tipos, etc. Foi analisado também que, esta mesma categoria de organizações, teve a maior quantidade de dados violados neste período, chegando a mais de 5 bilhões de dados.

Também foi observado que, nos anos iniciais, havia um intervalo de tempo bem maior entre os ataques do que nos anos finais, evidenciando que a frequência dos ataques se tornou maior com o passar dos anos. É muito comum, inclusive, que ocorram mais de um ataque no mesmo dia.

Quanto aos modelos, no ARIMA os valores ajustados relativos a estimação dos modelos testados apresentaram, em sua maioria, valores que acompanham a curva dos dados reais mas de maneira desnivelada. O modelo GARCH indica, além da previsão, um intervalo de confiança e, observando os dados separados para a comparação com a previsão, fica indiscutível que a maioria dos valores estão dentro deste intervalo, sugerindo assim, uma previsão de volatilidade em conformidade com a realidade.

Já o ajustamento exponencial, testado para o caso simples e com tendência, apresentou p-valor maior do que 0,05 apenas no segundo caso mas, comparando os dados com os valores estimados, foi indiscutível que não estavam em conformidade. Por outro lado, o valor ajustado

no caso simples conseguiu acompanhar melhor as curvas dos dados reais e apresentou limites de acordo com os extremos apresentados.

Por fim, comparando todos os resultados obtidos, o modelo considerados melhor ajustado aos dados observados foi o ARIMA (0,1,2) por apresentar a menor soma de erro quadrático médio das previsões (488,1422) entre os modelos analisados.

Como sugestão de trabalho futuro, sugere-se que um estudo nesse sentido seja realizado com dados para o Brasil. Nos últimos anos, de acordo com relatórios publicados, o Brasil foi um dos países que mais sofreu com esses ataques. Os casos têm aumentado consideravelmente e, por isso, seria interessante essa análise, visto que não existe nenhuma nessa direção. Uma outra sugestão é analisar o risco dos ataques através da volatilidade de forma mais aprofundada.

REFERÊNCIAS

AGRAFIOTIS, I. *et al.* A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. **Journal of Cybersecurity**, v. 4, n. 1, p. 1–15, 2018. ISSN 20572093. 4

BALTAR, B. D. P. Metodologia. 2009. 9

BONNER, L. Cyber Risk: How the 2011 Sony Data Breach and the Need for Cyber Risk Insurance Policies Should Direct the Federal Response to Rising Data Breaches. **Washington University Journal of Law & Policy**, v. 40, p. 257–277, 2012. Disponível em: <http://openscholarship.wustl.edu/cgi/viewcontent.cgi?article=1581{&}context=law{_}journal{_}l>. 4, 6

CHATFIELD, B. T. If we look at technology over very long timescales, our definition of what it is transforms, and as Tom Chatfield argues, it also displays a form of evolution entwined with our own. p. 1–9, 2021. 3

COHEN, R. D. *et al.* An investigation of cyber loss data and its links to operational risk. **Journal of Operational Risk**, v. 14, n. 3, p. 1–25, 2019. ISSN 17552710. 5

FELIPE, B.; LINS, E. A evolução da Internet : uma perspectiva histórica. **Caderno ASLEGIS**, v. 48, p. 11–45, 2013. 3

FINKELSTEIN, M. E.; FINKELSTEIN, C. PRIVACIDADE e LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS. **Revista de Direito Brasileira**, v. 23, n. 9, p. 284, 2020. ISSN 2237-583X. 3, 6, 7

GORDON, L. A. *et al.* A framework for using insurance for cyber-risk management. **Communications of the ACM**, v. 46, n. 3, p. 81–85, 2003. ISSN 00010782. 6

HERATH, H. S.; HERATH, T. C. Copula-based actuarial model for pricing cyber-insurance policies. **Workshop on the Economics of Information Security**, v. 2, n. 1, p. 7–20, 2011. Disponível em: <http://weis2007.econinfosec.org/papers/24.pdf>. 5

JEKLIN, A. æ¸ç¡No Title No Title No Title. n. July, p. 1–23, 2016. 10

KSHETRI, N. The evolution of cyber-insurance industry and market: An institutional analysis. **Telecommunications Policy**, Elsevier Ltd, v. 44, n. 8, sep 2020. ISSN 03085961. 5, 6

LESSIG, L. The Architecture of Privacy. In: **Taiwan Net 98 Conference**. [S.l.: s.n.], 1998. 5

MANWORREN, N. *et al.* Why you should care about the Target data breach. **Business Horizons**, "Kelley School of Business, Indiana University", v. 59, n. 3, p. 257–266, 2016. ISSN 00076813. Disponível em: <http://dx.doi.org/10.1016/j.bushor.2016.01.002>. 5, 6

MORETTIN, P.; TOLOI, C. **Analise de Séries temporais**. 1987. 538 p. 10, 11, 12

PANDA, S. *et al.* Cyber-Insurance: Past, Present and Future. 2021. 5

PENG, C. *et al.* Modeling multivariate cybersecurity risks. **Journal of Applied Statistics**, Taylor & Francis, v. 45, n. 15, p. 2718–2740, 2018. ISSN 13600532. Disponível em: <https://doi.org/02664763.2018.1436701>. 1, 2

PETERS, G. *et al.* Understanding Cyber-Risk and Cyber-Insurance. **SSRN Electronic Journal**, 2018. 1, 4

POPPENSIEKER, T.; RIEMENSCHNITTER, R. A new posture for cybersecurity in a networked world. **McKinsey&Company**, n. March, 2018. Disponível em: <https://www.mckinsey.com/business-functions/risk/our-insights/a-new-posture-for-cybersecurity-in-a-networked-world>. 5

SHIM, W. An Analysis of Information Security Management Strategies in the Presence of Interdependent Security Risk. **Asia Pacific Journal of Information Systems**, v. 22, n. 1, p. 79–101, 2012. 4

TALESH, S. A. Data Breach, Privacy, and Cyber Insurance: How Insurance Companies Act as âCompliance Managersâ for Businesses. **Law and Social Inquiry**, v. 43, n. 2, p. 417–440, 2018. ISSN 17474469. 3, 4, 5, 6

ZELLER, G.; SCHERER, M. **A comprehensive model for cyber risk based on marked point processes and its application to insurance**. Springer Berlin Heidelberg, 2021. ISSN 21909741. ISBN 0123456789. Disponível em: <https://doi.org/10.1007/s13385-021-00290-1>. 6

APÊNDICE - CÓDIGO DO R PARA A MODELAGEM

```
#####################################################################
#                       TCC: CRIME CIBERNÉTICO                      #
#                                                                   #
#                     11 de dezembro de 2021                        #
#                       Aluno: Lais Muniz                           #
#                  Orientação: Renata Alcoforado                    #
#                                                                   #
#####################################################################

###################### Carregando os Pacotes ######################

library(lubridate)
library(moments)
library(urca)
library(aTSa)
library(ggplot2)
library(lmtest)
library(stats)
library(forecast)
library(UnitCircle)
library(mvtnorm)
library(rugarch)
library(tseries)
library(dplyr)
```

```
######################## Leitura dos Dados ########################

##Análise dos dados

#Número de ataques em cada tipo de organização
freq = table(base$`Type of organization`)

#BSF BSO BSR EDU GOV MED NGO
#52 183  68  51  19  92   5

max(base$`Total Records`)

base %>%
group_by(base$`Type of organization`) %>%
summarise(sum(`Total Records`))
#Ou
coluna <- group_by(base, base$`Type of organization`)
summarise(coluna,sum(`Total Records`))

freq2 = table(base$City)
max(freq2)
#San Francisco - 58

#Processamento dos dados

## Coluna de datas dos incidentes colocada dentro da variável data
data = Base_base_hacking$`Date Made Public`
data[0]= data[1]

##Cálculo do intervalo de tempo entre os incidentes
inter = c()
for (i in 0:length(data)) {
inter[i] = as.numeric(data[i+1] - data[i])
}

## Série temporal e plot
##Transformação dos intervalos em série temporal e Separação dos
#dados da modelagem e da previsão(10% finais)
```

```
tempos = as.ts(inter[1:423])
real= as.ts(inter[424:469])
realdata = na.omit(real)
Dias = tempos
plot.ts(Dias, main = "Intervalo de tempo entre os ataques")

## Estatísticas dos dados
summary(tempos)
var(tempos)
sd(tempos)
skewness(tempos)
kurtosis(tempos)

#MODELO ARIMA

## Teste de Dickey-Fuller para verificar se a série é estacionária
#pelo p-value

## Série estacionária pois p-valor é menor do que 0,05
adf.test(tempos, nlag = 5)

## Correlogramas Fac e FACP
acf(tempos,lag = 40)
pacf(tempos,lag = 40)
ggtsdisplay(tempos)

## Diferenciação da série
dif1 = diff(tempos)
plot.ts(dif1, main="Intervalo de tempo entre os ataques-diferenciado")
adf.test(dif1)
acf(dif1, lag = 40, main = "FAC - Série diferenciada")
pacf(dif1, lag = 40, main = "FACP - Série diferenciada")
ggtsdisplay(dif1)
checkresiduals(dif1)

##Função que retorna parâmetros eficientes para a modelagem
fit <- auto.arima(tempos)
summary(fit) #ARIMA(1,1,2)
```

```
acf(tempos)
acf(diff(tempos))
auto.arima(tempos,ic="bic") #ARIMA(1,1,1)


#Modelagem
modelo2 <- Arima(tempos,order=c(1,1,2))
modelo2 #AIC =3441.95
modelo2.pred = predict(modelo2, 46)
pred2 = as.vector(modelo2.pred$se, mode="numeric")
Box.test(resid(modelo2), type=  "Ljung-Box")
checkresiduals(modelo2)

modelo6 <- Arima(tempos,order=c(1,1,1))
modelo6 #AIC =3439.97
modelo6.pred = predict(modelo6, 46)
pred6 = as.vector(modelo6.pred$se, mode="numeric")
Box.test(resid(modelo6), type=  "Ljung-Box")
checkresiduals(modelo6)

#Gráfico: real data x models
plot.ts(realdata)
lines(modelo1$fitted, col = "green")
lines(modelo6$fitted, col="red")
legend(1, 100, legend=c("Intervalo de tempo real", "Valor
Ajustado - ARIMA(1,1,2)", "Valor Ajustado - ARIMA(1,1,1)"),
col=c("black", "blue", "red", "yellow", "orange", "purple", "red"),
cex=0.6, text.font=2, lwd=1)

#ERRO QUADRÁTICO MÉDIO DAS PREVISÕES!!

#ARIMA 1,1,2
er2 = (realdata - pred2)^2
erro2 = sum(er2)/46
erro2  #492.3689

#ARIMA 1,1,1
er6 = (realdata - pred6)^2
erro6 = sum(er6)/46
```

```
erro6  #491.9916

#OUTROS MODELOS

##Modelo7
modelo7 <- Arima(tempos,order=c(0,1,1))
modelo7 #AIC = 3462.93
modelo7.pred = predict(modelo7, 46)
pred7 = as.vector(modelo7.pred$se, mode="numeric")
Box.test(resid(modelo7), type=  "Ljung-Box") #0.0005278
checkresiduals(modelo7)
#Ljung-box = 0.0005278

##Modelo8
modelo8 <- Arima(tempos,order=c(1,1,0))
modelo8 #AIC = 3526.73
modelo8.pred = predict(modelo8, 46)
pred8 = as.vector(modelo8.pred$se, mode="numeric")
Box.test(resid(modelo8), type=  "Ljung-Box") #0.04953
checkresiduals(modelo8)
#Ljung-box = 0.04953

##Modelo10
modelo10 <- Arima(tempos,order=c(2,1,1))
modelo10 #AIC = 3441.96
modelo10.pred = predict(modelo10, 46)
pred10 = as.vector(modelo10.pred$se, mode="numeric")
Box.test(resid(modelo10), type=  "Ljung-Box") #0.9346
checkresiduals(modelo10)
#Ljung-box = 0.9346

##Modelo11
modelo11 <- Arima(tempos,order=c(2,1,0))
modelo11 #AIC = 3491.29
modelo11.pred = predict(modelo11, 46)
pred11 = as.vector(modelo11.pred$se, mode="numeric")
Box.test(resid(modelo11), type=  "Ljung-Box") #0.3282
checkresiduals(modelo11)
#Ljung-box = 0.3282
```

```
##Modelo12
modelo12 <- Arima(tempos,order=c(0,1,2))
modelo12 #AIC = 3441.62
modelo12.pred = predict(modelo12, 46)
pred12 = as.vector(modelo12.pred$se, mode="numeric")
Box.test(resid(modelo12), type=  "Ljung-Box") #0.7458
checkresiduals(modelo12)
#Ljung-box = 0.7458

##Modelo13
modelo13 <- Arima(tempos,order=c(2,1,2))
modelo13 #AIC = 3443.02
modelo13.pred = predict(modelo13, 46)
pred13 = as.vector(modelo13.pred$se, mode="numeric")
Box.test(resid(modelo13), type=  "Ljung-Box") #0.9413
checkresiduals(modelo13)
#Ljung-box = 0.9413

#NOVO GRÁFICO

#Gráfico: real data x models
plot.ts(realdata)
lines(modelo2$fitted, col = "green")
lines(modelo6$fitted, col="red")
lines(modelo10$fitted, col="pink")
lines(modelo11$fitted, col="orange")
lines(modelo12$fitted, col="purple")
lines(modelo13$fitted, col="blue")
legend(1, 115, legend=c("Intervalo de tempo real", "Valor Ajustado -
ARIMA(1,1,2)", "Valor Ajustado - ARIMA(1,1,1)", "Valor Ajustado -
ARIMA(2,1,1)", "Valor Ajustado - ARIMA(2, 1, 0)", "Valor Ajustado -
ARIMA(0, 1, 2)", "Valor Ajustado - ARIMA(2,1,2)"),
col=c("black", "green", "red", "pink", "orange", "blue", "purple"),
cex=0.6, text.font=2, lwd=1)

#EQM

#ARIMA 2,1,1
```

```
er10 = (realdata - pred10)^2
erro10 = sum(er10)/46
erro10  #492.2757

#ARIMA 2,1,0
er11 = (realdata - pred11)^2
erro11 = sum(er11)/46
erro11  #1175.822

#ARIMA 0,1,2  ----> MELHOR AJUSTADO AOs DADOS!
er12 = (realdata - pred12)^2
erro12 = sum(er12)/46
erro12  #488.1422

#ARIMA 2,1,2
er13 = (realdata - pred13)^2
erro13 = sum(er13)/46
erro13  #491.4423


#MODELO GARCH

#Especificando e ajustando o modelo
mod.spec=ugarchspec()
mod.fit = ugarchfit(tempos, spec=mod.spec)
mod.fit

#Var e resíduo
var = mod.fit@fit$var
res2 = (mod.fit@fit$residuals)^2

#Plot dos resíduos e da volatilidade da série
plot(res2, type = "l")
lines(var, col = "red")

#Previsão da série e plot
predict = ugarchforecast(mod.fit, n.ahead = 46)
plot(predict)
```

```
#Análise de diagnóstico
plot(mod.fit,which="all")
summary(mod.fit)

#Dados reais x Ajustados
plot(realdata)
lines(mod.fit@fit@fitted.values)

#Resíduo e Volatilidade para os Últimos 20 ataques
var_t = c(tail(var, 20), rep(NA, 10))
res2_t = c(tail(res2, 20), rep(NA, 10))
plot(res2_t, type = "l")
lines(var_t, col = "red")

#Coluna de previsões
pred = predict@forecast$seriesFor

#Erro Quadrático Médio

#GARCH(1,1)
er.g = (realdata - pred)^2
er.g
erro.g = sum(er.g)/46
erro.g #567.8893



#ALISAMENTO EXPONENCIAL

#Alisamento Exponencial Simples
aes = HoltWinters(tempos, beta = FALSE, gamma=FALSE)
aes
Box.test(resid(aes), type=  "Ljung-Box")
aes.pred = predict(aes, n.ahead = 46, prediction.interval=TRUE)
aes.pred

#Alisamento Exponencial de Holt com tendência
aeh = HoltWinters(tempos, gamma=FALSE)
aeh
```

```
Box.test(resid(aeh), type=  "Ljung-Box")
aeh.pred = predict(aeh, n.ahead = 46, prediction.interval=TRUE)
aeh.pred

#aeh
plot.ts(realdata, ylim=c(min(aeh.pred[,3]), max(aeh.pred[,2])),
xlab="Ataques", ylab="Intervalo de dias")
lines((aeh$fitted[,1]), col="red")
points(1:46, aeh.pred[,1], type="l", col="blue")
points(1:46, aeh.pred[,2], col="blue", lty=3, type="l")
points(1:46, aeh.pred[,3], col="blue", lty=3, type="l")
legend(1, 110, legend=c("Dados Observados", "Valor ajustado -
Alisamento Exponencial Holt-Winters"),
col=c("black", "red"), cex=0.6, text.font=2, lwd=1)

#real data x aeh
plot.ts(realdata, xlab="Ataques", ylab="Intervalo de Dias")
#lines(aes$fitted[1:46], col="blue")
lines(aeh$fitted[1:46], col="red")
legend(1, 110, legend=c("Dados Observados", "Valor ajustado -
Alisamento Exponencial de Holt-Winters"),
col=c("black", "red"), cex=0.6, text.font=2, lwd=1)

#Erro quadrático médio

#Holt
holt = as.numeric(aeh.pred[ ,1])
er1.2 = (holt - realdata)^2
er1.2
erro1.2 = sum(er1.2)/46
erro1.2  #1958.1
```

A Figura 5.0.1 apresenta todas as saídas da análise de diagnóstico do modelo GARCH(1,1).

Figura 5.0.1 – Plot - GARCH (1,1)