Rev. Elet. Gestão e Serviços
V.14, n.1, Jan./Jun. 2023

ARTIGO ORIGINAL
ORIGINAL ARTICLE

PREVISÃO DE SÉRIES TEMPORAIS DE AÇÕES COM DADOS SINTÉTICOS GERADOS COM TIMEGAN

FORECASTING STOCK TIME SERIES WITH SYNTHETIC DATA GENERATED WITH TIMEGAN

Mauricio Cruz Vidotto[1]
Pedro Paulo Balbi[2]
Eli Hadad Junior[3]

Resumo

Prever o comportamento de ativos no mercado financeiro sempre foi um grande desafio devido à sua imprevisibilidade. Cada vez mais técnicas de aprendizado automático são desenvolvidas e utilizadas para esta finalidade. Neste estudo foi utilizada a causalidade de Granger, para verificar se entre duas séries de preços de ações, cointegradas, e, com relação de causalidade, um modelo gerativo pode aprender com os dados reais e produzir séries sintéticas que poderão ser utilizadas para previsão. Para tanto, utilizou-se a estrutura TimeGAN, que permite combinar a versatilidade das Redes Adversárias Ge- negativas não-supervisionadas com o controle sobre a dinâmica temporal condicional, proporcionada pelos modelos auto-regressivos supervisionados. Foi observado que o volume de dados utilizado impacta no tempo de treinamento e o modelo aprende a dinâmica temporal de dados das séries, sendo capaz de gerar dados sintéticos de aparência realística e capturando as tendências de curto prazo. Os dados foram avaliados utilizando-se as técnicas PCA e t-SNE, e a precisão, utilizando o erro percentual médio absoluto. Os resultados obtidos demonstraram que a previsão de valores futuros a partir de séries de dados sintéticas éuma abordagem possível. Os valores sintéticos previstos a partir da séries sintéticas apresentaram um erro percentual médio absoluto próximo ao das previsões utilizando as séries reais confirmando que a TimeGAN aprendeu a dinâmica temporal das séries de dados.

Palavras-chave: Estrutura TimeGAN; técnicas PCA e t-SNE; séries de dados sintéticas.

Abstract

The financial market is notoriously volatile, making it difficult to ac- curately predict asset behavior. To combat this challenge, machine learning techniques are becoming increasingly prevalent. Our study utilized Granger causality to determine if a generative model could learn from actual data to produce synthetic series for use in predictions. Specifically, we investigated the causal relationship between two cointegrated stock price series using the Time-GAN structure, which combines unsupervised Generative Adversarial Networks with supervised auto-regressive models to control conditional temporal dyna- mics. We found that training time was impacted by the volume of data used, and that the model successfully learned the temporal dynamics of the data series. The synthetic data produced closely resembled real data and captured short- term trends, as confirmed by evaluations using PCA and t-SNE techniques and mean absolute percentage error measurements. Our results demonstrate that synthetic data series can be a viable approach for predicting future values, with TimeGAN effectively learning the temporal dynamics of the data series.

Key-words: Time-GAN structure; PCA and t-SNE techniques; synthetic data series.

[1]Pós-Graduação em Engenharia Elétrica e Computação-Universidade Presbiteriana Mackenzie (UPM). E-mail: 72100222@makenzista.com.br

[2]Professor do Programa de Pós-Graduação em Engenharia Elétrica e Computação, da Universidade Presbiteriana Mackenzie, São Paulo. Professor convidado do Programa de Doutorado em Engenharia de Sistemas Complexos, Universidade Adolfo Ibáez, Santiago, Chile. DPhil em Ciências Cognitivas pela Universidade de Sussex, Inglaterra. E-mail: pedrob@mackenzie.br

[3]Pós-Doutorado em Economia - Sistemas de Projeção de taxas de câmbio - FGV - Fundação Getúlio Vargas (CAPES 7) - Em andamento Doutorado em Adm. de Empresas - Finanças Estratégicas - Universidade Presbiteriana Mackenzie. E-mail: eli.hadad@mackenzie.br

Artigo recebido em: 10 de outubro de 2023. Artigo aceito em 17 de outubro de 2023.

INTRODUÇÃO

Prever o comportamento de ações no mercado financeiro é um desejo tão antigo quanto a existência do próprio mercado. Investidores utilizam-se de diversas ferramentas para auxiliarem o processo de tomada de decisões de compra e venda de ativos financeiros, como a análise fundamentalista e a análise técnica. A finalidade é entender como está a saúde financeira de uma companhia, detectar as mudanças estruturais sobre oferta e demanda e as tendências de mercado. No mercado financeiro, mesmo as pequenas flutuações de índices podem levar a grandes variações de preços nos ativos. A possibilidade de anteciparmos estes movimentos para tomada de decisões futuras nos permitiria operar de forma mais segura nas bolsas de valores. Fama (1965) apresentou um estudo sobre como as cotações das ações respondem a um evento. Destacou o fato de que é muito difícil prever os movimentos dos preços dos ativos no curto prazo porque os mercados incorporam rapidamente qualquer informação relevante sobre os preços. Entretanto, alguns críticos de seu trabalho acreditam que com modelos adequados é possível antecipar estes movimentos. Neste contexto, o uso de técnicas de aprendizagem de máquina podem fornecer *insights* relevantes, complementando os métodos tradicionais de pesquisa.

Este artigo demonstra o uso da estrutura TimeGAN (Yoon et al. 2019) para geração de dados temporais sintéticos realistas e utiliza-los para previsão de preços futuros. O artigo está organizado da seguinte forma na seção2 tem-se a revisão sistemática, na seção 3 o referencial teórico, na seção 4 a estrutura TimeGAN, na seção 5 a metodologia adotada, na seção 6 os resultados e a conclusão na seção 7.

2. REVISÃO SISTEMATICA

Como forma de analisar a contribuição deste artigo para preencher lacunas de conhecimentos, ainda não exploradas por estudos anteriores, realizou-se a análise bi-biométrica e a revisão sistemática da literatura, no período de 1 de janeiro de 1945 a 22 de agosto de 2023, com amostra final de 23 artigos. A análise bibliométrica refere-se à análise quantitativa, que se desenvolve em meio à contagem de frequências e cocitaçõ es. A revisão sistemática uma análise qualitativa, que considera a correlação entre temas mais significativos, mas ainda pouco estudados pela academia.

O estoque de pesquisas analisadas é originário das base de dados *Web of Science (WoS)*, *Science Direct*, *Web of Science*, *Taylor & Francis* e *Google Scholar* e, tanto a análise bibliométrica quanto a revisão sistemática, utilizaram os softwares *R* (linguagem de programação estatística e gráfica), *RStudio* (interface de programação R), *Biblioshiny*

(análises de mapeamento científico) e *VOSViewer* (construção e visualização de redes bibliométricas).

A revisão da literatura foi realizada com a identificação de métodos para a previsão do preço de ações com uso de redes adversariais generativas (*GANs*).

A pesquisa foi realizada em 7 etapas descritas a seguir. As etapas 1 a 5 atendem tanto a` metodologia da análise bibliométrica quanto da revisão sistemática, e as etapas 6 e 7 referem-se a` revisão sistemática.

Etapa 1: Escolha da base de dados. Os artigos da amostra são obtidos a partir da *WoS*, a base de dados de citações líder no mundo, que indexa os periódicos mais citados em suas respectivas áreas, possuindo atualmente mais de 9.000 periódicos indexados.

Etapa 2: Uso de parâmetros iniciais de pesquisa a partir da *WoS*. Para o período de 1º de janeiro de 1945 a 23 de agosto de 2023. Inicialmente, são identificados 92 artigos com base em variações das palavras-chave *generative models AND stock*. Na sequência, são realizadas exclusões por meio da aplicação de filtros na própria *WoS*, resultando numa amostra intermediária de 45 artigos, conforme Tabela 1.

Sinal	Descrição	Número de artigos
(+)	Palavras-chave iguais a: "*generative* models AND stock"	92
(-)	Tipo de documento: diferente de "*articles*"	30
(-)	Idioma: diferente de "*english*"	0
(-)	Categorias da *WoS* diferentes de "Computer Science Information Systems", "Com- puter Science Artificial Intelligence", "Engineering Electrical Eletronic", "Com- puter Science Software Engineering", "Computer Science Interdisciplinary Appli- cations", "Computer Science Theory Methods", "Management", "Operations Re- search Management Science", "Business", "Business Finance", "Mathematics In- terdisciplinary Applications", "Engineering Multidisciplinary", "Mathematics Ap- plied", "Statistics Probability"	17
(=)	Amostra intermediaria	45
(-)	Exclusão de artigos não relacionados	22
(=)	Amostra final	23

Tabela 1. Evolucão da amostra por meio de filtros da WoS

Etapa 3: Exclusão de artigos não relacionados. Após a leitura inicial do resumo, introdução e conclusão dos artigos – a fim de verificar se eles estão de acordo com o tema definido - são excluídos 22 dos 45 artigos da amostra intermediária. Os motivos para essas exclusões, são não estarem relacionados a redes adversariais generativas (*GANs*), mas especificamente a outros tipos de modelos generativos (10), análise de sentimentos (2), gestão moderna do risco financeiro (2), publicidade *online* (2), impressão 3D (1), indústria automobilística (1), manipulação do preço de ações (1), reconhecimento de superfícies (1), recursos marítimos (1) e teoria dos jogos (1).

Etapa 4: Criação do banco de dados e coleta de artigos. Os 23 artigos da amostra final são obtidos por meio das seguintes bases de pesquisa acadêmica: *Science Direct*, *Web of Science*, *Taylor & Francis* e *Google Scholar*. A partir disso, são coletadas as seguintes informações para a captura dos dados gerais do artigo: ttítulo, nome do autor, instituição afiliada e país de origem dos autores/pesquisadores, nome do periódico, volume e número da edição, página inicial e final, ano da publicação, país de origem dos dados e número de anos de dados da amostra, palavras-chave, Identificador Digital de Objetos (DOI), *Journal of Economic Literature* (JEL) e número de citações de artigos na base de dados *WoS*.

Etapa 5: Análise bibliométrica. Por meio dos softwares *R*, *RStudio*, *Biblioshiny* e *VOSviewer*, são analisados os dados dos artigos para a elaboração e a análise das tabelas e mapas de relacionamento / co-citação.

Etapa 6: Leitura e codificação dos artigos. Trata-se da identificação dos objetivos, amostra, métodos e contribuições dos artigos. Além disso, eles são classificados e codificados em categorias e subcategorias estruturadas.

Etapa 7: Revisão sistemática. Após a codificação da matriz de (sub) categorização a partir dos 23 artigos da amostra final, érealizada a contagem de frequência das subcategorias, de maneira a facilitar a identificação das lacunas de conhecimento. A partir disso, essas lacunas são comparadas com as subcategorias, avenidas para estudos futuros, a fim de se obter hiatos passíveis de novos estudos.

De acordo com a figura 1, a amostra final écomposta por 23 artigos, distribuídos entre os anos de 2018 e 2023, obtidos a partir da base *WoS*, sendo possível notar um grande

aumento do interesse por parte dos pesquisadores nesse período de cinco anos, acentuadamente a partir de 2021 onde éidentificada uma média de aproximadamente sete publicações de artigos por ano, que tratam especificamente sobre tema *generative adversar ial network* destinada a previsões.

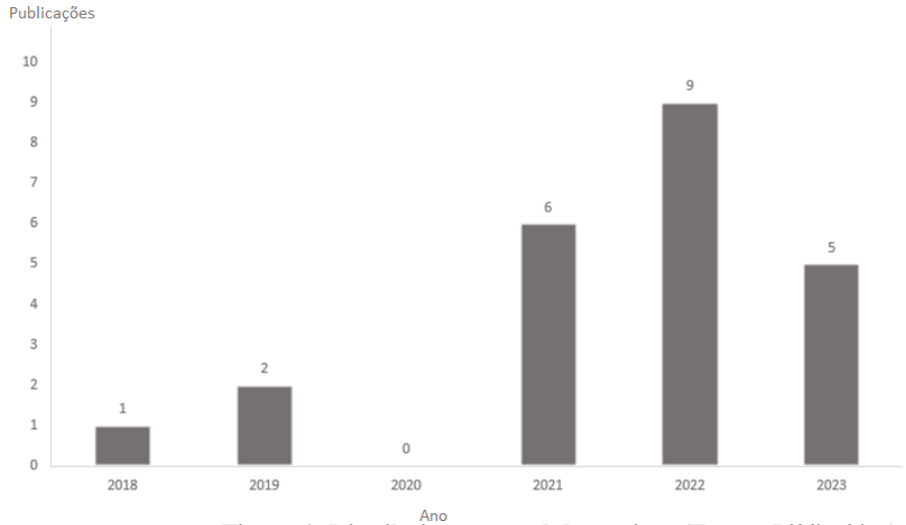

Figura 1. Distribuic¸ã o anual dos artigos (Fonte: *Biblioshiny*)

A figura 2 apresenta o mapa de concorrências das palavras-chave mais utilizadas nos artigos da amostra final. Destacam-se as palavras generative adversarial network, artificial intelligence, generative model, machine learning e deep learning. O tamanho dos nós indica maior relevância dessas palavras nos artigos e a espessura das linhas indica a força da ligação entre elas, enquanto as cores destacam os grupos.

Considerando-se a escassez de pesquisas sobre o tema, este estudo realiza uma análise bibliométrica e a revisão sistemática da literatura, a respeito da previsão do preço de ações com uso do modelo *generative adversarial network (GAN)*. Como resultado, há a identificação de lacunas de conhecimento a serem preenchidas por uma proposta de futura agenda de pesquisas relacionadas a esse tema.

A amostra inicial écomposta por 92 artigos, que após a adoção dos critérios de exclusão mencionados acima, éreduzida para 23 estudos finais, obtidos na base de da- dos *WoS*. A análise bibliométrica apresenta dados quantitativos por meio de gráficos e mapas de relacionamento. Essas verificações ocorrem através dos softwares *RStudio*, *Bi- blioshiny* e *VOSviewer*. Por sua vez, a revisão sistemática identifica a frequência das (sub) categorias

definidas para a amostra final. Sua verificação permite a percepção de quais combinações de subcategorias são viáveis para investigações futuras.

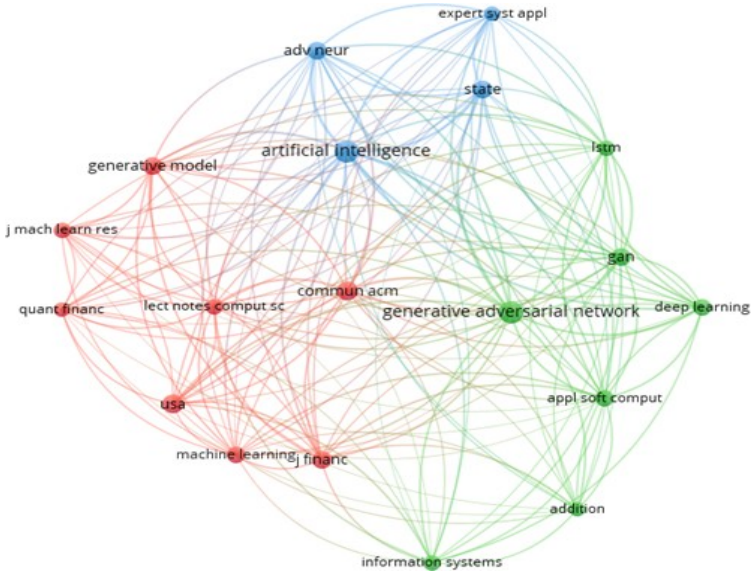

Figura 2. Mapa de coocorrências das palavras-chave (Fonte: *VosViewer*)

Como épossível depreender através dos resultados obtidos através da análise bibliomémtrica e revisão sistemática da literatura, este artigo traz resultados pioneiros ao apresentar o conceito de causalidade preditiva proposto por Granger com objetivo de analisar por meio de duas séries de tempo que apresentam correlação e causalidade se épossível modelar a rede adversarial generativa para aprender o comportamento de preços reais e dessa maneira gerar dados sintéticos que poderão ser utilizados para previsões.

Além disso, o artigo traz como inovação o uso de dados em alta frequência do mercado brasileiro, utilizando séries de dados dos ativos Indice Bovespa (IBOV) publicados a cada 30 segundos, gerando um total de 517.909 observações e o ativo Petrobrás (PETR3), que são publicados a cada negócio registrado, resultando em um total de 20.503.734 observações. Quantidades volumosas de dados observadas como essas, permitem análises que garantem nníveis mais altos de significância estatística.

3. REFERENCIAL TEÓRICO

As Redes Generativas Adversárias (GAN) foram citadas pela primeira vez por (Goodfellow *et al*., 2014) quando apresentou esta nova abordagem utilizando técnicas de

aprendizado profundo para gerar imagens sintéticas. Em sua forma mais simples, ela écomposta por duas redes neurais, uma geradora e uma discriminadora, que competem entre si e são capazes de analisar, capturar e copiar as variações de um conjunto de da- dos. O objetivo do gerador éenganar o discriminador gerando dados sintéticos realistas que tenham a mesma distribuição dos dados de treinamento. O gerador será treinado para gerar amostras falsas com o objetivo de maximizar a probabilidade de o discriminador cometer um erro. O discriminador por sua vez tenta distinguir entre as amostras falsas e verdadeiras. Dessa forma, a operação em conjunto tem um enorme potencial, pois pode levar ao aprendizado de como se imitar qualquer distribuição de dados. Segundo (Goodfellow *et al*., 2014), com tempo e recursos computacionais suficientes, espera-se que o discriminador e gerador entrem em equilíbrio.

Desde então, muitos avanços ocorreram na sua utilização. Sabe-se que as redes GAN são muito boas para o processamento de imagens, sendo possível avaliar a qualidade das imagens geradas visualmente. Entretanto, a avaliação de informações que não são imagens ainda éum desafio não solucionado.

Um dos benefícios das redes GAN é a geração de dados sintéticos realísticos que seguem a distribuição da amostra de dados e mantém as suas características intrínsecas (Lee; Seok 2021). Estes dados sintéticos podem ser utilizados para treinar novos modelos com um conjunto de dados inicial mínimo, o que elimina parte do viés de treinamento em um cenário com quantidade limitada de dados. Assim, os dados sintéticos surgem como uma forma de solucionar desafios em diversas áreas ajudando a impulsionar projetos inovadores.

A pesquisa de (Esteban et al. 2017) propôs o uso de uma GAN Recorrente (RGAN) e uma GAN Condicional Recorrente (RCGAN) para produzir séries temporais multidimensionais com valores realistas para aplicação a dados médicos. Na área de energia, (Fekri *et al*., 2019) utilizou uma RGAN para gerar dados de consumo de energia realistas aprendendo com dados reais.

No contexto de previsão de preços, muitos estudos utilizam os modelos GAN para aumentar o número de amostras a serem utilizadas no treinamento do modelo para melhoraro desempenho e reduzir a aleatoriedade no preço das ações mas, acabam tendo que o problema de *overfitting* durante o treinamento (Xu *et al*., 2022;Wu et al., 2022).

Um dos motivos pelo qual as redes generativas não são muito utilizadas para predição se deve a` dificuldade de se configurar adequadamente o conjunto de hiper-parâmetros

(Polamuri et al., 2022). Uma alternativa para resolução deste problema foi proposta por (He;Kita, 2021) que utilizaram um algorítimo genético para determinar os hiper-parâmetros adequados e tamanho do conjunto de dados de entrada para seu modelo GAN.

Segundo (Liu et al., 2022), o aprendizado adversário écomumente utilizado para previsão de quedas e aumentos de preço, entretanto, não existem muitos estudos sobre previsão de tendências de preços para séries temporais.

Em alguns casos de previsão de preços de ações, a rede adversária éutilizada para imitar o processo de aprendizagem de um operador que, progressivamente aprende a fazer previsões melhores a partir dos eventos reais (Staffini, 2022) para prever a evolução do preço. Aumentar o número de amostras também éuma abordagem adotada por alguns pesquisadores para melhorar a precisão do modelo de previsão (Abraham, 2021).

Entre as diversas abordagem utilizadas para previsão temos as pesquisas de (Zhou *et al*., 2018;Kumar *et al*., 2021) que utilizaram redes neurais do tipo *Long Short- Term Memory* (LSTM) e *Convolution Neural Networks* (CNN) para treinamento adversário para previsão do mercado de ações de alta frequência e, (Lin *et al*., 2022) que utilizou uma abordagem similar para previsão de risco de crédito. O modelo de (Asgarian *et al*., 2023) utilizou como gerador redes LSTM e como discriminadores redes CNN para previsão do preço de fechamento de ações e, (Staffini, 2022) propôs o uso de uma Rede Adversarial Generativa Convolucional Profunda (DCGAN) para prever a evolução do preço das ações utilizando dados históricos de transações e um modelo LSTM-GAN para prever movimentos de alta e baixa de ações atingindo precisão de 60%.

Em (Yoon *et al*., 2019) uma promissora estrutura foi proposta para lidar com séries de tempo denominada *Time-Series Generative Adversarial Network*, ou simplesmente *TimeGAN*. Esta estrutura tem como objetivo contabilizar correlações temporais combinando treinamento supervisionado e não-supervisionado. Para que isto ocorra, o *TimeGAN* incorpora a natureza autorregressiva das séries temporais, combinando a perda adversária não-supervisionada em sequências reais e sintéticas sincronizadas por meio de uma perda supervisionada passo-a-passo em relação aos dados originais. O objetivo érecompensar o modelo por aprender a distribuição sobre as transições de um ponto no tempo, para o próximo valor nos dados históricos.

De acordo com (Plesner *et al*., 2019), o *TimeGAN* éum dos primeiros modelos utilizando redes neurais adversárias que produzem dados sintéticos de séries temporais de

alta confiança porque considera a dependência do tempo nos dados de treinamento. Relembrando (Goodfellow *et al*., 2014), a GAN foi proposta para gerar dados a partir de uma distribuição alvo. Desta forma, se elas podem gerar dados de mercado realistas, eles podem ser utilizados por modelos de aprendizagem de máquina com o intuito de prever um comportamento futuro.

A pesquisa de (Zhang *et al*., 2022) utilizou-se do *TimeGAN* para aumentar os da- dos através da geração de dados sintéticos para melhorar a precisão de um modelo de previsão de aquecimento em subestações de energia elétrica, e (Li *et al*., 2022) apresentou um modelo híbrido denominado *HTSCG* utilizando *TimeGAN* com o objetivo de expandir históricos de energia fotovoltaica para melhorar a precisão da previsão e, assim, melhorar a qualidade de geração de energia para dias nublados e chuvosos.

Para prever falhas de disco de armazenamento de dados e reduzir custos em data centers, (Hai *et al*., 2022) utilizaram uma rede neural *GRU*, capaz de se adaptar ao impacto de longas sequências de dados, e uma rede adversária *TimeGAN*, capaz de lidar com o desequilíbrio de dados. Experimentalmente, obtiveram uma taxa média de detecção de falha de 95% e uma taxa de alarmes falsos de 0,2% na previsão das falhas.

Uma rede *TimeGAN* também foi utilizada por (Shangguan *et al*., 2023) para aprender a distribuição dos dados e gerar dados sintéticos para prever o desgaste do diâmetro da roda de trens devido a` fadiga de materiais. Assegurar que os trens operem sem falhas éde suma importância para a segurança de todo o sistema ferroviário. Os resultados experimentais obtidos demonstraram que as amostras geradas estavam bem próximas da distribuição real e apresentavam baixos erros de predição de degradação. Quando um estudo é feito, o objetivo é desenvolver uma fórmula para fazer previsões sobre a variável dependente, com base nos valores observados das variáveis independentes. Em uma análise causal, as variáveis independentes são consideradas como causa da variável dependente.

Por este motivo, em pesquisas é comum nos depararmos com a frase “correlação não implica causalidade”. A correlação éum indicador estatístico da relação entrevariáveis e descreve uma associação entre tipos de variáveis. Já a causalidade significa que mudanças em uma variável provocam mudanças na outra e existe uma relação de causa e efeito entre elas. Assim, podemos afirmar que ”correlação não implica causalidade, mas causalidade sempre implica correlação”.

O mesmo conceito éutilizado neste estudo para avaliar a utilidade de uma série para previsão de outra quando existe uma relação de causalidade entre elas, ou seja, as séries se cointegram. (Granger, 1969) propôs uma definição estatística de causalidade entre processos estocásticos, baseados na chamada cointegração para diferenciar e combinar análise das flutuações de curto prazo e tendências de longo prazo. A ideia principal édeterminar se uma série temporal éum fator e oferece informações úteis na previsão de outra série temporal.

A ideia não éalgo novo e, a pesquisa de (Apte *et al*., 2021) utilizou-se do conceito de causalidade preditiva proposto por Granger (1969) para detecção de anomalias. Os autores formularam a ideia de séries temporais causalmente anômalas e propuseram algorítmicos para identificá-las em um determinado conjunto de dados utilizando a causalidade, observando as relações causais entre as séries e identificando a série como anômala quando essa causalidade é violada. No estudo avaliaram anomalias em 3 conjuntos de dados que continham respectivamente temperatura de motores elétricos, indicadores de desenvolvimento mundial e dados de pessoas com doença de Parkinson.

Assim, duas séries de ativos serão escolhidas considerando que existe uma relação de causalidade serão selecionados e, uma vez constatada a existência de causalidade entre elas, as duas séries serão utilizadas em conjunto com uma estrutura *TimeGAN* que gerara séries sintéticas para os ativos que sejam úteis para aplicação em um modelo para previsão de preços.

1. A ESTRUTURA TIMEGAN

Como proposto por (Yoon *et al*., 2019), a estrutura *TimeGAN* combina uma rede adversária com um autoencoder (figura 3). Ela é composta por quatro componentes: uma função de incorporação, uma função de recuperação, um gerador de sequência e um discriminador de sequência. Os dois primeiros componentes fazem parte do autoencoder e são responsáveis por permitir reconstruções precisas dos dados originais e suas representações latentes.

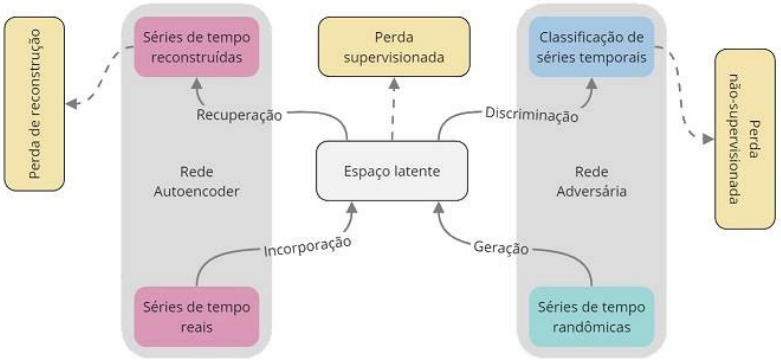

Figura 3. Componentes da estrutura TimeGAN

No autoencoder, a rede de incorporação éresponsável por reduzir a dimensionalidade do espaço de aprendizagem adversário e fornecer o espaço latente. A rede adversária opera dentro deste espaço e a dinâmica latente de dados reais e sintéticos será sincronizada por uma perda supervisionada, introduzida pelo treinamento passo-a-passo utilizando os dados originais. Dessa forma encoraja-se o modelo a capturar as distribuições condicionais nos dados. Como consequência, o modelo se beneficia, a` medida em que extrai mais informações dos dados de treinamento. Desta forma, ao invés de ter de despender esforço sobre, se cada dado éreal ou sintético, o modelo pode aprender com a dinâmica de transição das sequências reais.

Uma rede supervisora capturara a distribuição condicional de tempo dentro dos dados usando os dados originais como uma supervisão. A perda supervisionada éminimizada pelo treinamento conjunto das redes de incorporação e geração, de forma que o espaço latente não sirva apenas para promover a eficiência dos parâmetros mas também esteja condicionado a facilitar o aprendizado das relações temporais do gerador, guiando o aprendizado adversário. Resumidamente, usando um espaço de aprendizado incorporado, o *TimeGAN* estimula a rede a imitar a estrutura dos dados de treinamento.

Com o treinamento conjunto dos componentes da rede autoencoder e da rede adversária o modelo aprende simultaneamente a codificar recursos, gerar representações e iterar ao longo do tempo.

2. METODOLOGIA E DADOS

Após a observação de comportamento de alguns ativos, dois foram selecionados considerando a similaridade de comportamento entre eles. Foram utilizadas séries de dados dos ativos Indice Bovespa (IBOV) e Petrobrás (PETR3) obtidas do processo de difusão de dados de mercado da B3 denominado UMDF (do acrônimo em inglês Unified Market Data Feed) para o período compreendido entre 08/04/2019 e 30/12/2021.

A plataforma UMDF fornece informações de eventos que ocorrem para todos os ativos negociados na B3, como o livro de ofertas e suas atualizações, negócios realizados, estados de instrumentos, dados estatísticos e serviços para sincronização e recuperação de mensagens. A plataforma ébaseada no protocolo FIX (do acrônimo em inglês *Finan- cial Information eXchange*) com compressão FAST (do acrônimo em inglês *FIX Adapted for STreaming*) para otimização do consumo de banda, e os dados são publicados via multicast UDP, ou seja, o dado étransmitido ao mesmo tempo para todos os destinatários.

A dinâmica de publicação de dados édiferente para cada ativo pois esta relacionada- a intensidade de negociação de cada um deles. Considerando apenas as séries utilizadas, os preços de IBOV são publicados a cada 30 segundos e de PETR3 a cada negócio realizado. Para o período analisado foram capturados 517,909 e 20,503,734 objetos-, respectivamente.

Foi utilizado o Modelo Vetorial Autorregressivo (VAR) conforme estabelecido no artigo seminal de (Sims 1980). No VAR, o valor-p éusado para testar se uma variável defasada tem um efeito significativo em outra variável no sistema. Se o valor-p for menor ou igual a um nível de significância escolhido, geralmente 0.05, rejeitamos a hipótese nula de que não há efeito e concluímos que há evidência de uma relação causal entre as duas variáveis.

O valor-p foi introduzido por (Fisher, 1922) para dar alguma formalidade a` análise dos dados coletados em suas investigações científicas. (Neyman; Pearson, 1933) posteriormente modificaram para um procedimento pelo qual um desvio observado com valor-p menor que 5%, levaria a` rejeição da hipótese, com aceitação em contrário. Assim, tradicionalmente, o valor de corte para rejeitar a hipótese nula éde 0.05, o que significa que, quando não há nenhuma diferença, um valor tão extremo para a estatística de teste éesperado em menos de 5% das vezes.

O ambiente de teste foi montado no Google Colab (Google 2018) utilizando uma máquina virtual com 12 processadores Intel(R) Xeon(R) CPU @ 2.20GHz, 83.5 GB de RAM, 40.0 GB de GPU (NVIDIA A100) e 166.8 GB de disco.

O estudo foi realizado nas etapas a seguir: preparação dos dados, teste de estacionariedade, teste de cointegração, estimação do VAR, teste de causalidade, treinamento e geração da série temporal sintética e avaliação da previsão.

2.1. PREPARAÇÃO DOS DADOS

Os arquivos foram gerados, um para cada ativo, a partir da captura de dados da difusão UMDF e não contém informações nulas, inválidas ou ausentes. Os arquivos forma processados e os dados dos ativos foram colapsados minuto-a-minuto e armazenados em um único arquivo contendo a data e a hora, e os preços de cada um dos ativos ao término de cada minuto. Em cenários onde, no intervalo de tempo, nenhum novo negócio tenha sido realizado, o valor do minuto anterior foi utilizado. Os objetos foram organizados em ordem ascendente por data e hora, e o conjunto de dados final contém 273,673 objetos e 3 atributos, delimitados pelo caractere *ponto e vírgula*. A tabela 2 contém uma amostra dos cinco últimos registros do conjunto de dados.

Data Hora	Valor IBOV	Valor PETR3
2021-12-30 17:46:00	104,882.95	30.61
2021-12-30 17:47:00	104,851.36	30.62
2021-12-30 17:48:00	104,873.23	30.60
2021-12-30 17:49:00	104,885.85	30.61
2021-12-30 18:12:00	104,822.44	30.70

Tabela 2. Amostra dos 5 u últimos registros do conjunto de dados

2.2. CAUSALIDADE E ESTACIONARIDADE

A hipótese nula para o teste de Causalidade éque, dadas duas séries temporais, elas não estão relacionadas, o que implica que o coeficiente dos modelos autorregressivos ézero. Logo, se o teste produzir uma estatística com valor-p significativamente pequeno (< 0.05), podemos rejeitar a hipótese nula e especular que pode haver relação entre as séries temporais.

Para realizar o teste de causalidade énecessário que as séries sejam estacionárias, ou seja, elas se comportam de forma aleatória ao longo do tempo ao redor de uma média constante. Séries temporais que possuem tendência e/ou sazonalidade não são estacionárias. Para identificar se há presença significativa de tendência nas séries temporais das variáveis foi utilizado o teste de Dickey-Fuller Aumentado, ADF (do acrônimo em inglês *Augmented Dickey-Fuller*) (Dickey; Fuller, 1979). Este teste permite saber se há presença de tendência nas séries temporais por meio de um teste de hipótese, portanto, verificar se a série éestacionária ou não. primeiro passo consiste na seleção adequada de defasagens (*lags)*, o que foi feito utilizando o comando *Vector Autoregressive Specification Order Criterion* (VARSOC) no Stata. Este comando retorna a função de probabilidade (LL, log-*likelihood*) e o teste de razão de probabilidade (LR, *likelihood-ratio test*), além de quatro critérios de informação, o *Final Prediction Error* (FPE), o critério de Akaike (AIC), o *Hannan Quinn Information Criterion* (HQIC) e o *Schwarz Bayesian Information Criterion* (SBIC).

O resultado obtido esta disponível na tabela 3 e, o número ótimo de defasagens do sistema éigual a 4, confirmado pelos testes LR, FPE, AIC, HQIC, SBIC.

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-3.0e+06				2.5e+09	27.3052	27.3052	27.3053
1	-854991	4.3e+06	4	0.000	8.45532	7.81055	7.81063	7.81083
2	-841338	27306	4	0.000	7.46412	7.68586	7.686	7.68633
3	-840887	902.19	4	0.000	7.4337	7.68178	7.68197	7.68244
4	-840824	127.11*	4	0.000	7.42965*	7.68123*	7.68148*	7.68208*

Tabela 3. Seleção do número timo de defasagens

Na sequência, realiza-se o teste de estacionariedade de Dickey-Fuller, nas séries, para verificar a presença de raízes unitárias. Os resultados são apresentados nas tabelas 4 e 5.

	Teste estatístico	Valor Crítico (1%)	Valor Crítico (5%)	Valor Crítico (10%)
Z(t)	-1.741	-3.960	-3.410	-3.120
MacKinnon aproximado valor-p para Z(t) = 0.7326				

Tabela 4. Teste ADF para IBOV

Teste estatístico Valor Crítico (1%) Valor Crítico (5%) Valor Crítico (10%)

Z(t)	-1.680	-3.960	-3.410	-3.120
MacKinnon aproximado valor-p para Z(t) = 0.7596				

Tabela 5. Teste ADF para PETR3

Existem evidências estatísticas suficientes, ao nnível de significância de 5%, para rejeitar-se a hipótese nula de estacionariedade das séries IBOV e PETR3.

Como existe a presença de raiz unitária em cada série, foi realizada a diferenciação de cada uma. A diferenciação éa o resultado da diferença entre um valor no tempo T, menos o valor anterior, no tempo T-1, que retira a possibilidade de que o coeficiente do termo T-1, seja maior que 1 e tenha efeito explosivo no comportamento da série.

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

Após a diferenciação os testes ADF são refeitos nas séries diferenciadas. Isto énecessário porque o teste ADF foi desenvolvido para detectar a presença de uma raiz unitária, mas a série pode possuir mais do que uma.

Com o que foi evidenciado nas tabelas 6 e 7, em suas primeiras diferenças, as séries se tornam estacionárias (figura 4).

	Teste estatístico	Valor Crítico (1%)	Valor Crítico (5%)	Valor Crítico (10%)
Z(t)	-200.116	-3.960	-3.410	-3.120
MacKinnon aproximado valor-p para Z(t) = 0.0000				

Tabela 6. Teste ADF para IBOV em sua primeira diferença¸a

	Teste estatístico	Valor Crítico (1%)	Valor Crítico (5%)	Valor Crítico (10%)
Z(t)	-213.147	-3.960	-3.410	-3.120
MacKinnon aproximado valor-p para Z(t) = 0.0000				

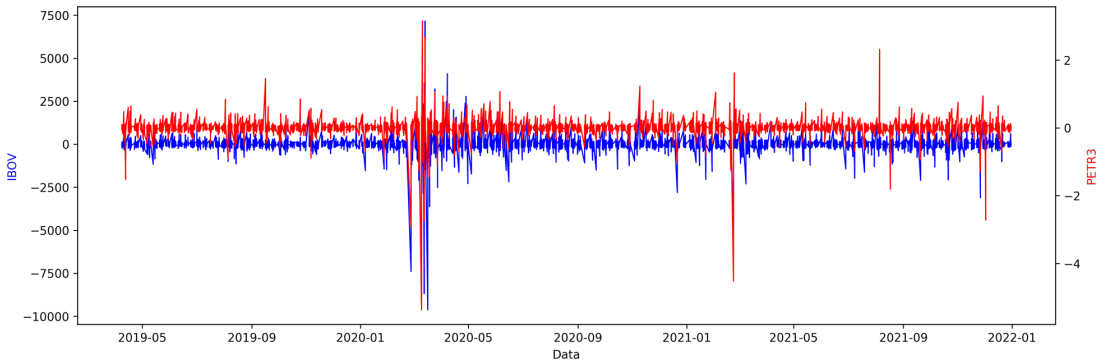

Tabela 7. Teste ADF para PETR3 em sua primeira diferença¸a

Figura 4. Séries apos a primeira diferenciação

3. MODELO VETORIAL AUTORREGRESSIVO (VAR) E CORRELAÇÃO RESIDUAL

De acordo com (Sims 1980) o VAR éum modelo de estimação de um sistema de séries de tempo para duas ou mais séries. Se as séries são estacionárias em nnível escolhemos de forma empírica o número adequado de defasagens utilizando o critério de informação Akaike (AIC) (Akaike, 1974). O AIC éuma métrica que mensura a qualidade de um modelo estatístico. Além disso, éde fácil avaliação, pois o melhor modelo éo que possuir o menor valor do AIC.

Se as séries não forem estacionárias, o modelo VAR não pode ser utilizado pois não vai estimar corretamente as relações entre as séries. Neste caso, o Modelo de Correção de Erro Vetorial (VEC), que é um caso especial do VAR, éo mais recomendado para estimar os coeficientes das variáveis que são estacionárias em suas diferenças.

Assim, considerando o menor valor do teste AIC, obtido para IBOV e PETR3, AIC = 7.68123, com um número de lags igual a 4, a matriz de correlação de resíduos obtida indica que as variáveis possuem um relacionamento fraco ou correlação positiva fraca ($0.25 < r = 0.4870 < 0.50$).

A análise dos resíduos éimportante para identificação de observações discrepantes. Foi utilizada a estatística de Durbin-Watson (Durbin; Watson, 1950), que éum teste estatístico que compara a presença de autocorrelação nos resíduos de uma regressão, para verificar a presença de autocorrelação para defasagem igual a 1 nos resíduos (erros de previsão) de uma análise de regressão. Caso existam, o modelo precisara ser alterado. O valor desta estatística pode variar entre 0 e 4 e, quanto mais próximo do valor 2, pode mos dizer que não há correlação significativa. Se o valor estiver mais próximo de 0 existe uma correlação positiva e de 4 uma correlação negativa. Para o caso em estudo o valor obtido no teste foi ”2.0” o que indica que não existe correlação significativa.

3.1. COINTEGRAÇÃO

O teste de cointegração éum método estatístico usado para testar a interdependênciaentre duas ou mais séries temporais não estacionárias no longo prazo ou para um período especificado. Sua função éidentificar parâmetros de longo prazo ou equilíbrio para duas ou mais séries. A cointegração ocorre quando duas ou mais séries temporais não estacionárias têm um equilíbrio de longo prazo e se movem juntas de tal forma que sua combinação linear resulta em uma série temporal estacionária.

Para verificar se as existe cointegração entre as series diferenciadas foi utilizado o comando VECRANK no Stata, que produz estatísticas usadas para determinar o número de equações de cointegração em um modelo vetorial de correção de erros (VECM). O resultado obtido consta na tabela 8 na qual, r indica a classificação máxima, ER a estatística de traço e LL a função de probabilidade (log-likelihood).

r	Paraˆmetros	LL	Autovalor	ER	Valor Crítico (5%)
0	14	-893527.57	.	1.06e+05	15.41
1	17	-863667.75	0.23874	4.60e+04	3.76
2	18	-840667.62	0.18951		

Tabela 8. Verificação de cointegração entre IBOV e PETR3 em sua primeira diferença

Como a estatística de traço em r=0 de 1.06e+05 excede seu valor crítico de 15.41, rejeitamos a hipótese nula de nenhuma equação de cointegração. Da mesma forma, como a estatística de traço em r=1 de 4.60e+04 excede seu valor crítico de 3.76, rejeitamos a hipótese nula de que existe uma ou menos equações de cointegração. Em suma, a hipótese nula de não cointegração foi rejeitada e a hipótese de que existe pelo menos uma equação de cointegração, não pode ser rejeitada.

O sistema VEC foi estimado usando as 4 defasagens indicadas previamente pelos testes e, em cada equação foram utilizados os dados em sua primeira diferença. Como resultado temos uma equação de cointegração que, na equação de IBOV, ao nnível de 1% (valor-p=0.0000) esta corrigindo a trajetória em 49.37% (-0.4937) em cada período de tempo.

Na equação de PETR3, ao nnível de 1% (valor-p=0.0000) esta corrigindo a trajetória de IBOV em 0,02% (-0.00028) contribuindo pouco para o equilíbrio do sistema. A equação de

cointegração esta funcionando ao nnível de 1%, entretanto, observa-se que para o sistema se corrigir (tender a` média de longo prazo) para cada 1 ponto de IBOV, são necessários 2993.825 pontos no sentido contrário e a constante indica a velocidade de correção do sistema, neste caso de 18.58 períodos de tempo.

O Teste de Causalidade de Granger apresentou os resultados contidos na tabela 9. O Teste evidencia uma causalidade bidirecional, na qual a hipótese de não causalidade érejeitada ao nnível de 1%, para as duas variáveis. Isto implica que a variável IBOV éútil para prever PETR3 considerando os valores passados de ambas as séries.

	IBOV-PETR3		PETR3-IBOV	
F(4, 273659)	5455.59		44.39	
Prob ¿ F	0.0000		0.0000	
chi2(4)	21823.09	(assintotico)	177.57	(assintotico)
Prob ¿ chi2	0.0000	(assintotico)	0.0000	(assintotico)

Tabela 9. Resultados do Teste de Causalidade de Granger

3.2. GERAÇÃO DE DADOS SINTÉTICOS

A estrutura *TimeGAN* foi utilizada para a geração de dados sintéticos realísticos que serão utilizados na etapa de previsão. Os quatro componentes foram compostos por redes neurais do tipo *Long Short-Term Memory* (LSTM). Este tipo de arquitetura de rede neural recorrente écapaz de aprender dependências de longo prazo em problemas de predição de sequência o que as torna ideais para classificar e prever séries temporais de duração desconhecida.

Inicialmente os dados foram normalizados com ”MinMaxScaler” para que os atributos ficassem em uma mesma escala, entre 0 e 1. Em seguida foram dividi-los em treinamento e teste na proporção 80% (218,937 objetos) e 20% (54,735 objetos), respectivamente. Os dados de treinamento foram divididos em 218,913 janelas móveis com sequências sobrepostas de 24 pontos de dados, conforme recomendação em (Yoon *et al.*, 2019). A Tabela 10 contém os valores dos hiper parâmetros utilizados para o treinamento.

Batch Size	128
Iterações	10.000
Hidden Dimensions	24
Número de camadas	3

Tabela 10. Hiperparâmetros para o treinamento do TimeGAN.

Cada um dos componentes da estrutura *TimeGAN* foi implementado com uma rede neural com 3 camadas LSTM (*Long Short Term Memory*), cada uma com 24 nos (*hidden dimension*), e com uma camada de *Dropout* (0.2) entre elas a fim de melhorar a generalização da rede e reduzir o efeito de *overfitting*.

O *Dropout* (Hinton *et al*., 2012) éuma técnica que reduz co-adaptações complexas de neurônios, já que um neurônio não pode confiar na presença de outros neurônios em particular. Desta forma, éforçado a aprender recursos mais robustos que são úteis em conjunto com muitos subconjuntos aleatórios diferentes dos outros neurônios. Pensando em uma rede onde o objetivo érealizar previsões, o *Dropout* éuma forma de garantir que o modelo seja robusto para a perda de qualquer evidência individual.

Para a geração das sequências aleatórias com distribuição uniforme no intervalo entre 0 e 1 foi utilizada a função *np.random.uniform* da biblioteca *NumPy* (Harris *et al*., 2020).

Em sintonia com (Yoon *et al*., 2019), o treinamento foi realizado em 3 etapas, cada uma com 10.000 iterações, a fim de se obter as perdas descritas a seguir:

- Perda de reconstrução, que se refere ao autoencoder (função de incorporação e função de recuperação) e compara o quão bem foi a reconstrução dos dados sintéticos em relação aos dados reais;
- Perda supervisionada que é responsável por capturar o quão bem o gerador se aproxima do próximo passo de tempo no espaço latente; e
- Perda não-supervisionada que refere a` relação entre as redes geradora e discriminadora (Minimax *Game*).

Na primeira etapa, o autoencoder foi treinado com os dados reais para reconstrução ideal. O cálculo da perda de reconstrução utilizou o erro médio quadrático, MSE (da sigla em inglês *Mean Squared Error*), para compreensão de quão perto a linha de regressão esta dos pontos previstos (entrada versus saída). Na etapa seguinte treinou-se o supervisor usando os dados da sequência real para capturar o comportamento temporal das informações históricas. A perda supervisionada também foi calculada com MSE. A etapa final consistiu no treinamento combinado dos quatro componentes, minimizando todas as três funções de perda mencionadas. A Tabela 11 contém as perdas resultantes do processo de treinamento conjunto onde as colunas são respectivamente:

- D loss: perda do discriminador
- G loss U: perda não-supervisionada
- G loss S: perda supervisionada
- G loss V: perda de correspondência de momento utilizada para melhorar a diversidade das amostras geradas
- E loss T0: perda de reconstrução

A perda total do gerador correspondera a` soma das perdas G loss S, G loss U e G loss V. (Yoon *et al*., 2019)

Iterac¸a˜o	D_loss	G_loss_U	G_loss_S	G_loss_V	E_loss_T0
0	2.0499	0.7133	0.0007	0.2382	0.0511
1,000	1.1766	1.9932	0.0002	0.0534	0.0054
2,000	0.8949	2.0971	0.0001	0.0185	0.0056
3,000	1.3904	1.7088	0.0001	0.0353	0.0040
4,000	1.0272	1.6578	0.0001	0.0251	0.0042
5,000	1.0289	2.0222	0.0000	0.0112	0.0048
6,000	1.4026	3.1482	0.0001	0.0109	0.0037
7,000	1.2813	1.8424	0.0000	0.0373	0.0078
8,000	1.3888	1.9451	0.0001	0.0212	0.0048
9,000	1.4121	1.5733	0.0000	0.0260	0.0036
10,000	1.4023	1.5376	0.0000	0.0244	0.0058

Tabela 11. Perdas do treinamento conjunto da estrutura TimeGAN

Para avaliação da utilidade dos dados gerados foi utilizada uma rede neural simples com apenas uma camada LSTM. A rede foi inicialmente treinada e com dados reais de treinamento (80%) e validado com os dados reais de teste (20%). Em seguida, o modelo foi treinado com os dados sintéticos gerados pelo modelo e validado com os mesmos dados reais de teste.

4. RESULTADOS

O treinamento conjunto do modelo consumiu 45 minutos e, a geração de séries sintéticas 7 horas e 56 minutos. Os resultados foram avaliados em relação a quantidade, qualidade e utilidade. Para exemplificar, o índice de uma das 218,914 janelas de dados contendo 24 valores foi aleatoriamente selecionado (e.g. 90,270) e, os valores sintéticos gerados pelo modelo foram comparados com os valores reais. A figura 5 mostra esta comparação apresentando no eixo x o índice da janela e no eixo y o valor.

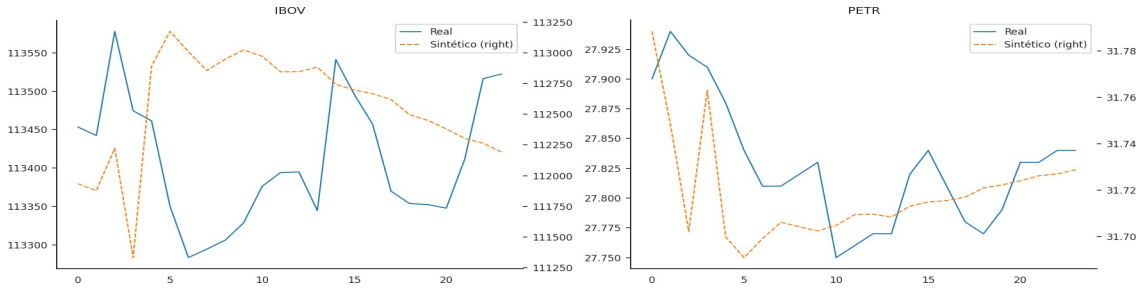

Figura 5. Comparação entre os dados reais e sintéticos.

Visualmente épossível realizar uma primeira avaliação comparando os dados sintéticos com os dados reais. Naturalmente, essa não éuma forma eficaz de avaliação, mas nos fornece uma boa ideia de como esta a qualidade dos dados gerados. A análise foi aplicada a um subconjunto contendo 2,000 janelas de dados selecionadas aleatoriamente.

Para esta análise visual dois métodos foram utilizados, a Análise de Componentes Principais, PCA (da sigla em inglês *Principal Component Analysis*) o *T-distributed Stochastic Neighbor Embedding* (t-SNE).

A PCA éum método estatístico multivariado que foi introduzido por Pearson (Pearson, 1901) com o objetivo de identificar padrões ocultos no conjunto de dados e reduzir dimensão removendo ruídos. No entanto, por conta da alta sensibilidade do PCA a` presença de *outliers* também foi aplicada a técnica de t-SNE, que sabe lidar melhor com esse problema e émenos sensível.

Uma maneira de comparar os resultados de PCA e t-SNE éobservar como eles preservam a estrutura dos dados. Em geral, o t-SNE émelhor em preservar a estrutura local, enquanto o PCA émelhor em preservar a estrutura global. Os gráficos da figura 6 indicam em uma análise visual que o resultado obtido épromissor e a qualidade dos dados sintéticos éboa pois seguem os padrões dos dados reais. Isto écaracterizado pelo fato de vermos uma boa sobreposição entre os pontos de dados sintéticos e reais. Para comparar os resultados de PCA e t-SNE de forma quantitativa, foi utilizada a métrica *Trustworthiness Score* para comparar as similaridades das representações de baixa dimensões produzidas. Os valores obtidos para PCA e t-SNE são respectivamente 1.000 e 0.999. O *Trustworthiness Score* varia de 0 a 1, com valores mais altos indicando melhor preservação das semelhanças pareadas. Uma pontuação de confiabilidade de 1 indica a preservação perfeita das

semelhanças entre pares. Isto indica que os dados produzidos são úteis para serem utilizadas em previsões.

Após o treinamento do modelo foi gerado um conjunto de dados sintéticos com 218,914 objetos que foram utilizados para previsão. Utilizando uma rede neural com uma camada LSTM para simulação, o modelo foi treinado com os dados reais de treinamento

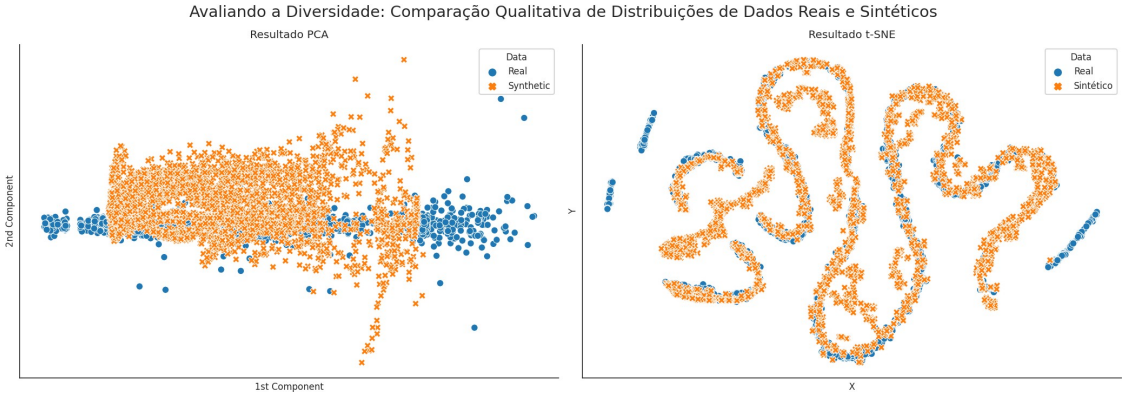

Figura 6. Comparação qualitativa das distribuições de dados reais e sintéticos.

e validado com os dados reais de teste. Em seguida o modelo foi treinado com os dados sintéticos e validado com os dados reais de teste. O objetivo éavaliar a quantidade e utilidade dos dados sintéticos gerados para predição. O resultado obtido esta disponível na tabela 12.

	r2	MAE	MRLE	MAPE
Real	0.999113	0.003670	0.000016	0.003065
Sintético	0.989465	0.008419	0.000270	0.008434

Tabela 12. Comparação de resultados de treinamento

O Erro Percentual Médio Absoluto, MAPE (da sigla em inglês *Mean Absolute Percentage Error*), obtido foi 0.008434 e corresponde à média de todos os erros percentuais absolutos entre os valores previstos e reais são obtidos. Em geral, um valor de MAPE inferior a 20% éconsiderado bom para previsão de séries temporais, o que indicaria que, em média, as previsões durante todo o período de tempo estavam menos de 20% distantes dos valores reais. Além disso, em geral, um valor de MAE menor indica que o modelo émelhor em fazer previsões precisas.

5. CONCLUSÃO

O propósito desta pesquisa foi a de verificar se a utilização de uma arquitetura *TimeGAN*, era viável na previsão de valores futuros de séries financeiras. Foram utiliza- dos dados dos ativos Ibovespa (IBOV) e Petrobras (PETR3). Foram realizados os testes de presença de raiz unitária em cada série, escolha de melhor número de lags, teste de cointegração, estimação VEC e teste de causalidade. O modelo VEC mostrou-se funcional e com causalidade bidirecional entre as duas séries de tempo. Isto indica que o comportamento passado éútil na previsão do comportamento futuro. Na sequência, foi estimada a *TimeGAN* utilizando as duas séries, com geração das séries sintéticas que serão utilizadas para a previsão dos dois ativos.

Esta arquitetura mostrou-se efetiva pois ela aprendeu a dinâmica temporal de da- dos das séries e foi capaz de gerar dados sintéticos de aparência realística. Em um cenário, no qual os dados iniciais são insuficientes para uma previsão mais precisa, a utilização de dados sintéticos para ampliar o conjunto de dados reais, pelos resultados obtidos, demonstrou-se uma boa alternativa.

A fase de treinamento conjunta é a etapa mais demorada e é impactada principal- mente pela quantidade de dados utilizados.

Os resultados são promissores na utilização da arquitetura para projeção de dados de séries de tempo financeiras.

As possibilidades de pesquisas futuras, dando continuação a esta, são inúmeras, pois pode-se variar os hiper-parâmetros, a frequência dos dados, os tamanhos de amostra, número de iterações, sem comprometer a qualidade dos dados sintéticos, aumentando a robustez do modelo.

A B3 possui grande movimentação financeira e muita volatilidade, sendo um lugar ideal para testes de modelos generativos, que podem trabalhar conjuntamente com os modelos econométricos para obtenção de mais bem resultados de projeção de preços.

RECONHECIMENTOS

P.P.B. obrigado as agências brasileiras CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nnível Superior) para o projeto STIC-AmSud (CAMA) no. 88881.694458/2022-

01 e CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) para a bolsa de pesquisa PQ 303356/2022-7.

Referências

ABRAHAM, J. B. Improving stock price prediction with gan-based data augmenta- tion. *Indonesian Journal of Artificial Intelligence and Data Mining*, 4(1):1–10. 2021.

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. 1974.

APTE, M., VAISHAMPAYAN, S., PALSHIKAR, G. K. Detection of causally anoma- lous time-series. *International Journal of Data Science and Analytics*, 11(2):141–153. 2021.

ASGARIAN, S., GHASEMI, R., MOMTAZI, S. Generative adversarial network for sentiment-based stock prediction. *Concurrency and Computation:* Practice and Expe- rience, 35(2):e7467. 2023.

DICKEY, D. A. FULLER, W. A. Distribution of the estimators for autoregres- sive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431. 1979.

DURBIN, J. WATSON, G. S. Testing for serial correlation in least squares regres- sion: I. *Biometrika*, 37(3/4):409–428. 1950.

ESTEBAN, C., HYLAND, S. L., RÄTSCH, G.. *Real-valued (medical) time series generation with recurrent conditional gans*. 2017.

FAMA, E. F. The behavior of stock-market prices. *The Journal of Business*, 38(1):34–105. 1965.

FEKRI, M. N., GHOSH, A. M., GROLINGER, K. Generating energy data for machine learning with recurrent generative adversarial networks. *Energies*, 13(1):130. 2019.

Fisher, R. A. On the mathematical foundations of theoretical statistics. *Philo- sophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368. 1922.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COUR- VILLE, A., AND BENGIO, Y. Generative adversarial nets. Advances in neural information processing systems, 27. 2014.

GOOGLE . Google colaboratory. https://colab.research.google.com/. 2018.

GRANGER, C. W. Investigating causal relations by econometric models and cross- spectral methods. Econometrica: journal of the Econometric Society, p 424–438, 1969.

HAI, Q., ZHANG, S., LIU, C., AND HAN, G. Hard disk drive failure prediction based on gru neural network. In: 2022 IEEE/CIC International Conference on Communicati- ons in China (ICCC), p. 696–701, 2022.

HARRIS, C. R., MILLMAN, K. J., VAN DER WALT, S. J., GOMMERS, R., VIRTANEN, P., COURNAPEAU, D., WIESER, E., TAYLOR, J., BERG, S., SMITH, N. J., KERN, R., PICUS, M., HOYER, S., VAN KERKWIJK, M. H., BRETT, M., HALDANE, A., DEL RIO, J. F., WIEBE, M., PETERSON, P., GERARD- MARCHANT, P., SHEPPARD, K.,

REDDY, T., WECKESSER, W., ABBASI, H., GOHLKE, C., AND OLIPHANT, T. E. Array programming with NumPy. Nature, 585(7825):357– 362. 2020.

HE, B.; KITA, E. The application of sequential generative adversarial networks for stock price prediction. The Review of Socionetwork Strategies, 15. 2021.

HINTON, G. E., SRIVASTAVA, N., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. Improving neural networks by preventing co-adaptation of feature detectors. CoRR, abs/1207.0580. 2012.

KUMAR, A., ALSADOON, A., P.W.C, P., ABDALLA, S., RASHID, T., PHAM, D., AND TRAN QUOC VINH, N. Generative adversarial network (gan) and enhanced root mean square error (ermse): Deep learning for stock price movement prediction. 2021.

LEE, M.; SEOK, J. Estimation with uncertainty via conditional generative ad- versarial networks. Sensors, 21(18):6194. 2021.

LI, Q., ZHANG, X., MA, T., LIU, D., WANG, H., HU, W. A multi-step ahead photovoltaic power forecasting model based on timegan, soft dtw-based k-medoids clustering, and a cnn-gru hybrid neural network. Energy Reports, 8:10346–10362. 2022.

LIN, S., LIU, D., HUANG, H. Credit default swap prediction based on generative adversarial networks. Data Technologies and Applications. Publisher Copyright: © 2022, Emerald Publishing Limited. 2022.

LIU, L., PEI, Z., CHEN, P., GAO, Z., GAN, Z., FENG, K. An effective gan-based multi-classification approach for financial time series. In: Qian, Z., Jabbar, M., and Li, X., editors, Proceeding of 2021 International Conference on Wireless Communi- cations, Networking and Applications, p. 1100–1107, Singapore. Springer Nature Singapore. 2022.

NEYMAN, J.; PEARSON, E. On the problem of the most efficient tests of statistical. 1933.

PEARSON, K. Liii. on lines and planes of closest fit to systems of points in space.The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,2(11):559–572. 1901.

PLESNER, J., ZHANG, Y., WANG, X., WANG, J., LI, X. Fetgan: Federated time- series generative adversarial network. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019.

POLAMURI, S. R., SRINIVAS, D. K., KRISHNA MOHAN, D. A. Multi-model gene-rative adversarial network hybrid prediction algorithm (mmgan-hpa) for stock market prices prediction. Journal of King Saud University - Computer and Information Sciences, 34(9):7433–7444. 2022.

SHANGGUAN, A., XIE, G., FEI, R., MU, L., HEI, X. Train wheel degradation generation and prediction based on the time series generation adversarial network. Re- liability Engineering System Safety, 229:108816. 2023.

SIMS, C. A. Macroeconomics and reality. Econometrica: journal of the Econo- metric Society, p. 1–48. 1980.

STAFFINI, A. Stock price forecasting by a deep convolutional generative adversarial network. Frontiers in Artificial Intelligence, 5. 2022.

WU, J.-L., TANG, X.-R., AND HSU, C.-H. A prediction model of stock market tra- ding actions using generative adversarial network and piecewise linear representation approaches. Soft Comput., 27(12):8209–8222. 2022.

XU, H., CAO, D., LI, S. A self-regulated generative adversarial network for stock price movement prediction based on the historical price and tweets. Knowledge- Based Systems, 247:108712. 2022.

YOON, J., JARRETT, D., VAN DER SCHAAR, M.. Time-series generative adversa- rial networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., dAlche-Buc, F., Fox, E., and Garnett, R., editors, Advances in Neural Information Processing Systems, v. 32. Curran Associates, Inc. 2019.

ZHANG, Y., ZHOU, Z., LIU, J., YUAN, J. Data augmentation for improving heating load prediction of heating substation based on timegan. Energy, 260:124919. 2022.

ZHOU, X., PAN, Z., HU, G., TANG, S., ZHAO, C. Stock market prediction on high- frequency data using generative adversarial nets. Mathematical Problems in En- gineering, 2018:1–11. 2018.