Geração de dados sintéticos longitudinais a partir de estruturas causais

Alessandro Silva Angeruzzi

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2026

Alessandro Silva Angeruzzi

Geração de dados sintéticos longitudinais a partir de estruturas causais

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Marcelo Keese Albertini

Uberlândia
2026

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

A587 Angeruzzi, Alessandro Silva, 1982-
2026 Geração de dados sintéticos longitudinais a partir de estruturas causais [recurso eletrônico] / Alessandro Silva Angeruzzi. - 2026.

Orientador: Marcelo Keese Albertini.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2026.41
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Albertini, Marcelo Keese,1984-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. III. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 02/2026, PPGCO				
Data:	19 de Janeiro de 2026	Hora de início:	13:51	Hora de encerramento:	15:20
Matrícula do Discente:	12322CCP001				
Nome do Discente:	Alessandro Silva Angeruzzi				
Título do Trabalho:	Geração de dados sintéticos longitudinais a partir de estruturas causais				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Ciência de Dados				
Projeto de Pesquisa de vinculação:	CNPq 306795/2022-1				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Bruno Augusto Nassif Travençolo - FACOM/UFU, Luís Alvaro de Lima Silva - CT/UFSM e Marcelo Keese Albertini - FACOM/UFU, orientador do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Luís Alvaro de Lima Silva - Santa Maria/RS. O outro membros da banca e o aluno(a) participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Marcelo Keese Albertini, apresentou a Comissão Examinadora e o(á) candidato(a), agradeceu a presença do público, e concedeu ao(á) Discente a palavra para a exposição do seu trabalho.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o(á) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato(a):

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Bruno Augusto Nassif Travençolo**, **Professor(a) do Magistério Superior**, em 20/01/2026, às 08:00, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Keese Albertini**, **Professor(a) do Magistério Superior**, em 30/01/2026, às 11:44, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Luís Alvaro de Lima Silva**, **Usuário Externo**, em 30/01/2026, às 18:27, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6975665** e o código CRC **1F2DC2B3**.

Referência: Processo nº 23117.000785/2026-56 SEI nº 6975665

Dedico este trabalho à minha família: à minha esposa Alessandra, que me incentivou e me apoiou em toda esta jornada, e à minha filha Julia, sempre compreensiva nos momentos em que precisei dizer: "Desculpa, o papai não pode agora, estou estudando."

Agradecimentos

Agradeço, antes de tudo, à minha esposa Alessandra, por seu amor, apoio e paciência em todos os momentos desta jornada. Sua presença constante tornou possível cada etapa deste trabalho. À minha filha Julia, que trouxe leveza aos dias mais intensos e renovou minha motivação sempre que precisei.

Ao meu irmão Rafael, agradeço pelo apoio gráfico e por contribuir diretamente para a qualidade visual das imagens desta dissertação.

Registro minha profunda gratidão ao meu orientador Prof. Dr. Marcelo K. Albertini, pela orientação dedicada, pelas discussões construtivas e pela confiança no desenvolvimento desta pesquisa.

Aos membros da banca, agradeço pela gentileza em aceitar o convite e pela disponibilidade em contribuir com este trabalho.

À Universidade Federal de Uberlândia e à FACOM, que foram minha casa desde a graduação, deixo meu sincero reconhecimento por todo o ambiente de formação, acolhimento e desenvolvimento acadêmico que encontrei ao longo desses anos.

Aos professores que tive o privilégio de encontrar nessa trajetória, agradeço pelos ensinamentos, pelo incentivo e por despertarem em mim o interesse pela pesquisa e pelo rigor científico.

Por fim, agradeço ao Programa de Pós-Graduação em Ciência da Computação, pelo apoio institucional e por proporcionar um ambiente fértil para a pesquisa e para o avanço do conhecimento.

"Tudo o que temos de decidir é o que fazer com o tempo que nos é dado."
Gandalf, O Senhor dos Anéis: A Sociedade do Anel, J. R. R. Tolkien.

Resumo

A inferência causal busca identificar relações de causa e efeito, indo além da correlação ao estimar como os resultados mudariam sob diferentes condições. Essa capacidade de prever desfechos contrafactuais é fundamental em aplicações reais — como medicina, finanças e ciências sociais — onde decisões confiáveis dependem de uma compreensão causal dos fenômenos.

A avaliação sistemática de modelos de inferência causal ainda é limitada pela escassez de conjuntos de dados de referência cujos mecanismos subjacentes sejam completamente conhecidos. Nesse cenário, esta dissertação apresenta o *Causal Synthetic Data Generator* (CSDG), uma ferramenta de código aberto capaz de gerar dados longitudinais sintéticos governados por estruturas causais, com dinâmicas autorregressivas explícitas.

O CSDG permite controle detalhado sobre a intensidade dos efeitos das variáveis, das intervenções no tratamento e dos níveis de ruído, oferecendo uma plataforma flexível e independente de domínio para experimentação e avaliação de algoritmos de aprendizado causal. A formalização proposta baseia-se em equações estruturais causais autorregressivas, que integram relações de causa e efeito com dependências temporais, possibilitando a geração de cenários factuais e contrafactuais sob diferentes estruturas causais.

Com o objetivo de validar a consistência dos dados gerados, este trabalho inclui uma etapa de análise quantitativa e testes em uma tarefa de previsão de resultados. As avaliações realizadas demonstram que os dados sintéticos preservam propriedades esperadas de correlação e resposta às intervenções e os resultados obtidos evidenciam a coerência causal e o realismo estatístico dos dados gerados, confirmando a adequação do CSDG como ferramenta de *benchmarking* e validação de modelos causais.

O código fonte do CSDG encontra-se disponível no repositório <https://github.com/angeruzzi/causal-synthetic-data-gen>.

Palavras-chave: *Benchmarks.* Inferência Causal. Dados Longitudinais. Geração de Dados Sintéticos. Séries Temporais.

Abstract

Causal inference seeks to identify cause-and-effect relationships, going beyond correlation by estimating how outcomes would change under different conditions. This ability to predict counterfactual outcomes is fundamental in real-world applications — such as medicine, finance, and the social sciences — where reliable decisions depend on a causal understanding of phenomena.

The systematic evaluation of causal inference models is still limited by the scarcity of benchmark datasets whose underlying mechanisms are fully known. In this context, this dissertation presents the Causal Synthetic Data Generator (CSDG), an open-source tool capable of generating synthetic longitudinal data governed by causal structures with explicit autoregressive dynamics.

The CSDG enables detailed control over the strength of variable effects, treatment interventions, and noise levels, offering a flexible and domain-independent platform for the experimentation and evaluation of causal learning algorithms. The proposed formalization is based on autoregressive structural causal equations, integrating cause-effect relationships with temporal dependencies and allowing the generation of both factual and counterfactual scenarios under different causal structures.

To validate the consistency of the generated data, this work includes a quantitative analysis stage and experiments on a prediction task. The evaluations show that the synthetic data preserve expected properties of correlation and response to interventions, and the results demonstrate both causal coherence and statistical realism, confirming the suitability of the CSDG as a benchmarking and validation tool for causal models.

The CSDG source code is publicly available at <https://github.com/angeruzzi/causal-synthetic-data-gen>.

Keywords: Benchmarks. Causal Inference. Longitudinal Data. Synthetic Data Generation. Time Series.

Lista de ilustrações

Figura 1 – Exemplos de Diagramas de Trajetória de Wright 32
Figura 2 – Exemplo de Coeficiente de Trajetória 32
Figura 3 – Factual x Contrafactual 35
Figura 4 – Estruturas Causais Gerais 38
Figura 5 – DAG de Estrutura Causal Direta 50
Figura 6 – Exemplo gerado a partir da estrutura Direta com efeito não linear . . . 51
Figura 7 – DAG de Estrutura Causal em Cadeia 51
Figura 8 – Exemplo gerado a partir da estrutura Cadeia com efeito não linear . . 52
Figura 9 – DAG de Estrutura Causal Confundidor 53
Figura 10 – Exemplo gerado a partir da estrutura Confundidor com efeito não linear 54
Figura 11 – Exemplo gerado de dado contrafactual 55
Figura 12 – Verificação de Pré-tendência 65
Figura 13 – Curvas $ATE(t)$ — Pontual, Direta, Linear 66
Figura 14 – Curvas $ATE(t)$ — Gradual, Direta, Linear 66
Figura 15 – Curvas $ATE(t)$ — Contínua, Direta, Linear 67
Figura 16 – Curvas $ATE(t)$ — Pontual, Direta, Não Linear - G2 68
Figura 17 – Curvas $ATE(t)$ — Gradual, Confundidor, Não Linear - G2 69
Figura 18 – Curvas $ATE(t)$ — Gradual, Cadeia, Não Linear - G2 70
Figura 19 – Curvas $ATE(t)$ — Contínua, Confundidor, Não Linear - G2 71
Figura 20 – Curvas $ATE(t)$ — Contínua, Cadeia, Não Linear - G2 71
Figura 21 – Curvas $ATE(t)$ — Gradual, Confundidor, Não Linear - G3 72
Figura 22 – Curvas $ATE(t)$ — Gradual, Cadeia, Não Linear - G3 72
Figura 23 – Curvas $ATE(t)$ — Contínua, Confundidor, Não Linear - G3 73
Figura 24 – Curvas $ATE(t)$ — Contínua, Cadeia, Não Linear - G3 74
Figura 25 – Monotonicidade por tipo - Direta, Linear - G1 75
Figura 26 – Monotonicidade por tipo - Direta, Não Linear - G3 76
Figura 27 – Monotonicidade por tipo - Confundidor, Não Linear - G3 76
Figura 28 – Monotonicidade por tipo - Cadeia Não Linear - G3 77

Figura 29 – Monotonicidade por faixa - Pontual, Direta, Não Linear - G3 78
Figura 30 – Monotonicidade por faixa - Gradual, Direta, Não Linear - G3 79
Figura 31 – Monotonicidade por faixa - Contínua, Direta, Não Linear - G3 79
Figura 32 – Monotonicidade por faixa - Contínua, Cadeia, Não Linear - G3 80
Figura 33 – Correlação e Δr — Contínua, Confundidor, Linear - G1 81
Figura 34 – Agregação temporal de Δr por tipo — Direta, Linear - G1 82
Figura 35 – Correlação e $\Delta\xi$ — Contínua, Direta, Não Linear - G2 83
Figura 36 – Correlação e $\Delta\xi$ — Contínua, Confundidor, Não Linear - G2 84
Figura 37 – Correlação e $\Delta\xi$ — Contínua, Cadeia, Não Linear - G2 84
Figura 38 – Agregação temporal de $\Delta\xi$ por tipo — Direta, Não Linear - G2 85
Figura 39 – Agregação temporal de $\Delta\xi$ por tipo — Cadeia, Não Linear - G2 86
Figura 40 – Correlação e $\Delta\xi$ — Contínua, Direta, Não Linear - G3 87
Figura 41 – Correlação e $\Delta\xi$ — Contínua, Confundidor, Não Linear - G3 87
Figura 42 – Correlação e $\Delta\xi$ — Contínua, Cadeia, Não Linear - G3 88
Figura 43 – Agregação temporal de $\Delta\xi$ por tipo — Direta, Não Linear - G3 89
Figura 44 – Agregação temporal de $\Delta\xi$ por tipo — Cadeia, Não Linear - G3 89
Figura 45 – Histórico e Horizonte de Predição 92
Figura 46 – Arquitetura *Encoder-Decoder* 93

Lista de tabelas

Tabela 1 – Componentes de um Modelo Causal Estrutural (SCM) 34
Tabela 2 – Abordagens de geração de dados sintéticos com estrutura causal 41
Tabela 3 – Abordagens de geração de dados sintéticos longitudinais 43
Tabela 4 – Parâmetros definidos no arquivo de configuração YAML do CSDG . . . 56
Tabela 5 – Estrutura dos arquivos gerados: chaves, formas e descrição. 59
Tabela 6 – Parâmetros globais de geração dos dados. 63
Tabela 7 – Descrição dos *datasets* de análise. 63
Tabela 8 – Dados médios de Correlação e Δr - G1 81
Tabela 9 – Dados médios de Correlação e $\Delta\xi$ - G2 83
Tabela 10 – Dados médios de Correlação e $\Delta\xi$ - G3 86
Tabela 11 – Parâmetros de geração dos dados da prova de conceito. 91
Tabela 12 – Hiperparâmetros utilizados nas redes neurais *baseline*. 93
Tabela 13 – RMSE da predição factual 94
Tabela 14 – PEHE das predições factual e contrafactual 95

Lista de siglas

ANOVA Análise de Variância

ARIMA *Auto-Regressive Integrated Moving Average*

ATE *Average Treatment Effect*

BART *Bayesian Additive Regression Trees*

CATE *Conditional Average Treatment Effect*

CFRNet *Counterfactual Regression Network*

CRN *Counterfactual Recurrent Network*

CSDG *Causal Synthetic Data Generator*

DAG *Directed Acyclic Graph*

DONALD *Dortmund Nutritional and Anthropometric Longitudinally Designed Study*

G-Net *G-Computation for Counterfactual Prediction*

GANITE *Generative Adversarial Nets in Individualized Treatment Effects*

GANs *Generative Adversarial Nets*

GNNs *Graph neural network*

GRU *Gated Recurrent Unit*

ITE *Individual Treatment Effect*

LSTM *Long Short-Term Memory*

MITRA *Mixed Synthetic Priors for Enhancing Tabular Foundation Models*

MTS *Multivariate Time Series*

NPZ *NumPy Zip*

PEHE *Precision in Estimation of Heterogeneous Effect*

RCT *Randomized Controlled Trials*

RMSE *Root Mean Square Error*

RMSN *Recurrent Marginal Structural Networks*

RNN *Recurrent Neural Network*

SCM *Structural Causal Model*

TARNet *Treatment-Agnostic Representation Network*

TMLE *Targeted Maximum Likelihood Estimation*

VAMBN *Variational Autoencoder Modular Bayesian Network*

VAR *Vector Auto-Regressive*

VARMA *Vector Auto Regressive Moving Average*

Lista de símbolos

$Y_i(1)$, $Y_i(0)$ Resultados potenciais do indivíduo i, com e sem tratamento.

$\hat{Y}_i(1)$, $\hat{Y}_i(0)$ Resultados estimados por um modelo para os dois mundos.

$Y_{i,t}^{(f)}$ Resultado factual do indivíduo i no tempo t.

$Y_{i,t}^{(cf,k)}$ Resultado contrafactual do indivíduo i no tempo t para o cenário k.

$\hat{Y}_{i,t}^{(f)}$ Predição factual do modelo no tempo t.

$\hat{Y}_{i,t}^{(cf,k)}$ Predição contrafactual no tempo t para o cenário k.

T_t Tratamento factual no tempo t.

Y_t Resultado factual no tempo t.

X_t Covariável, confundidor ou mediador, factual no tempo t.

$T_t^{(cf,k)}$ Tratamento contrafactual no tempo t para o cenário k.

$Y_t^{(cf,k)}$ Resultado contrafactual no tempo t para o cenário k.

$X_t^{(cf,k)}$ Covariável contrafactual no tempo t para o cenário k.

t Índice temporal da série.

S Número total de instantes temporais da série.

t_{int} Instante da intervenção no tratamento.

$\mathcal{H}$ Horizonte de tempos pós-intervenção ($\{\ \mathrm{t} \mid t \geq t_{\text{int}}\}$).

τ Defasagem pós-intervenção ($\tau = t - t_{\text{int}}$).

Φ Coeficientes autorregressivos (dependência temporal).

β Coeficientes de efeito causal entre variáveis, por exemplo β_{XY}.

δ Parâmetro (dose) da intervenção no tratamento.

ε Termo de ruído aleatório.

$f(\cdot)$, $g(\cdot)$ Funções de complexidade causal (linear ou não linear).

$r_f(t)$ Correlação, em geral Pearson, $T \leftrightarrow Y$ factual no tempo t

$r_{cf}^{(k)}(t)$ Correlação, em geral Pearson, $T \leftrightarrow Y$ contrafactual k no tempo t.

Δr Mudança de associação, em geral Pearson, entre factual e contrafactual.

$\xi_f(t)$ Correlação de Chatterjee $T \leftrightarrow Y$ factual no tempo t.

$\xi_{cf}^{(k)}(t)$ Correlação de Chatterjee $T \leftrightarrow Y$ contrafactual k no tempo t.

$\Delta\xi$ Mudança de associação, Chatterjee, entre factual e contrafactual.

N Número de indivíduos.

Sumário

1 INTRODUÇÃO 25
1.1 Hipótese 27
1.2 Contribuições 28
1.3 Organização da Dissertação 29

2 FUNDAMENTAÇÃO TEÓRICA 31
2.1 Inferência Causal 31
2.2 Resultados Potenciais e Contrafactuais 35
2.3 *Potential Outcomes Framework* 36
2.4 Estruturas Causais 37
2.5 Dados Longitudinais 39
2.6 Dados Sintéticos Causais 40
2.7 Métricas 42
2.7.1 **Efeito médio de Tratamento (ATE) 43**
2.7.2 **Mudança de associação (Δr) 44**
2.7.3 ***Root Mean Square Error* (RMSE) 46**
2.7.4 ***Precision in Estimation of Heterogeneous Effect* (PEHE) . . . 46**
2.8 Considerações Finais 47

3 *CAUSAL SYNTHETIC DATA GENERATOR* 49
3.1 Composição da Estrutura de Geração 49
3.2 Estruturas Causais Implementadas 50
3.2.1 **Estrutura Causal Direta 50**
3.2.2 **Estrutura Causal Cadeia 51**
3.2.3 **Estrutura Causal Confundidor 52**
3.3 Cenários Contrafactuais 54
3.4 Geração dos Dados 56
3.4.1 **Considerações finais 60**

4 ANÁLISE DO CSDG 61
4.1 Método de Análise 61
4.2 Dados e cenários analisados 62
4.3 Resultados — Efeito médio (ATE) 64
4.3.1 **Pré-tendências** 64
4.3.2 **Curvas temporais do ATE** 64
4.3.3 **Monotonicidade por tipo** 74
4.3.4 **Monotonicidade por faixas** 77
4.4 Resultados — Mudança de associação (Δr) 80
4.4.1 **Grupo G1 - *Datasets* Lineares** 80
4.4.2 **Grupo G2 - *Datasets* Não Lineares** 82
4.4.3 **Grupo G3 - *Datasets* Não Lineares** 85
4.5 Conclusão da Análise 90

5 APLICAÇÃO: PROVA DE CONCEITO COM APRENDIZADO TEMPORAL CAUSAL 91
5.1 Dados Utilizados 91
5.2 Modelos 92
5.3 Resultados 94
5.3.1 **Predição factual** 94
5.3.2 **Predição contrafactual** 94
5.3.3 **Síntese e perspectivas** 95

6 CONCLUSÃO 97
6.1 Principais Contribuições 97
6.2 Trabalhos Futuros 98
6.3 Contribuições em Produção Bibliográfica 99
6.4 Agradecimentos adicionais 99

REFERÊNCIAS 101

Capítulo 1

Introdução

A inferência causal tem como objetivo identificar relações de causa e efeito entre variáveis, estimando o impacto de uma variável sobre outra. Diferentemente da correlação, que capta apenas associações estatísticas, a inferência causal permite prever desfechos contrafactuais, cenários hipotéticos que descrevem o que teria ocorrido sob diferentes condições ou intervenções. Essa capacidade é crucial em áreas como medicina, economia, finanças e ciências sociais, onde a compreensão meramente correlacional não é suficiente para fundamentar decisões confiáveis (CHENG et al., 2022).

Os ensaios controlados aleatorizados (*Randomized Controlled Trials* - RCTs) são amplamente reconhecidos como o padrão-ouro para a inferência causal. Contudo, sua aplicação muitas vezes é inviável devido a restrições financeiras, éticas ou logísticas. Diante disso, métodos baseados em dados observacionais tornam-se essenciais.

Modelos de aprendizado de máquina têm alcançado resultados expressivos em tarefas de previsão, contudo, em sua forma tradicional, são essencialmente associativos — capturam correlações, mas não distinguem causa e efeito. Essa limitação compromete sua aplicabilidade em cenários nos quais decisões dependem de estimativas contrafactuais, como tratamentos médicos (KADDOUR et al., 2025).

Nos últimos anos surgiram diversos modelos de *deep learning* que tentam implementar uma modelagem causal, porém, treinar e avaliar tais métodos representam um desafio significativo. Em dados reais, os mecanismos geradores subjacentes são desconhecidos e os contrafactuais verdadeiros, os possíveis resultados em condições que não ocorreram, são, por definição, não observáveis (RUBIN, 1974).

Com isso temos uma escassez de dados reais e *benchmarks* padronizados para avaliação rigorosa de métodos de inferência causal. Essa lacuna constitui um dos principais gargalos da área (KADDOUR et al., 2025) e é particularmente crítica em contextos longitudinais, onde os efeitos acumulam-se ao longo do tempo e a validação exige dados com dependências temporais realistas e relações causais conhecidas (CHENG et al., 2022).

Dados sintéticos surgem como uma alternativa promissora para contornar essa limitação, pois permitem controle experimental sobre a estrutura causal, o ruído e a comple-

xidade das relações entre variáveis — algo frequentemente inviável em cenários reais — criando cenários controlados.

Um exemplo recente no uso de dados sintéticos para o treino é o MITRA (ZHANG et al., 2025), um *Tabular Foundation Model* desenvolvido pela Amazon, que demonstrou que modelos podem alcançar forte poder de generalização mesmo quando pré-treinados exclusivamente com dados sintéticos baseados em estruturas causais. Esse resultado reforça o papel estratégico dos geradores sintéticos no avanço de métodos de aprendizado, ao evidenciar que o controle sobre as relações causais embutidas nos dados impacta diretamente o desempenho em tarefas reais de previsão e análise.

Apesar desse potencial, os geradores existentes ainda apresentam limitações relevantes. Abordagens baseadas em dados observacionais reais priorizam a preservação de propriedades estatísticas e a proteção de privacidade (BUN et al., 2024; KÜHNEL et al., 2024), mas não fornecem *ground truth* causal para testes comparativos. Por outro lado, bibliotecas de simulação como o CausalTables.jl oferecem suporte a intervenções em modelos estruturais (BALKUS; HEJAZI, 2025), porém restritas a dados estáticos, sem considerar dependências autorregressivas ou a evolução temporal de tratamentos e resultados. Assim, permanece uma lacuna quanto a *benchmarks* sintéticos que combinem estrutura causal explícita, dependência temporal e cenários contrafactuais sob diferentes regimes de intervenção.

Este trabalho é motivado, portanto, em suprir a escassez de dados e *benchmarks* para treino e validações de modelos de inferência causal no contexto longitudinal. Partindo da premissa de que a ausência de conjuntos de dados longitudinais com estrutura causal conhecida limita a validação empírica de modelos de inferência causal, torna-se essencial, assim, investigar ferramentas para a geração de dados controlados, com estrutura causal explícita e ruído parametrizável, que sirvam como base experimental para o avanço de algoritmos de aprendizado causal.

Para enfrentar esse desafio, propõe-se o *Causal Synthetic Data Generator* (CSDG), uma ferramenta projetada para gerar dados longitudinais sintéticos com controle explícito sobre relações causais, dinâmica temporal, ruído e intervenções. Ao tornarem-se observáveis tanto resultados factuais quanto contrafactuais, o CSDG possibilita a construção de cenários reprodutíveis e parametrizáveis para *benchmarking* de algoritmos de inferência causal em dados longitudinais.

O objetivo geral deste trabalho foi investigar um método para a geração de dados sintéticos longitudinais a partir de estruturas causais conhecidas, denominado CSDG, visando assim, contribuir com uma ferramenta capaz de construir cenários reprodutíveis e controlados, destinados especificamente ao *benchmarking* de modelos de inferência causal em dados longitudinais.

De forma mais específica, buscou-se:

(i) investigar como diferentes estruturas causais — direta, em cadeia e com confundidor

— podem ser formalizadas e simuladas em um contexto dinâmico;

(ii) gerar cenários factuais e contrafactuais reprodutíveis, controlando explicitamente os parâmetros de causalidade, ruído e tipo de intervenção;

(iii) avaliar a coerência estatística e causal dos dados gerados por meio de métricas formais, como o efeito médio do tratamento (ATE) e a mudança de associação (Δr);

(iv) analisar, como prova de conceito, o desempenho de modelos de aprendizado temporal causal aplicados a esses dados sintéticos, comparando a capacidade preditiva factual e contrafactual de diferentes arquiteturas.

Suprir a ausência de *benchmarks* padronizados para validar métodos de inferência causal em cenários longitudinais, envolveu três desafios principais. O primeiro diz respeito à formulação de equações estruturais capazes de representar relações causais sob dependência temporal, garantindo que os efeitos propagassem ao longo da série de modo consistente com o comportamento esperado das variáveis. Para isso, foram adotados modelos autorregressivos explícitos, nos quais o resultado presente depende tanto de seus valores passados quanto das variáveis causais associadas. O segundo concentrou-se em assegurar o realismo estatístico dos dados simulados, equilibrando controle experimental com variabilidade natural por meio da parametrização de funções lineares e não lineares, além da introdução de ruído aleatório configurável. Por fim, foi necessário estabelecer mecanismos para quantificar a qualidade causal dos dados gerados de forma independente do modelo preditivo utilizado, o que motivou a definição de um protocolo de avaliação baseado em métricas formais, além da aplicação prática do conjunto gerado em uma prova de conceito com modelos de aprendizado temporal.

A mitigação desses desafios permitiu estabelecer uma base sólida para a geração de dados longitudinais com estrutura causal conhecida, viabilizando a avaliação sistemática da consistência estatística e dos efeitos das intervenções simuladas.

Como resultado, o CSDG foi projetado para produzir dados com coerência causal e comportamentos estatísticos desejáveis, configurando um ambiente experimental para avaliação da precisão factual e contrafactual de algoritmos de inferência causal. Assim, este trabalho estabelece uma base para estudos sistemáticos de modelos causais aplicados a dados longitudinais.

1.1 Hipótese

A hipótese central desta pesquisa é que o controle explícito da estrutura causal, da dinâmica temporal e das intervenções permite que o gerador proposto — o CSDG —

produza cenários sintéticos reprodutíveis e confiáveis, com variabilidade factual e contrafactual adequada para o *benchmarking* de modelos de inferência causal em dados longitudinais.

Parte-se do pressuposto de que equações estruturais autorregressivas podem representar de maneira consistente dinâmicas causais ao longo do tempo, resultando em dados que preservam propriedades esperadas de correlação temporal, efeitos médios coerentes e respostas monotônicas às intervenções. Assume-se, ainda, que esses dados são suficientemente informativos para a avaliação da capacidade de modelos de aprendizado temporal em estimar efeitos causais com precisão, tanto em cenários factuais quanto contrafactuais.

A validação experimental dessa hipótese fundamenta-se na análise estatística e na aplicação prática dos dados gerados, demonstrando que o CSDG pode atuar como ferramenta de *benchmarking* para métodos de inferência causal longitudinal.

Com base nessa hipótese e na metodologia proposta, este trabalho oferece contribuições que avançam tanto no desenvolvimento de ferramentas quanto na avaliação de métodos causais em dados longitudinais.

1.2 Contribuições

As contribuições deste trabalho concentram-se no desenvolvimento e avaliação de um gerador de dados sintéticos longitudinais com estrutura causal explícita. Primeiramente, propõe-se uma base formal para simulação a partir de Equações Estruturais Causais Autorregressivas, integrando dependências temporais e relações de causa e efeito em um mesmo modelo. Com esse fundamento, foi formulado e implementado o CSDG, um gerador capaz de produzir dados longitudinais com diferentes graus de complexidade funcional e parâmetros controláveis de ruído e intervenção, abrangendo tanto cenários factuais quanto contrafactuais.

Adicionalmente, estabelece-se um protocolo sistemático de avaliação da coerência causal dos dados produzidos, combinando análises estatísticas, métricas de dependência e experimentos com modelos de aprendizado temporal. Essa abordagem permitiu caracterizar quantitativamente a qualidade dos dados sintéticos, com base em propriedades desejáveis, como variabilidade, previsibilidade e consistência causal — mensuradas por métricas como efeito médio do tratamento (ATE) e mudança de associação (Δr). Como prova de conceito, foi realizada uma análise comparativa do desempenho de modelos lineares e redes neurais recorrentes na estimação de efeitos causais em cenários sintéticos controlados, explorando tanto predições factuais quanto contrafactuais.

Por fim, este trabalho discute diretrizes para a reprodutibilidade e padronização de experimentos com dados sintéticos no contexto de inferência causal temporal, reforçando o papel do CSDG como ferramenta aberta e adequada para *benchmarking* de métodos causais. Dessa forma, a dissertação contribui conceitualmente e tecnicamente para a área,

ao disponibilizar um ambiente experimental controlado para investigação e validação de algoritmos de aprendizado causal em dados longitudinais.

1.3 Organização da Dissertação

Para apresentar de forma estruturada o desenvolvimento e os resultados desta pesquisa, a dissertação está estruturada em seis capítulos. O Capítulo 1 apresenta o contexto da pesquisa, destacando a motivação, os objetivos, a hipótese e as contribuições alcançadas. No Capítulo 2, são discutidos os fundamentos teóricos sobre inferência causal, resultados potenciais, estruturas causais e dados longitudinais, além de uma revisão dos principais trabalhos correlatos. O Capítulo 3 descreve a proposta metodológica, detalhando a formulação do gerador de dados sintéticos CSDG, a estrutura causal adotada, os cenários factuais e contrafactuais e o processo de geração dos dados longitudinais. Em seguida, o Capítulo 4 apresenta a análise quantitativa dos dados simulados, incluindo as métricas utilizadas e os resultados referentes à coerência estatística e causal dos cenários. O Capítulo 5 traz a aplicação prática do gerador, por meio de uma prova de conceito com modelos de aprendizado temporal, avaliando sua capacidade de previsão tanto factual quanto contrafactual. Por fim, o Capítulo 6 sintetiza as conclusões, discute limitações e aponta direções para pesquisas futuras, além de relatar as contribuições acadêmicas decorrentes deste trabalho.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos que embasam a presente pesquisa, abordando os principais conceitos e métodos relacionados à inferência causal, resultados potenciais, estruturas causais e dados longitudinais. O objetivo é oferecer uma visão abrangente da evolução da área, desde os modelos clássicos até as abordagens recentes baseadas em aprendizado profundo e geração de dados sintéticos.

2.1 Inferência Causal

Em *Correlation and Causation*, Wright (1921) apresentou métodos para distinguir correlações espúrias de relações causais, estabelecendo as bases da área de inferência causal. A correlação é uma medida estatística que descreve um grau de associação entre duas variáveis — frequentemente restrita à relação linear no caso da correlação de Pearson. No entanto, correlação não implica causalidade; duas variáveis podem apresentar associação sem que uma seja a causa de outra, resultando em uma correlação espúria entre elas. Ao identificar uma relação em que uma variável, a causa, influencia diretamente a outra, o efeito, temos uma relação de causalidade.

Wright (1921) também introduziu a técnica estatística de *Path Analysis*, ou diagramas de trajetória, para analisar relações causais entre um conjunto de variáveis; a técnica é uma extensão da regressão múltipla e permite decompor correlações em componentes diretos e indiretos.

Com os diagramas de trajetória (Figura 1) temos uma ferramenta poderosa para a representação e interpretação de relações causais complexas, em que setas são utilizadas para indicar uma relação causal direta de uma variável para outra. Foram introduzidos também os Coeficientes de Trajetória (Figura 2), que quantificam a influência direta de uma variável sobre a outra em um modelo causal.

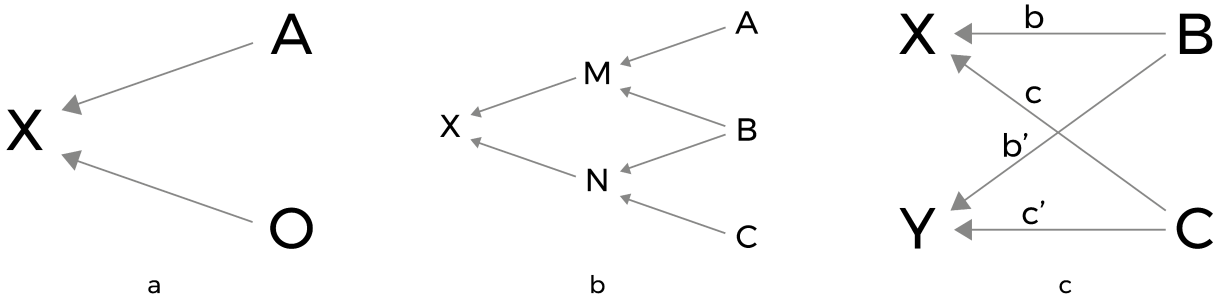

Figura 1 – Exemplos de Diagramas de Trajetória de Wright. Em (a) temos um diagrama simples de causa e efeito, onde X é determinado por A e O; em (b) temos um sistema em que a variável X é determinada por M e N, que são determinadas de forma independente por A e C e são relacionadas por B; em (c) temos um diagrama que mostra que duas variáveis X e Y estão relacionadas por ambas serem determinadas pelas variáveis independentes B e C. Fonte: Wright (1921).

Neyman (1923) introduziu conceitos fundamentais para o planejamento de experimentos, incluindo a distribuição de amostragem e a importância da aleatorização. A distribuição de amostragem é essencial para compreender a variabilidade dos estimadores, o que permite realizar inferências estatísticas válidas. Neyman também destacou a importância da aleatorização na condução de experimentos, pois ela garante a comparabilidade entre os grupos experimentais, eliminando vieses e confundidores. Vieses referem-se a distorções sistemáticas que podem levar a resultados errôneos, enquanto confundidores são variáveis externas que podem influenciar tanto a variável independente quanto a dependente, distorcendo a verdadeira relação entre elas. A aleatorização assegura que quaisquer diferenças observadas entre os grupos possam ser atribuídas aos tratamentos aplicados e não a fatores externos.

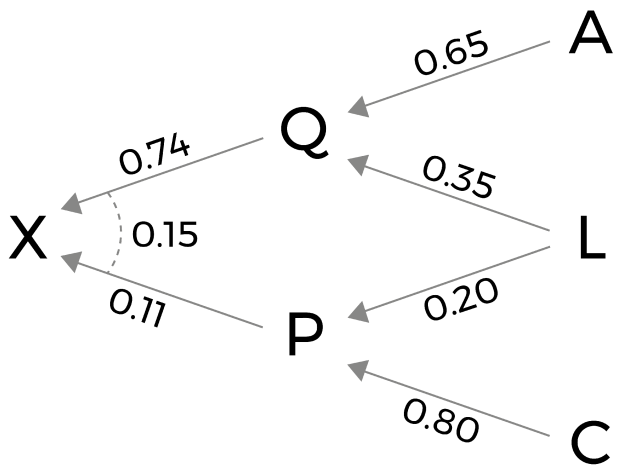

Figura 2 – Exemplo de Coeficiente de Trajetória. Os coeficientes de trajetória aplicado em um diagrama quantificam a relação causal entre as variáveis. Fonte: Wright (1921)

Os conceitos trazidos por Neyman foram fundamentais para a criação do *Randomized Controlled Trial* (RCT), que foi formalizado mais tarde por Fisher (1935) e tornou-se o

padrão de uso, principalmente em ensaios clínicos da área de saúde, mas também na pesquisa científica em geral para se estabelecer causalidade devido ao seu rigor metodológico.

O RCT é um tipo de estudo científico projetado para testar a eficácia de uma intervenção ou tratamento, como por exemplo a administração de um analgésico, e avaliar o resultado, que neste caso seria a eficácia do medicamento contra uma dor de cabeça. Embora o termo tratamento seja frequentemente associado a contextos médicos, ele pode ser aplicado em diversos cenários. Por exemplo, na educação, pode-se avaliar a eficácia de um novo método de ensino (tratamento) e observar se há uma melhora nas notas dos alunos (resultado). Em ciências sociais, pode-se investigar se uma determinada política (tratamento) melhora as condições de vida de uma população (resultado).

Uma base do RCT é o estabelecimento de um grupo de tratamento e um grupo de controle. O grupo de tratamento recebe a intervenção a ser avaliada, enquanto o grupo de controle não recebe a intervenção ou recebe um tratamento padrão ou placebo. Isso permite a comparação dos resultados entre os grupos e a verificação da eficácia da intervenção.

A aleatorização é crucial no RCT, pois assegura que os indivíduos sejam distribuídos de forma aleatória entre os grupos para evitar o viés de seleção. Outro conceito importante é o mascaramento (*blinding*), que pode ser simples-cego, onde os indivíduos não sabem a qual grupo pertencem, ao tratamento ou controle, evitando o viés de expectativa, ou duplo-cego, onde tanto os participantes quanto os pesquisadores desconhecem a distribuição dos grupos, prevenindo o viés de observação e interpretação.

Por fim, os dados do experimento devem ser analisados por meio de métodos estatísticos, como a Análise de Variância (ANOVA), para determinar se os efeitos do tratamento são significativos em comparação ao controle (MONTGOMERY, 2017).

Entre os principais desafios na utilização do RCT estão seus altos custos (RUBIN, 1974). Além disso, questões éticas frequentemente surgem, especialmente quando é necessário deixar de aplicar um tratamento potencialmente benéfico a um grupo controle. Em outros casos, a natureza do problema pode impossibilitar a organização dos grupos de forma adequada, por exemplo, em situações de desastres naturais como terremotos ou inundações, onde não é possível escolher aleatoriamente quais comunidades serão expostas ao evento para criar grupos de controle, ou mesmo em situações históricas, ou de políticas amplamente aplicadas em uma população. Essas limitações requerem a consideração cuidadosa de alternativas metodológicas, como estudos observacionais bem planejados, que possam fornecer evidências causais robustas em contextos onde o RCT não é viável.

Rubin (1974) preocupou-se em desenvolver um método para estimar efeitos causais tanto em estudos randomizados quanto não randomizados, formalizando a ideia de resultados potenciais e contrafactuais, que serão explorados na Seção 2.2.

Pearl (2009) trouxe uma contribuição significativa e uma abordagem moderna aos problemas de Inferência Causal, como o *Directed Acyclic Graph* (DAG), o *Structural*

Causal Model (SCM) e o *do-calculus*.

Os DAGs são grafos acíclicos usados para representar graficamente as relações de causa e efeito entre variáveis; cada nó no grafo representa uma variável e as arestas direcionadas indicam a direção da causalidade, ou seja, a influência de uma variável sobre a outra. Diferentemente dos diagramas de trajetória de Wright modelando estritamente relações de causalidade direta e de forma unidirecional, não permitindo ciclos e garantindo clareza na direção dos efeitos.

O SCM foi desenvolvido como uma extensão do DAG, fornecendo uma base matemática para representar e analisar relações causais. Os Modelos Causais Estruturais têm origem nos modelos de equações estruturais desenvolvidos em áreas como genética, econometria e ciências sociais (PETERS; JANZING; SCHÖLKOPF, 2017). O ferramental causal proposto por Pearl incorporou a esses modelos uma semântica causal explícita, baseada em grafos direcionados, intervenções e contrafactuais (PEARL, 2009).

Formalmente, conforme Peters, Janzing e Schölkopf (2017), um Modelo Causal Estrutural pode ser definido como a tupla

$$\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U})). \tag{1}$$

Cada variável endógena $V_i \in \mathbf{V}$ é determinada por uma equação estrutural da forma

$$V_i := f_i(PA_i, U_i), \tag{2}$$

onde $U_i \in \mathbf{U}$ corresponde ao termo exógeno associado e PA_i denota o conjunto de pais causais de V_i no grafo causal, definido como

$$PA_i := \{\, V_j \in \mathbf{V} \mid V_j \to V_i \,\}. \tag{3}$$

A Tabela 1 resume os principais componentes que constituem um Modelo Causal Estrutural.

Tabela 1 – Componentes de um Modelo Causal Estrutural (SCM)

Componente	**Descrição**
$\mathbf{U}$	Variáveis exógenas (fatores não observados)
$\mathbf{V}$	Variáveis endógenas (variáveis observáveis)
$\mathbf{F}$	Conjunto de equações estruturais
$P(\mathbf{U})$	Distribuição conjunta dos ruídos exógenos

Dessa forma, os SCMs combinam a representação gráfica com uma base quantitativa, permitindo a simulação de intervenções (*do-operations*) e a análise de cenários contrafactuais.

O *do-calculus* é uma ferramenta para manipular expressões causais e derivar conclusões sobre intervenções. Ele estabelece um conjunto de regras para transformar probabilidades condicionais observacionais em probabilidades intervencionais.

2.2 Resultados Potenciais e Contrafactuais

O conceito de resultados potenciais foi estabelecido por Rubin (1974) e são os possíveis resultados que podem ser observados em um indivíduo durante um experimento. Por exemplo, em um experimento para avaliar um medicamento para o tratamento de hipertensão, podemos observar o resultado caso o paciente receba o medicamento ou o resultado caso o paciente não receba o medicamento. Formalmente dizemos que $Y_i(1)$ é o resultado que seria observado se o paciente "i" recebesse o tratamento e $Y_i(0)$ caso não recebesse.

Apesar de podermos observar diferentes resultados entre os indivíduos de um experimento, podemos observar apenas um único resultado por indivíduo. Ou seja, ou o indivíduo recebeu o tratamento $Y_i(1)$ ou o indivíduo não recebeu o tratamento $Y_i(0)$. Este é um problema fundamental na inferência causal, pois não conseguimos observar todos os resultados potenciais no mesmo indivíduo. Nos estudos randomizados, esse problema é contornado garantindo a comparabilidade entre os grupos de tratamento e controle. Em estudos não randomizados, técnicas como pareamento, estratificação e ajuste por pontuação de propensão são usadas para criar grupos comparáveis.

O resultado potencial não observado em um indivíduo durante o experimento é chamado de contrafactual. Na Figura 3 podemos observar que, para o indivíduo que recebeu o tratamento $T = 1$ temos o resultado factual observado $Y(1)$, o cenário contrafactual seria o resultado caso não tivesse recebido o tratamento $T = 0$, ou seja, $Y(0)$.

Figura 3 – Factual x Contrafactual. Exemplificação do cenário Factual T=1, paciente recebe o tratamento, que resulta em Y(1) e do cenário contrafactual T=0, paciente não recebe o tratamento, que resultado em Y(0).

Ao tornarem-se observáveis tanto $Y_i(0)$ quanto $Y_i(1)$, é possível avaliar de forma direta a capacidade dos algoritmos em estimar efeitos causais - seja em nível individual, por meio do *Individual Treatment Effect* (ITE), nível médio com o *Average Treatment Effect* (ATE) ou condicional com o *Conditional Average Treatment Effect* (CATE) (CHENG et al., 2022).

2.3 *Potential Outcomes Framework*

Rubin (1978) foi pioneiro em criar um método baseado em estruturas de resultados potenciais, trazendo o conceito de *Potential Outcomes Framework*, que passou a ser amplamente utilizado em pesquisas subsequentes para novos métodos que se propõem à análise de causalidade e estimativa de contrafactuais em estudos observacionais e experimentais em diversas áreas e contextos.

Alguns métodos não-paramétricos foram propostos, como o de Xu, Xu e Saria (2016), que apresenta uma abordagem bayesiana não-paramétrica para estimar curvas de resposta ao tratamento a partir de séries temporais esparsas. As curvas de resposta ao tratamento representam a relação funcional entre diferentes intensidades de tratamento e seus respectivos resultados potenciais. Cada ponto da curva corresponde ao valor que o desfecho apresentaria sob uma intervenção hipotética de intensidade específica, permitindo analisar como o efeito causal varia de acordo com a dose ou trajetória do tratamento.

Schulam e Saria (2017) propõem modelos contrafactuais para suporte a decisões confiáveis, destacando a importância de modelos robustos e interpretáveis na inferência causal. Soleimani, Subbaswamy e Saria (2017) discutem modelos de resposta ao tratamento para raciocínio contrafactual com intervenções contínuas, ou seja, intervenções que variam continuamente ao longo do tempo. Curth e Schaar (2021) propõem a estimativa não-paramétrica de efeitos de tratamento heterogêneos, oferecendo uma abordagem teórica sólida que se traduz em algoritmos de aprendizado eficientes.

Os métodos não-paramétricos oferecem uma abordagem flexível para a estimativa de efeitos de tratamento sem assumir uma forma funcional específica para os dados e são especialmente úteis quando os dados são esparsos ou apresentam estruturas complexas.

Quanto às propostas de métodos paramétricos, temos vários exemplos. Laan e Rubin (2006) introduzem o aprendizado por máxima verossimilhança direcionada (*Targeted Maximum Likelihood Estimation* - TMLE), um método paramétrico para inferência causal que combina modelagem estatística robusta com a flexibilidade de aprendizado de máquina. Chipman, George e McCulloch (2010) apresentam as Árvores de Regressão Aditiva Bayesiana (*Bayesian Additive Regression Trees* - BART), uma metodologia poderosa para inferência causal que utiliza técnicas bayesianas para construir modelos de previsão robustos e interpretáveis. Johansson, Shalit e Sontag (2016) discutem o aprendizado de representações para inferência contrafactual, propondo um método paramétrico que melhora a precisão da estimativa de efeitos de tratamento a partir de dados observacionais. Kuzmanovic, Hatt e Feuerriegel (2023) tratam da estimativa de efeitos de tratamento médio condicional com informações de tratamento ausentes, abordando um desafio comum em dados observacionais e propondo soluções paramétricas robustas.

Os métodos paramétricos assumem uma forma funcional específica para os dados e oferecem uma estrutura mais rígida para a modelagem de efeitos causais. Eles são úteis quando se tem uma boa compreensão da estrutura dos dados.

Com uma abordagem diferente das clássicas, modelos de *deep learning* têm sido utilizados em vários trabalhos de inferência causal também, permitindo a modelagem de relações complexas em grandes conjuntos de dados. Johansson, Shalit e Sontag (2016) propõem técnicas de aprendizado de representações que utilizam redes neurais profundas para inferência contrafactual, destacando a capacidade do *deep learning* em lidar com dados complexos e de alta dimensionalidade. Yoon, Jordon e Schaar (2018) introduziram a *Generative Adversarial Nets in Individualized Treatment Effects* (GANITE), uma abordagem inovadora que utiliza *Generative Adversarial Nets* (GANs) para estimar efeitos de tratamento individualizados, demonstrando a eficácia das técnicas de *deep learning* na inferência causal. Shalit, Johansson e Sontag (2017) apresentaram a *Treatment-Agnostic Representation Network* (TARNet) e a *Counterfactual Regression Network* (CFRNet), redes neurais profundas projetadas para inferência contrafactual, mostrando como o *deep learning* pode ser aplicado para melhorar a precisão das estimativas de efeitos de tratamento.

Os métodos de *deep learning* permitem a modelagem de relações complexas em grandes conjuntos de dados, capturando padrões e estruturas latentes que outras abordagens podem não detectar. Estas técnicas são especialmente úteis para dados complexos e de alta dimensionalidade.

Um contexto comum na área médica é a ocorrência de uma sequência de tratamentos ao longo de um período (ALLAM et al., 2021). Para a estimativa contrafactual desse tipo temos como exemplos a *Recurrent Marginal Structural Networks* (RMSN) (LIM; ALAA; SCHAAR, 2018), a *Counterfactual Recurrent Network* (CRN) (BICA et al., 2020) e a *G-Computation for Counterfactual Prediction* (G-Net) (LI et al., 2021).

Ainda neste contexto, temos o *Causal Transformer* (MELNYCHUK; FRAUEN; FEUERRIEGEL, 2022), que inova ao implementar uma arquitetura baseada em *Transformer*, que empilha blocos de processamento, sendo que cada camada possui mecanismos de *cross-attention* paralelos que associam diferentes entradas de dados longitudinais - tratamentos, resultados e covariáveis - variáveis no tempo, além de propor uma nova função de perda chamada *Counterfactual Domain Confusion loss* (CDC *loss*).

Nichani, Damian e Lee (2024) demonstraram em seu trabalho sobre aprendizado de estruturas causais que, através do gradiente descendente, um *Transformer* simplificado de duas camadas codifica o grafo causal latente na primeira camada de atenção a partir de sequências de dados gerados por cadeias de Markov, trazendo a perspectiva de que *Transformers* são capazes de recuperar estruturas causais.

2.4 Estruturas Causais

Desde os primeiros estudos de Wright (1921), diagramas de trajetória passaram a ser utilizados para representar de forma gráfica relações causais entre variáveis. No entanto, os

avanços propostos por Judea Pearl, por meio dos DAGs e dos SCMs, que se consolidou uma estrutura formal e poderosa para a representação e análise de sistemas causais complexos (PEARL, 2000).

Adotamos neste trabalho definições consolidadas na literatura de inferência causal, segundo as quais diferentes tipos de variáveis desempenham papéis específicos nas relações causais. Variáveis mediadoras são aquelas que se interpõem entre a causa (tratamento) e o efeito (resultado), atuando como o mecanismo pelo qual a intervenção exerce sua influência. As confundidoras são variáveis que afetam simultaneamente tanto o tratamento quanto o resultado, podendo gerar associações espúrias que distorcem a estimativa do efeito causal real. Essas variáveis, muitas vezes não observáveis, costumam aparecer em diagramas causais como causas comuns de ambas as variáveis principais.

É fundamental também compreender as estruturas de junção causal (PEARL, 2018), que servem como base para a análise de dependências e independências condicionais. Como exemplificado na Figura 4, na estrutura de cadeia (A $\rightarrow$ B $\rightarrow$ C), a variável intermediária B transmite o efeito de A para C, e condicionar em B rompe essa dependência. Na bifurcação (A $\leftarrow$ B $\rightarrow$ C), B é um fator comum que influencia A e C; ao condicionar em B, elimina-se a correlação espúria entre elas. Por fim, na configuração de colisor (A $\rightarrow$ B $\leftarrow$ C), A e C influenciam conjuntamente B, e, diferentemente dos casos anteriores, condicionar em B — ou em qualquer um de seus descendentes — introduz uma dependência artificial entre A e C.

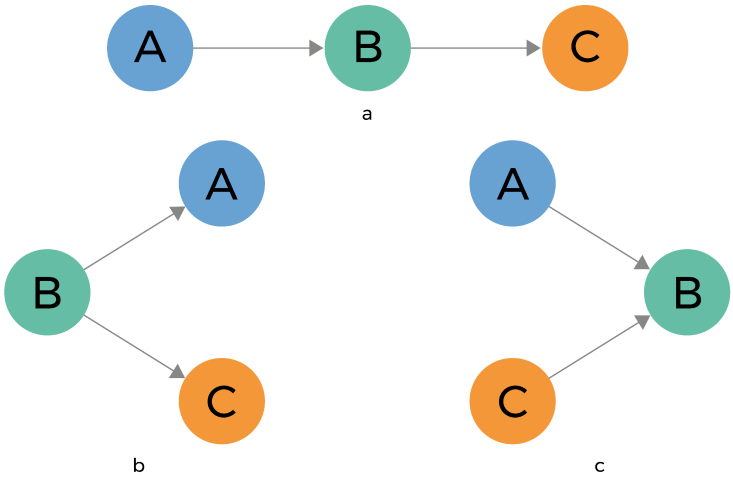

Figura 4 – Estruturas Causais Gerais. Na DAG (a) exemplificação de uma estrutura de cadeia, em (b) estrutura de bifurcação e em (c) uma estrutura de colisão.

Em contextos longitudinais, as relações causais podem ser representadas não apenas por grafos estáticos, mas também por equações estruturais com dependência temporal. Nesses casos, cada variável é influenciada tanto por suas causas contemporâneas quanto por valores passados de si mesma ou de outras variáveis do sistema, caracterizando uma estrutura causal autorregressiva. Esse formalismo permite a simulação de processos dinâmicos e constitui a base da modelagem adotada neste trabalho.

2.5 Dados Longitudinais

Os dados longitudinais são medições repetidas de uma mesma unidade em diferentes pontos de um domínio ordenado. Dentro desta categoria, o tipo mais comum encontrado são as séries temporais, que se referem tipicamente a observações sequenciais de uma única unidade ao longo do tempo, e os dados de painel, que envolvem observações de múltiplas unidades nos mesmos ou diferentes períodos de tempo (DIGGLE et al., 2002).

A análise de dados longitudinais possui um longo histórico de aplicação em diversas áreas de estudo e essa estrutura rica de dados permite investigar, por exemplo, a dinâmica temporal de fenômenos, modelar trajetórias individuais e analisar os efeitos de variáveis que mudam ao longo do tempo, controlando para heterogeneidade não observada entre as unidades, tema central em delineamentos experimentais clássicos e modelos mistos (MONTGOMERY, 2017).

Um dos modelos estatísticos fundamentais aplicados em séries temporais univariadas, que frequentemente servem como base para modelos mais complexos, é o modelo *Autoregressive Integrated Moving Average* (ARIMA). Desenvolvido extensivamente por Box e Jenkins (1970), um modelo ARIMA descreve uma variável dependente como uma função de seus próprios valores passados (componente autorregressivo), dos erros de previsão passados (componente de médias móveis) e da diferenciação dos dados para torná-los estacionários (componente integrado).

A característica autorregressiva de séries temporais é um fator importante considerado nos dados de diversos domínios, como em economia na análise de variáveis macroeconômicas e em finanças na modelagem de retornos, volatilidade e preços de ativos (ENDERS, 2010).

Outro contexto de estudo também crescente que envolve séries temporais é o *Multivariate Time Series* (MTSs). Em MTSs são analisadas as correlações existentes entre séries temporais de duas ou mais variáveis, como por exemplo a análise da relação entre variáveis macroeconômicas, onde podem ser aplicadas ferramentas estatísticas como o *Vector Auto-Regressive* (VAR) e o *Vector Auto Regressive Moving Average* (VARMA) (LÜTKEPOHL, 2005).

A área de *forecasting* com MTSs tem testemunhado um aumento significativo de trabalhos explorando modelos de *deep learning* avançados, como *Graph Neural Networks* (GNNs) e arquiteturas baseadas em *Transformers*, visando aprimorar a acurácia das previsões (MENDIS; WICKRAMASINGHE; MARASINGHE, 2024).

Apesar das diversas ferramentas existentes para abordar problemas com dados longitudinais, ao tratarmos de análise causal temos desafios únicos, como a necessidade de modelar a dependência temporal e controlar a heterogeneidade não observada. Métodos tradicionais como Diferença em Diferenças e Variáveis Instrumentais em Painel têm sido amplamente utilizados e o aumento recente de pesquisas tem permitido o desenvolvimento de novos métodos estatísticos para estimar efeitos causais, particularmente

com intervenções binárias, usando observações de múltiplas unidades ao longo do tempo (ARKHANGELSKY; IMBENS, 2024).

Soluções de *deep learning* também têm sido exploradas em problemas de *causal discovery* em dados longitudinais (KADDOUR et al., 2025), que é o processo de aprender relações de causa e efeito entre variáveis diretamente a partir de dados, sem conhecimento prévio da estrutura causal, com isso oferecendo novas perspectivas para a área.

2.6 Dados Sintéticos Causais

Nos últimos anos, o uso de dados sintéticos para treinamento de modelos tem ganhado destaque especialmente em cenários onde há limitação de dados rotulados e custos elevados de coleta e anotação. Um avanço recente nesse sentido é o MITRA (ZHANG et al., 2025), um *Tabular Foundation Model* desenvolvido no laboratório de pesquisa da Amazon Web Services, que foi pré-treinado exclusivamente com dados sintéticos gerados a partir de uma mistura de *priors* projetados para maximizar diversidade, desempenho e distintividade; um *prior* é uma distribuição geradora ou um mecanismo gerador de dados propriamente dito, que será utilizado para fornecer os dados de treino para um modelo. O estudo demonstra que a eficácia de modelos tabulares depende não apenas da arquitetura, mas, sobretudo, das propriedades estatísticas e estruturais dos geradores utilizados no pré-treino do modelo. Mecanismos geradores baseados em *Structural Causal Models* (SCM) apresentaram desempenho superior em tarefas reais e contribuíram para melhor capacidade de generalização, reforçando a relevância de modelos com estrutura causal explícita para representar distribuições observadas em aplicações práticas.

Apesar de o foco do MITRA não ser a disponibilização dos dados sintéticos gerados, ele traz um ponto diretamente relacionado ao objetivo do CSDG: ao colocar a curadoria de geradores sintéticos como fator central no desenvolvimento de modelos, o trabalho evidencia a importância de investigar como diferentes dinâmicas causais e complexidades funcionais podem impactar o desempenho final do modelo em tarefas especializadas.

De forma similar, o CausalTables.jl (BALKUS; HEJAZI, 2025) representa um avanço relevante na simulação de dados sintéticos com estrutura causal explícita; essa biblioteca desenvolvida para a linguagem Julia oferece funcionalidades para simulação e armazenamento de dados causais em formato tabular, disponibilizando uma interface para definição de modelos causais via SCMs, suporte para operações de intervenção e acesso a quantidades de *ground truth*, como distribuições condicionais e estimativas causais do tipo ATE. Essas funcionalidades tornam a ferramenta uma importante base para o desenvolvimento e avaliação controlada de estimadores causais em dados tabulares. Contudo, assim como em outras abordagens de geração de dados sintéticos, o CausalTables.jl restringe-se a cenários estáticos, sem incorporar dependência temporal entre observações ou relações dinâmicas entre tratamento e resultado. Dessa forma, apesar de sua utilidade para a ava-

liação de métodos causais em dados não sequenciais, ele não contempla efeitos acumulados ou contrafactuais sequenciais, aspectos centrais em aplicações longitudinais. A metodologia proposta neste trabalho complementa esse ecossistema ao possibilitar a simulação de dados longitudinais com estrutura causal explícita, permitindo o estudo de algoritmos de inferência causal em cenários realistas onde intervenções, resultados e covariáveis evoluem ao longo do tempo.

A Tabela 2 sintetiza essa comparação, destacando a lacuna existente no estado da arte quanto ao suporte simultâneo a estrutura causal e dinâmica temporal.

Tabela 2 – Comparação entre abordagens de geração sintética com estrutura causal.

Característica	**MITRA** (ZHANG et al., 2025)	**CausalTables.jl** (BALKUS; HEJAZI, 2025)	**CSDG** (este trabalho)
Tipo de dado gerado	Tabular estático	Tabular estático	Séries longitudinais
Estrutura causal explícita	Sim (via *priors* diversos)	Sim (via SCMs)	Sim (via SCM autorregressivo)
Dependência temporal	Não	Não	Sim
Evolução de tratamento e resposta	Não	Não	Sim
Cenários contrafactuais sequenciais	Não	Parcial	Sim
Controle de ruído e da função de efeito	Parcial	Parcial	Sim
Acesso a *ground truth* causal	Parcial	Sim (condicional e estimandos)	Sim (factual e contrafactual)
Aplicação principal	Pré-treinamento e generalização	Avaliação de estimadores	*Benchmarking*
Domínios exemplares	Aprendizado tabular	Causal Tabular	Causal Longitudinal

A tabela destaca a lacuna existente ao que se refere a geração de cenários sintéticos com estrutura causal, aspecto que o CSDG busca preencher ao permitir intervenções e efeitos acumulados ao longo do tempo para avaliação de métodos de inferência causal em dados longitudinais.

Além das abordagens baseadas em estrutura causal sintética, trabalhos recentes têm explorado a geração de dados longitudinais a partir de dados reais, com foco em utilidade estatística e preservação de privacidade.

No estudo de Bun et al. (2024) vemos uma proposta de geração de dados sintéticos longitudinais a partir de dados médicos reais de pacientes, ao passo que o objetivo é preservar a privacidade dos indivíduos, capturando propriedades estatísticas importantes dos dados que devem ser mantidas e assegurando consistência temporal das trajetórias individuais em estudos observacionais. Apesar desses avanços, tais métodos não modelam explicitamente relações de causa e efeito, nem permitem o controle paramétrico de

intervenções e contrafactuais — limitações que restringem seu emprego em *benchmarking* causal.

Kühnel et al. (2024) também abordam a geração de dados sintéticos longitudinais a partir de dados reais, com foco em estudos nutricionais. Os autores utilizam o método *Variational Autoencoder Modular Bayesian Network* (VAMBN) e o estendem com uma camada de uma rede *Long Short-Term Memory* (LSTM) para modelar dados longitudinais do estudo *Dortmund Nutritional and Anthropometric Longitudinally Designed Study* (DONALD), que é um estudo alemão que acompanha informações sobre dieta e saúde de crianças ao longo do tempo. Essa extensão com LSTM permite ao modelo capturar características de autorregressão nos dados, melhorando a reprodução de dependências temporais. O trabalho de Kühnel et al. (2024) compartilha com a proposta deste trabalho o interesse na geração de dados sintéticos longitudinais; porém, da mesma forma que em Bun et al. (2024), diferenciam-se por focar na preservação de propriedades estatísticas para reproduzir análises do mundo real, enquanto este trabalho concentra-se na geração de dados para avaliação de algoritmos de inferência causal. As diferenças dessas abordagens estão explicitadas na Tabela 3.

Em síntese, embora existam iniciativas relevantes tanto na geração de dados sintéticos para pré-treinamento quanto na construção de bases longitudinais sintéticas para preservação de privacidade, permanece sem solução um componente crucial: a geração de dados longitudinais com mecanismo causal explícito e controle de intervenções para validação empírica de estimadores causais.

Nesse contexto, o presente trabalho diferencia-se ao oferecer a capacidade de especificar mecanismos causais autorregressivos, gerando dados que preservam tanto a dinâmica temporal quanto a coerência causal subjacente, aspectos essenciais para avaliação rigorosa de algoritmos de inferência causal em dados longitudinais.

2.7 Métricas

Em inferência causal, é essencial quantificar não apenas o impacto de uma intervenção, mas também como essa intervenção altera a associação estatística entre tratamento (T) e resultado (Y). Para isso, adotamos duas métricas fundamentadas no arcabouço dos resultados potenciais: o *Average Treatment Effect* (ATE) (RUBIN, 1974), que mensura o efeito causal médio em nível populacional, e a variação de associação Δr (MENG; ROSENTHAL; RUBIN, 1992), que captura mudanças estruturais na relação $\mathrm{T} \leftrightarrow \mathrm{Y}$ após intervenções no tratamento.

Neste trabalho também avaliamos a capacidade de modelos em estimar trajetórias factuais e contrafactuais ao longo do tempo. Nesse contexto, utilizamos o RMSE para mensurar a acurácia preditiva no cenário factual, e o *Precision in Estimation of Heterogeneous Effect* (PEHE) (HILL, 2011), que avalia a qualidade da estimativa de efeitos individuais do tratamento.

Tabela 3 – Comparação entre abordagens de geração de dados sintéticos longitudinais.

Característica	**Bun et al. (2024)** **Kühnel et al. (2024)**	**CSDG** (este trabalho)
Origem / mecanismo gerador	gerados a partir de bases observacionais reais	gerados a partir de estruturas causais autorregressivas
Finalidade principal	Privacidade de dados e preservação de propriedades estatísticas	Avaliação causal (*benchmarking*) de estimadores
Modelo causal explícito	Não	Sim
Controle de intervenções	Não	Sim
Cenários contrafactuais	Não	Sim
Dependência temporal	Sim	Sim
Reprodutibilidade experimental	Limitada	Alta
Acesso ao *ground truth* causal	Não	Sim

A tabela contrasta métodos longitudinais baseados em dados reais, focados em privacidade e preservação de propriedades estatísticas, com a abordagem sintética causal proposta neste trabalho, que combina dinâmica temporal com mecanismos causais explícitos para avaliação de métodos de inferência causal.

Nesta seção, apresentamos as métricas citadas que serão utilizadas ao longo deste trabalho para a avaliação dos dados sintéticos gerados e dos modelos aplicados sobre esses dados.

2.7.1 Efeito médio de Tratamento (ATE)

O efeito médio do tratamento (ATE) quantifica, em média, a influência do tratamento T sobre o resultado Y em uma população. Seu cálculo parte do conceito fundamental de Efeito Individual do Tratamento (ITE), definido no arcabouço de resultados potenciais como:

$$\text{ITE}_i = Y_i(1) - Y_i(0), \tag{4}$$

em que $Y_i(1)$ representa o resultado observado ou simulado para o indivíduo i sob o tratamento, e $Y_i(0)$ o resultado contrafactual caso o mesmo indivíduo não tivesse recebido o tratamento. Em dados observacionais reais, apenas um desses resultados é observável; nos dados sintéticos gerados pelo CSDG, ambos são conhecidos, o que permite estimar diretamente o ITE e, por consequência, o ATE.

O ATE é então definido como a média dos efeitos individuais:

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)), \tag{5}$$

onde N é o número de indivíduos. Essa métrica reflete a diferença média esperada no resultado entre os cenários com e sem intervenção e constitui a medida clássica de efeito

causal em nível populacional.

Para os dados longitudinais gerados, o cálculo é realizado em cada instante temporal t e após a intervenção t_{int}, considerando-se as séries de resultados factuais e contrafactuais. Assim, obtemos um perfil temporal do efeito médio, representado como ATE(t), que permite observar a evolução do impacto do tratamento ao longo do tempo.

Ainda no nosso contexto, é necessário realizar o cálculo para cada cenário contrafactual k e tempo t. Definimos, portanto, o efeito médio no corte transversal como:

$$\text{ATE}(t)^{(k)} = \mathbb{E}_i\Big[\, Y_{i,t}^{(cf,k)} - Y_{i,t}^{(f)} \,\Big] \tag{6}$$

onde o termo à direita representa a média, sobre os indivíduos i, da diferença entre os resultados contrafactuais e factuais no tempo t para o cenário k.

Os valores de $\text{ATE}(t)^{(k)}$ são posteriormente agregados por tipo de intervenção no tratamento e intensidade (δ), possibilitando a comparação entre diferentes regimes de tratamento e níveis de efeito. A partir de $\text{ATE}(t)^{(k)}$, calculamos três medidas complementares:

a) o ATE médio pós-intervenção no cenário k:

$$\overline{\text{ATE}}^{(k)} = \frac{1}{|\mathcal{H}|} \sum_{t \in \mathcal{H}} \text{ATE}(t)^{(k)} \tag{7}$$

b) o ATE final, correspondente a $t = S - 1$;

c) o ATE acumulado (soma dos efeitos no período pós-intervenção), utilizados como visões complementares da magnitude do efeito.

Avaliamos a relação dose–resposta por tipo de intervenção por meio da correlação de *Spearman* entre a intensidade da intervenção (δ) e o efeito médio pós-intervenção ($\overline{\text{ATE}}^{(k)}$). Essa escolha decorre do fato de Spearman medir a monotonicidade da relação entre duas variáveis de maneira não paramétrica, independentemente da forma funcional. Como o CSDG pode gerar tanto padrões monotônicos quanto não monotônicos — especialmente em funções não lineares como senoides — a correlação de Spearman é apropriada por capturar relações ordinais mesmo quando há inversões locais.

2.7.2 Mudança de associação (Δr)

Em dados longitudinais, a relação entre variáveis pode ser analisada sob duas dimensões complementares: a temporal, que descreve a dependência entre observações sucessivas de uma mesma unidade, e a transversal, que expressa a associação entre diferentes unidades observadas em um mesmo instante. Conforme discutido por Shen et al. (2023), essas duas perspectivas — denominadas regressão horizontal e regressão vertical — representam formas equivalentes de avaliação de inferência causal em painéis, diferenciando-se apenas quanto à fonte de variação explorada: a primeira utiliza padrões de correlação no

tempo para o mesmo indivíduo, enquanto a segunda baseia-se em padrões de correlação transversal entre indivíduos no mesmo tempo.

Nesta dissertação, o cálculo de $r(t) = \text{corr}(T_t, Y_t)$ adota precisamente essa perspectiva transversal, permitindo examinar como a associação contemporânea entre tratamento e resultado se comporta ao longo do tempo. Essa abordagem está em consonância com a noção de correlação transversal apresentada por Baltagi, Kao e Peng (2016), que enfatizam a importância de considerar a dependência cruzada entre unidades em um mesmo período para garantir inferências válidas em modelos de painel com correlação serial.

A diferença $\Delta r(t)$ expressa, assim, a variação na dependência entre T e Y após a intervenção, revelando possíveis mudanças no grau de associação causal. Dessa forma, a métrica $\Delta r(t)$ é interpretada aqui como um indicador dinâmico da intensidade e estabilidade da associação causal entre tratamento e resultado.

Para quantificar como as intervenções alteram a associação entre T e Y no conjunto de indivíduos, calculamos as correlações do cenário factual r_f e dos cenários contrafactuais r_{cf} no corte transversal, para cada tempo t:

$$r_f(t) = \text{corr}\Big(T_t^{(f)}, Y_t^{(f)}\Big), \qquad r_{cf}^{(k)}(t) = \text{corr}\Big(T_t^{(cf,k)}, Y_t^{(cf,k)}\Big) \tag{8}$$

e definimos a mudança na correlação Δr causada pelas intervenções como:

$$\Delta r^{(k)}(t) = r_{cf}^{(k)}(t) - r_f(t), \tag{9}$$

onde $r_f(t)$ e $r_{cf}^{(k)}(t)$ são calculados sobre as mesmas unidades amostrais, caracterizando correlações dependentes no sentido discutido por Meng, Rosenthal e Rubin (1992). Assim, $\Delta r^{(k)}(t)$ expressa a variação na dependência $T \leftrightarrow Y$ no tempo t, entre os cenários factual e contrafactual k. A série $\Delta r^{(k)}(t)$ permite identificar os momentos em que a associação $T \leftrightarrow Y$ se altera em função do tratamento.

Ao agregamos por tipo de intervenção - pontual, gradual ou contínua - e intensidade δ, obtemos medidas resumidas como a média no pós-intervenção definida na Equação 10:

$$\overline{\Delta r}^{(k)} = \frac{1}{|\mathcal{H}|} \sum_{t \in \mathcal{H}} \Delta r^{(k)}(t) \tag{10}$$

Essa medida sintetiza o efeito da intervenção na dependência contemporânea entre T e Y, permitindo comparar regimes de intervenção e estruturas causais.

Nos cenários com efeitos causais lineares, a correlação de Pearson é a escolha natural para quantificar a dependência entre T_t e Y_t, pois mede precisamente a força da relação linear entre duas variáveis aleatórias. Sua interpretação é direta, simétrica e compatível com os efeitos lineares impostos pelo gerador, justificando seu uso exclusivo nesses casos.

Nos cenários com efeitos causais não lineares utilizamos a correlação de Chatterjee ξ (CHATTERJEE, 2021). O coeficiente ξ é uma medida de dependência monotônica não paramétrica, ele quantifica quão bem X ordena os valores de Y e possui propriedades

cruciais para este trabalho: (i) vale 0 se e somente se T e Y são independentes; (ii) vale 1 se e somente se Y é função de T, não necessariamente linear; (iii) é invariante a transformações monotônicas marginais; (iv) detecta relações monotônicas mesmo em presença de forte distorção não linear.

Essas características tornam ξ particularmente adequado para os cenários do CSDG, nos quais a dependência entre tratamento e resultado pode assumir formas altamente não lineares, preservando apenas a monotonicidade estrutural.

Bucher e Dette (2024) apresentam críticas ao método e demonstram que o coeficiente ξ não é contínuo sob convergência fraca e que, por consequência, testes de independência e intervalos de confiança uniformes baseados em ξ tendem a ter comportamento estatístico inadequado. Esses resultados, porém, não se aplicam ao uso adotado nesta dissertação, pelas seguintes razões: (i) Neste trabalho ξ é empregado como medida descritiva de associação em dados sintéticos, não como estatística de teste. (ii) A estrutura causal verdadeira é conhecida e controlada pelo gerador, afastando o uso inferencial criticado no artigo. (iii) A aplicação foca exclusivamente em mudanças de associação factual–contrafactual. Isso se enquadra exatamente no tipo de dependência monotônica que ξ foi projetado para medir. (iv) Os cenários são gerados de modo determinístico e controlado, evitando os casos patológicos usados por Bucher e Dette (2024).

Portanto, mesmo reconhecendo essas limitações teóricas, a correlação de Chatterjee continua sendo a métrica ideal para mensurar alterações de associação nos cenários não lineares construídos pelo CSDG.

2.7.3 *Root Mean Square Error* (RMSE)

A qualidade das previsões nos cenários factuais, sem intervenção no tratamento, foi avaliada utilizando o RMSE, visto ser uma métrica comum para avaliação neste contexto (CHENG et al., 2022), calculado entre os valores de referência Y e os preditos $\hat{Y}$ ao longo do horizonte de previsão. Fixamos a origem do horizonte no instante de intervenção t_{int} e expressamos o tempo por $\tau = t - t_{\text{int}}$, com $\tau \geq 0$. O RMSE factual por defasagem é dado por:

$$\text{RMSE}^{(f)}(\tau) \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y^{(f)}_{i,\,t_{\text{int}}+\tau} - \hat{Y}_{i,\,t_{\text{int}}+\tau}\right)^2}, \qquad \tau \geq 0 \tag{11}$$

2.7.4 *Precision in Estimation of Heterogeneous Effect* (PEHE)

Para os cenários contrafactuais, utilizamos o PEHE (HILL, 2011), que avalia o erro na estimativa do efeito individual do tratamento comparando, para cada indivíduo, a diferença entre os resultados sob tratamento e não tratamento com a diferença que o modelo estima para esses dois mundos. Na forma clássica, sem índice temporal, o PEHE é definido por:

$$\epsilon_{\text{PEHE}} \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big\{\big[\,Y_i(1)-Y_i(0)\,\big] \;-\; \big[\,\hat{Y}_i(1)-\hat{Y}_i(0)\,\big]\Big\}^2 \tag{12}$$

Como o problema é longitudinal, introduzimos a noção de ITE ao longo do tempo, indexando os resultados por defasagem $\tau = t - t_{\text{int}}$ em relação ao instante da intervenção t_{int}. Para cada $\tau \geq 0$ (pós-intervenção), definimos o ITE verdadeiro do indivíduo i no instante $t_{\text{int}} + \tau$ como:

$$\text{ITE}_i(\tau) \;=\; Y^{(cf,k)}_{i,\,t_{\text{int}}+\tau} \;-\; Y^{(f)}_{i,\,t_{\text{int}}+\tau} \tag{13}$$

onde $Y^{(f)}$ é o resultado factual (sem intervenção aplicada naquele indivíduo e tempo) e $Y^{(cf,k)}$ é o resultado contrafactual sob o cenário k (tipo de intervenção e dose δ). O ITE estimado pelo modelo é:

$$\widehat{\text{ITE}}_i(\tau) \;=\; \hat{Y}^{(cf,k)}_{i,\,t_{\text{int}}+\tau} \;-\; \hat{Y}^{(f)}_{i,\,t_{\text{int}}+\tau} \tag{14}$$

Com isso, obtemos a versão temporal do PEHE por defasagem τ, escrita diretamente em termos dos ITEs:

$$\epsilon_{\text{PEHE}}(\tau) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big\{\text{ITE}_i(\tau) \;-\; \widehat{\text{ITE}}_i(\tau)\Big\}^2, \qquad \tau \geq 0 \tag{15}$$

2.8 Considerações Finais

Este capítulo apresentou os fundamentos teóricos que sustentam o desenvolvimento do CSDG e os experimentos conduzidos neste trabalho. Iniciamos revisando os princípios da Inferência Causal, destacando desde as primeiras formulações de Wright e Neyman até a consolidação moderna dos modelos causais estruturais (SCMs) e dos diagramas acíclicos dirigidos (DAGs). Em seguida, discutimos o arcabouço dos Resultados Potenciais e o *Potential Outcomes Framework*, enfatizando sua relevância para definir efeitos contrafactuais e formalizar estimandos como ITE, ATE e CATE, essenciais para análise causal tanto teórica quanto empírica.

A seção seguinte apresentou um panorama das principais abordagens para estimação de efeitos causais em dados observacionais, abrangendo métodos não paramétricos, modelos estatísticos clássicos, técnicas bayesianas, métodos baseados em árvores e avanços recentes que empregam arquiteturas de *deep learning*, com destaque para modelos recorrentes e abordagens baseadas em *Transformers*. Abordamos também modelos dedicados à predição contrafactual ao longo do tempo, como RMSN, CRN, G-Net e o *Causal Transformer*, destacando seus avanços e limitações.

Posteriormente, discutimos diferentes estruturas causais, evidenciando como distintos padrões estruturais afetam a propagação de efeitos e a interpretação causal. Essa discussão forneceu a base conceitual necessária para compreender o comportamento dinâmico

adotado no CSDG. Na sequência, apresentamos a natureza dos dados longitudinais e sintetizamos como dados reais e sintéticos diferem em termos de controle causal, dependência temporal e disponibilidade de *ground truth*, destacando a importância de mecanismos autorregressivos para cenários realistas.

Finalmente, detalhamos as métricas utilizadas ao longo da pesquisa: o ATE e a mudança de associação Δr, que permitem avaliar diretamente os efeitos das intervenções nos sistemas causais simulados; e as métricas RMSE e PEHE, fundamentais para mensurar a capacidade preditiva e a qualidade das estimativas contrafactuais dos modelos. A escolha das correlações de Spearman, Pearson e Chatterjee foi justificada conforme a natureza estrutural dos cenários — respectivamente testes de monotonicidade, efeitos lineares e não lineares — assegurando coerência entre o mecanismo gerador e a medida estatística empregada.

Em conjunto, os elementos apresentados neste capítulo estabelecem o arcabouço conceitual necessário para compreender o processo de geração dos dados sintéticos e os métodos utilizados para avaliá-los. Eles constituem o alicerce teórico sobre o qual os resultados experimentais do restante da dissertação foi desenvolvido e interpretado.

Capítulo 3

Causal Synthetic Data Generator

A proposta apresentada neste trabalho consiste em um gerador que simula fontes de dados com relação causal entre si a partir de equações estruturais causais autorregressivas que modelam o comportamento dessas variáveis ao longo do tempo.

As estruturas causais disponíveis para geração foram definidas com base em padrões disponíveis na literatura, e a saída do gerador é composta pelas séries sintéticas que incluem tratamentos e resultados e que podem incluir covariáveis e contrafactuais.

3.1 Composição da Estrutura de Geração

As variáveis T e Y são modeladas como processos autorregressivos, nos quais o valor atual depende do valor observado no período anterior, controlado por um coeficiente temporal Φ e um termo de erro aleatório. Essa modelagem captura a dinâmica temporal comum em séries longitudinais reais:

$$Y_t = \Phi_Y Y_{t-1} + \dots \tag{16}$$

As funções $f(\cdot)$ e $g(\cdot)$ determinam a natureza da relação entre variáveis. Quando a relação é linear, temos $f(x) = x$. Para relações não lineares, podem ser utilizadas funções como quadrática x^2, oscilatória $\sin(x)$ ou logarítmica $\log(1 + x^2)$ com crescimento suave, definidas por parametrização ou atribuídas aleatoriamente no momento da geração dos dados.

Todos os termos de erro ε são amostrados de uma distribuição probabilística definida pelo usuário, podendo ser Normal ou Uniforme, conforme parametrização do gerador. Assim, a distribuição do ruído é selecionada externamente e aplicada de forma independente a cada iteração:

$$\varepsilon \sim \begin{cases} \mathcal{N}(0, \sigma^2), & \text{se } \texttt{noise\_dist} = \texttt{"normal"}, \\ \text{Uniform}(a, b), & \text{se } \texttt{noise\_dist} = \texttt{"uniform"}. \end{cases} \tag{17}$$

Onde a e b são, respectivamente, os limites inferior e superior da distribuição uniforme, enquanto σ^2 representa a variância da distribuição normal. Esses parâmetros são definidos no arquivo de configuração e determinam a amplitude ou a dispersão do ruído aplicado. Em ambos os casos, o ruído é amostrado de forma independente para cada indivíduo, variável e instante temporal.

3.2 Estruturas Causais Implementadas

A seguir, são descritas as estruturas causais implementadas, cada uma com suas respectivas equações geradoras. As variáveis de tratamento (T) e resultado (Y) possuem em comum em todas as estruturas uma dinâmica temporal autorregressiva, porém são influenciadas de forma diferente.

3.2.1 Estrutura Causal Direta

O tratamento (T) não é influenciado por outras variáveis no modelo, enquanto o resultado (Y) é influenciado diretamente por T, caracterizando a estrutura causal $T \rightarrow Y$ como visto na Figura 5. Um exemplo que poderia ser modelado por essa estrutura seria o efeito da dose de um medicamento (T) sobre a pressão arterial (Y), sem considerar outros fatores.

Figura 5 – DAG de Estrutura Causal Direta.

O comportamento do tratamento nessa estrutura é modelado pela Equação 18. Nela, o coeficiente temporal Φ_T atua sobre o valor do tratamento no passo anterior, determinando o grau de dependência temporal do processo. O termo de erro ε_{T_t} representa as variações aleatórias não explicadas pelo modelo.

O resultado é descrito pela Equação 19. O primeiro termo, ponderado por Φ_Y, reflete a componente autorregressiva de Y_t, enquanto o segundo termo expressa a influência causal do tratamento sobre o resultado, sendo β_{TY} o coeficiente que controla a intensidade dessa relação e $f(T_t)$ a função que define a complexidade do efeito do tratamento. Por fim, o termo de erro ε_{Y_t} representa o ruído associado à variável de resultado.

$$T_t = \Phi_T T_{t-1} + \varepsilon_{T_t} \tag{18}$$

$$Y_t = \Phi_Y Y_{t-1} + \beta_{TY} f(T_t) + \varepsilon_{Y_t} \tag{19}$$

A estrutura causal do tipo Direta gera duas séries, uma de tratamento e outra de resultado como pode ser visto no exemplo na Figura 6.

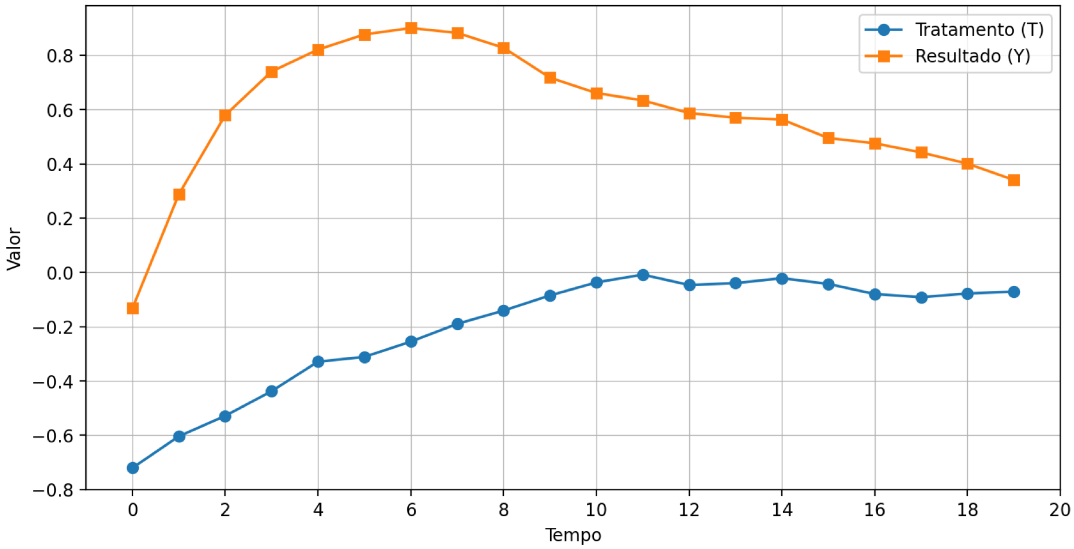

Figura 6 – Exemplo gerado a partir da estrutura Direta com efeito não linear.

3.2.2 Estrutura Causal Cadeia

Assim como na estrutura anterior, o tratamento (T) não sofre outras influências, porém o tratamento influencia o resultado (Y) por meio de uma covariável mediadora (X), formando a estrutura causal $T \rightarrow X \rightarrow Y$ vista na Figura 7. Como exemplo, podemos ter o efeito da prática de atividade física (T) sobre o nível de colesterol (Y), mediado pela perda de peso (X).

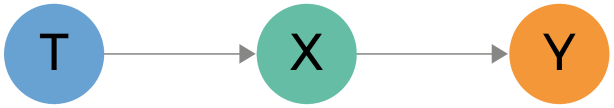

Figura 7 – DAG de Estrutura Causal em Cadeia.

Nesta estrutura, o tratamento descrito na Equação 20 também evolui no tempo de forma autorregressiva, com o coeficiente temporal Φ_T capturando a persistência de T_t a partir do valor imediatamente anterior, enquanto o termo de erro ε_{T_t} representa as variações aleatórias não modeladas.

A covariável mediadora X_t é determinada pelo tratamento contemporâneo, conforme a Equação 21. O coeficiente β_{TX} controla a intensidade do efeito causal de T_t sobre X_t, enquanto $f(T_t)$ define a complexidade da associação. O termo de erro ε_{X_t} agrega as variações não modeladas em X_t.

O resultado Y_t combina uma componente autorregressiva, ponderada por Φ_Y, com a influência do tratamento mediada por X_t, como expresso na Equação 22. Nessa equação,

β_{XY} quantifica a força do efeito de X_t sobre Y_t e $f(X_t)$ descreve a complexidade dessa associação, enquanto ε_{Y_t} resume as perturbações aleatórias do resultado.

$$T_t = \Phi_T T_{t-1} + \varepsilon_{T_t} \tag{20}$$

$$X_t = \beta_{TX} f(T_t) + \varepsilon_{X_t} \tag{21}$$

$$Y_t = \Phi_Y Y_{t-1} + \beta_{XY} f(X_t) + \varepsilon_{Y_t} \tag{22}$$

Na estrutura causal Cadeia temos três séries, a de tratamento, a de resultado e a da covariável no papel de mediadora, como pode ser visto no exemplo da Figura 8.

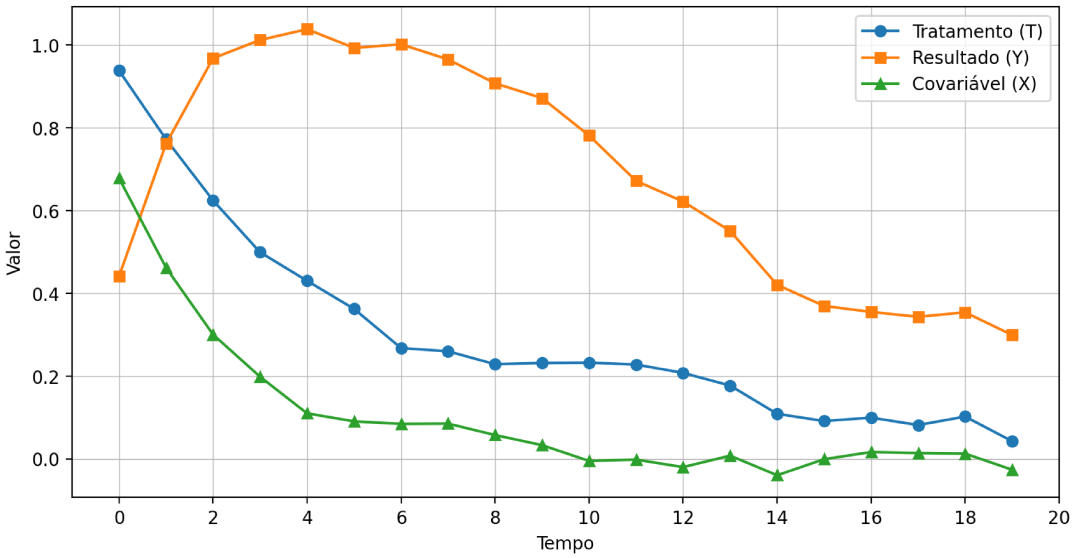

Figura 8 – Exemplo gerado a partir da estrutura Cadeia com efeito não linear.

3.2.3 Estrutura Causal Confundidor

Nesta estrutura, uma variável (X), gerada aleatoriamente e de forma independente a partir de uma distribuição uniforme, atua como um confundidor, influenciando tanto o tratamento (T) quanto o resultado (Y), introduzindo um viés de confundimento na associação observada entre essas variáveis. O tratamento também influencia diretamente o resultado, resultando na estrutura $X \to T \to Y$, $X \to Y$ como visto na Figura 9. Um exemplo seria o nível de estresse diário (X), que pode influenciar tanto a decisão de praticar exercícios (T) quanto a qualidade do sono (Y), sendo que a qualidade do sono também é influenciada pela prática de exercícios.

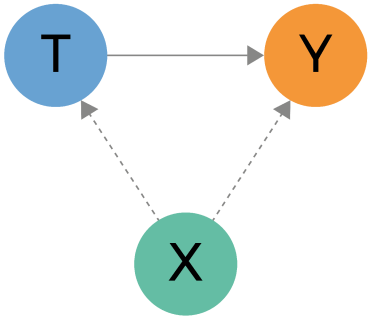

Figura 9 – DAG de Estrutura Causal Confundidor.

A variável confundidora X_t é gerada conforme a Equação 23, sendo amostrada de uma distribuição uniforme $U(a, b)$, o que representa uma fonte externa de variação independente no tempo.

O tratamento T_t, descrito pela Equação 24, possui componente autorregressivo ponderado por Φ_T, refletindo sua persistência temporal, e um termo adicional $\beta_{XT} f(X_t)$ que expressa a influência causal exercida por X_t sobre o tratamento. O termo de erro ε_{T_t} representa as variações aleatórias não explicadas pelo modelo.

O resultado Y_t, definido pela Equação 25, combina quatro componentes: (i) uma parte autorregressiva ponderada por Φ_Y; (ii) o efeito do confundidor X_t, controlado por β_{XY} e mediado pela função $f(X_t)$; (iii) o efeito direto do tratamento T_t, controlado por β_{TY} e modelado pela função $g(T_t)$; e (iv) o termo de erro ε_{Y_t}, que representa o ruído associado à variável de resultado.

$$X_t \sim U(a, b) \tag{23}$$

$$T_t = \Phi_T T_{t-1} + \beta_{XT} f(X_t) + \varepsilon_{T_t} \tag{24}$$

$$Y_t = \Phi_Y Y_{t-1} + \beta_{XY} f(X_t) + \beta_{TY} g(T_t) + \varepsilon_{Y_t} \tag{25}$$

Na estrutura Confundidor também temos três séries, a de tratamento, a de resultado e a da covariável que agora assume o papel de confundidora, como pode ser visto no exemplo da Figura 10.

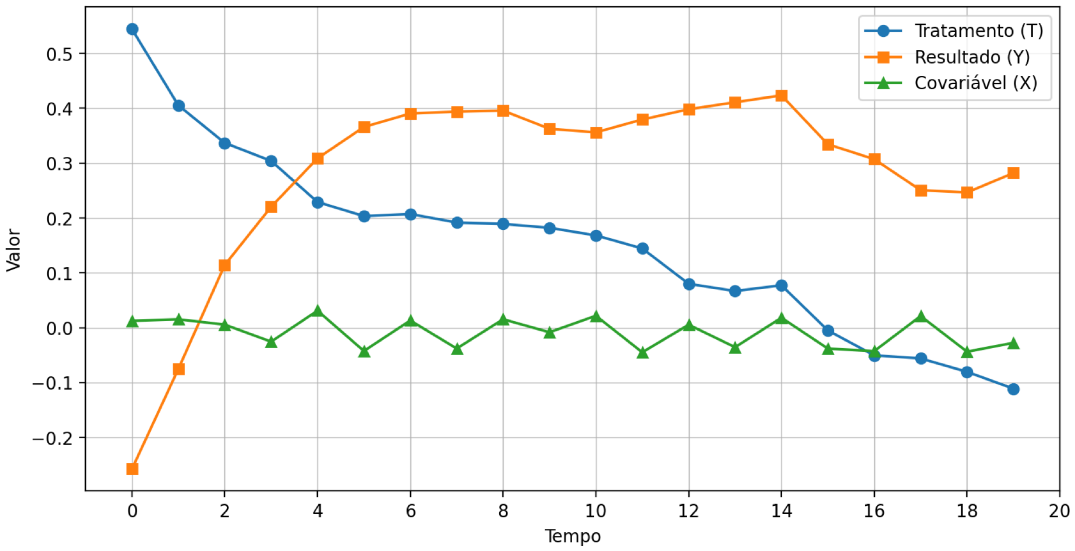

Figura 10 – Exemplo gerado a partir da estrutura Confundidor com efeito não linear.

3.3 Cenários Contrafactuais

Para permitir a avaliação de métodos em cenários contrafactuais, o gerador simula mudanças no tratamento a partir de um ponto de intervenção no tempo t_{int}, criando trajetórias alternativas de tratamento e resultado.

Na estrutura causal Direta por exemplo, as variáveis contrafactuais são geradas a partir da Equação 26 para o tratamento e da Equação 27 para o resultado:

$$T_t^{cf} = \Phi_T T_{t-1}^{cf} + \varepsilon_{T_t} + \delta_t \tag{26}$$

$$Y_t^{cf} = \begin{cases} Y_t, & \text{se } t < t_{int} \\ \Phi_Y Y_{t-1}^{cf} + \beta_{TY} f(T_t^{cf}) + \varepsilon_{Y_t}, & \text{se } t \geq t_{int} \end{cases} \tag{27}$$

O resultado contrafactual Y_t^{cf} é idêntico ao resultado factual até o ponto de intervenção, sendo alterado apenas a partir de t_{int}, quando o tratamento passa a ser modificado pela intervenção δ_t. Um exemplo de dado contrafactual gerado pode ser visto na Figura 11.

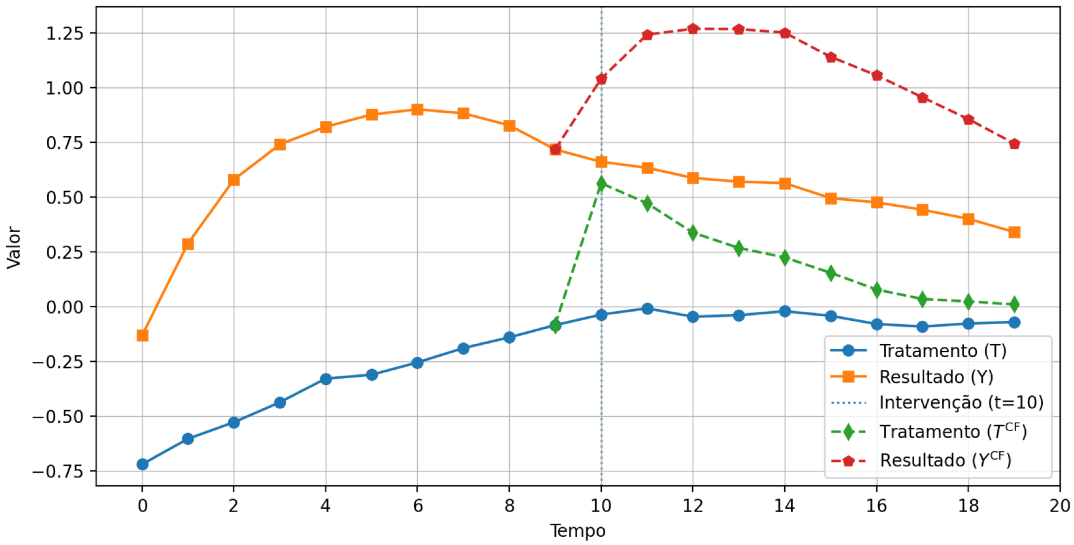

Figura 11 – Exemplo gerado de dado contrafactual. Cenário contrafactual de um indivíduo de uma estrutura Direta; em t_{int}=10 é realizada uma intervenção no Tratamento, gerando a partir desse ponto um tratamento e um resultado contrafactual que divergem do dado factual.

O gerador suporta diferentes tipos de intervenção, descritos a seguir.

Intervenção Pontual. A intervenção é aplicada apenas no instante t_{int}, onde δ assume um valor α que é a intensidade da intervenção a ser aplicada; nos demais períodos δ é 0, não sendo então aplicadas novas intervenções:

$$\delta_t = \begin{cases} \alpha, & \text{se } t = t_{int} \\ 0, & \text{se } t \neq t_{int} \end{cases} \tag{28}$$

Intervenção Contínua. A intervenção é aplicada a partir de t_{int} continuamente, ou seja, δ assume o valor α a partir de t_{int} até o fim da série:

$$\delta_t = \begin{cases} 0, & \text{se } t < t_{int} \\ \alpha, & \text{se } t \geq t_{int} \end{cases} \tag{29}$$

Intervenção Gradual. A intervenção é aplicada gradualmente a partir de t_{int}, ou seja, inicialmente δ assume uma fração de α, e a cada período é incrementado, atingindo valor máximo de α apenas no fim da série:

$$\delta_t = \begin{cases} 0, & \text{se } t < t_{int} \\ \alpha \frac{t - t_{int}}{k}, & \text{se } t_{int} \leq t \end{cases} \tag{30}$$

3.4 Geração dos Dados

O processo de geração dos dados sintéticos no CSDG é controlado por meio de um arquivo de configuração no formato YAML, que define todos os parâmetros necessários para a execução do gerador. Essa abordagem declarativa garante reprodutibilidade e flexibilidade, permitindo que diferentes cenários sejam simulados apenas informando o arquivo de configuração no momento da execução, sem necessidade de alterações no código-fonte. Os parâmetros disponíveis para configuração no arquivo estão listados na Tabela 4.

Tabela 4 – Parâmetros definidos no arquivo de configuração YAML do CSDG

Parâmetro	**Descrição**
name	Referência do *dataset* na saída
seed	Semente de aleatoriedade
n	Número de indivíduos (séries independentes)
t	Comprimento temporal das séries
structure_type	Estrutura causal (*direct*, *chain*, *confounder*)
nonlinear	Uso de funções não lineares nas relações causais
normalize	Normalização temporal (*z-score*)
effects.ty	Função de efeito de T → Y
effects.tx	Função de efeito de T → X
effects.xt	Função de efeito de X → T
effects.xy	Função de efeito de X → Y
phi_T	Persistência temporal do tratamento
phi_Y	Persistência temporal do resultado
betas.ty	Coeficiente causal T → Y
betas.tx	Coeficiente causal T → X
betas.xy	Coeficiente causal X → Y
betas.xt	Coeficiente causal X → T
*noise*_dist	Tipo da distribuição do ruído (*uniform* ou *normal*)
sigma_T	Desvio padrão do ruído de T (se normal)
sigma_Y	Desvio padrão do ruído de Y (se normal)
sigma_X	Desvio padrão do ruído de X (se normal)
noise_T_range	Intervalo do ruído de T (se uniforme)
noise_Y_range	Intervalo do ruído de Y (se uniforme)
noise_X_range	Intervalo do ruído de X (se uniforme)
confounder_mode	Tipo de geração da covariável X (*iid* ou *ar1*)
phi_X	Persistência temporal de X (se *confounder_mode*=ar1)
interventions	Lista de intervenções (tipo, tempo e intensidade)
split.train	Proporção destinada ao treino
split.val	Proporção destinada à validação
split.test	Proporção destinada ao teste

O arquivo é composto por blocos de parâmetros agrupados conforme suas finalidades:

- **Dados principais:** definem o nome do *dataset* (`name`), o número de indivíduos (`n`) e o comprimento das séries (`t`). O campo `structure_type` seleciona a estrutura causal desejada, *direct*, *chain* ou *confounder*. Os parâmetros `nonlinear` e

normalize controlam, respectivamente, a inclusão de não-linearidades nas equações e a normalização dos valores gerados.

- **Funções de efeito (`effects`):** os termos `ty`, `tx`, `xy` e `xt` indicam quais funções serão utilizadas na aplicação do efeito entre as variáveis. As opções são: *linear* - linear -, *quadratic* - quadrado - , *sine* - seno - e *logarithmic* - logaritmo natural. As funções especificadas são utilizadas caso *nonlinear=True* , se não forem especificadas são sorteadas aleatoriamente.

- **Coeficientes autorregressivos:** os termos `phi_T` e `phi_Y` correspondem aos coeficientes autorregressivos das variáveis de tratamento e resultado. Caso não sejam especificados, são sorteados aleatoriamente dentro de intervalos predefinidos, garantindo variabilidade entre instâncias.

- **Coeficientes causais (`betas`):** determinam as forças de causalidade entre as variáveis, variando conforme a estrutura selecionada. Por exemplo, em uma estrutura Direta, apenas o parâmetro `ty` (efeito de T_t sobre Y_t) é utilizado, enquanto nas estruturas *chain* e *confounder* os termos adicionais `tx`, `xy` e `xt` são ativados conforme o grafo causal correspondente.

- **Configuração dos ruídos:** o parâmetro `noise_dist` define a distribuição probabilística dos termos de erro ($\varepsilon_{T_t}, \varepsilon_{Y_t}, \varepsilon_{X_t}$). Pode-se escolher entre distribuição uniforme ou normal, com intensidade controlada pelos parâmetros `sigma` ou pelos intervalos `noise_range`. Essa configuração influencia diretamente o sinal-ruído (SNR) do sistema gerado.

- **Modo de geração do confundidor:** quando a estrutura envolve uma variável confundidora X_t, o parâmetro `confounder_mode` define se a série será independente e identicamente distribuída (`iid`) ou autorregressiva de primeira ordem (`ar1`). O coeficiente `phi_X` controla a persistência temporal nesse último caso.

- **Intervenções:** o bloco *interventions* especifica o conjunto de intervenções aplicadas ao tratamento T_t. Cada elemento define o tipo de intervenção (`pontual`, `gradual` ou `contínua`), o instante de aplicação (`t_interv`) e a intensidade da mudança (`delta_T`). Essa parametrização permite explorar o comportamento do sistema sob diferentes magnitudes e durações de intervenção, o que é fundamental para avaliar a resposta causal e a relação dose–efeito.

- **Divisão dos dados:** a chave *split* define as proporções destinadas aos conjuntos de treino, validação e teste, garantindo consistência experimental nas etapas de modelagem e análise.

Durante a execução, o gerador lê o arquivo YAML, inicializa o estado aleatório conforme a semente (***seed***), instancia as equações estruturais conforme a estrutura causal definida e aplica as intervenções configuradas.

Os dados gerados pelo CSDG são armazenados no formato *NumPy Zip* (NPZ), um contêiner compactado que agrupa múltiplos arrays nomeados e metadados em um único arquivo. Esse formato oferece (i) eficiência de entrada e saída (I/O), (ii) compressão transparente via `zip`, e (iii) portabilidade direta dentro do ecossistema Python, sendo suportado nativamente por bibliotecas como NumPy, SciPy, Scikit-learn, PyTorch e TensorFlow. Dessa forma, cada partição do conjunto de dados — ***train***, ***val*** e ***test*** — é salva como um arquivo `.npz` contendo tanto os arrays principais (***treatments***, ***outcomes***, ***covariates***) quanto suas versões contrafactuais e respectivos metadados de configuração.

Além de sua eficiência e compatibilidade, o NPZ preserva os tipos numéricos e as formas (***shape***) dos *arrays*, evitando ambiguidades comuns em formatos tabulares como CSV, em que pode ocorrer perda de precisão, ordenação ou cabeçalhos inconsistentes. A estrutura em chaves nomeadas também simplifica o acesso seletivo a subconjuntos específicos de dados — como séries factuais, contrafactuais ou variáveis auxiliares — mantendo, no mesmo contêiner, os parâmetros e condições experimentais responsáveis pela geração do *dataset*. Essa característica é fundamental para garantir reprodutibilidade e rastreabilidade dos experimentos.

Por fim, embora o NPZ seja um formato nativo do NumPy, sua utilização é amplamente compatível com outras linguagens e ambientes científicos. Em Julia, pode ser lido diretamente via o pacote `NPZ.jl`; em R ou MATLAB, o acesso pode ser realizado por meio de *bindings* Python, como `reticulate` e *MATLAB Engine*. Em cenários onde o suporte direto não está disponível, os arquivos podem ser convertidos de forma simples para formatos intercambiáveis como CSV ou *Parquet*, preservando a estrutura e o conteúdo dos dados originais.

Convenção de nomes

Os arquivos são gravados seguindo o padrão:

```
{name}_{structure}_{nonline}_{t}p_{n}n_{split}.npz
```

onde:

- ***name***: identificador do *dataset* (ex.: `synthetic1`);
- ***structure***: estrutura causal (***direct***, ***chain***, ***confounder***);
- `nonlin`: indicação de linearidade (***linear*** ou ***nonlinear***);
- `t`: número de passos temporais, com sufixo `p` (ex.: `20p` $\Rightarrow t = 20$);

- `n`: número de indivíduos, com sufixo `n` (ex.: `250n` $\Rightarrow n = 250$);
- *`split`*: partição (*`train`*, *`val`* ou *`test`*).

Esse esquema torna o arquivo auto-descritivo: apenas pelo nome é possível identificar a estrutura causal, a presença de não-linearidades, os tamanhos (n, t) e a partição.

Estrutura interna

A função de geração retorna um dicionário cujas chaves são serializadas no `NPZ`. A Tabela 5 resume os principais campos, suas formas (dimensionalidade) e significado. O primeiro bloco contém as chaves com os dados gerados e o segundo os metadados de geração.

Tabela 5 – Estrutura dos arquivos gerados: chaves, formas e descrição.

Chave	**Forma (*shape*)**	**Descrição**
`treatments`	(n, t)	Tratamento factual T.
`outcomes`	(n, t)	Resultado factual Y.
`covariates`	(n, t)	Covariável X;
`treatments_cf`	(k, n, t)	Tratamento contrafactual.
`outcomes_cf`	(k, n, t)	Resultado contrafactual.
`covariates_cf`	(k, n, t)	Covariável contrafactual.
`interventions`	lista de tuplas	Especificação das intervenções.
`phi_T`, `phi_Y`, `phi_X`	decimal	Coeficientes autorregressivos.
`betas_ty`, *`betas_tx`*, *`betas_xt`*, *`betas_xy`*	decimal	Coeficientes causais.
`noise_dist`	*string*	Distribuição dos ruídos.
`sigma_T`, `sigma_Y`, `sigma_X`	decimal	Intensidades do ruído para *`normal`*.
`noise_T_range`, *`noise_Y_range`*, *`noise_X_range`*	pares (ℓ, u)	Intervalos do ruído para *`uniform`*.
`structure_type`	*string*	Estrutura causal.
`normalize`	booleanos	Sinaliza normalização.
`nonlinear`	booleanos	Sinaliza não linearidade nos efeitos.
`effect_ty`, *`effect_tx`*, *`effect_xt`*, *`effect_xy`*	*string*	Funções de efeito de cada relação causal.
`confounder_mode`	*string*	`iid` ou `ar1`.
`seed`	inteiro	Semente para reprodutibilidade.

Observações importantes:

- k é o número de intervenções definidas em `config["interventions"]`; cada fatia `[i, :, :]` em `_cf` corresponde a um cenário distinto.

- ❏ As formas `(n, t)` e `(k, n, t)` facilitam operações vetorizadas para métricas por indivíduo ou por cenário.
- ❏ Quando *`normalize=true`*, a normalização é aplicada de modo consistente entre factual e contrafactual.

3.4.1 Considerações finais

O gerador proposto neste capítulo consolida uma abordagem formal, flexível e reprodutível para a criação de dados sintéticos longitudinais com estrutura causal explícita. Ao integrar dependências temporais autorregressivas, efeitos causais contemporâneos e diferentes formas funcionais entre tratamento, resultado e covariáveis, o CSDG permite simular, de forma controlada, cenários compatíveis com dinâmicas reais observadas em processos longitudinais. As três estruturas causais implementadas — Direta, Cadeia e Confundidor — abrangem padrões fundamentais da literatura, possibilitando investigar efeitos diretos, mediadores e confundidores ao longo do tempo. Exemplos gráficos apresentados ao longo do capítulo ilustram que o gerador produz trajetórias realistas e coerentes com as propriedades esperadas de cada estrutura, conforme visto nas Figuras 6, 8 e 10.

A capacidade de gerar trajetórias factuais e contrafactuais, aliada ao suporte para diferentes tipos de intervenção - pontual, gradual e contínua - estabelece um ambiente experimental rico para avaliar algoritmos de inferência causal sob múltiplas condições estruturais. O uso de funções de efeito configuráveis, coeficientes autorregressivos e intensidades de intervenção permite ajustar a complexidade do sistema simulado, tornando o gerador aplicável tanto a estudos exploratórios quanto a experimentos metodológicos mais exigentes.

A seção dedicada ao processo de geração descreveu ainda o papel do arquivo de configuração YAML, que oferece uma interface declarativa e reprodutível para controle de todos os elementos do gerador, desde coeficientes autoregressivos e funções de efeito até ruídos estruturais e faixas de variabilidade. Essa abordagem amplia a flexibilidade experimental, permitindo que cenários diversos sejam produzidos sem necessidade de alterações diretas no código.

Em síntese, essas contribuições reforçam o papel do CSDG como uma ferramenta aberta, reprodutível e metodologicamente fundamentada para experimentação causal em contextos longitudinais. O gerador permite a criação de cenários realistas com *ground truth* causal conhecido e oferece um ambiente controlado para testar e comparar algoritmos de aprendizado causal. Ao integrar estruturas causais explícitas, dinâmicas autorregressivas e diferentes regimes de intervenção, o CSDG estabelece uma base consistente para estudos que exigem controle rigoroso das relações causais e do comportamento temporal das variáveis.

Capítulo 4

Análise do CSDG

Neste capítulo temos como objetivo demonstrar que o CSDG produz dados que respeitam princípios de consistência causal, realismo estatístico e reprodutibilidade de efeitos esperados sob manipulação das variáveis.

Buscou-se observar empiricamente o comportamento dinâmico dos efeitos do tratamento sob diferentes tipos e intensidades de intervenção, e foram verificados se as propriedades estatísticas e causais dos cenários simulados se comportam conforme o esperado a partir das estruturas definidas.

Essa etapa é essencial para validar a coerência do gerador e confirmar se as relações causais impostas pelas estruturas e pelos parâmetros se traduzem em padrões observáveis de efeito médio e de associação entre variáveis ao longo do tempo.

4.1 Método de Análise

Analisamos os diferentes cenários de dados sintéticos gerados pelo CSDG sob duas perspectivas complementares: (i) o efeito médio do tratamento sobre o resultado (ATE) e (ii) a mudança de associação entre tratamento e resultado provocada pelas intervenções (Δ_r). O objetivo é verificar se as estruturas causais e funções de efeito especificadas no gerador se manifestam de forma consistente nas propriedades observadas dos dados.

Em todas as análises, alinhamos o tempo em $\tau = t - t_{\text{int}}$, definindo a defasagem temporal relativa à intervenção, de modo que $\tau = 0$ representa o instante da intervenção, $\tau < 0$ indica períodos anteriores e $\tau > 0$ períodos posteriores.

Padronizamos T e Y por tempo utilizando *z-score*, onde para cada instante t subtraímos a média e dividimos pelo desvio-padrão calculados sobre todos os indivíduos. Essa etapa garante comparabilidade entre períodos, elimina efeitos de escala e assegura que as medidas de dependência reflitam exclusivamente a covariação relativa entre indivíduos.

As análises foram estratificadas por tipo de intervenção e por intensidade aplicada (δ). No estudo da mudança de associação Δr, utilizamos métricas de dependência escolhidas de acordo com o caráter funcional dos cenários. Correlação de Pearson foi aplicado

aos cenários lineares, por medir com precisão a variação na componente linear da associação e correlação de Chatterjee foi aplicado aos cenários não lineares, por responder adequadamente às transformações introduzidas pelos mecanismos funcionais empregados.

A mudança de associação foi avaliada comparando-se diretamente as curvas de dependência factual e contrafactual ao longo da defasagem temporal, permitindo observar como a intervenção altera a relação entre T e Y em cada cenário. Apresentamos também a média temporal $\mathbb{E}[\Delta r(\tau)]$ acompanhada de intervalos de confiança obtidos por reamostragem, a fim de sintetizar o comportamento agregado após a intervenção sempre que essa representação contribui para a interpretação dos resultados.

Por convenção, valores positivos de $\Delta r^{(k)}(t)$ indicam aumento da associação $T \leftrightarrow Y$ no contrafactual em comparação ao factual, enquanto valores negativos indicam redução dessa dependência.

4.2 Dados e cenários analisados

A avaliação conduzida neste capítulo baseia-se em um conjunto ampliado de dados sintéticos gerados pelo CSDG, contemplando múltiplas combinações estruturais e funcionais, com intuito de investigar como diferentes mecanismos causais e formas de não-linearidade influenciam o comportamento das métricas de interesse. Foram construídos nove conjuntos de dados, variando-se (i) a estrutura causal subjacente, (ii) a complexidade funcional dos efeitos entre as variáveis e (iii) o tipo de função não linear empregada.

Todos os *datasets* compartilham a mesma configuração factual: cada indivíduo possui uma série temporal de $T = 20$ períodos, com intervenção aplicada em $t_{\text{int}} = 10$, e população total de $N = 250$ indivíduos. Para cada indivíduo são gerados $k = 30$ cenários contrafactuais, correspondentes às combinações dos três tipos de intervenção com dez intensidades δ variando de 0,1 a 1,0 em incrementos de 0,1.

Os parâmetros de geração foram mantidos fixos entre todos os conjuntos de dados, conforme a Tabela 6, de modo a isolar o efeito das estruturas causais e das funções de efeito.

Os coeficientes temporais foram definidos como $\phi_T = \phi_Y = 1.0$, produzindo séries totalmente persistentes. Essa escolha elimina a influência de componentes autorregressivas e faz com que qualquer variação observada ao longo do horizonte decorra exclusivamente das relações causais especificadas. Dessa forma, as diferenças entre cenários passam a refletir diretamente os mecanismos estruturais e as funções de efeito, e não dinâmicas temporais intrínsecas às séries.

Além disso, os coeficientes β_{TY} e β_{TX} foram mantidos iguais em todas as estruturas, assegurando que o tratamento conserve a mesma força causal na relação com Y e com X, independentemente da estrutura causal.

Tabela 6 – Parâmetros globais de geração dos dados empregados nos nove *datasets*.

Coeficientes temporais	$\phi_T = 1.0$, $\phi_Y = 1.0$
Coeficientes causais	$\beta_{TY} = 1.2$, $\beta_{TX} = 1.2$, $\beta_{XY} = 0.9$, $\beta_{XT} = 0.6$
Ruído (distribuição)	*uniform*
Intervalos do ruído	$T, X, Y : [-0.1, 0.1]$

Os nove *datasets* foram organizados em três grupos conforme descrito na Tabela 7. O grupo 1 (G1) utiliza relações estritamente lineares entre as variáveis e os grupos 2 (G2) e 3 (G3) empregam funções não lineares. No grupo 2 (G2), o efeito do tratamento é intensificado por funções quadráticas aplicadas a $T \to Y$ e $T \to X$, enquanto a combinação $X \to T$ com função logarítmica introduz confundimento ao tratamento. No grupo 3 (G3), a função senoide associada ao tratamento produz efeitos oscilatórios, permitindo avaliar cenários em que o efeito causal varia ao longo do tempo. Em ambos os grupos não lineares, o efeito $X \to Y$ é modelado pela função logarítmica, comprimindo a resposta de Y às variações de X; na estrutura Cadeia, isso reduz a influência mediada do tratamento, enquanto na estrutura Confundidor modera a influência direta de X sobre Y.

Tabela 7 – Descrição dos nove *datasets* de análise, agrupados por estrutura causal, tipo de complexidade e grupo funcional.

Dataset	**Estrutura**	**Complexidade**	**Grupo**
Direta–Linear	Direta	Linear	
Confundidor–Linear	Confundidor	Linear	G1
Cadeia–Linear	Cadeia	Linear	
Direta–Não Linear (G2)	Direta	Não linear	
Confundidor–Não Linear (G2)	Confundidor	Não linear	G2
Cadeia–Não Linear (G2)	Cadeia	Não linear	
Direta–Não Linear (G3)	Direta	Não linear	
Confundidor–Não Linear (G3)	Confundidor	Não linear	G3
Cadeia–Não Linear (G3)	Cadeia	Não linear	

Funções de efeito aplicadas em cada um dos grupos:

Grupo 1 (G1): $T \to X$, $T \to Y$, $X \to T$, $X \to Y$ = Linear.

Grupo 2 (G2): $T \to X$, $T \to Y$ = Quadrática ; $X \to T$, $X \to Y$ = Logarítmica.

Grupo 3 (G3): $T \to X$, $T \to Y$ = Senoidal ; , $X \to T$, $X \to Y$ = Logarítmica.

Essa configuração experimental permite observar, de maneira sistemática, como a estrutura causal, a complexidade e as funções de efeito influenciam o comportamento das métricas discutidas nas próximas seções, incluindo o ATE, a mudança de associação Δr e demais indicadores analisados ao longo deste capítulo.

4.3 Resultados — Efeito médio (ATE)

Nesta seção analisamos o efeito médio dos tratamentos (ATE), caracterizando de que forma os cenários contrafactuais modificam os resultados e como essas alterações variam entre tipos e intensidades de intervenção. A análise contempla quatro aspectos complementares: a verificação das pré-tendências para estabelecer a equivalência das trajetórias iniciais; a descrição das curvas temporais do ATE no período pós-intervenção; o estudo da monotonicidade associada ao tipo de intervenção; e a avaliação da monotonicidade por faixas de intensidade, com foco na modulação introduzida pelos diferentes valores de δ.

4.3.1 Pré-tendências

Antes de analisar os efeitos pós-intervenção, é fundamental verificar a ausência de tendência prévia entre os cenários factual e contrafactual.

Em estudos de inferência causal, a hipótese de pré-tendência nula estabelece que, para $\tau < 0$, as trajetórias dos resultados devem evoluir de forma paralela, de modo que eventuais diferenças observadas após a intervenção possam ser atribuídas ao tratamento, e não a discrepâncias pré-existentes.

No contexto deste trabalho, avaliamos a pré-tendência por meio do comportamento do efeito médio alinhado no tempo, esperando que, no período pré-intervenção, esse valor seja próximo de zero. Desvios sistemáticos de zero no pré-intervenção indicariam que factual e contrafactual não são equivalentes antes do tratamento — o que caracteriza um viés indesejado na construção das séries.

A Figura 12 ilustra o caso de referência Direta–Linear. Observa-se que, no período pré-intervenção, as médias de $ATE(t)$ permanecem próximas de zero, confirmando a ausência de pré-tendência. O mesmo comportamento foi verificado nos demais cenários analisados, reforçando a consistência das séries simuladas e garantindo que as diferenças observadas no pós-intervenção reflitam exclusivamente o impacto causal das intervenções.

4.3.2 Curvas temporais do ATE

Após confirmada a ausência de pré-tendência, analisamos a evolução temporal do efeito médio $ATE(t)$, alinhado em $\tau = t - t_{\text{int}}$, com o objetivo de caracterizar a dinâmica do impacto das intervenções ao longo do tempo. As curvas representam a diferença média entre os resultados factual e contrafactual em cada instante, permitindo identificar respostas imediatas, efeitos acumulados e a evolução temporal do impacto do tratamento em cada cenário.

As Figuras 13–24 apresentam as curvas temporais de $ATE(t)$ estratificadas por tipo de intervenção, estrutura causal e complexidade funcional. Em cada gráfico, as linhas representam diferentes intensidades de intervenção (δ). O período $\tau = -1$ corresponde ao

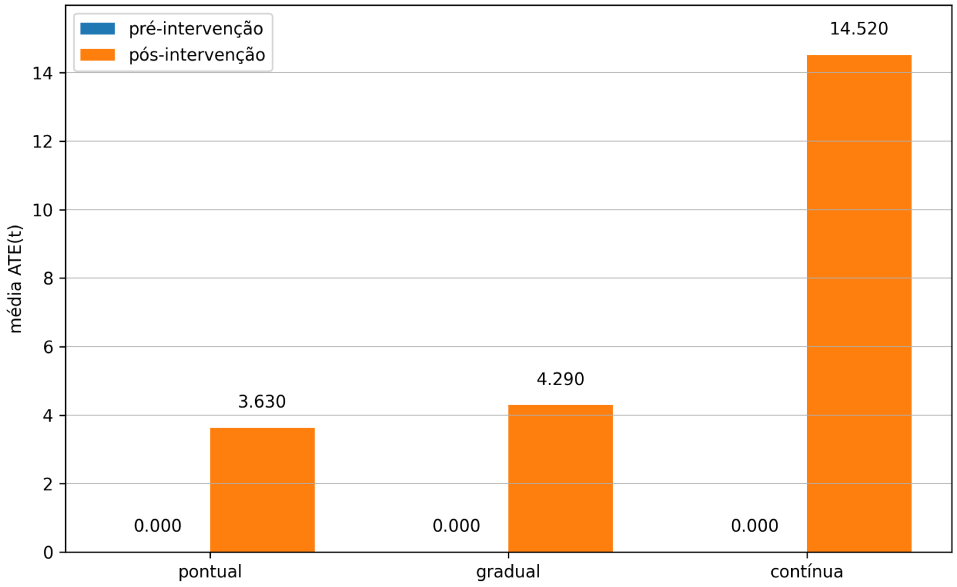

Figura 12 – Verificação de Pré-tendência. Na verificação de pré-tendência no cenário Direta–Linear, as médias de $ATE(t)$ permanecem próximas de zero antes da intervenção ($\tau < 0$), indicando ausência de viés temporal entre os cenários factual e contrafactual.

instante imediatamente anterior à intervenção, enquanto a linha tracejada azul em $\tau = 0$ marca o momento exato da aplicação do tratamento contrafactual.

Datasets Lineares - Grupo 1

Nos *datasets* do grupo G1, com efeitos lineares, as curvas de $ATE(t)$ apresentam comportamentos muito semelhantes entre as três estruturas causais, sendo que as diferenças mais relevantes aparecem entre os tipos de intervenção. Assim, as Figuras 13 a 15, referentes à estrutura Direta–Linear, são representativas do padrão observado também nas demais estruturas lineares.

Na intervenção pontual, que pode ser vista na Figura 13, observa-se um aumento imediato do $ATE(t)$ logo após $\tau = 0$. A partir desse ponto, o efeito cresce de maneira aproximadamente linear ao longo do tempo, refletindo a propagação direta do choque aplicado ao tratamento. Há distinção clara entre as intensidades de intervenção: curvas com maior δ apresentam inclinações proporcionalmente maiores, mantendo uma separação ordenada e monotônica entre si.

Nos cenários com intervenção gradual, observado na Figura 14, a resposta ao tratamento é discreta nos períodos iniciais e se intensifica ao longo do tempo, resultando em uma curva de crescimento acelerado. Esse comportamento decorre do acúmulo progressivo dos incrementos no tratamento, produzindo uma trajetória suavemente curvada, característica de um crescimento polinomial, aproximadamente quadrático. Assim como na intervenção pontual, observa-se uma separação clara entre as intensidades de inter-

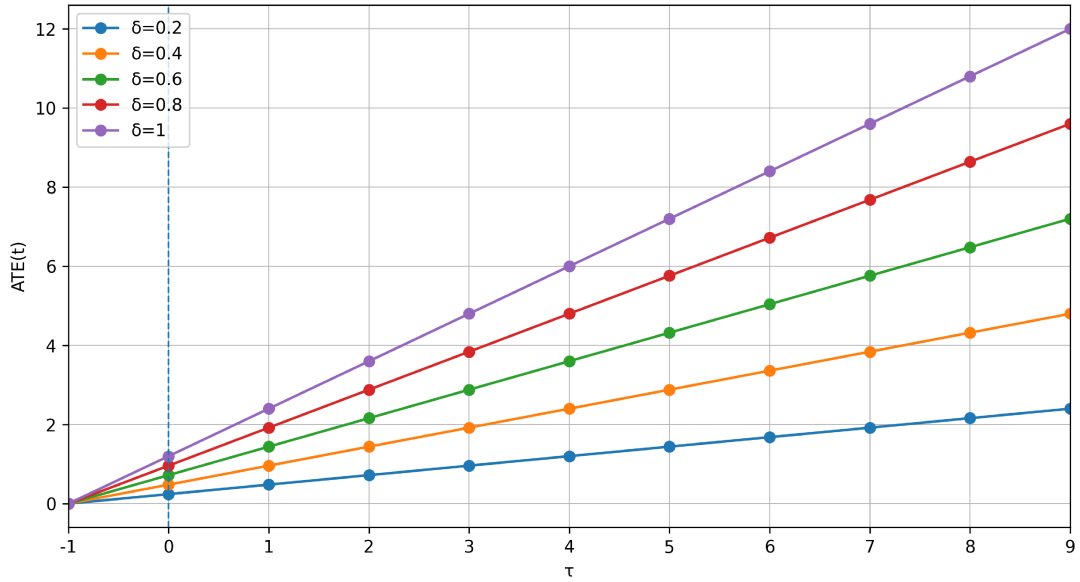

Figura 13 – Curvas $ATE(t)$ — intervenção Pontual, estrutura Direta, Linear - G1. Para cinco intensidades de intervenção δ distintas.

venção, mas com valores finais mais elevados em razão do maior volume cumulativo de tratamento aplicado ao longo do tempo.

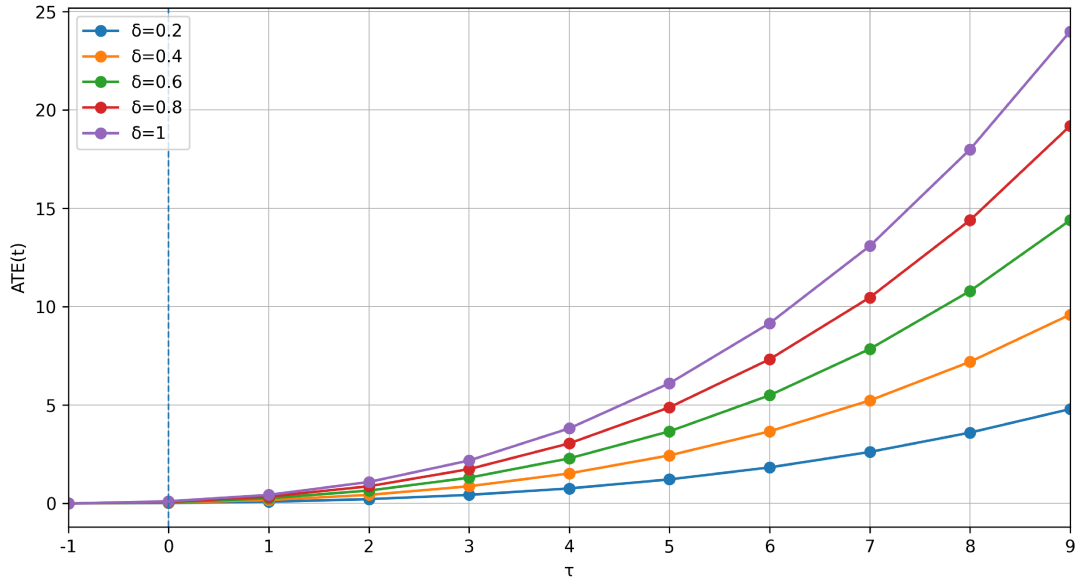

Figura 14 – Curvas $ATE(t)$ — intervenção Gradual, estrutura Direta, Linear - G1. Para cinco intensidades de intervenção δ distintas.

No cenário de intervenção contínua, ilustrado na Figura 15, observa-se um crescimento acelerado do $ATE(t)$ ao longo do tempo. Esse formato curvo decorre do efeito cumulativo do tratamento, que é aplicado em todos os períodos pós-intervenção, de modo que cada incremento adicional propaga-se ao longo da dinâmica autorregressiva do resultado. O padrão resultante é uma trajetória suavemente curvada, típica de um crescimento polinomial, mais acentuada que na intervenção gradual devido ao maior volume acumulado

de tratamento. Assim como nos demais cenários, as curvas exibem separação clara entre as intensidades de intervenção, com valores finais proporcionalmente maiores para doses mais elevadas.

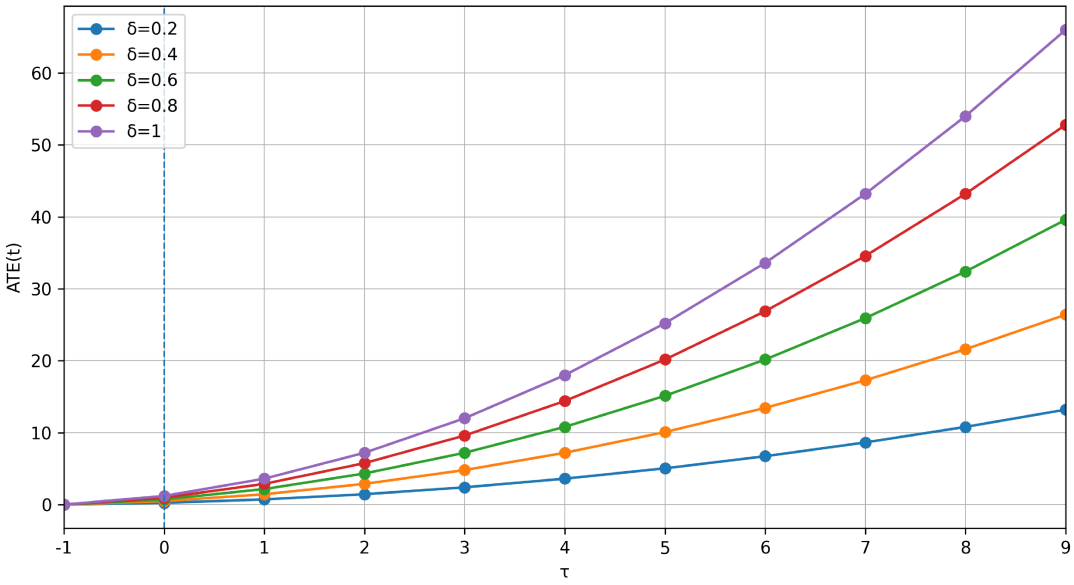

Figura 15 – Curvas $ATE(t)$ — intervenção Contínua, estrutura Direta, Linear - G1. Para cinco intensidades de intervenção δ distintas.

Datasets Não Lineares - Intervenção Pontual

A Figura 16 apresenta as curvas ATE representativa dos cenários não lineares sob intervenção pontual. Apesar das diferenças de funções de efeito entre os grupos G2 e G3, observou-se um padrão comum em todos os experimentos, inclusive entre as estruturas causais distintas: as curvas temporais de ATE ao longo de τ exibem uma trajetória essencialmente linear em todos os cenários analisados, variando apenas a inclinação e com ATEs finais distintos.

Embora as funções que mapeiam o tratamento para o resultado possam ser quadráticas ou senoidais, o esperado para a intervenção pontual nestes cenários é que as trajetórias contrafactuais evoluam como uma versão consistentemente inclinada da trajetória factual. Como o ATE representa a diferença entre essas duas trajetórias, o resultado é um incremento aproximadamente constante a cada período, o que gera uma curva linear no tempo. Em termos práticos, o ATE reflete a acumulação incremental produzida pela intervenção, mesmo quando a relação $T \to Y$ é não linear, deslocamentos repetidos e de baixa amplitude em T tendem a produzir diferenças quase lineares entre as trajetórias factual e contrafactual.

Embora a linearidade seja um padrão comum, surgem diferenças claras entre grupos e estruturas. Os cenários do grupo G2, que utiliza efeitos quadráticos, apresentam ATEs finais mais elevados e curvas mais inclinadas em comparação aos cenários do grupo G3,

que utiliza efeitos senoidais. Isso ocorre porque a função quadrática amplifica incrementos no tratamento conforme o valor de T cresce, fazendo com que a intervenção produza diferenças progressivamente maiores ao longo do tempo. No caso das funções senoidais, o efeito marginal é suavizado devido ao comportamento oscilatório do seno, o que limita a intensidade acumulada do impacto da intervenção.

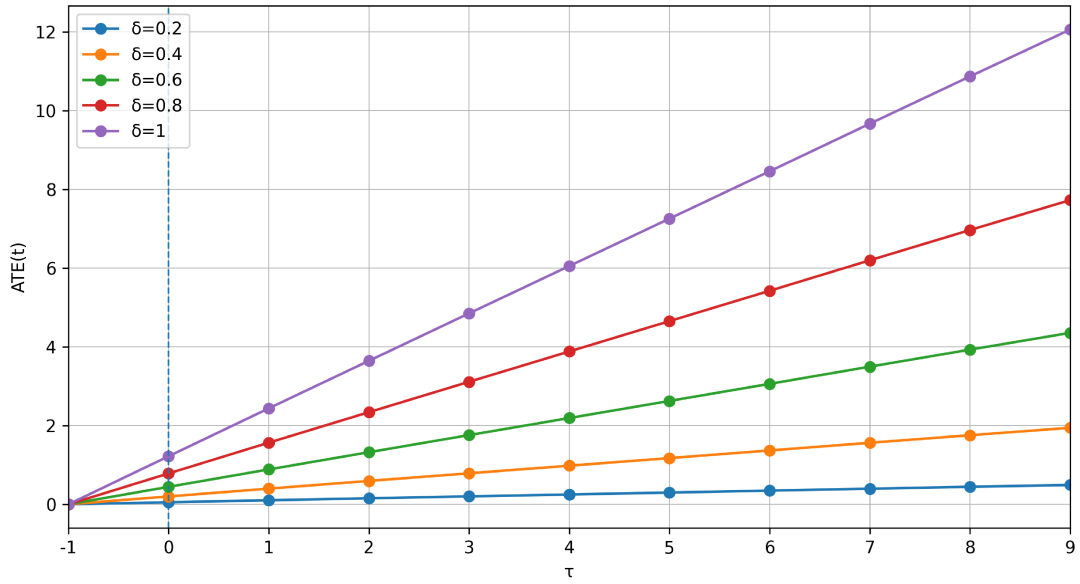

Figura 16 – Curvas $ATE(t)$ — intervenção Pontual, estrutura Direta, Não Linear - G2. Para cinco intensidades de intervenção δ distintas.

Além dessas diferenças entre grupos, há também um padrão consistente entre estruturas causais. A estrutura Confundidora apresenta os maiores valores de ATE, seguida pela estrutura Direta, enquanto a estrutura em Cadeia apresenta os menores valores. Essas diferenças refletem o papel da covariável X em cada cenário. Na estrutura Confundidora, o tratamento e o resultado compartilham uma influência comum, o que intensifica simultaneamente dois caminhos causais e aumenta o efeito acumulado. Na estrutura Direta, o tratamento afeta apenas o resultado, com efeitos de magnitude intermediária. Na estrutura em Cadeia, o efeito de T é mediado por X antes de alcançar Y, o que suaviza a propagação da intervenção e reduz o efeito incremental por período.

Em conjunto, os resultados mostram que, embora a forma funcional influencie a magnitude dos efeitos e a estrutura causal determine a eficiência com que a intervenção é transmitida, a dinâmica temporal de ATE sob intervenção pontual permanece linear em todos os cenários analisados. Esse comportamento é explicado pela natureza incremental e persistente do deslocamento introduzido no tratamento, que se traduz em diferenças acumuladas uniformemente ao longo do tempo.

***Datasets* Não Lineares - Grupo G2 - Intervenção Gradual e Contínua**

Nos cenários não lineares com intervenção gradual do grupo G2 apresentados nas Figuras 17 e 18, observa-se que o efeito médio do tratamento cresce de forma acelerada ao longo do tempo, formando curvas de crescimento acelerado e formato convexa. Esse comportamento decorre do acúmulo progressivo da intervenção no tratamento, uma vez que os aumentos sucessivos geram incrementos cada vez maiores no resultado quando combinados com funções quadráticas e logarítmicas que amplificam a propagação temporal do efeito.

Os ATEs das estruturas direta e confundidora, Figura 17, possuem comportamentos muito similares, pois o efeito dominante ocorre pela ação direta do tratamento sobre o resultado, ainda que a confundidora inclua um caminho adicional via variável latente. A estrutura em cadeia, Figura 18, apresenta valores finais inferiores e inclinações reduzidas, resultado da suavização introduzida pelo caminho mediado, no qual o tratamento modifica primeiro a variável intermediária e só então influencia o resultado, produzindo uma acumulação mais lenta do efeito ao longo do tempo.

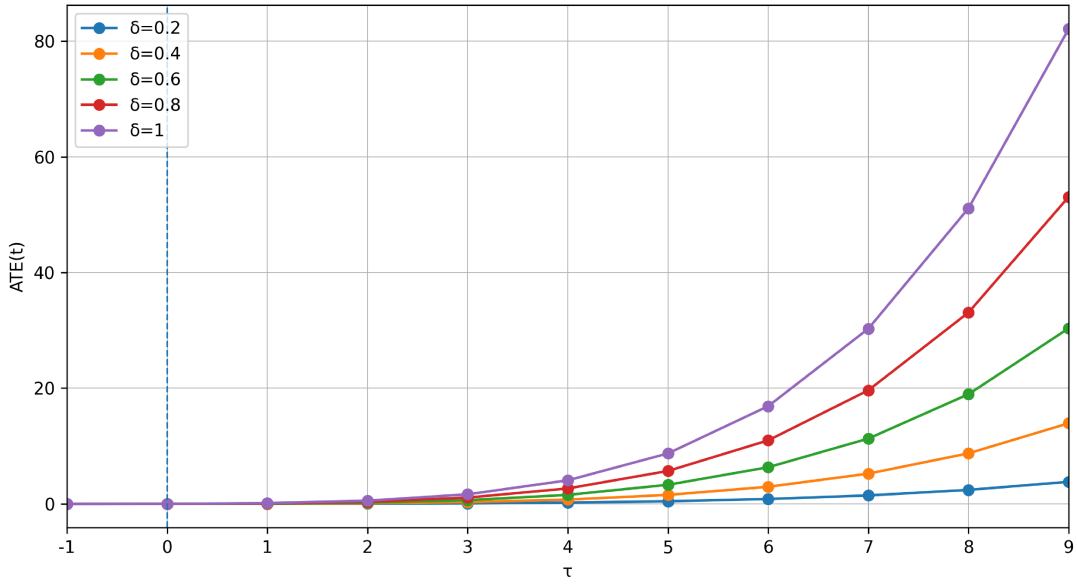

Figura 17 – Curvas $ATE(t)$ — intervenção Gradual, estrutura Confundidor, Não Linear - G2. Para cinco intensidades de intervenção δ distintas.

Nos cenários com intervenção contínua descritos nas Figuras 19 e 20, o comportamento altera-se de forma marcante. Como o tratamento recebe incrementos sucessivos a cada período, o efeito sobre o resultado passa a ser acumulado de maneira progressiva, e as curvas tornam-se aceleradas e crescentemente convexas. Esse crescimento mais rápido ocorre porque o acúmulo das intervenções no tratamento amplifica o impacto das funções quadráticas associadas ao caminho que conecta $T \rightarrow Y$ no grupo dois, o que produz valores finais de ATE significativamente mais elevados.

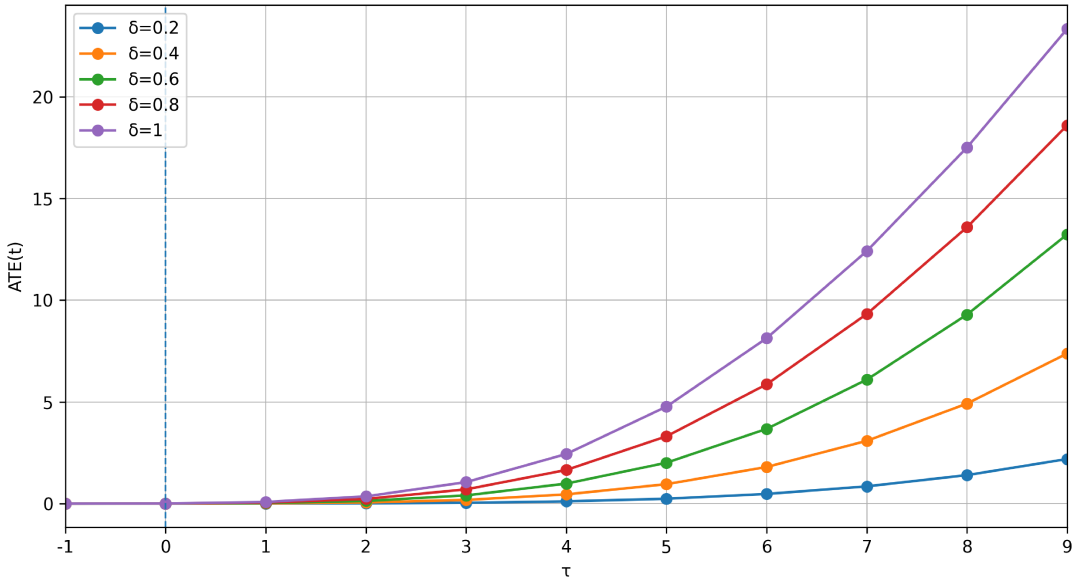

Figura 18 – Curvas $ATE(t)$ — intervenção Gradual, estrutura Cadeia, Não Linear - G2. Para cinco intensidades de intervenção δ distintas.

Assim como no caso da intervenção pontual, as estruturas direta e confundidora, Figura 19, exibem curvas de ATE quase idênticas, refletindo novamente a predominância do efeito direto do tratamento sobre o resultado. A estrutura em cadeia, Figura 20, continua apresentando valores consideravelmente menores, pois a presença do efeito logarítmico em sequência ao tratamento reduz o efeito deste ao longo do sistema e mantém a evolução do ATE mais suave e menos inclinada. Dessa forma, a diferença entre as três estruturas fica mais acentuada quando a intervenção é contínua, uma vez que o acúmulo prolongado das intervenções potencializa a separação entre trajetórias amplificadas diretamente e trajetórias amortecidas pela mediação.

Datasets Não Lineares - Grupo G3 - Intervenção Gradual e Contínua

Nos cenários não lineares com intervenção gradual do grupo G3, ilustrados nas Figuras 21 e 22, observa-se que o efeito médio do tratamento não cresce de forma contínua ou acelerada ao longo do tempo, mas exibe trajetórias inicialmente ascendentes que, após certo ponto, tendem à desaceleração e até à reversão, indicando um comportamento de natureza oscilatória. Esse padrão surge devido ao efeito senoidal do tratamento, fazendo com que o impacto de incrementos graduais não seja constante ao longo do tempo: conforme o sistema atravessa regiões de maior ou menor sensibilidade da função seno, o acúmulo do efeito pode se intensificar ou dissipar, interrompendo a monotonicidade esperada em intervenções crescentes.

A estrutura confundidora sob intervenção gradual, apresentada na Figura 21, representa também o comportamento da estrutura direta, uma vez que ambas exibem curvas de efeito médio muito semelhantes ao longo de todo o horizonte temporal. O efeito domi-

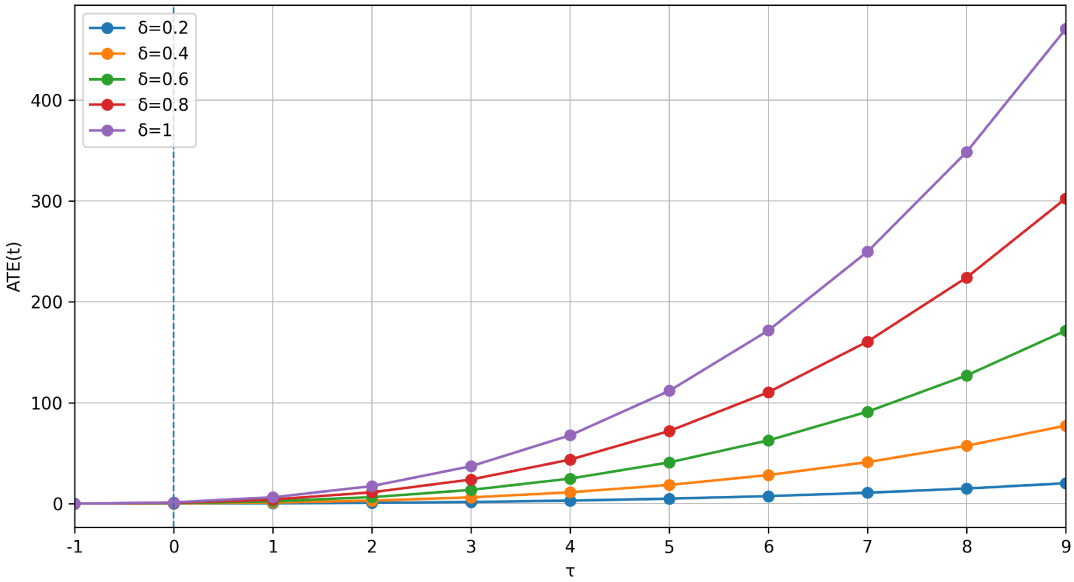

Figura 19 – Curvas $ATE(t)$ — intervenção Contínua, estrutura Confundidor, Não Linear - G2. Para cinco intensidades de intervenção δ distintas.

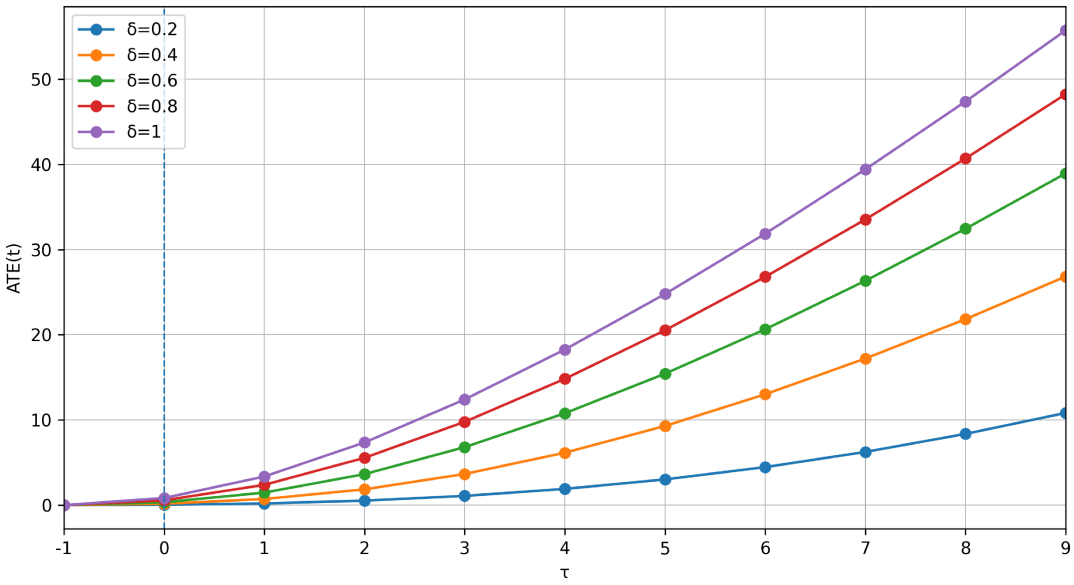

Figura 20 – Curvas $ATE(t)$ — intervenção Contínua, estrutura Cadeia, Não Linear - G2. Para cinco intensidades de intervenção δ distintas.

nante nesses casos é transmitido pelo caminho principal do tratamento até o resultado, de modo que as oscilações introduzidas pela função senoidal produzem perfis temporais semelhantes. As curvas crescem nos primeiros períodos, mas perdem inclinação à medida que o tratamento atravessa regiões menos sensíveis da função seno, entrando em fases de estabilização ou declínio, mais visível em δ maiores, o que explica a ausência de um crescimento convexo persistente.

A estrutura em cadeia com intervenção gradual, mostrada na Figura 22, apresenta um comportamento distinto e aparentemente mais instável. Nesse caso, o tratamento afeta inicialmente a variável intermediária, que por sua vez influencia o resultado, de modo que

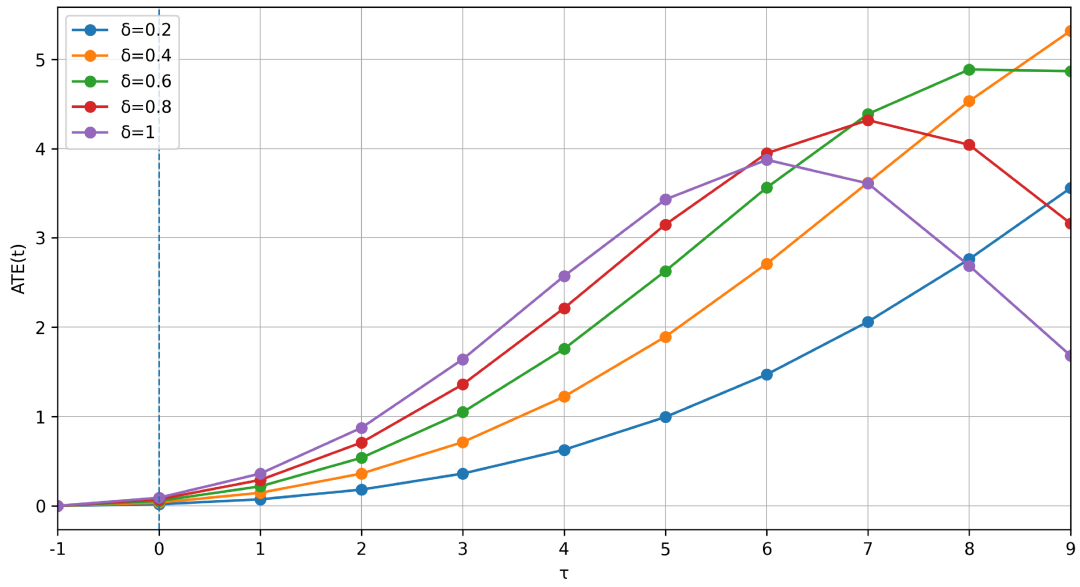

Figura 21 – Curvas $ATE(t)$ — intervenção Gradual, estrutura Confundidor, Não Linear - G3. Para cinco intensidades de intervenção δ distintas.

o efeito passa por duas transformações não lineares sucessivas. Essa dupla passagem intensifica a alternância entre regiões de sensibilidade alta e baixa, produzindo curvas mais irregulares e com inversões mais acentuadas, nas quais o aumento do tratamento pode resultar em reduções temporárias do ATE. Assim, mesmo sob doses crescentes de intervenção, o efeito acumulado pode oscilar ou mesmo diminuir em determinados períodos, refletindo a natureza faseada da função senoidal e a desaceleração adicional introduzida pela função logarítmica.

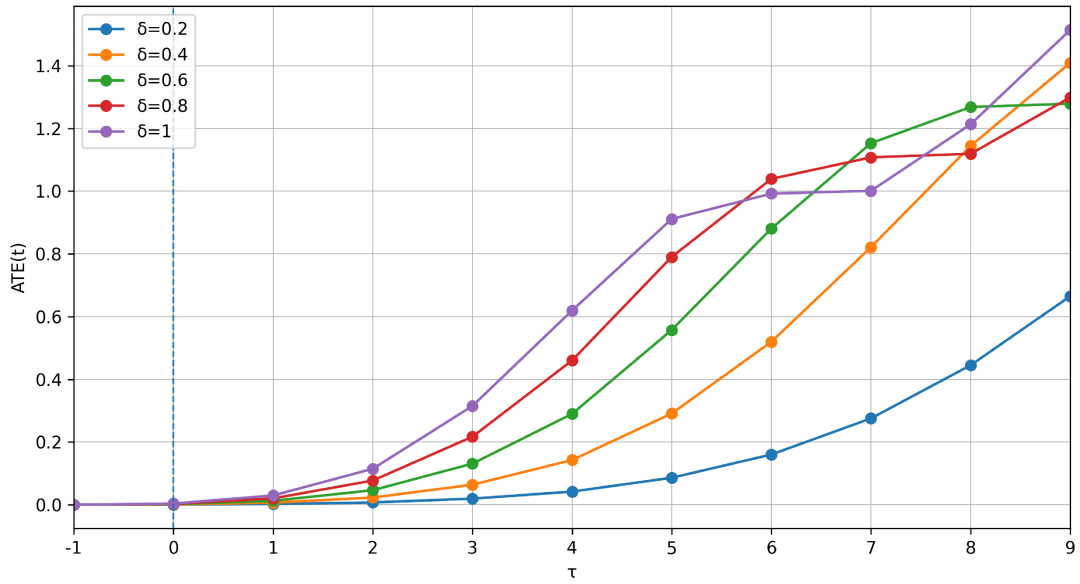

Figura 22 – Curvas $ATE(t)$ — intervenção Gradual, estrutura Cadeia, Não Linear - G3. Para cinco intensidades de intervenção δ distintas.

Nos cenários de intervenção contínua, apresentados nas Figuras 23 e 24, o tratamento

cresce de forma ininterrupta ao longo do tempo devido às intervenções, mas o efeito médio resultante mantém o mesmo padrão qualitativo observado no caso gradual. Embora as curvas exibam um crescimento mais prolongado nos períodos iniciais, o sistema rapidamente alcança regiões da função senoidal em que o efeito marginal do tratamento diminui, visto que a sensibilidade da função varia de acordo com sua fase. Essa oscilação na resposta impede a sustentação de um crescimento convexo e causa trechos de desaceleração mesmo sob uma intervenção continuamente crescente.

A estrutura confundidora sob intervenção contínua, mostrada na Figura 23, novamente representa também a estrutura direta, dada a semelhança entre suas curvas ATE. O comportamento observado reflete a ação direta do tratamento sobre o resultado, modulada pelas oscilações senoidais e pela suavização logarítmica, produzindo curvas que sobem de maneira ordenada nos primeiros períodos, mas que reduzem sua inclinação à medida que o tratamento adentra regiões de menor sensibilidade.

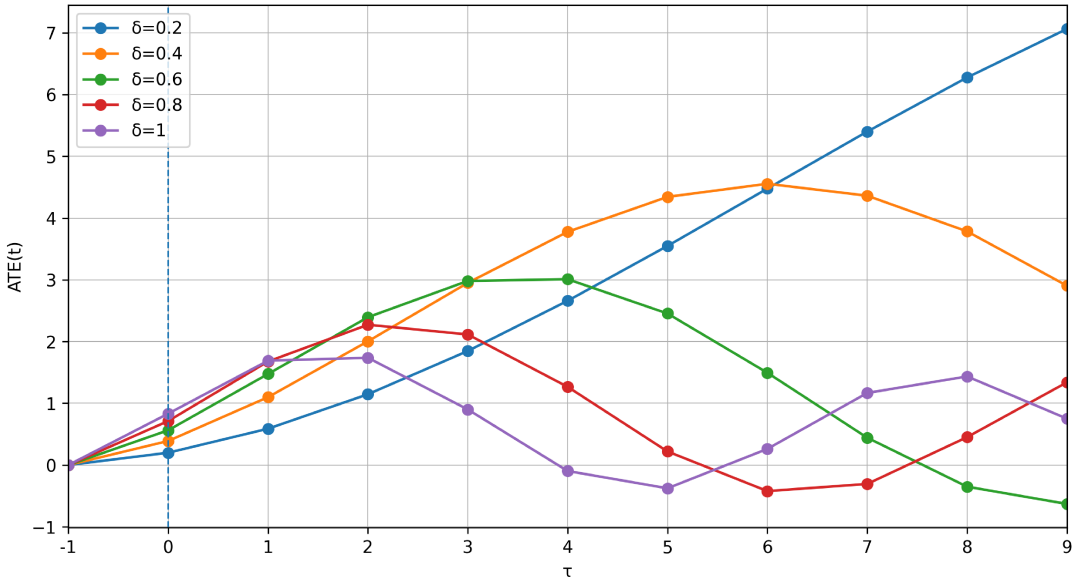

Figura 23 – Curvas $ATE(t)$ — intervenção Contínua, estrutura Confundidor, Não Linear - G3. Para cinco intensidades de intervenção δ distintas.

Por fim, a estrutura em cadeia com intervenção contínua, ilustrada na Figura 24, apresenta comportamento distinto dos demais cenários. Embora as curvas ATE exibam tendência de crescimento, o efeito do tratamento é modulado pela função senoidal no caminho para a variável intermediária e depois desacelerado pela função logarítmica na transição desta para o resultado. O efeito acumulado perde monotonicidade e tende a exibir oscilações mais pronunciadas. Assim, mesmo com a intervenção contínua elevando o tratamento de forma estável ao longo do tempo, o ATE pode apresentar regiões de crescimento fraco, estabilização ou declínio, refletindo a interação entre as funções não lineares e o efeito mediado da estrutura em cadeia.

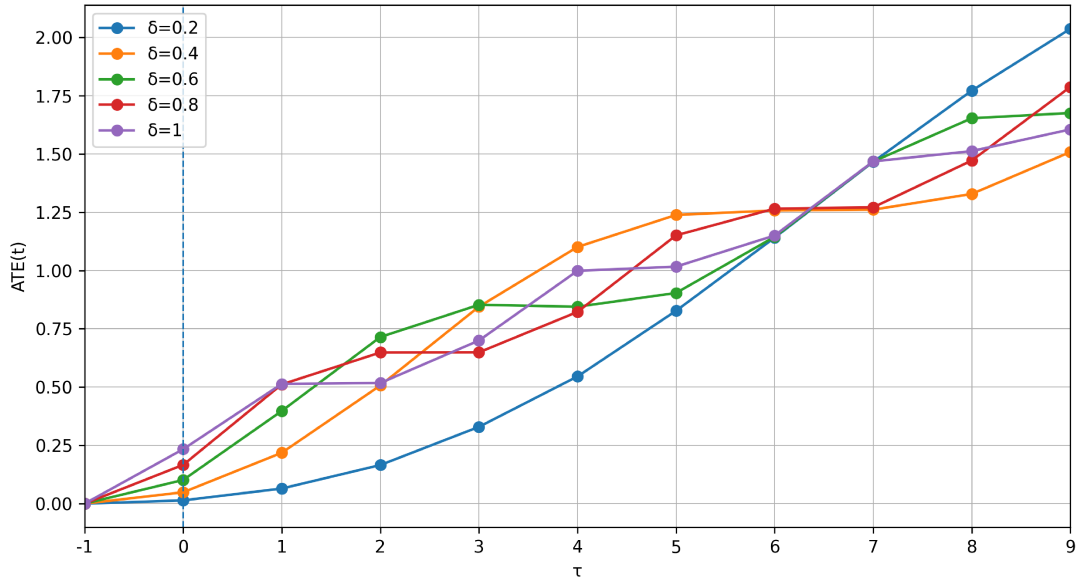

Figura 24 – Curvas $ATE(t)$ — intervenção Contínua, estrutura Cadeia, Não Linear - G3. Para cinco intensidades de intervenção δ distintas.

4.3.3 Monotonicidade por tipo

A relação dose–resposta foi examinada para cada tipo de intervenção por meio da correlação de Spearman entre a intensidade δ e o efeito médio pós-intervenção $\overline{ATE}^{(k)}$. Essa medida permite avaliar se aumentos na dose de intervenção estão associados a efeitos também crescentes, caracterizando uma relação monotônica positiva entre intervenção e resultado.

Coeficientes próximos de zero indicam a ausência de uma relação ordenada entre dose e efeito, sugerindo que variações em δ não se traduzem em mudanças sistemáticas no ATE. Coeficientes negativos revelam um padrão inverso, no qual intensidades maiores da intervenção tendem a gerar efeitos menores, caracterizando uma resposta oposta à esperada em um regime monotônico crescente.

A análise da monotonicidade revelou um comportamento bastante estável nos grupos G1 e G2, cujo padrão é bem representado pelo cenário mostrado na Figura 25. Em todos esses casos, a correlação de Spearman entre δ e o ATE médio permanece perfeita e positiva para os três tipos de intervenções, com coeficiente igual a um e valor de p próximo de zero, indicando uma relação estritamente monotônica em que aumentos na intensidade da intervenção produzem aumentos ordenados no efeito médio. Mesmo na intervenção contínua, onde as curvas apresentam maior variabilidade temporal, o coeficiente permanece positivo e estatisticamente significativo, o que demonstra que, apesar das diferenças no formato das trajetórias ao longo do tempo, a relação entre dose e efeito preserva a consistência da ordenação monotônica.

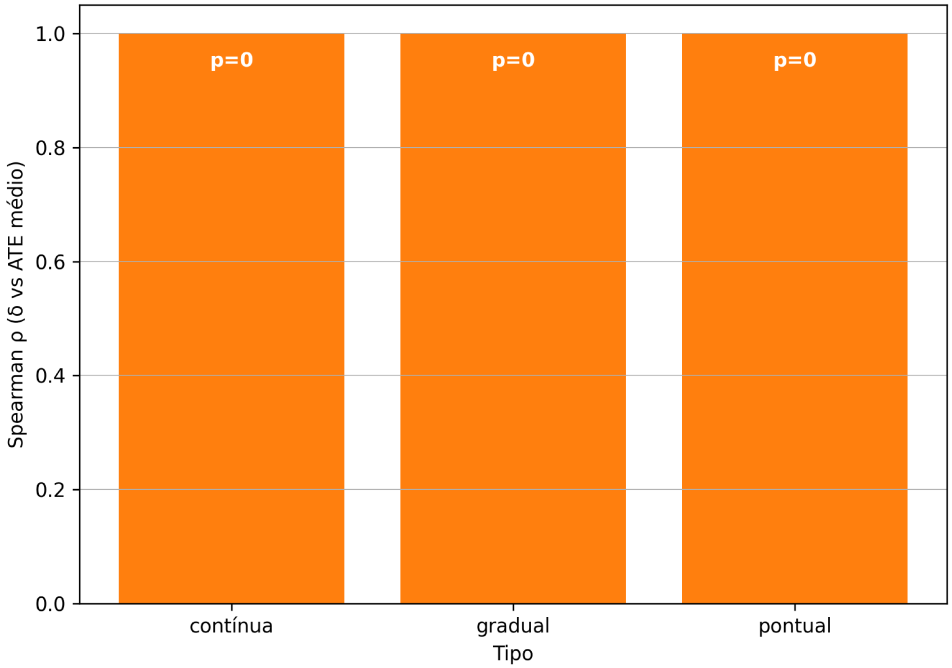

Figura 25 – Monotonicidade por tipo de intervenção na estrutura Direta, Linear - G1. Cada barra representa o coeficiente de Spearman entre a intensidade da intervenção δ e o ATE médio para o respectivo tipo de intervenção. O valor indicado no topo de cada barra corresponde ao p-valor do teste de monotonicidade, que avalia a significância estatística da relação monotônica entre δ e o ATE.

No grupo G3, as funções não lineares senoidais produzem um comportamento qualitativamente distinto do observado nos grupos anteriores.

A estrutura direta, ilustrada na Figura 26, e a estrutura confundidora, apresentada na Figura 27, exibem correlação positiva e praticamente perfeita no cenário pontual, e de correlação elevada no cenário gradual. Esses resultados mostram que, apesar das oscilações locais do ATE ao longo do tempo, a ordenação global dos efeitos permanece alinhada ao aumento de δ nesses dois tipos de intervenção. A intervenção contínua rompe esse padrão, tendo uma correlação fortemente negativa e estatisticamente significativa, indicando que, sob dinâmica senoidal, o aumento progressivo de δ não preserva a monotonicidade e altera de maneira substancial o comportamento global do efeito médio.

A estrutura em cadeia, apresentada na Figura 28, exibe padrão distinto daquele das estruturas direta e confundidora do grupo G3 e aproxima-se do observado nos grupos G1 e G2. A relação entre δ e o ATE médio permanece monotônica e positiva nos três tipos de intervenção. Nos cenários pontual e gradual, a monotonicidade é perfeita, com coeficiente igual a um e p-valor zero. No cenário contínuo, embora o coeficiente seja menor, observa-se correlação positiva e estatisticamente significativa, indicando preservação da ordenação dos efeitos. Nesse caso, a variável mediadora suaviza as oscilações introduzidas pelas funções não lineares, reduzindo a intensidade da monotonicidade sem provocar inversão do sinal. A estrutura em cadeia, portanto, modula o impacto das não linearidades e impede a perda total de monotonicidade observada na estrutura direta do grupo G3.

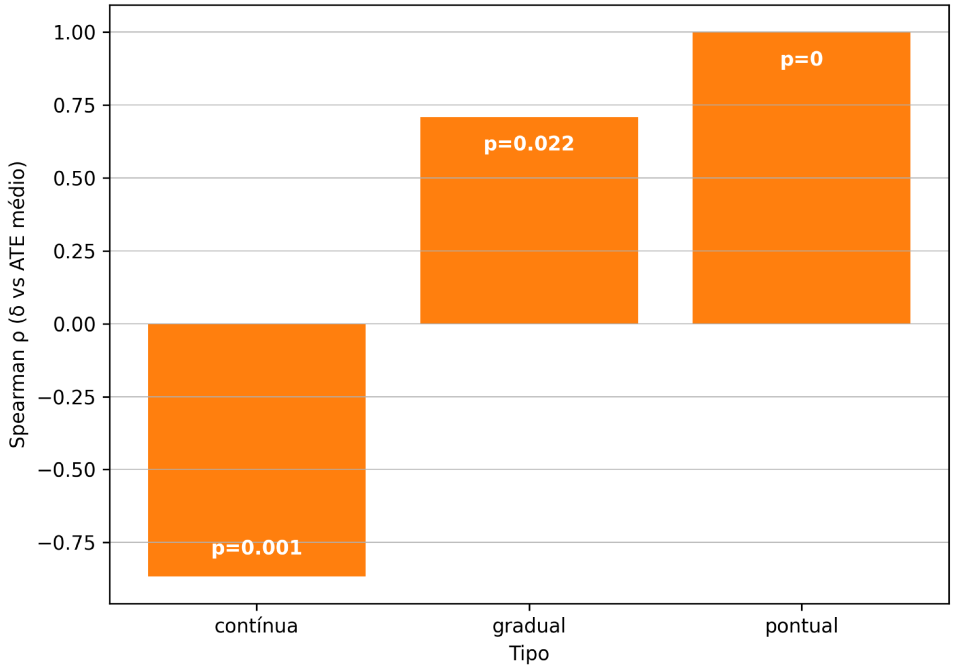

Figura 26 – Monotonicidade por tipo de intervenção, estrutura Direta, Não Linear - G3. Cada barra representa o coeficiente de Spearman entre a intensidade da intervenção δ e o ATE médio para o respectivo tipo de intervenção. O valor indicado no topo de cada barra corresponde ao p-valor do teste de monotonicidade, que avalia a significância estatística da relação monotônica entre δ e o ATE.

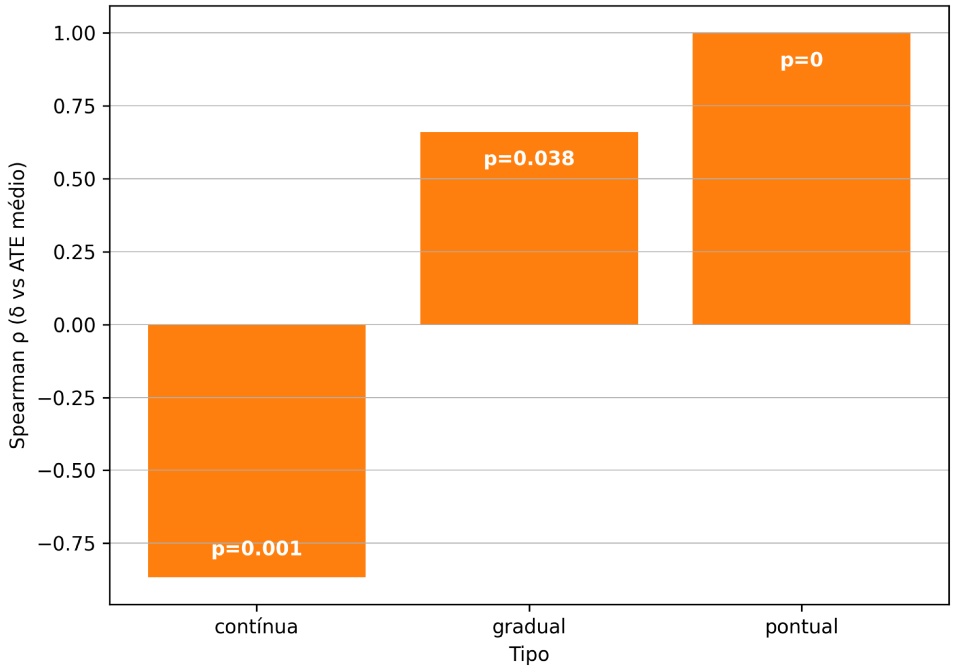

Figura 27 – Monotonicidade por tipo de intervenção, estrutura Confundidor, Não Linear - G3. Cada barra representa o coeficiente de Spearman entre a intensidade da intervenção δ e o ATE médio para o respectivo tipo de intervenção. O valor indicado no topo de cada barra corresponde ao p-valor do teste de monotonicidade, que avalia a significância estatística da relação monotônica entre δ e o ATE.

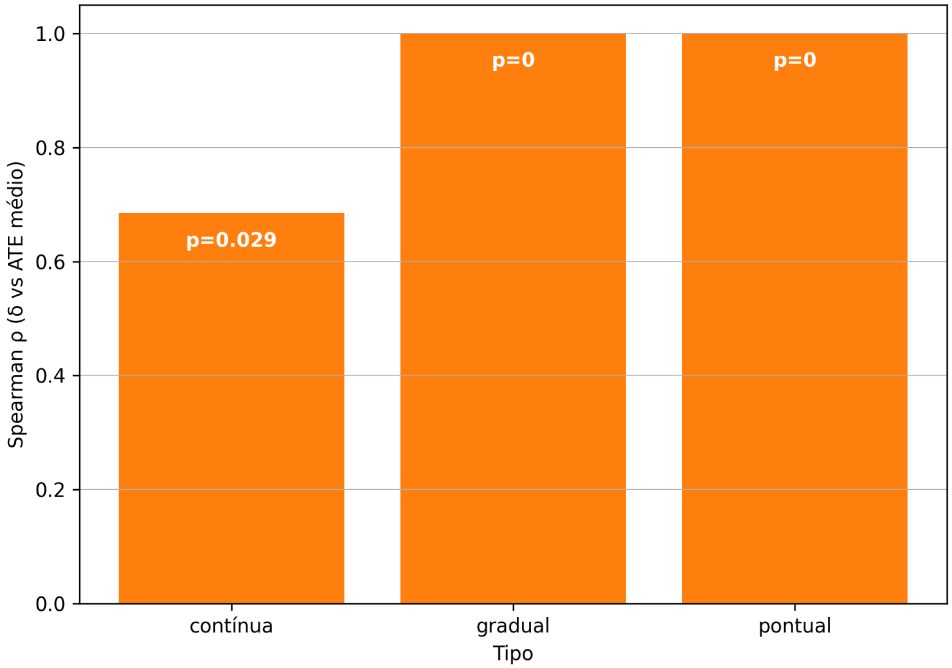

Figura 28 – Monotonicidade por tipo de intervenção, estrutura Cadeia, Não Linear - G3. Cada barra representa o coeficiente de Spearman entre a intensidade da intervenção δ e o ATE médio para o respectivo tipo de intervenção. O valor indicado no topo de cada barra corresponde ao p-valor do teste de monotonicidade, que avalia a significância estatística da relação monotônica entre δ e o ATE.

4.3.4 Monotonicidade por faixas

Avaliamos o comportamento local da relação dose–resposta por meio da correlação de Spearman calculada em faixas discretas de intensidade. Essa abordagem, denominada análise *piecewise*, permite identificar variações não lineares, regiões de saturação do efeito, inversão de tendência ou mudanças na sensibilidade ao tratamento nas quais incrementos adicionais da dose deixam de produzir aumentos proporcionais no efeito médio. Para isso, o domínio das doses foi dividido em dois intervalos sucessivos, um para doses menores ou iguais a 0,5 e outro para doses iguais ou superiores a 0,6, sendo as correlações locais calculadas separadamente em cada faixa. Essa estratégia possibilita verificar se a resposta ao tratamento preserva uma relação monotônica em todo o espectro de intensidades ou se há pontos em que o efeito se estabiliza ou muda de direção.

A maioria dos cenários apresentou comportamento consistente com uma relação monotônica positiva em ambas as faixas de intensidade, como ilustrado na Figura 29 no cenário pontual direto não linear. Nesse padrão predominante, tanto para doses baixas quanto altas, a correlação local permanece elevada e estatisticamente significativa, indicando que incrementos de dose geram aumentos sistemáticos no efeito médio independentemente da intensidade. A divisão por faixas, portanto, não altera a conclusão qualitativa da análise global de monotonicidade, sugerindo que a resposta ao tratamento é estável e continua crescendo mesmo quando a dose entra em regiões superiores do domínio.

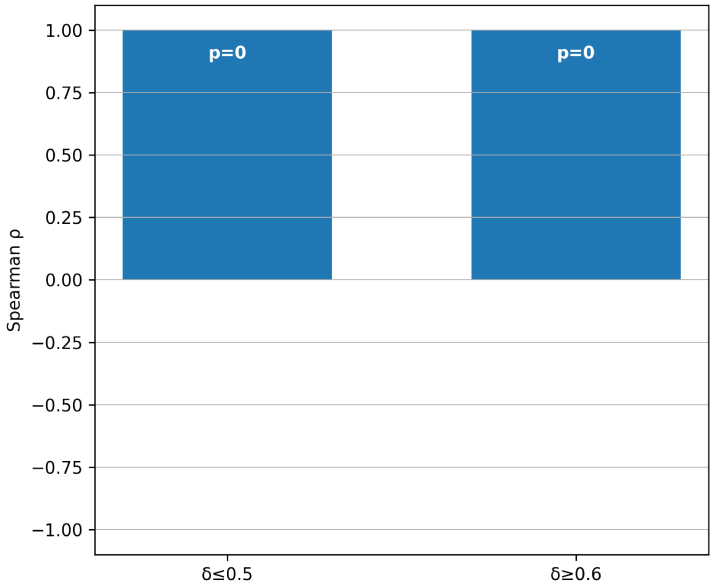

Figura 29 – Monotonicidade por faixa de intensidade de intervenção pontual, estrutura Direta, Não linear - G3. Calculado em duas faixas, a primeira para o intervalo de intensidades de tratamento δ de 0.1 a 0.5 e a segunda para o intervalo de 0.6 a 1.0. O valor p indicado em cada barra corresponde ao p-valor do teste, que avalia a significância estatística da relação monotônica entre e o ATE e a faixa correspondente.

Entretanto, alguns cenários não lineares no grupo G3 evidenciaram que essa relação monotônica perde-se parcialmente em intensidades mais altas. Nos cenários gradual direto e gradual confundidor, representados na Figura 30, observa-se que a faixa de doses mais baixas mantém correlação positiva, embora fraca e não significativa, enquanto a faixa de doses acima de 0,6 apresenta correlação negativa estatisticamente significativa. Esse comportamento indica que, após certo limiar, aumentos sucessivos de dose deixam de elevar o efeito médio e passam a produzir respostas decrescentes, sugerindo uma mudança na sensibilidade do sistema possivelmente associada à combinação entre o efeito senoidal no tratamento e a saturação local induzida pela estrutura das funções não lineares.

Um padrão semelhante, porém mais pronunciado, aparece nos cenários contínuo direto e contínuo confundidor, nos quais a correlação local é fortemente negativa na faixa de doses mais altas, como demonstrado na Figura 31. Apesar de doses baixas ainda apresentarem tendência positiva, o regime de doses elevadas reverte completamente o comportamento esperado de uma relação dose–resposta, com aumentos de intensidade gerando efeitos médios consistentemente menores. Esse resultado reforça a presença de regiões de inversão decorrentes das oscilações senoidais, que tornam-se mais relevantes quando a intervenção cresce continuamente ao longo do tempo.

Por fim, o cenário contínuo em cadeia, mostrado na Figura 32, também apresenta perda de monotonicidade em doses altas, embora preserve uma correlação positiva estatisticamente significativa na faixa de doses menores. A combinação entre o efeito senoidal na primeira etapa da propagação e o efeito logarítmico na etapa seguinte faz com que incrementos de dose gerem respostas positivas apenas até certo ponto, após o qual a sen-

sibilidade marginal se reduz de forma acentuada, levando a correlações negativas na faixa superior. Essa interação mediada acentua a instabilidade típica das funções não lineares e amplia a propensão à inversão de tendência em doses maiores.

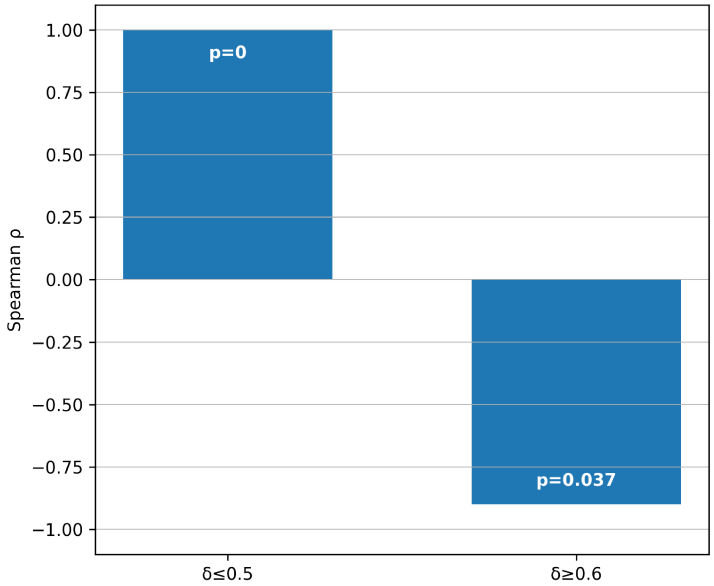

Figura 30 – Monotonicidade por faixa de intensidade de intervenção gradual, estrutura Direta, Não linear - G3. Calculado em duas faixas, a primeira para o intervalo de intensidades de tratamento δ de 0.1 a 0.5 e a segunda para o intervalo de 0.6 a 1.0.

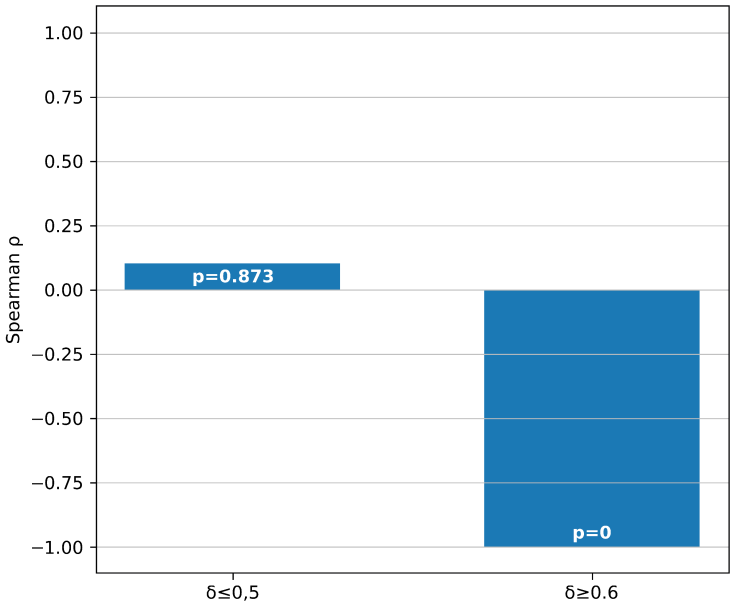

Figura 31 – Monotonicidade por faixa de intensidade de intervenção contínua, estrutura Direta, Não Linear - G3. Calculado em duas faixas, a primeira para o intervalo de intensidades de tratamento δ de 0.1 a 0.5 e a segunda para o intervalo de 0.6 a 1.0. O valor p indicado em cada barra corresponde ao p-valor do teste, que avalia a significância estatística da relação monotônica entre e o ATE e a faixa correspondente.

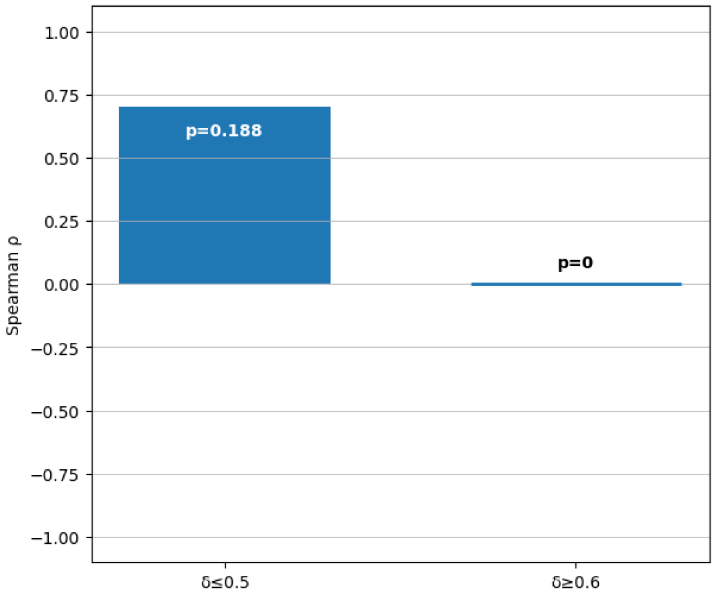

Figura 32 – Monotonicidade por faixa de intensidade de intervenção contínua, estrutura Cadeia, Não linear - G3. Calculado em duas faixas, a primeira para o intervalo de intensidades de tratamento δ de 0.1 a 0.5 e a segunda para o intervalo de 0.6 a 1.0. O valor p indicado em cada barra corresponde ao p-valor do teste, que avalia a significância estatística da relação monotônica entre e o ATE e a faixa correspondente.

4.4 Resultados — Mudança de associação (Δr)

Nesta seção apresentamos os resultados referentes à mudança na associação entre tratamento (T) e resultado (Y), quantificada por Δr, que compara a dependência observada nos cenários factual e contrafactual ao longo do tempo. Como detalhado na Seção 4.1, utilizamos correlação de Pearson nos cenários lineares e correlação de Chatterjee nos não lineares. Por essa razão, adotamos a notação $\Delta r(t)$ de forma geral e para os casos em que a dependência é medida por Pearson, enquanto $\Delta\xi(t)$ é utilizada quando a associação é avaliada pela correlação de Chatterjee. Assim, a métrica expressa, conforme o cenário, a mudança linear ou a mudança monotônica na relação $T \leftrightarrow Y$ em cada instante t.

Para complementar essa visão global, analisamos também o comportamento temporal de $\mathbb{E}[\Delta r(\tau)]$, com bandas de confiança obtidas por reamostragem, o que evidencia a persistência ou dissipação dos efeitos ao longo do horizonte pós-intervenção.

No restante desta seção, descrevemos como as diferentes estruturas causais, complexidades funcionais e tipos de intervenção influenciam o sinal, a magnitude e a dinâmica temporal de Δr, destacando padrões consistentes e exceções observadas nas nove configurações analisadas.

4.4.1 Grupo G1 - *Datasets* Lineares

Nos *datasets* do grupo G1, que adotam relações estritamente lineares entre as variáveis, foram obtidos coeficientes de correlação de Pearson elevados e praticamente idênticos nos

cenários factual e contrafactual em todas as estruturas, como mostrado na Tabela 8. Os valores médios de Δr permanecem extremamente próximos de zero, indicando que a intervenção — independentemente do tipo — não altera de forma mensurável a dependência linear entre T e Y.

Tabela 8 – Médias temporais das correlações de Pearson dos dados factuais ($\overline{r_f}$), contrafactuais ($\overline{r_{cf}}$) e das diferenças ($\overline{\Delta r}$) ao longo dos períodos pós-intervenção, no grupo G1.

Estrutura	$\overline{r_f} \pm dp$	$\overline{r_{cf}} \pm dp$	$\overline{\Delta r} \pm dp$
Direta	0.97 ± 0.004	0.97 ± 0.004	$(7{,}23 \times 10^{-10}) \pm (2{,}80 \times 10^{-9})$
Cadeia	0.98 ± 0.005	0.98 ± 0.005	$(7{,}23 \times 10^{-10}) \pm (2{,}80 \times 10^{-9})$
Confundidor	0.98 ± 0.005	0.98 ± 0.005	$(-6{,}91 \times 10^{-10}) \pm (2{,}76 \times 10^{-9})$

$\overline{r_f}$: Média dos $r_f(\tau)$ da curva de correlação factual do *dataset*, N=10 ($\tau = 10$). $\overline{r_{cf}}$: Média dos $r_{cf}^{(k)}(\tau)$ das curvas de correlação contrafactual, N=300 ($k = 30$, $\tau = 10$).

A Figura 33 exemplifica esse comportamento para a estrutura Confundidora. No painel inferior, observa-se que a diferença $\Delta r(t)$ oscila apenas na ordem de 10^{-8}, de modo que as curvas $r_f(t)$ e $r_{cf}(t)$ exibidas no painel superior tornam-se praticamente sobrepostas. Esse padrão confirma que, em sistemas totalmente lineares, a intervenção modifica os valores das séries, mas não afeta a covariação relativa entre indivíduos no corte transversal.

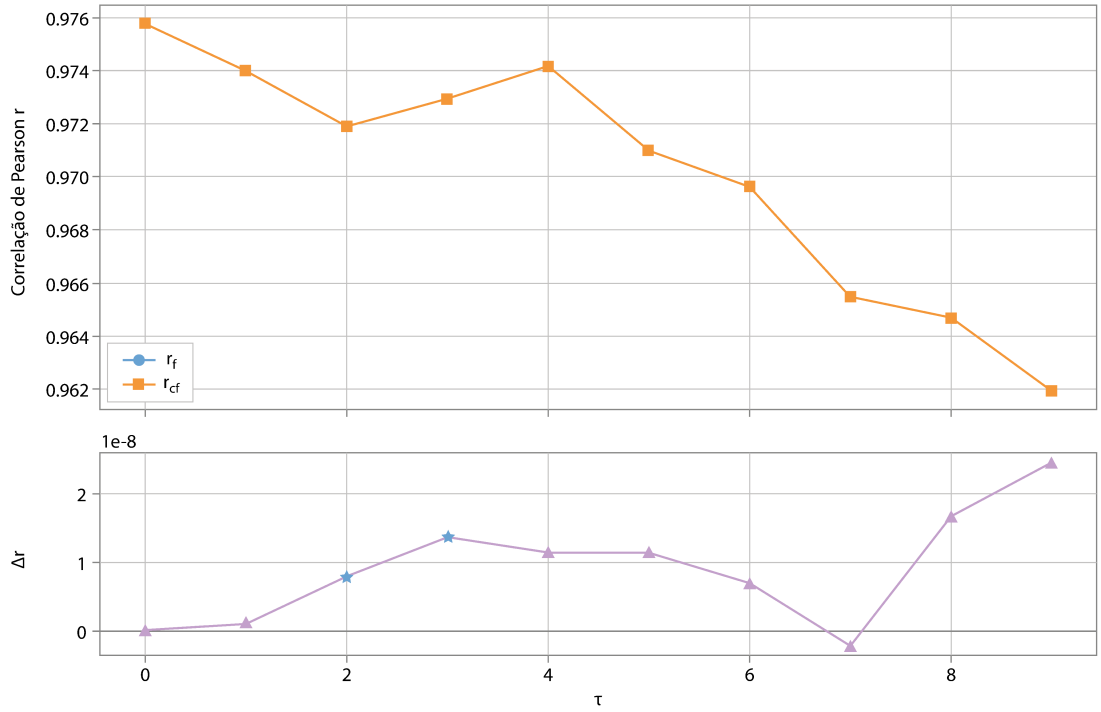

Figura 33 – Correlação e Δr — Contínua, Confundidor, Linear, $\delta = 1.0$ - G1. Painel superior: Curvas de r_f e r_{cf} ao longo de τ. Painel inferior: diferença $\Delta r(\tau)$. Observa-se que $\Delta r(\tau)$ permanece próximo de zero, com escala da ordem de 10^{-8}, indicando mudança mínima na associação.

A Figura 34 apresenta a evolução temporal das médias de $\Delta r(\tau)$ com intervalos de confiança via *bootstrap*. As curvas oscilam muito próximo de zero e os intervalos frequentemente cruzam o eixo nulo, reafirmando que a associação entre T e Y permanece estável ao longo de todo o período pós-intervenção. Pequenas oscilações positivas e negativas refletem apenas variações estocásticas, sem qualquer tendência ou sinal de alteração estrutural.

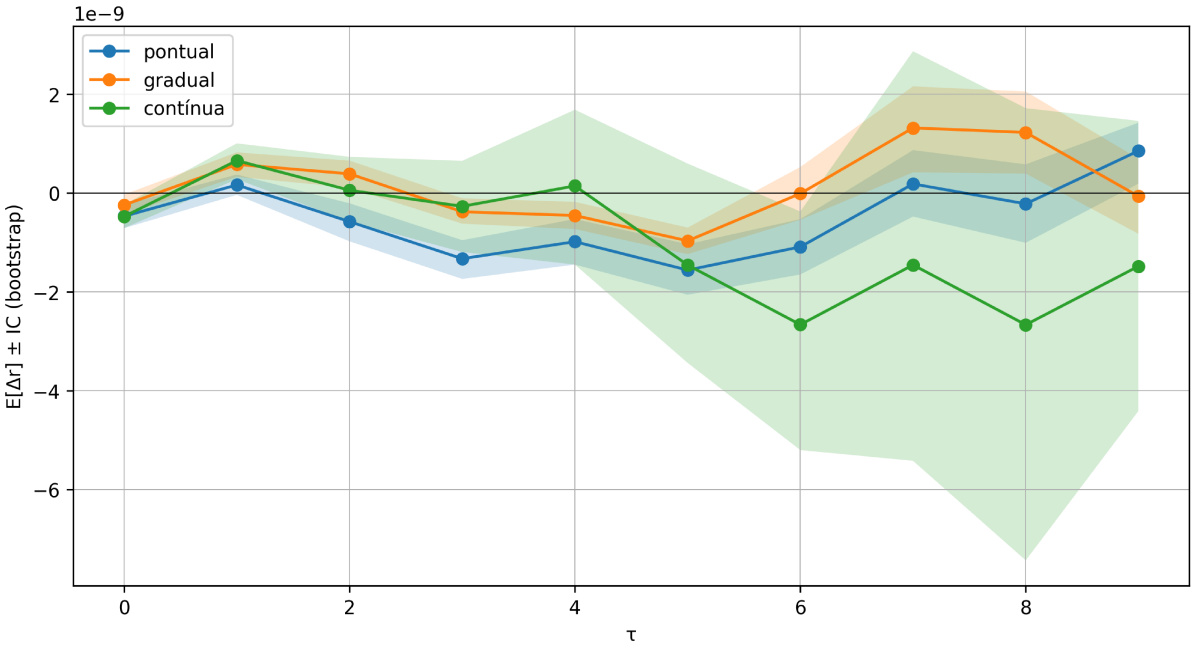

Figura 34 – Agregação temporal de Δr por tipo de intervenção — Direta, Linear - G1. Curvas médias $\mathbb{E}[\Delta r(\tau)]$ com intervalos de confiança obtidos por *bootstrap*. As três tipologias de intervenção apresentam comportamento oscilante em torno de zero, com variações na ordem de 10^{-9}, sem tendência significativa.

4.4.2 Grupo G2 - *Datasets* Não Lineares

Os *datasets* do grupo G2 empregam funções quadráticas para modelar os efeitos do tratamento sobre as variáveis e funções logarítmicas para os caminhos mediados ou de confundimento. Esse conjunto de transformações introduz dependências monotônicas não lineares, cuja intensidade aumenta à medida que o tratamento se afasta de sua região basal. Por essa razão, utilizamos a correlação de Chatterjee, apropriada para quantificar dependências não lineares.

A Tabela 9 apresenta as médias temporais das correlações factuais e contrafactuais ao longo dos períodos pós-intervenção. Observa-se que, em todas as estruturas, a dependência factual situa-se em níveis moderados $\overline{\xi_f} \in [0.52, 0.60]$. No contrafactual, verifica-se uma tendência média de maior dependência $\overline{\xi_f} \in [0.62, 0.68]$, produzindo diferenças médias $\overline{\Delta\xi} \approx 0{,}09$–$0{,}10$. Contudo, os desvios-padrão relativamente elevados revelam que esse aumento não ocorre de maneira uniforme, refletindo heterogeneidade substancial entre cenários e períodos.

Tabela 9 – Médias em τ (períodos pós-intervenção) das correlações de Chatterjee dos dados factuais ($\overline{\xi_f}$), contrafactuais ($\overline{\xi_{cf}}$) e diferenças ($\overline{\Delta\xi}$), no grupo G2.

Estrutura	$\overline{\xi_f} \pm dp$	$\overline{\xi_{cf}} \pm dp$	$\overline{\Delta\xi} \pm dp$
Direta	0.59 ± 0.022	0.68 ± 0.093	0.09 ± 0.099
Cadeia	0.52 ± 0.017	0.62 ± 0.068	0.10 ± 0.069
Confundidor	0.60 ± 0.033	0.68 ± 0.092	0.09 ± 0.102

$\overline{\xi_f}$: Média dos $\xi_f(\tau)$ da curva de correlação factual do *dataset*, N=10 ($\tau = 10$). $\overline{\xi_{cf}}$: Média dos $\xi_{cf}^{(k)}(\tau)$ das curvas de correlação contrafactual, N=300 ($k = 30$, $\tau = 10$).

As Figuras 35–37 ilustram exemplos representativos das correlações das três estruturas. Para fins de visualização, foram selecionados os casos de maior amplitude entre as replicações, onde as diferenças $\Delta\xi(\tau)$ atingem seus valores mais elevados após a intervenção, geralmente associados a intervenções contínuas com valores elevados de δ. Em contraste com o comportamento linear do grupo anterior, observa-se aqui uma separação clara entre as curvas factual e contrafactual, com $\xi_{cf}(\tau)$ crescendo de forma mais acentuada após a intervenção. A diferença $\Delta\xi(\tau)$ apresenta trajetória ascendente e estabiliza-se em valores entre 0,15 e 0,30, indicando que o tratamento intensifica progressivamente a dependência monotônica entre tratamento e resultado.

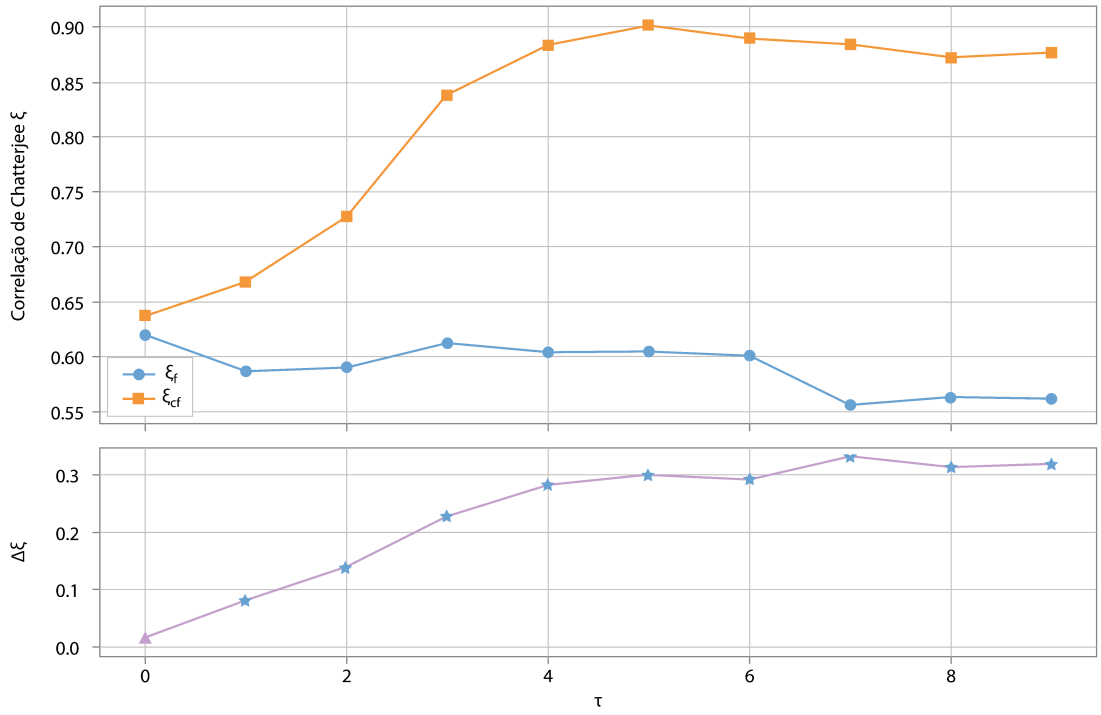

Figura 35 – Correlação e $\Delta\xi$ — Contínua, Direta, Não Linear, $\delta = 1.0$ - G2. Painel superior: Curvas de ξ_f e ξ_{cf} ao longo de τ. Painel inferior: diferença $\Delta\xi(\tau)$. Observa-se um aumento consistente da dependência no cenário contrafactual, com $\Delta\xi(t)$ crescendo após a intervenção, refletindo o efeito quadrático do tratamento..

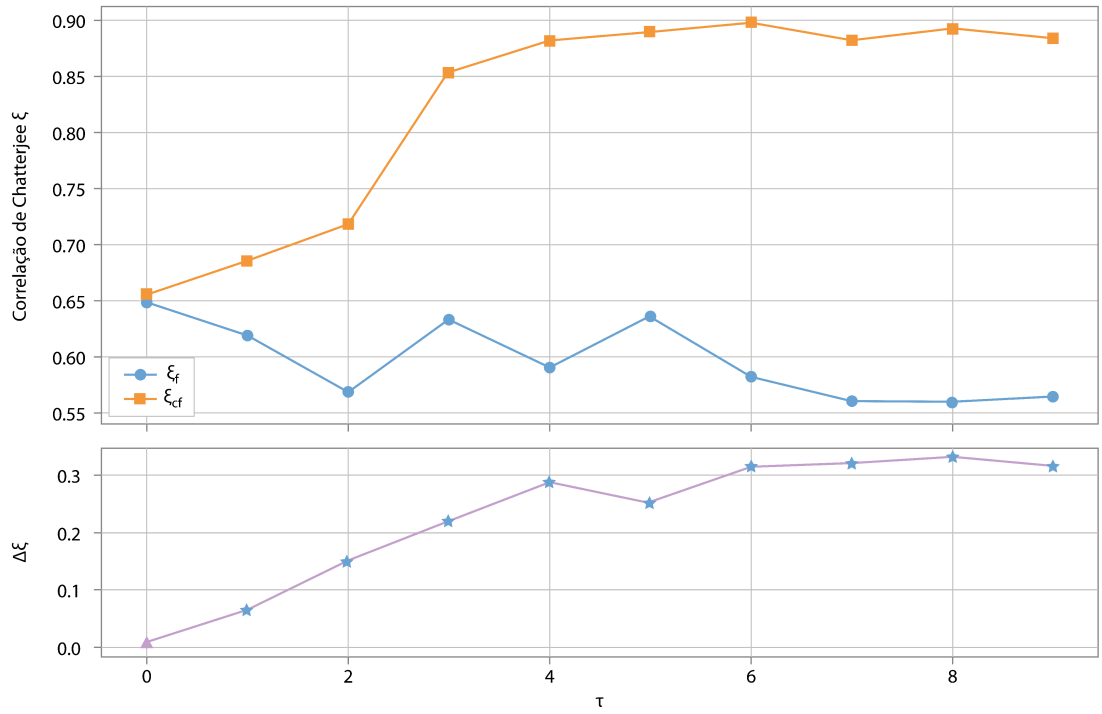

Figura 36 – Correlação e $\Delta\xi$ — Contínua, Confundidor, Não Linear, $\delta = 1.0$ - G2. Painel superior: Curvas de ξ_f e ξ_{cf} ao longo de τ. Painel inferior: diferença $\Delta\xi(\tau)$. A estrutura com confundidor exibe amplificação monotônica semelhante à da estrutura direta, com $\Delta\xi(\tau)$ crescendo de forma progressiva após a intervenção.

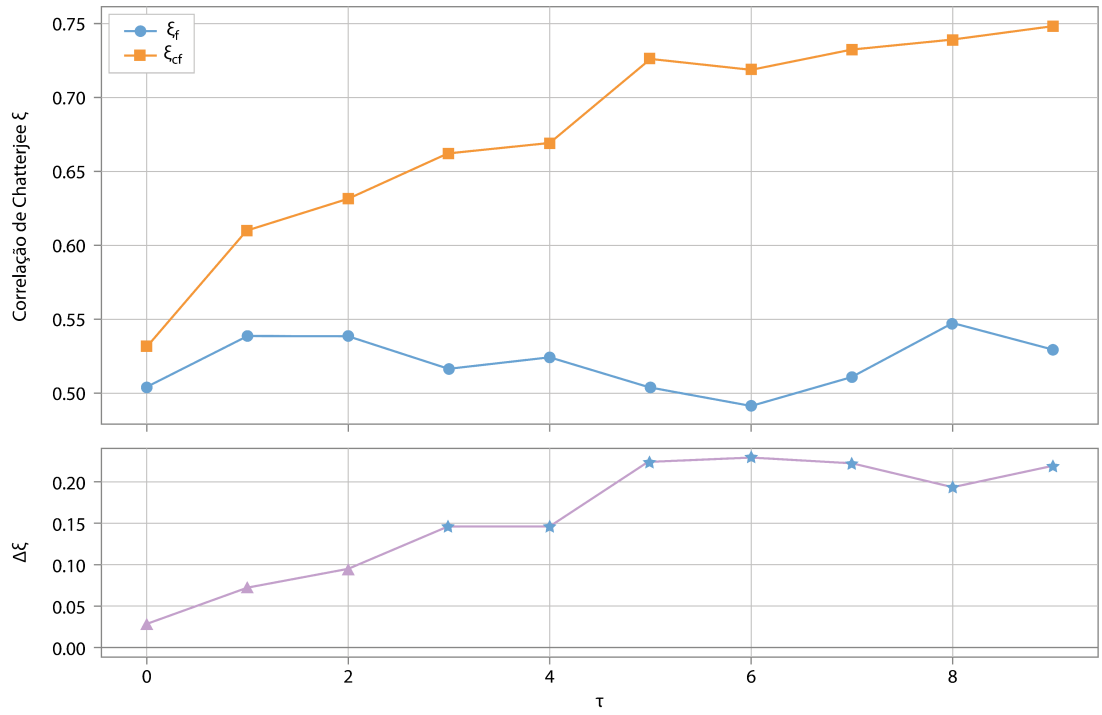

Figura 37 – Correlação e $\Delta\xi$ — Contínua, Cadeia, Não Linear, $\delta = 0.4$ - G2. Painel superior: ξ_f e ξ_{cf} ao longo de τ. Painel inferior: $\Delta\xi(\tau)$. O aumento da dependência é menos acentuado nos primeiros instantes, mas torna-se pronunciado a partir de $\tau \approx 5$, refletindo a mediação logarítmica no caminho $T \to X \to Y$.

A agregação temporal sintetiza os padrões observados nos cenários. As estruturas Direta e Confundidor, com representação na Figura 38, exibem comportamento muito semelhante: em ambas, as curvas médias $\mathbb{E}[\Delta\xi(\tau)]$ crescem de forma gradual ao longo da defasagem, mantendo intervalos de confiança relativamente compactos e trajetórias suaves. Esse paralelismo é consistente com o fato de que, nas duas estruturas, o efeito quadrático atua diretamente sobre o resultado, produzindo aumento progressivo da dependência monotônica.

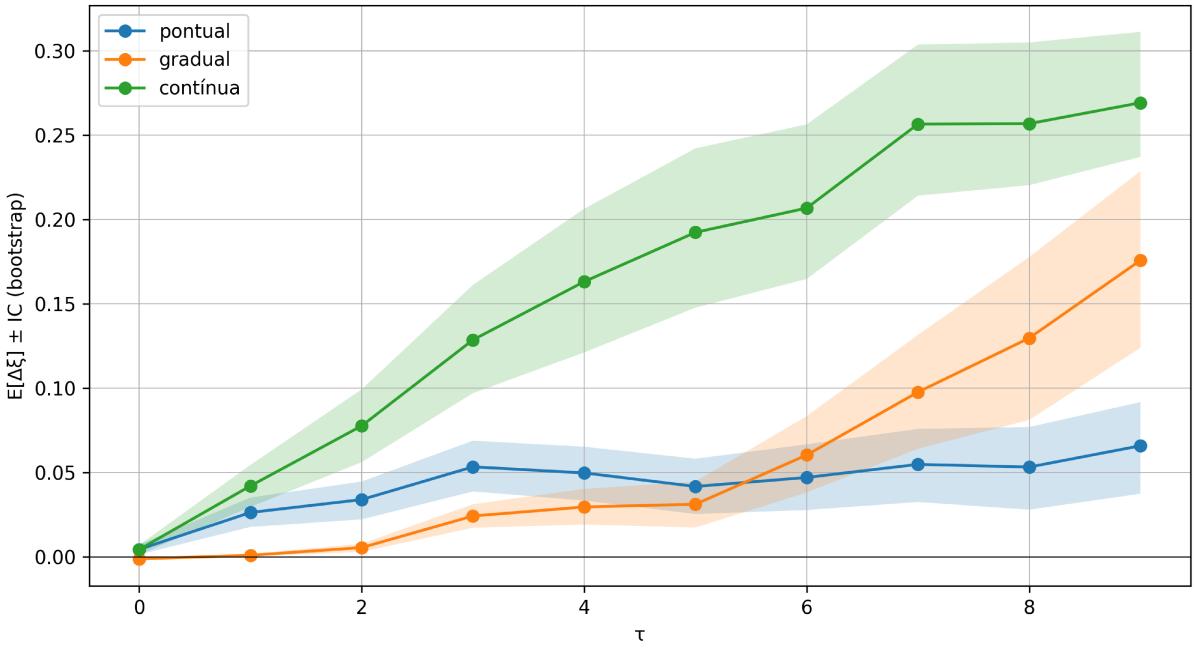

Figura 38 – Agregação temporal de $\Delta\xi$ por tipo de intervenção — Direta, Não Linear - G2. As curvas exibem crescimento progressivo da dependência monotônica, com trajetórias suaves e intervalos de confiança relativamente compactos. A intervenção contínua produz efeitos maiores, seguida da gradual e da pontual, atingindo valores acima de 0.25 em τ mais elevados.

A estrutura Cadeia, observada na Figura 39, apresenta crescimento mais irregular e intervalos de confiança mais amplos, refletindo a maior heterogeneidade introduzida pela propagação do efeito via variável mediadora X. Em todos os casos, a intervenção contínua é a mais pronunciada, alcançando valores superiores a 0.20–0.25 no final da série, seguida pela gradual e pela pontual, evidenciando o caráter cumulativo da função quadrática.

Esses resultados demonstram que, quando o sistema possui relações quadráticas e logarítmicas, intervenções em T não alteram apenas a média do resultado, mas modificam substancialmente sua ordenação entre os indivíduos — e esse efeito é capturado de maneira consistente pela correlação de Chatterjee.

4.4.3 Grupo G3 - *Datasets* Não Lineares

O Grupo G3 combina efeitos senoidais do tratamento com efeitos mediados modelados por função logarítmica. Essa combinação produz padrões temporais mais dinâmicos, com

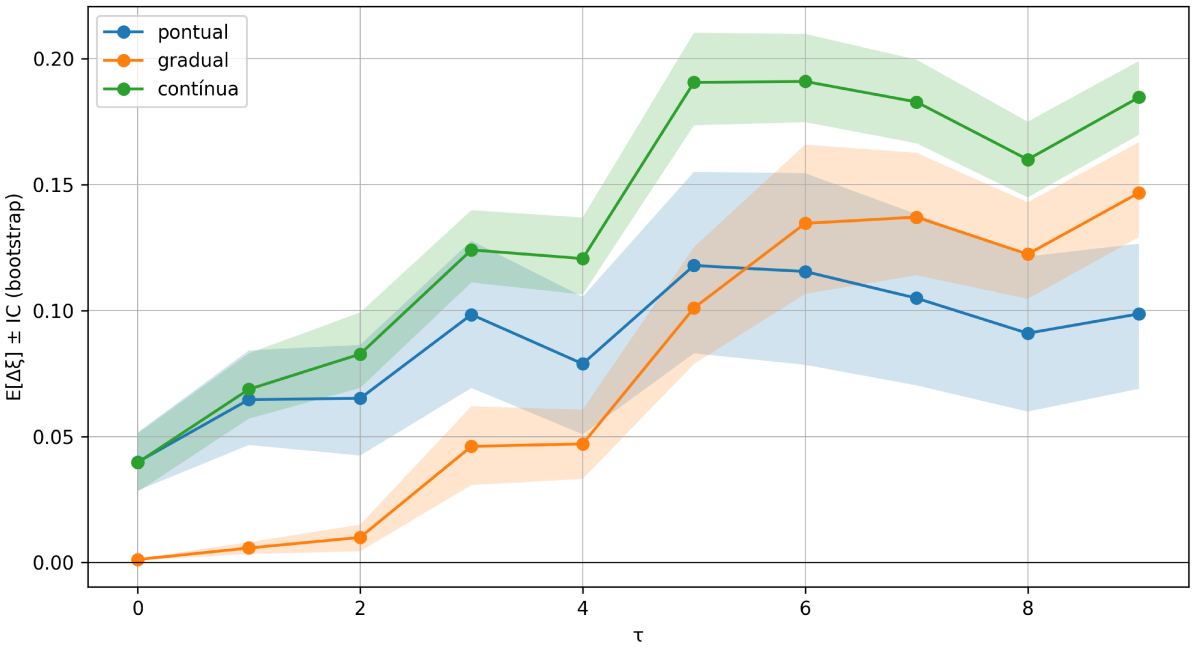

Figura 39 – Agregação temporal de $\Delta\xi$ por tipo de intervenção — Cadeia, Não Linear - G2. As curvas exibem crescimento mais irregular e ICs mais amplos do que nos casos Direta e Confundidor, refletindo a heterogeneidade adicional introduzida pela mediação via X. A intervenção contínua permanece a mais intensa, alcançando valores próximos a 0.20, seguida da gradual e da pontual.

oscilações na associação monotônica entre tratamento e resultado, e respostas comprimidas nos caminhos mediados. A Tabela 10 resume as médias temporais das correlações factual, contrafactual e da diferença $\Delta\xi$, revelando que todas as estruturas apresentam uma tendência de redução média da associação factual–contrafactual.

Tabela 10 – Médias em τ (períodos pós-intervenção) das correlações de Chatterjee dos dados factuais ($\overline{\xi_f}$), contrafactuais ($\overline{\xi_{cf}}$) e diferenças ($\overline{\Delta\xi}$), no grupo G3.

Estrutura	$\overline{\xi_f} \pm dp$	$\overline{\xi_{cf}} \pm dp$	$\overline{\Delta\xi} \pm dp$
Direta	0.79 ± 0.014	0.76 ± 0.058	-0.04 ± 0.051
Cadeia	0.58 ± 0.013	0.51 ± 0.072	-0.06 ± 0.069
Confundidor	0.79 ± 0.013	0.75 ± 0.063	-0.04 ± 0.057

$\overline{\xi_f}$: Média dos $\xi_f(\tau)$ da curva de correlação factual do *dataset*, N=10 ($\tau = 10$). $\overline{\xi_{cf}}$: Média dos $\xi_{cf}^{(k)}(\tau)$ das curvas de correlação contrafactual, N=300 ($k = 30$, $\tau = 10$).

Nas Figuras 40–42, observa-se que as curvas contrafactuais tendem a decair mais rapidamente do que as factuais ao longo de τ, sendo essa diferença mais pronunciada nas estruturas Cadeia e Confundidor. Em todos os casos, $\Delta\xi(t)$ apresenta tendência negativa, crescendo em magnitude com τ, o que indica perda gradual da associação monotônica $T \leftrightarrow Y$ após a intervenção.

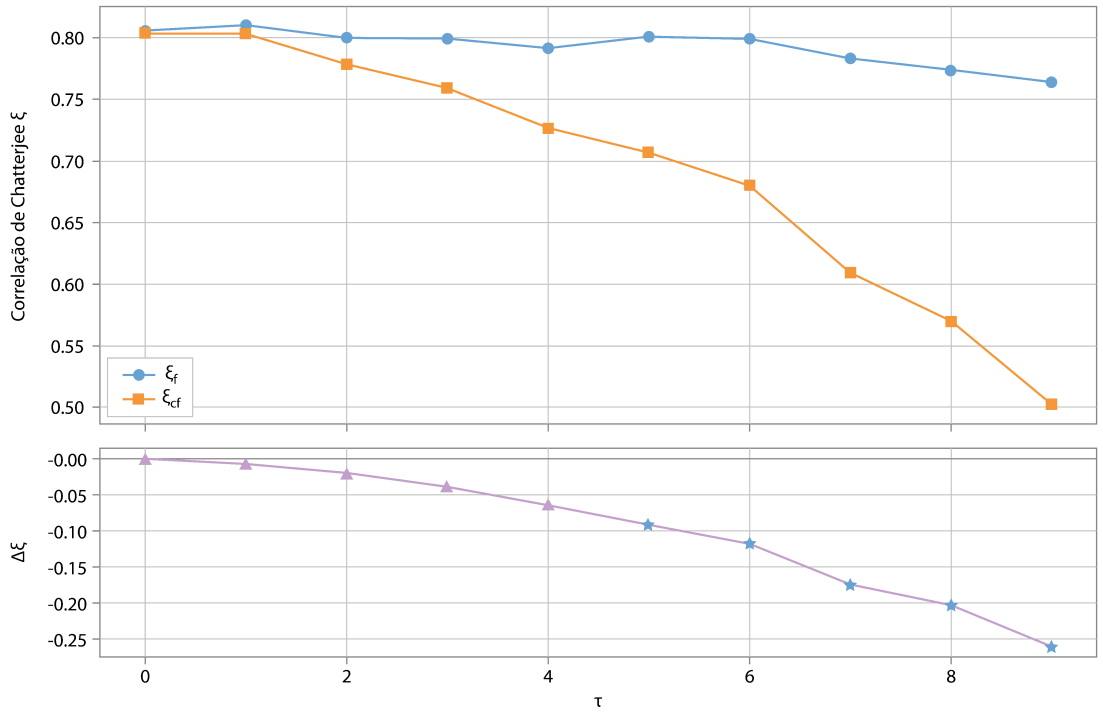

Figura 40 – Correlação e $\Delta\xi$ — Contínua, Direta, Não Linear, $\delta = 0.4$ - G3. Painel superior: Curvas de ξ_f e ξ_{cf} ao longo de τ. Painel inferior: diferença $\Delta\xi(\tau)$. O factual mantém correlações elevadas e estáveis, o contrafactual exibe queda moderada ao longo de τ. A diferença $\Delta\xi(\tau)$ torna-se gradualmente negativa, mas permanece de baixa magnitude, indicando impacto limitado da intervenção.

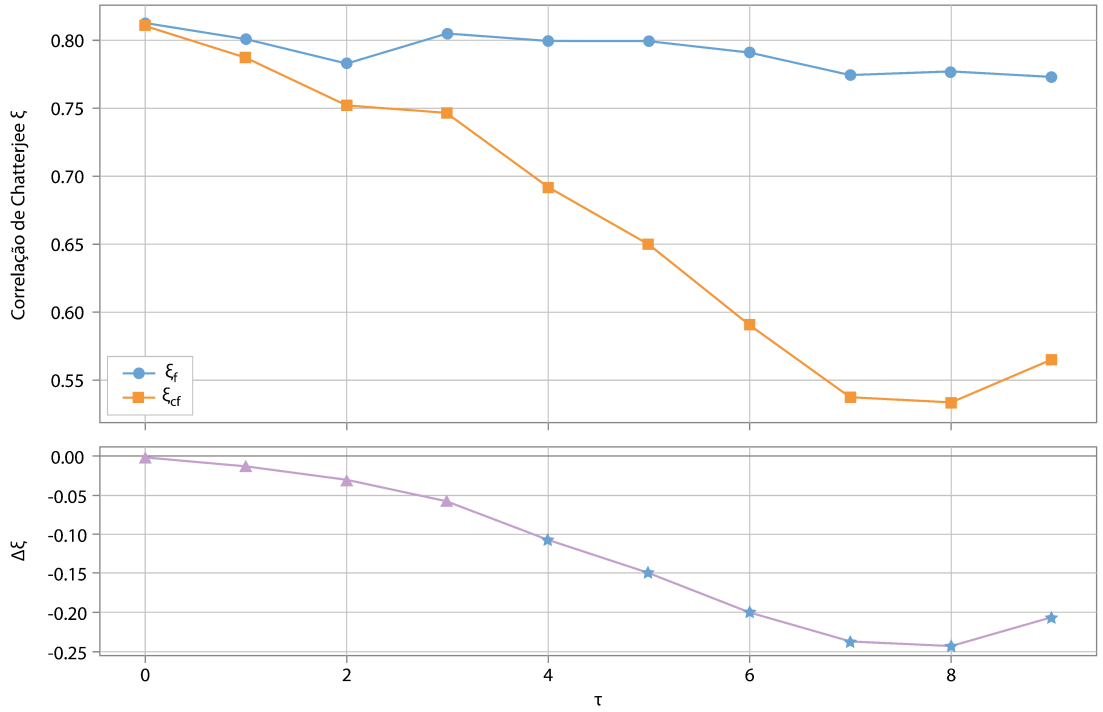

Figura 41 – Correlação e $\Delta\xi$ — Contínua, Confundidor, Não Linear, $\delta = 0.5$ - G3. Painel superior: Curvas de ξ_f e ξ_{cf} ao longo de τ. Painel inferior: diferença $\Delta\xi(\tau)$. O factual apresenta queda suave, o contrafactual sofre redução mais acentuada, refletindo o efeito amplificador do confundidor logarítmico. A diferença $\Delta\xi(\tau)$ cresce negativamente com τ, indicando perda consistente da associação.

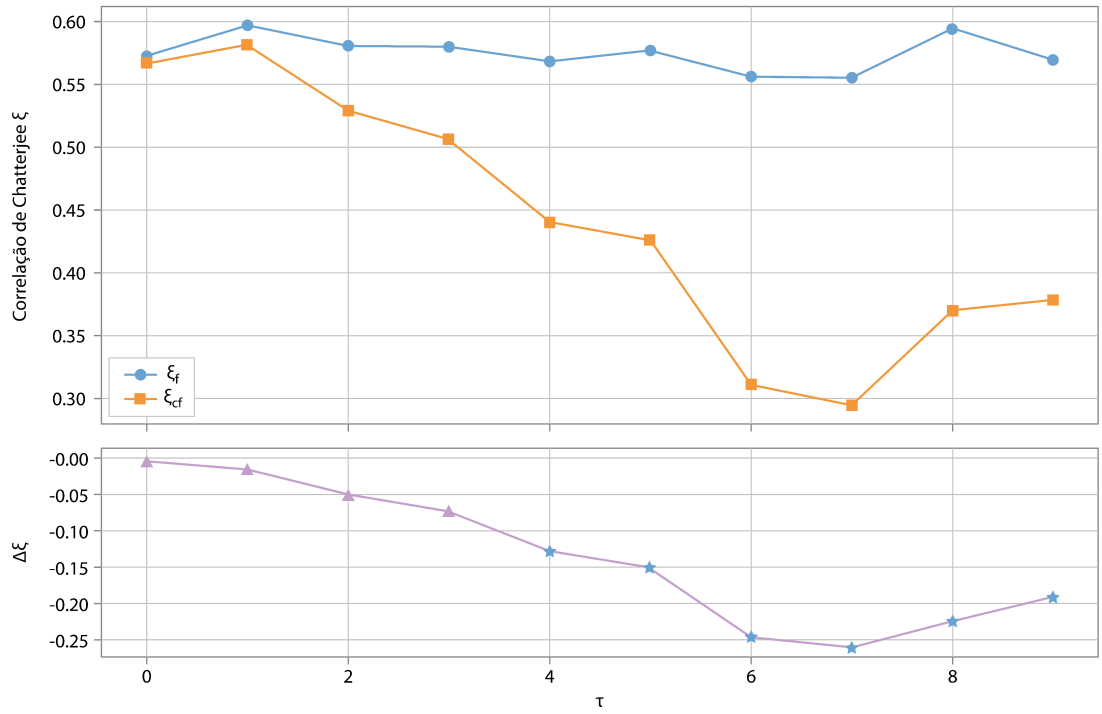

Figura 42 – Correlação e $\Delta\xi$ — Contínua, Cadeia, Não Linear, $\delta = 0.3$ - G3. Painel superior: Curvas de ξ_f e ξ_{cf} ao longo de τ. Painel inferior: diferença $\Delta\xi(\tau)$. A mediação via X intensifica o efeito da intervenção: ξ_{cf} decai mais rapidamente e atinge os valores mais baixos entre as estruturas. A diferença $\Delta\xi(\tau)$ apresenta o declive negativo mais pronunciado, sugerindo forte sensibilidade aos efeitos acumulados da intervenção.

As Figuras 43 e 44 mostram as curvas médias $\mathbb{E}[\Delta\xi(\tau)]$ e seus intervalos de confiança. Nos três cenários estruturais, observa-se um padrão consistente: intervenções contínuas produzem quedas progressivas de $\Delta\xi$, intervenções graduais apresentam declínio intermediário, enquanto intervenções pontuais permanecem próximas de zero em toda a defasagem.

A estrutura Confundidora apresenta comportamento muito semelhante ao da estrutura Direta, reforçando que, sob modulação senoidal, apenas intervenções prolongadas são capazes de alterar a dependência monotônica de forma sistemática. O comportamento mais acentuado ocorre na estrutura Cadeia, na qual o declínio de $\Delta\xi(\tau)$ é mais rápido e profundo, aproximando-se de um padrão quase linear até $\tau \approx 8$, refletindo a amplificação causada pelo caminho mediado $T \to X \to Y$.

Esses resultados mostram que, em sistemas com efeitos senoidais e compressão logarítmica, intervenções em T não apenas deslocam o resultado ao longo do tempo, mas também alteram progressivamente a ordenação relativa entre os indivíduos — sobretudo em intervenções graduais e contínuas — um comportamento capturado de forma sensível pela correlação de Chatterjee.

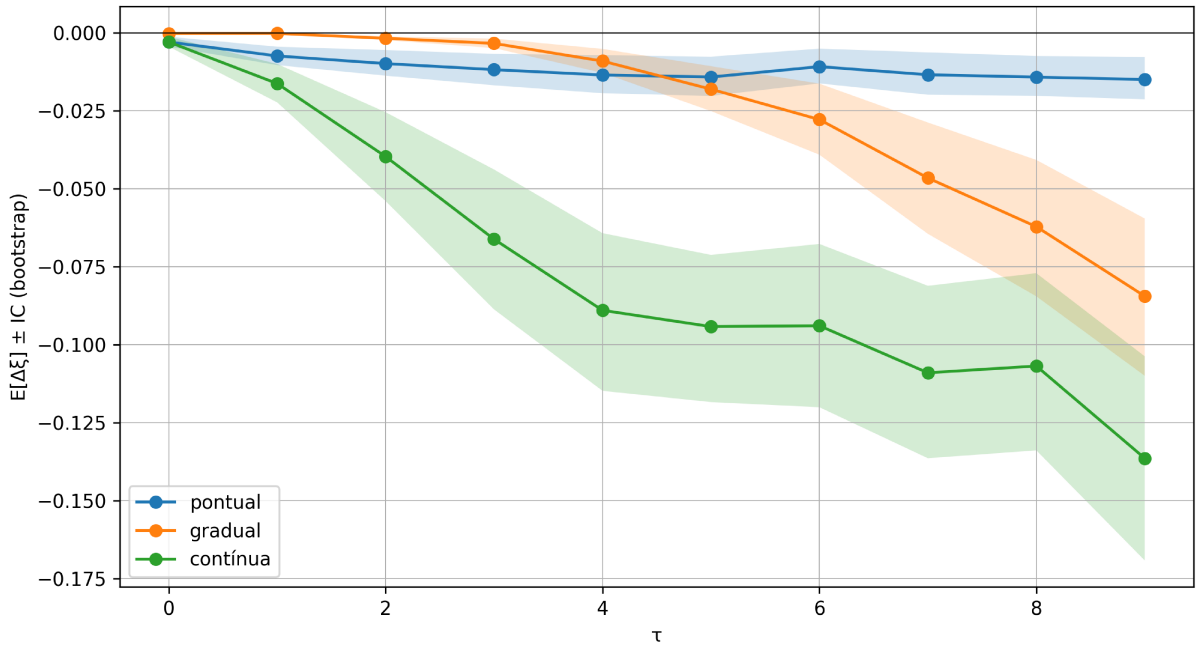

Figura 43 – Agregação temporal de $\Delta\xi$ por tipo de intervenção — Direta, Não Linear — G3. Nas intervenções pontuais permanece próximas de zero ao longo de toda a defasagem, enquanto nas intervenções graduais apresenta leve declínio. Nas intervenções contínuas exibe reduções mais claras, ainda que moderadas, com intervalos de confiança estreitos indicando baixa variabilidade entre replicações.

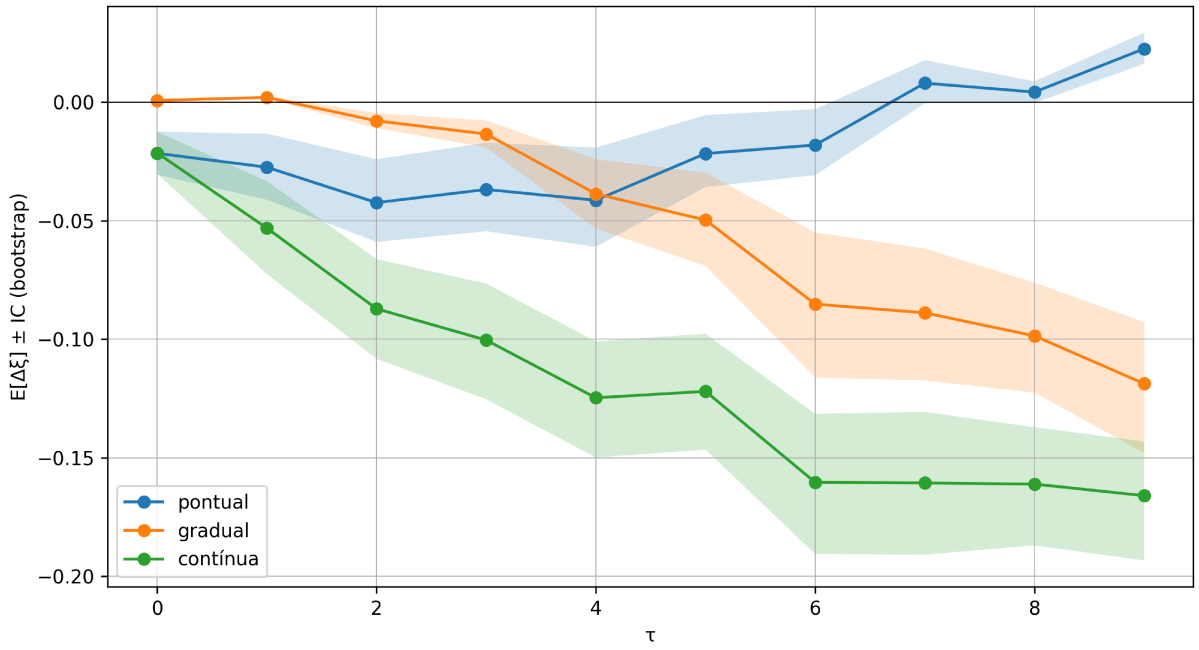

Figura 44 – Agregação temporal de $\Delta\xi$ por tipo de intervenção — Cadeia, Não Linear — G3. Esta estrutura apresenta as reduções mais intensas de $\Delta\xi$ dentre as três, especialmente sob intervenções contínuas, que exibem declínio quase linear ao longo da defasagem. O caminho mediado $T \to X \to Y$ amplifica o impacto das oscilações senoidais aplicadas ao tratamento, resultando em efeitos acumulativos mais pronunciados.

4.5 Conclusão da Análise

A análise conduzida ao longo deste capítulo evidenciou que os dados sintéticos produzidos pelo CSDG reproduzem de maneira consistente os comportamentos esperados para cada um dos cenários causais definidos. As medidas avaliadas — efeitos médios, curvas contrafactuais, correlações factuais e contrafactuais e deltas de correlação — apresentaram padrões compatíveis tanto com a estrutura causal quanto com o tipo de intervenção aplicada, confirmando que o gerador é capaz de produzir dinâmicas coerentes com os mecanismos que lhe deram origem.

Nos cenários lineares, observou-se um comportamento mais estável, com variações suaves ao longo do tempo e diferenças reduzidas entre factual e contrafactual. Esses cenários servem, portanto, como um *baseline* natural, útil para validar implementações, calibrar métodos e verificar se métricas e testes são sensíveis às mudanças mais elementares na estrutura dos dados. A previsibilidade desses cenários também facilita identificar comportamentos anômalos, pois qualquer variação inesperada tende a indicar inconsistências no processo de geração ou análise, e não um efeito legítimo do cenário.

Os cenários não lineares introduziram oscilações mais expressivas, mudanças acentuadas na associação entre tratamento e resultado e efeitos pós-intervenção mais persistentes. Nessas condições, as diferenças entre factual e contrafactual tornam-se mais pronunciadas e heterogêneas, refletindo maior sensibilidade às interações entre variáveis e aos graus de complexidade especificados nos parâmetros. Tais cenários representam um desafio mais realista e exigente para modelos de inferência causal e, por isso, são especialmente adequados para *benchmarking*, testes de robustez e avaliação da capacidade de capturar efeitos dinâmicos e não lineares.

A intervenção gradual e a contínua destacaram-se por amplificar diferenças temporais, sobretudo em cadeias e cenários com confundidores. Isso demonstra que o gerador é capaz de representar modificações acumuladas no tempo, algo essencial para estudos longitudinais realistas. As intervenções pontuais mantiveram proporções mais baixas de tempos significativos, como esperado em perturbações de curta duração.

De forma geral, os resultados mostram que o CSDG oferece variabilidade suficiente de estruturas, intensidades de efeitos e regimes de intervenção, permitindo que diferentes níveis de complexidade sejam explorados de acordo com o objetivo experimental. A coerência observada entre o comportamento empírico dos dados e os mecanismos causais que os geraram confirma a adequação do gerador como ferramenta de estudo, simulação e *benchmark* para métodos de inferência causal em séries longitudinais.

Assim, concluímos que a análise apresentada valida tanto a confiabilidade quanto a flexibilidade do CSDG, reforçando sua utilidade como base experimental para o desenvolvimento, comparação e avaliação de modelos que buscam estimar efeitos causais dinâmicos em contextos com diferentes formas de não linearidade, interação e intervenção.

Capítulo 5

Aplicação: Prova de Conceito com Aprendizado Temporal Causal

Este capítulo apresenta um experimento como prova de conceito de uso dos dados gerados pelo CSDG. O experimento proposto é baseado no aprendizado de estrutura causal a partir de dados observacionais históricos e estimativa de resultados potenciais a partir de tratamentos futuros definidos, ou seja, modelos de resultados potenciais. Compara-se o desempenho de modelos de regressão linear - tomado como uma *baseline* - e redes neurais recorrentes na predição factual e contrafactual, utilizando as métricas RMSE e PEHE. Assim, a prova de conceito busca demonstrar a utilidade dos dados sintetizados pelo CSDG para a avaliação de algoritmos de resultados potenciais, destacando a capacidade do gerador em fornecer cenários adequados para comparação de métodos.

5.1 Dados Utilizados

Cada instância representa um indivíduo com as sequências temporais das variáveis de tratamento (T) e resultado (Y) ao longo de $t = 20$ períodos. As séries foram geradas utilizando a estrutura causal direta com relação linear e com os parâmetros definidos na Tabela 11. Uma intervenção pontual foi aplicada no tratamento no período $t = 10$, com intensidade $\delta_t = 0.5$, de modo que os resultados contrafactuais são gerados a partir do período $t_{int} = 10$.

Tabela 11 – Parâmetros de geração dos dados da prova de conceito.

Coeficientes temporais	$\phi_T = 0.8$, $\phi_Y = 0.7$
Coeficientes causais	$\beta_{ty} = 1.5$
Ruído (distribuição)	`uniform`
Intervalos do ruído	T: $[-0.1, 0.1]$, Y: $[-0.1, 0.1]$

Para fins de avaliação, dividimos a sequência de cada indivíduo em duas partes: histó-

rico e horizonte de predição. Definimos t_{h0} como o tempo que marca o início do horizonte, a origem a partir da qual as previsões são avaliadas; adotamos $t_{h0} = t_{\text{int}}$ para alinhar a avaliação ao marco da intervenção. Usamos a defasagem $\tau = t - t_{\text{int}}$, assim o histórico $t < t_{h0}$ corresponde a $\tau < 0$, e o horizonte $t \geq t_{h0}$ a $\tau \geq 0$, sendo $\tau = 0$ o primeiro passo do horizonte. No nosso experimento definimos $t_{\text{int}} = 10$, portanto o histórico vai de $t = 0$ a $t = 9$ e o horizonte de $t = 10$ a $t = 19$, e nos dados contrafactuais, o período pós-intervenção no tratamento coincide com o horizonte de predição. A Figura 45 exemplifica um indivíduo com o corte dos dados em histórico e horizonte de predição.

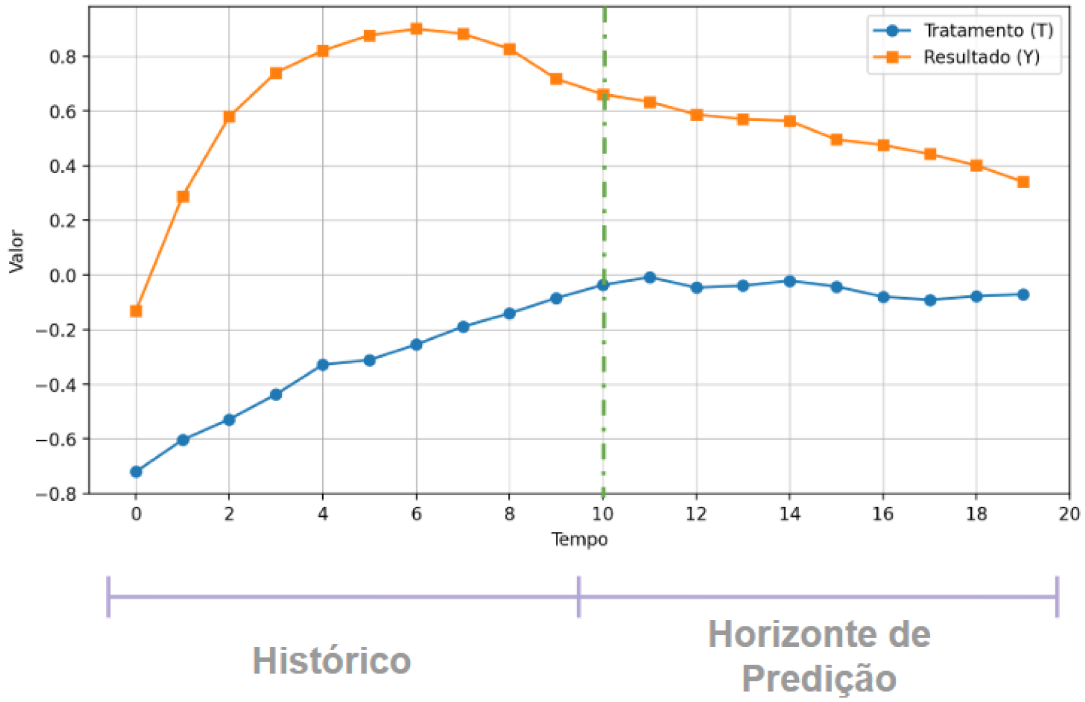

Figura 45 – Histórico e Horizonte de Predição. Exemplo de dado factual utilizado no experimento, onde temos $t_{h0} = 10$, sendo então o histórico definido pelo período $t = 0$ a $t = 9$ e o horizonte de predição de $t = 10$ a $t = 19$.

5.2 Modelos

Foram utilizados modelos de Regressão Linear e variantes de redes neurais recorrentes para modelar dependências temporais e causais em dados sequenciais. As redes utilizadas foram: *Recurrent Neural Network* (RNN) (ELMAN, 1990), *Long Short-Term Memory* (LSTM) (HOCHREITER; SCHMIDHUBER, 1997) e *Gated Recurrent Unit* (GRU) (CHO et al., 2014).

Todas as três redes neurais foram estruturadas na arquitetura *Encoder-Decoder*, conforme proposta para redes LSTM por Sutskever, Vinyals e Le (2014), com o objetivo de capturar a dinâmica temporal e as dependências causais presentes nas séries sintéticas geradas. Nesta arquitetura, o *encoder* recebe as sequências T, X e Y do histórico e gera uma representação latente da evolução temporal e da estrutura causal. Essa representação

é passada ao *decoder*, que utiliza as variáveis T e X do horizonte de previsão para estimar, de forma autorregressiva, os próximos valores de Y, como estão ilustrado na Figura 46.

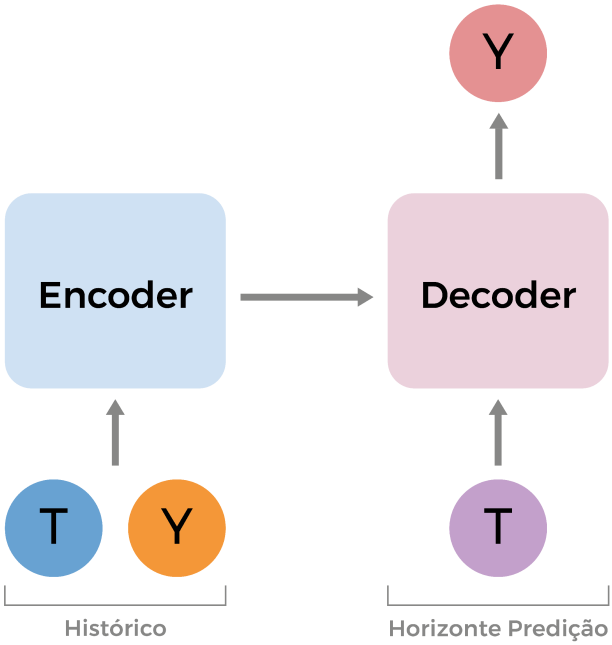

Figura 46 – Arquitetura *Encoder-Decoder*. Diagrama simplificado da arquitetura implementada para as redes neurais. A camada de *encoder* tem como entrada dados históricos de Tratamento e Resultado do indivíduo, o *decoder* recebe a repesentação gerada pelo *encoder* e o Tratamento futuro, gerando o resultado futuro como predição.

Os parâmetros de configuração das redes neurais pode ser consultado na Tabela 12. Para cada combinação de estrutura de dados sintéticos e arquitetura (RNN, GRU, LSTM), executamos 5 repetições independentes, variando a semente aleatória.

Tabela 12 – Hiperparâmetros utilizados nas redes neurais *baseline*.

Parâmetro	Valor
Dimensão oculta (d_h)	32
Número de camadas (ED)	4
Dropout	0.10
Otimizador	Adam
Taxa de aprendizado	0.01
Weight decay	0.0
Gradient clipping	1.0 (norma)
Batch size (treino)	250
Batch size (validação)	125
Número máximo de épocas	300
Early stopping	*patience* = 30

5.3 Resultados

Esta seção apresenta e discute os resultados obtidos na prova de conceito. Os experimentos tiveram como objetivo avaliar o desempenho de diferentes modelos de aprendizado temporal na predição factual e contrafactual, considerando métricas complementares de erro e consistência causal. As subseções a seguir estão organizadas conforme o tipo de avaliação realizada, seguidas de uma síntese interpretativa dos achados e suas implicações.

5.3.1 Predição factual

A Tabela 13 apresenta os resultados da predição factual em termos de RMSE. Observa-se que a Regressão Linear apresenta aumento progressivo de erro conforme o horizonte τ se distancia do tempo atual, o que indica sua limitação em capturar as dependências temporais presentes nos dados. Em contraste, as arquiteturas recorrentes (RNN, LSTM e GRU) mantêm desempenho estável mesmo em horizontes mais longos, refletindo maior capacidade de modelar relações dinâmicas e acumular informação ao longo do tempo.

Tabela 13 – RMSE da predição factual após 5 execuções para τ passos (média ± desvio padrão).

Modelo	$\tau = 0$	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
Regressão Linear	0,548	0,683	0,727	0,832	0,915	1,023
RNN	0,136 ± 0,007	0,162 ± 0,011	0,168 ± 0,010	0,171 ± 0,010	0,152 ± 0,010	0,145 ± 0,008
LSTM	0,127 ± 0,004	0,142 ± 0,002	**0,144 ± 0,002**	**0,152 ± 0,001**	**0,136 ± 0,001**	**0,132 ± 0,001**
GRU	**0,122 ± 0,002**	**0,140 ± 0,002**	0,145 ± 0,002	**0,152 ± 0,001**	0,139 ± 0,001	0,135 ± 0,001

Nota-se também que os modelos recorrentes apresentam desvios padrão pequenos, o que indica alta estabilidade entre execuções. As arquiteturas LSTM e GRU obtêm erros ligeiramente menores que a RNN simples, mas, dado o número de execuções e a proximidade entre as médias, não é possível afirmar estatisticamente que uma arquitetura supera consistentemente a outra em horizontes curtos. Assim, LSTM e GRU podem ser interpretadas como apresentando desempenho equivalente nesse cenário, ambas capturando de forma eficiente as relações temporais subjacentes ao processo factual.

5.3.2 Predição contrafactual

A Tabela 14 apresenta os resultados da métrica PEHE, que avalia a precisão das estimativas contrafactuais. Os valores mostram diferenças expressivas entre a abordagem estatística e as redes neurais. A Regressão Linear apresenta valores superiores a 1.0 em todos os horizontes, indicando incapacidade de capturar a heterogeneidade dos efeitos e o comportamento dinâmico dependente do histórico.

Em contraste, as arquiteturas recorrentes exibem PEHE muito baixas — da ordem de 10^{-2} — representando erros substancialmente inferiores e alcançando valores próximos de

zero a partir dos primeiros passos do horizonte de previsão, especialmente nas arquiteturas LSTM e GRU.

Tabela 14 – PEHE das predições factual e contrafactual após 5 execuções para τ passos (média ± desvio padrão).

Modelo	$\tau = 0$	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
Regressão Linear	1,048	1,257	1,339	1,314	1,224	1,102
RNN	0,023 ± 0,012	0,043 ± 0,025	0,049 ± 0,027	0,043 ± 0,024	0,032 ± 0,019	0,023 ± 0,013
LSTM	0,027 ± 0,015	**0,018 ± 0,008**	**0,015 ± 0,004**	**0,012 ± 0,003**	**0,008 ± 0,002**	**0,006 ± 0,002**
GRU	**0,017 ± 0,003**	0,021 ± 0,004	0,018 ± 0,004	0,014 ± 0,003	0,011 ± 0,002	0,009 ± 0,003

Entre as arquiteturas recorrentes, a GRU apresenta o menor erro em $\tau = 0$, mas a LSTM demonstra desempenho consistentemente superior para $\tau \geq 1$, com diferenças entre médias que superam os desvios padrão, indicando vantagem estatisticamente significativa mesmo com cinco execuções. Assim, a LSTM destacou-se, neste contexto, como a arquitetura mais eficaz para capturar efeitos individuais em horizontes mais distantes.

5.3.3 Síntese e perspectivas

De forma geral, as arquiteturas LSTM e GRU apresentaram os melhores resultados tanto em predição factual quanto contrafactual, confirmando a hipótese de que modelos com mecanismos de memória são mais adequados para capturar relações causais dinâmicas em séries longitudinais. Esses resultados sustentam a utilização do CSDG como ferramenta de avaliação controlada e reprodutível, capaz de evidenciar diferenças entre modelos sob condições experimentais conhecidas.

Como trabalhos futuros, há oportunidade de ampliar as análises para estruturas não lineares e cenários com confundimento, além de avaliar modelos baseados em *Transformers*, como o *Causal Transformer*, de modo a investigar a capacidade desses modelos em capturar dependências causais de longo prazo.

Capítulo 6

Conclusão

O método de geração de dados proposto neste trabalho permite a criação de dados longitudinais sintéticos com estrutura causal controlada, sendo útil para avaliação de algoritmos de inferência causal em diversos contextos. Entre as aplicações possíveis, destacam-se: aprendizado de estrutura causal, estimativa de efeitos médios e individuais, e simulação de resultados potenciais. O uso de covariáveis também é versátil, permitindo simular cenários com variáveis observáveis, ocultas ou com papel causal conhecido.

6.1 Principais Contribuições

As principais contribuições desta pesquisa estão associadas ao desenvolvimento de uma abordagem formal e experimental para a geração, análise e validação de dados sintéticos longitudinais com estrutura causal explícita. O trabalho apresenta avanços conceituais, metodológicos e empíricos que, em conjunto, consolidam um novo referencial para estudos de inferência causal temporal. A seguir, são descritas as contribuições centrais e os resultados que as sustentam.

Uso de Equações Estruturais Causais Autorregressivas. Foi proposto um formalismo matemático que integra dependências causais e temporais em um mesmo modelo, permitindo a representação explícita das relações entre tratamento (T_t), resultado (Y_t) e covariáveis (X_t) ao longo do tempo. Essa formulação, descrita no Capítulo 3, com as Equações 18–22, constitui a base teórica do gerador e diferencia-se de abordagens tradicionais por capturar simultaneamente efeitos contemporâneos e dinâmicos. Essa estrutura fornece controle sobre a intensidade das dependências autorregressivas (Φ) e dos efeitos causais (β), o que possibilita simulações mais realistas de processos longitudinais.

Desenvolvimento do gerador *Causal Synthetic Data Generator* (CSDG). A partir desse formalismo, foi implementado o CSDG, um gerador de dados sintéticos longitudinais com estrutura causal explícita, descrito no Capítulo 3. O

sistema permite definir o número de indivíduos, o tamanho das séries, o tipo de estrutura causal - Direta, Cadeia ou Confundidor - e o regime de intervenção - pontual, gradual ou contínua. Além disso, possibilita a geração de dados factuais e contrafactuais, permitindo o estudo empírico de cenários de intervenção controlada. Exemplos ilustrativos de dados gerados são apresentados nas Figuras 6, 8 e 10, demonstrando a flexibilidade do método em capturar diferentes padrões de causalidade e complexidade funcional. O CSDG está disponível no repositório <https://github.com/angeruzzi/causal-synthetic-data-gen> em licença *open-source*.

Definição de um protocolo de avaliação da coerência causal. Foi proposto um protocolo sistemático para quantificar a coerência causal dos dados gerados, combinando métricas de efeito médio do tratamento (ATE) e mudança de correlação (Δr) combinado com intervalos de confiança obtidos por *bootstrap*. Esse protocolo, detalhado na Seção 2.7 e aplicado no Capítulo 4 aos nove *datasets* gerados apresentados na Tabela 7, permitiu avaliar de forma padronizada o comportamento das intervenções e a consistência temporal dos efeitos.

Em conjunto, essas contribuições reforçam o papel do CSDG como uma ferramenta aberta, reprodutível e metodologicamente fundamentada para experimentação causal em contextos longitudinais. O gerador permite a criação de cenários realistas com *ground truth* causal conhecido e oferece um ambiente controlado para testar e comparar algoritmos de aprendizado causal.

6.2 Trabalhos Futuros

Um avanço relevante na estrutura do gerador seria permitir que o usuário informe uma estrutura causal personalizada por meio de uma matriz de adjacência, um SCM ou outra notação compatível. Essa abordagem tornaria o gerador ainda mais flexível, possibilitando a simulação de cenários específicos com múltiplos caminhos causais, tratamentos simultâneos ou estruturas híbridas, conforme as necessidades de diferentes experimentos.

A simulação de cenários com interferência entre indivíduos, que representa um dos casos mais desafiadores na área de inferência causal, também seria um adendo importante, permitindo avaliar métodos robustos a violações da suposição de estabilidade dos tratamentos.

A expansão experimental da prova de conceito, incorporando outras estruturas causais, relações não lineares de maior complexidade e cenários mistos de intervenções, pode não apenas tornar a avaliação do CSDG mais abrangente, como também estabelecer um *benchmark* útil para pesquisas futuras.

Outra direção promissora é o desenvolvimento de *benchmarks* temáticos, com características específicas de domínios reais, como avaliação de políticas públicas, análises econômicas ou modelagem da movimentação de ativos financeiros. Esses conjuntos sintéticos orientados ao domínio podem apoiar a comparação padronizada de modelos e algoritmos.

Os modelos avaliados no contexto de resultados potenciais também podem ser ampliados, incluindo adaptações de técnicas utilizadas em contextos análogos — como o *Causal Transformer* — a fim de investigar sua capacidade de capturar dependências causais de longo prazo em sistemas dinâmicos e heterogêneos.

Por fim, uma direção promissora consiste em expandir o escopo das análises realizadas, utilizando os dados gerados pelo CSDG para abordar outras classes de problemas de inferência causal, indo além do paradigma de resultados potenciais explorado neste trabalho. Em particular, os dados sintéticos podem ser empregados em estudos de identificação causal, cuja meta é determinar, a partir de um grafo causal ou de hipóteses estruturais, se e como um efeito causal pode ser identificado unicamente a partir da distribuição observada. Diferentemente da estimativa de contrafactuais — que parte da suposição de identificabilidade e busca predizer resultados potenciais individuais — os problemas de identificação focam em derivar expressões formais para quantidades causais por meio de regras de cálculo causal, como o *do-calculus*. Assim, um trabalho futuro relevante é investigar como modelos e algoritmos voltados à identificação causal se comportam quando avaliados em dados gerados pelo CSDG, especialmente em cenários com variáveis latentes, mediadores não observados ou estruturas parcialmente especificadas.

6.3 Contribuições em Produção Bibliográfica

Os resultados parciais desta dissertação originaram uma publicação científica apresentada no evento *KDMiLe 2025 – Symposium on Knowledge Discovery, Mining and Learning*, realizado em Fortaleza, Ceará, no contexto do Simpósio Brasileiro de Banco de Dados (SBBD 2025).

O artigo intitulado ***"Longitudinal Synthetic Data Generation from Causal Structures"*** apresenta a proposta inicial do gerador *Causal Synthetic Data Generator* (CSDG), descrevendo suas equações estruturais causais autorregressivas e o experimentos de validação. A publicação está disponível em acesso aberto no portal da Sociedade Brasileira de Computação (SBC) por meio do DOI: <https://doi.org/10.5753/kdmile.2025.247519> .

6.4 Agradecimentos adicionais

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio institucional concedido ao meu orientador, Prof. Dr. Marcelo Keese Albertini,

por meio da bolsa de Produtividade em Pesquisa (PQ - Pesquisador 2), Processo nº 306795/2022-1, conforme registro oficial do CNPq. Esse apoio tem sido fundamental para o avanço da pesquisa em Ciência da Computação e contribuiu diretamente para a realização deste trabalho.

Registro também meu agradecimento ao Programa de Pós-Graduação em Ciência da Computação (PPGCO) da Universidade Federal de Uberlândia, pelo apoio financeiro concedido para a participação em evento científico e apresentação do artigo decorrente desta pesquisa. Tal apresentação, descrita na Seção 6.3, foi essencial para a disseminação dos resultados, a troca de experiências com a comunidade acadêmica e o fortalecimento da formação científica durante o mestrado.

Referências

ALLAM, A. et al. Analyzing patient trajectories with artificial intelligence. **Journal of Medical Internet Research**, JMIR Publications, v. 23, n. 2, p. e29812, 2021. ISSN 1438-8871. DOI: 10.2196/29812. Disponível em: <https://doi.org/10.2196/29812>.

ARKHANGELSKY, D.; IMBENS, G. Causal models for longitudinal and panel data: a survey. **The Econometrics Journal**, v. 27, n. 3, p. C1–C61, 2024. DOI: 10.1093/ectj/utae014.

BALKUS, S.; HEJAZI, N. Causaltables.jl: Simulating and storing data for statistical causal inference in Julia. **Journal of Open Source Software**, v. 10, n. 106, p. 7580, 2025. DOI: 10.21105/joss.07580. Disponível em: <https://doi.org/10.21105/joss.07580>.

BALTAGI, B. H.; KAO, C.; PENG, B. Testing cross-sectional correlation in large panel data models with serial correlation. **Econometrics**, v. 4, n. 4, p. 44, 2016. DOI: 10.3390/econometrics4040044. Disponível em: <https://www.mdpi.com/2225-1146/4/4/44>.

BICA, I. et al. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In: **8th International Conference on Learning Representations (ICLR 2020)**. Addis Ababa, Ethiopia (virtual conference): OpenReview.net, 2020. DOI: 10.48550/arXiv.2002.04083. Disponível em: <https://doi.org/10.48550/arXiv.2002.04083>.

BOX, G. E. P.; JENKINS, G. M. **Time Series Analysis: Forecasting and Control**. San Francisco: Holden–Day, 1970. ISBN 9780816210947.

BUCHER, A.; DETTE, H. **On the lack of weak continuity of Chatterjee's correlation coefficient**. 2024. DOI: 10.48550/arXiv.2410.11418. Preprint, to appear in *Statistical Science*. Disponível em: <https://arxiv.org/abs/2410.11418>.

BUN, M. et al. Continual release of differentially private synthetic data from longitudinal data collections. **Proceedings of the ACM on Management of Data**, Association for Computing Machinery, v. 2, n. 2, p. 94:1–94:26, 2024. DOI: 10.1145/3651595. Disponível em: <https://doi.org/10.1145/3651595>.

CHATTERJEE, S. A new coefficient of correlation. **Journal of the American Statistical Association**, v. 116, n. 536, p. 2009–2022, 2021. DOI: 10.1080/01621459.2020.1758115.

CHENG, L. et al. Evaluation methods and measures for causal algorithms. **IEEE Transactions on Artificial Intelligence**, v. 3, p. 924–943, 2022. DOI: 10.1109/TAI.2022.3150264.

CHIPMAN, H. A.; GEORGE, E. I.; MCCULLOCH, R. E. Bart: Bayesian additive regression trees. **The Annals of Applied Statistics**, v. 4, n. 1, p. 266–298, 2010. DOI: 10.1214/09-AOAS285.

CHO, K. et al. On the properties of neural machine translation: Encoder–decoder approaches. In: **Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation**. Doha, Qatar: Association for Computational Linguistics, 2014. Disponível em: <https://api.semanticscholar.org/CorpusID:11336213>.

CURTH, A.; SCHAAR, M. van der. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In: **Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)**. Online: PMLR, 2021. Disponível em: <https://proceedings.mlr.press/v130/curth21a/curth21a.pdf>.

DIGGLE, P. J. et al. **Analysis of Longitudinal Data**. 2. ed. Oxford: Oxford University Press, 2002. ISBN 978-0-19-852484-0.

ELMAN, J. L. Finding structure in time. **Cognitive Science**, v. 14, n. 2, p. 179–211, 1990. DOI: 10.1207/s15516709cog1402$_1$.

ENDERS, W. **Applied Econometric Time Series**. 3. ed. Hoboken, New Jersey: John Wiley & Sons, 2010. ISBN 978-0-470-59030-4.

FISHER, R. A. **The Design of Experiments**. Edinburgh: Oliver and Boyd, 1935.

HILL, J. Bayesian nonparametric modeling for causal inference. **Journal of Computational and Graphical Statistics**, v. 20, n. 1, p. 217–240, 2011. DOI: 10.1198/jcgs.2010.08162.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

JOHANSSON, F.; SHALIT, U.; SONTAG, D. Learning representations for counterfactual inference. In: **Proceedings of the International Conference on Machine Learning**. PMLR, 2016. Disponível em: <https://proceedings.mlr.press/v48/johansson16.html>.

KADDOUR, J. et al. Causal machine learning: A survey and open problems. **Foundations and Trends in Machine Learning**, Now Publishers, v. 9, n. 1–2, p. 1–247, 2025. DOI: 10.1561/2400000052. Disponível em: <https://doi.org/10.1561/2400000052>.

KÜHNEL, L. et al. Synthetic data generation for a longitudinal cohort study–evaluation, method extension and reproduction of published data analysis results. **Scientific Reports**, Nature Publishing Group, v. 14, n. 1, p. 14412, 2024. DOI: 10.1038/s41598-024-62102-2. Disponível em: <https://doi.org/10.1038/s41598-024-62102-2>.

KUZMANOVIC, M.; HATT, T.; FEUERRIEGEL, S. Estimating conditional average treatment effects with missing treatment information. In: RUIZ, F.; DY, J.; MEENT, J.-W. van de (Ed.). **Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)**. Valencia, Spain: PMLR, 2023. v. 206, p. 746–766. Disponível em: <https://proceedings.mlr.press/v206/kuzmanovic23a.html>.

LAAN, M. J. van der; RUBIN, D. Targeted maximum likelihood learning. **The International Journal of Biostatistics**, v. 2, n. 1, p. Article 11, 2006. DOI: 10.2202/1557-4679.1043.

LI, R. et al. G-net: A deep learning approach to g-computation for counterfactual outcome prediction under dynamic treatment regimes. In: **Proceedings of Machine Learning for Health (ML4H)**. Virtual conference: PMLR, 2021. v. 158, p. 280–297. Disponível em: <https://proceedings.mlr.press/v158/li21a.html>.

LIM, B.; ALAA, A.; SCHAAR, M. van der. Forecasting treatment responses over time using recurrent marginal structural networks. In: **Advances in Neural Information Processing Systems 31 (NeurIPS 2018)**. Red Hook, NY, USA: Curran Associates, Inc., 2018.

LÜTKEPOHL, H. **New Introduction to Multiple Time Series Analysis**. Berlin: Springer, 2005. ISBN 978-3-540-40172-8.

MELNYCHUK, V.; FRAUEN, D.; FEUERRIEGEL, S. Causal transformer for estimating counterfactual outcomes. In: **Proceedings of the 39th International Conference on Machine Learning (ICML)**. Baltimore, MD, USA: PMLR, 2022. v. 162, p. 23613–23632. Disponível em: <https://proceedings.mlr.press/v162/melnychuk22a.html>.

MENDIS, K.; WICKRAMASINGHE, M.; MARASINGHE, P. Multivariate time series forecasting: A review. In: **Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition**. New York, NY, USA: Association for Computing Machinery, 2024. p. 1–9. DOI: 10.1145/3663976.3664241.

MENG, X.-L.; ROSENTHAL, R.; RUBIN, D. Comparing correlated correlation coefficients. **Psychological Bulletin**, v. 111, n. 1, p. 172–175, 1992. DOI: 10.1037/0033-2909.111.1.172.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 9. ed. Hoboken, NJ: John Wiley & Sons, 2017. Edição 9. ISBN 9781119113478.

NEYMAN, J. On the application of probability theory to agricultural experiments: Essay on principles. **Statistical Science**, v. 5, n. 4, p. 465–480, 1923.

NICHANI, E.; DAMIAN, A.; LEE, J. D. How transformers learn causal structure with gradient descent. In: **Proceedings of the 41st International Conference on Machine Learning**. PMLR, 2024. v. 235, p. 38018–38070. Disponível em: <https://proceedings.mlr.press/v235/nichani24a.html>.

PEARL, J. **Causality: Models, Reasoning, and Inference**. Cambridge: Cambridge University Press, 2000. ISBN 9780521773621.

_____. Causal inference in statistics: An overview. **Statistics Surveys**, v. 3, p. 96–146, 2009. ISSN 1935-7516. DOI: 10.1214/09-SS057.

_____. **The Book of Why: The New Science of Cause and Effect**. New York: Basic Books, 2018.

PETERS, J.; JANZING, D.; SCHÖLKOPF, B. **Elements of Causal Inference: Foundations and Learning Algorithms**. [S.l.]: MIT Press, 2017.

RUBIN, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. **Journal of Educational Psychology**, v. 66, n. 5, p. 688–701, 1974.

_____. Bayesian inference for causal effects: The role of randomization. **The Annals of Statistics**, v. 6, n. 1, p. 34–58, 1978.

SCHULAM, P.; SARIA, S. Reliable decision support using counterfactual models. In: **Advances in Neural Information Processing Systems 30 (NIPS 2017)**. Red Hook, NY, USA: Curran Associates, Inc., 2017. p. 1698–1709. DOI: 10.5555/3294771.3294933. Disponível em: <https://dl.acm.org/doi/10.5555/3294771.3294933>.

SHALIT, U.; JOHANSSON, F. D.; SONTAG, D. Estimating individual treatment effect: Generalization bounds and algorithms. In: **Proceedings of the 34th International Conference on Machine Learning (ICML)**. Sydney, Australia: PMLR, 2017. v. 70, p. 3076–3085. Disponível em: <https://proceedings.mlr.press/v70/shalit17a.html>.

SHEN, D. et al. Same root different leaves: Time series and cross-sectional methods in panel data. **Econometrica**, v. 91, n. 6, p. 2125–2154, 2023. DOI: 10.3982/ECTA21248.

SOLEIMANI, H.; SUBBASWAMY, A.; SARIA, S. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. In: **Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)**. Sydney, Australia: AUAI Press, 2017. Disponível em: <http://auai.org/uai2017/proceedings/papers/266.pdf>.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: **Advances in Neural Information Processing Systems 27**. Cambridge, MA, USA: Curran Associates, Inc., 2014. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2014/file/5a18e133cbf9f257297f410bb7eca942-Paper.pdf>.

WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, v. 20, n. 7, p. 557–585, 1921.

XU, Y.; XU, Y.; SARIA, S. A non-parametric bayesian approach for estimating treatment-response curves from sparse time series. In: DOSHI-VELEZ, F. et al. (Ed.). **Proceedings of the 1st Machine Learning for Healthcare Conference**. Los Angeles, CA, USA: PMLR, 2016. v. 56, p. 282–300. Disponível em: <https://proceedings.mlr.press/v56/Xu16.html>.

YOON, J.; JORDON, J.; SCHAAR, M. van der. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In: **International Conference on Learning Representations (ICLR)**. Vancouver, Canada: OpenReview.net, 2018. DOI: 10.3389/fgene.2020.585804. Disponível em: <https://doi.org/10.3389/fgene.2020.585804>.

ZHANG, X. et al. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In: **Proceedings of the 39th Conference on Neural Information Processing Systems**. San Diego, CA, USA: [s.n.], 2025. Poster presentation. Disponível em: <https://neurips.cc/virtual/2025/loc/san-diego/poster/115625>.