Longitudinal Synthetic Data Generation from Causal Structures

Alessandro S. Angeruzzi, Marcelo K. Albertini

Universidade Federal de Uberlândia, Brazil
alessandro.angeruzzi@ufu.br, albertini@ufu.br

Abstract. Robust assessment of temporal causal-inference models is hampered by the lack of benchmark datasets whose underlying mechanisms are fully known. We introduce the **Causal Synthetic Data Generator (CSDG)**, an open-source tool that creates longitudinal sequences governed by user-defined structural causal graphs with autoregressive dynamics. By allowing fine-grained control over confounding intensity, treatment policies, intervention timing, and noise, CSDG furnishes a flexible, domain-agnostic test-bed for stress-testing causal-learning algorithms. To demonstrate its utility, we generate synthetic cohorts for a one-step-ahead outcome-forecasting task and compare classical linear regression with encoder–decoder recurrent networks (vanilla RNN, LSTM, and GRU). The results reveal how predictive accuracy degrades as causal complexity increases, underscoring the need for models that explicitly exploit causal structure. Beyond forecasting, CSDG naturally extends to counterfactual data generation and bespoke causal graphs, paving the way for comprehensive, reproducible benchmarks across diverse application contexts.

The generator and reproducible experiments are available at `github.com/angeruzzi/causal-synthetic-data-gen`.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

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

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 - isto é, cenários hipotéticos que descrevem o que teria ocorrido sob diferentes condições ou intervenções. Essa capacidade é crucial em áreas como medicina, 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. Nesse contexto, métodos baseados em dados observacionais tornam-se essenciais. Avaliar tais métodos, no entanto, representa um desafio significativo, uma vez que os contrafactuais verdadeiros são, por definição, não observáveis [Rubin 1974].

Além disso, há uma escassez de dados reais e *benchmarks* padronizados que permitam a validação rigorosa desses métodos [Kaddour et al. 2022]. Segundo Cheng et al. [2022], essa lacuna é o principal gargalo da área, especialmente em contextos longitudinais, onde os efeitos se acumulam ou se transformam ao longo do tempo, exigindo dados que capturem essas dinâmicas temporais com precisão.

Nesse cenário, dados sintéticos surgem como uma alternativa promissora, ao possibilitar a geração de cenários com variações controladas de complexidade e estrutura causal - algo frequentemente inviável

com dados reais. No entanto, muitos dos geradores atualmente disponíveis falham em representar cenários realistas, sobretudo em contextos temporais e com variáveis contínuas [Cheng et al. 2022].

Este trabalho parte da premissa de que a ausência de conjuntos de dados longitudinais com estrutura causal conhecida limita a validação empírica de modelos de inferência causal. Propomos, portanto, o gerador CSDG como uma solução para esse problema. A hipótese subjacente é que, ao permitir o controle explícito sobre a estrutura causal, dinâmica temporal, intervenções e contrafactuais, o CSDG torna possível a construção de cenários reprodutíveis para benchmarking de modelos causais em séries temporais.

2. FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os conceitos fundamentais que embasam o desenvolvimento deste trabalho, incluindo as estruturas causais, os tipos de variáveis e junções causais, a aplicação e natureza dos dados longitudinais, bem como o conceito de resultados potenciais e cenários contrafactuais.

2.1 Estruturas Causais

Desde os primeiros estudos de Wright, diagramas de trajetória passaram a ser utilizados para representar de forma gráfica relações causais entre variáveis [Wright 1921]. No entanto, com os avanços propostos por Judea Pearl, por meio dos grafos direcionados acíclicos (DAGs - *Directed Acyclic Graph*s) e dos modelos causais estruturais (SCMs - *Structural Causal Models*), que se consolidou uma estrutura formal e poderosa para a representação e análise de sistemas causais complexos [Pearl 2000].

Os DAGs são grafos orientados e acíclicos cujos nós representam variáveis e as arestas direcionadas indicam relações causais diretas. Diferem dos diagramas de Wright por modelarem explicitamente a a causalidade unidirecional e por não permitirem ciclos, garantindo clareza na direção dos efeitos.

Os SCMs expandem os DAGs ao incorporar equações que descrevem como cada variável é determinada por suas causas diretas (pais no grafo) e por um termo de erro exógeno, geralmente considerado como ruído aleatório. 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.

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 (desfecho), atuando como o mecanismo pelo qual a intervenção exerce sua influência. As **confundidoras** são variáveis que afetam simultaneamente tanto o tratamento quanto o desfecho, 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.

Além disso, é fundamental compreender as estruturas de junção causal [Pearl 2018], que servem como base para a análise de dependências e independências condicionais. 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.

No gerador de dados proposto neste trabalho, tais relações são explicitamente codificadas nas estruturas utilizadas, garantindo um controle preciso sobre os mecanismos causais simulados.

2.2 Dados Longitudinais

Dados longitudinais referem-se a medições repetidas de uma ou mais variáveis ao longo de um domínio ordenado, geralmente o tempo. Essa estrutura permite investigar a evolução temporal de fenômenos, modelar trajetórias individuais e estimar os efeitos de variáveis que variam no tempo, ao mesmo tempo em que controla a heterogeneidade não observada entre unidades de análise. Os dois principais tipos de dados longitudinais são as séries temporais, que representam observações sequenciais de uma única variável ao longo do tempo, e os dados em painel, que englobam observações de múltiplas variáveis ou múltiplas unidades em diferentes momentos [Diggle et al. 2002].

Um modelo estatístico clássico para séries temporais univariadas é o modelo autorregressivo integrado de médias móveis (ARIMA), desenvolvido por Box e Jenkins [1970]. Esse modelo combina três componentes: autorregressão (valores passados da variável), média móvel (erros de previsão passados) e diferenciação (para lidar com a não-estacionaridade dos dados). A natureza autorregressiva das séries temporais é central em diversos domínios aplicados, como macroeconomia - na análise de inflação e PIB - e finanças - na modelagem de retornos e preços de ativos [Enders 2010].

Além das séries univariadas, o campo de séries temporais multivariadas (*Multivariate Time Series* - MTS) tem ganhado destaque. Nesse contexto, analisa-se a dinâmica conjunta de múltiplas variáveis temporais, explorando relações de dependência e causalidade entre elas. Ferramentas estatísticas como VAR (*Vector Autoregressive*) e VARMA (*Vector Autoregressive Moving Average*) são amplamente empregadas em análises desse tipo, especialmente em econometria [Lütkepohl 2005].

Nos últimos anos, observou-se um avanço significativo no uso de modelos de aprendizado profundo (*deep learning*) para previsão com MTSs. Arquiteturas como Redes Neurais Recorrentes (RNNs), *Transformers* e *Graph Neural Networks* (GNNs) têm sido exploradas para capturar padrões complexos e melhorar a acurácia preditiva [Mendis et al. 2024].

No contexto da inferência causal, dados longitudinais oferecem oportunidades e desafios únicos. A dependência temporal e a heterogeneidade não observada entre unidades exigem técnicas especializadas. Métodos tradicionais, como Diferença em Diferenças e Modelos com Variáveis Instrumentais para Dados em Painel, continuam sendo amplamente utilizados. No entanto, novas abordagens vêm sendo desenvolvidas para estimar efeitos causais de intervenções (especialmente binárias) ao longo do tempo em múltiplas unidades observacionais [Arkhangelsky and Imbens 2024].

Paralelamente, o uso de *deep learning* tem sido investigado na modelagem causal longitudinal. Abordagens como as *Statistical Recurrent Units* (SRUs) [Kaddour et al. 2022] propõem formas inovadoras de capturar dependências temporais em contextos causais, ampliando o leque de ferramentas disponíveis para a descoberta de relações de causa e efeito ao longo do tempo.

2.3 Resultados Potenciais e Contrafactuais

A formulação adotada neste trabalho fundamenta-se no *framework* de Resultados Potenciais (*Potential Outcomes Framework*) [Rubin 1974], amplamente utilizado na inferência causal moderna. De acordo com esse modelo, para cada indivíduo (ou instância observacional) i, existem dois desfechos possíveis: $Y_i(1)$, caso o indivíduo receba o tratamento, e $Y_i(0)$, caso não receba. No entanto, apenas um desses desfechos é observado na prática - aquele que corresponde à condição de tratamento efetivamente atribuída - enquanto o outro permanece contrafactual.

Essa impossibilidade de observar simultaneamente ambos os resultados potenciais para o mesmo indivíduo caracteriza o chamado problema fundamental da inferência causal, o qual representa uma limitação central na validação de modelos em cenários com dados observacionais reais.

Nesse cenário, a geração de dados sintéticos com contrafactuais conhecidos oferece uma solução eficaz. Ao tornar 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 (*Individual Treatment Effect* - ITE), médio (*Average Treatment Effect* - ATE) ou condicional (*Conditional Average Treatment Effect* - CATE) [Cheng et al. 2022].

A abordagem proposta neste trabalho, por meio do uso do *Causal Synthetic Data Generator* (CSDG), viabiliza a criação explícita de contrafactuais sintéticos com suporte a tratamentos contínuos, múltiplos períodos e intervenções em diferentes componentes de grafos causais. Essa flexibilidade permite construir cenários de teste abrangentes e realistas para validar o desempenho de modelos causais contemporâneos, como o *Causal Transformer* [Melnychuk et al. 2022], que operam em dados longitudinais complexos e dependem de representações estruturais e temporais ricas.

3. TRABALHOS CORRELATOS

Diversos trabalhos recentes têm explorado a geração de dados sintéticos longitudinais, embora com objetivos e metodologias distintas da proposta apresentada neste estudo.

Em Bun et al. [2024], os autores propõem a geração de dados sintéticos longitudinais a partir de registros médicos reais, com ênfase na preservação da privacidade e na manutenção das propriedades estatísticas dos dados originais. Embora represente uma contribuição relevante para aplicações em saúde, o foco deste trabalho difere substancialmente do nosso. Enquanto os autores buscam reproduzir as características estatísticas observadas nos dados empíricos, a proposta aqui apresentada concentra-se na simulação controlada de cenários causais, com estrutura subjacente conhecida, permitindo a avaliação rigorosa de algoritmos de inferência causal.

De forma semelhante, Kühnel et al. [2024] propõem a geração de dados longitudinais a partir de bases reais, utilizando o método VAMBN (rede bayesiana modular com codificação variacional) combinado com uma camada de redes neurais recorrentes para capturar dependências temporais em estudos nutricionais. Embora compartilhem o foco em dados longitudinais autorregressivos, tanto este trabalho quanto o de Bun et al. priorizam a fidelidade estatística aos dados reais, enquanto nossa abordagem busca a exploração e controle explícito da estrutura causal, fundamental para experimentos controlados e reprodutíveis com algoritmos preditivos e contrafactuais.

CausalTables.jl [Balkus and Hejazi 2025] é uma biblioteca para a linguagem Julia, voltada à simulação de dados causais tabulares com base em modelos causais estruturais, oferecendo suporte à geração de contrafactuais e acesso a estimativas de referência como o efeito médio do tratamento. Apesar de permitir a simulação de cenários com conhecimento causal, sua limitação reside no foco exclusivo em dados estáticos, sem suporte para a evolução temporal ou dependência autorregressiva. A proposta deste trabalho complementa essa abordagem ao gerar dados longitudinais e dinâmicos, que representam a evolução de tratamentos, desfechos e covariáveis ao longo do tempo - aspectos fundamentais para a avaliação de algoritmos causais em contextos temporais realistas.

4. PROPOSTA DE MÉTODO

A proposta deste trabalho consiste em um gerador que simula dados com relações causais definidas por equações estruturais autorregressivas, modelando o comportamento das variáveis ao longo do tempo. As estruturas causais foram baseadas em padrões da literatura e a saída do gerador inclui séries temporais sintéticas de tratamentos e desfechos, com suporte opcional a covariáveis e contrafactuais.

4.1 Notação

Nesta seção, a nomeclatura das variáveis principais são denotadas em relação ao tempo t: T_t representa o tratamento, Y_t o desfecho e X_t a covariável (confundidor ou mediador). Os coeficientes de ajuste são: Φ, que representa a dependência temporal (autoregressiva), e β, o efeito entre variáveis. Outros componentes são: ε, o termo de erro aleatório com distribuição normal, e δ, a intervenção aplicada ao

tratamento para gerar o cenário contrafactual. Por fim, as funções $f(\cdot)$ e $g(\cdot)$ definem a complexidade da relação, podendo ser lineares ou não lineares.

4.2 Composição das Estruturas de Geração

As variáveis T (tratamento) e Y (desfecho) são modeladas como processos autorregressivos de primeira ordem (AR(1)), em que o valor atual depende do valor anterior e de um termo de erro aleatório. Essa modelagem captura a dinâmica temporal típica de séries longitudinais reais.

As funções $f(\cdot)$ e $g(\cdot)$ determinam a natureza da relação entre variáveis. No caso linear, $f(x) = x$. Para equações não lineares, o gerador aplica funções quadráticas, senoides ou logarítmicas, atribuídas aleatoriamente.

O termo de erro ε é modelado como uma variável aleatória com distribuição normal de média zero e variância σ^2, definida por um parâmetro do gerador e sorteada a cada iteração:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

4.3 Estruturas Causais Implementadas

A seguir, são descritas as estruturas causais implementadas, cada uma com suas respectivas equações geradoras. As variáveis de Tratamento (T) e desfecho (Y) possuem uma dinâmica temporal autoregressiva em todas as estruturas, mas são influenciadas de maneiras distintas.

Relação Causal Direta (*Direct*): O tratamento (T) não é influenciado por outras variáveis no modelo, enquanto o desfecho (Y) é influenciado diretamente por T, caracterizando a estrutura causal $T \to Y$. Um exemplo que poderia ser modelado por esta estrutura seria o efeito da dose de um medicamento (T) sobre a pressão arterial (Y), sem considerar outros fatores.

$$T_t = \Phi_T \cdot T_{t-1} + \varepsilon_{T_t} \tag{2}$$

$$Y_t = \Phi_Y \cdot Y_{t-1} + \beta_{TY} \cdot f(T_t) + \varepsilon_{Y_t} \tag{3}$$

Cadeia Causal (*Chain*): Assim como na estrutura anterior, o tratamento (T) não sofre outras influências, mas afeta o desfecho (Y) por meio de uma covariável mediadora (X), formando a estrutura causal $T \to X \to Y$. Por exemplo, o efeito da prática de atividade física (T) sobre o nível de colesterol (Y), mediado pela perda de peso (X).

$$T_t = \Phi_T \cdot T_{t-1} + \varepsilon_{T_t} \tag{4}$$

$$X_t = \beta_{TX} \cdot f(T_t) + \varepsilon_{X_t} \tag{5}$$

$$Y_t = \Phi_Y \cdot Y_{t-1} + \beta_{XY} \cdot f(X_t) + \varepsilon_{Y_t} \tag{6}$$

Confundidor (*Confounder*): 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 desfecho (Y). T também influencia diretamente Y, resultando na estrutura $X \to T \to Y$ e $X \to Y$. 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) em cada dia.

$$X_t \sim U(a, b) \tag{7}$$

$$T_t = \Phi_T \cdot T_{t-1} + \beta_{XT} \cdot f(X_t) + \varepsilon_{T_t} \tag{8}$$

$$Y_t = \Phi_Y \cdot Y_{t-1} + \beta_{XY} \cdot f(X_t) + \beta_{TY} \cdot g(T_t) + \varepsilon_{Y_t} \tag{9}$$

4.4 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 desfecho. Na estrutura causal *Direct* por exemplo, as variáveis contrafactuais são geradas da seguinte forma:

$$T_t^{cf} = \Phi_T \cdot T_{t-1}^{cf} + \varepsilon_{T_t} + \delta_t \tag{10}$$

$$Y_t^{cf} = \begin{cases} Y_t, & \text{se } t < t_{int} \\ \Phi_Y \cdot Y_{t-1}^{cf} + \beta_{TY} \cdot f(T_t^{cf}) + \varepsilon_{Y_t}, & \text{se } t \geq t_{int} \end{cases} \tag{11}$$

O desfecho contrafactual Y_t^{cf} é idêntico ao desfecho 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.

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

Intervenção Pontual. A intervenção é aplicada apenas no instante t_{int}:

$$\delta_t = \begin{cases} \alpha, & \text{se } t = t_{int} \\ 0, & \text{se } t \neq t_{int} \end{cases} \tag{12}$$

Intervenção Contínua. A intervenção é aplicada a partir de t_{int} e persiste nos períodos seguintes:

$$\delta_t = \begin{cases} 0, & \text{se } t < t_{int} \\ \alpha, & \text{se } t \geq t_{int} \end{cases} \tag{13}$$

Intervenção Gradual. A intervenção é aplicada gradualmente a partir de t_{int}:

$$\delta_t = \begin{cases} 0, & \text{se } t < t_{int} \\ \alpha \cdot \frac{t - t_{int}}{k}, & \text{se} t_{int} \leq t \end{cases} \tag{14}$$

4.5 Geração de Dados

O processo de geração recebe como parâmetros: número de instâncias a serem geradas (n), comprimento das séries (t) e o tipo de estrutura causal (*Direct*, *Chain* ou *Confounder*). Os coeficientes Φ (autorregressivos) e β (efeitos causais) podem ser definidos ou aleatoriamente gerados.

Como saída, o gerador produz n conjuntos de séries de tamanho t, contendo o Tratamento (T) e o Desfecho (Y) para todas as estruturas, e a Covariável (X) quando aplicável (*Chain* e *Confounder*).

5. APLICAÇÃO: PROVA DE CONCEITO COM APRENDIZADO TEMPORAL CAUSAL

Este trabalho apresenta um experimento de prova de conceito baseado no aprendizado de estrutura causal a partir de dados observacionais históricos e na geração de resultados potenciais a partir de tratamentos futuros definidos.

Cada instância representa um indivíduo com as sequências temporais das variáveis de tratamento (T), covariável (X) e desfecho (Y) ao longo de $t = 20$ períodos. As séries foram geradas utilizando a estrutura causal *Direta* com relação *Linear*, coeficientes autorregressivos $\phi_T = 0{,}8$ e $\phi_Y = 0{,}7$, efeito causal $\beta = 1{,}5$, e ruídos amostrados de distribuições uniformes $\mathcal{U}(-0{,}1, 0{,}1)$. Uma intervenção

Pontual foi aplicada no tratamento no período $t = 10$, com intensidade $\delta_T = 0{,}5$, de modo que os desfechos contrafactuais são gerados a partir do período 11. Para fins de avaliação, a sequência de cada indivíduo foi dividida em duas partes: os 10 primeiros períodos compõem o histórico factual, enquanto os 10 últimos períodos constituem o horizonte de previsão.

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: RNN [Elman 1990], LSTM [Hochreiter and Schmidhuber 1997] e GRU [Cho et al. 2014].

Todas as três redes neurais foram estruturadas na arquitetura *Encoder-Decoder*, conforme originalmente proposta para redes LSTM [Sutskever et al. 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.

Tanto o encoder quanto o decoder das redes foram implementados com 4 camadas e 32 unidades ocultas em cada. Os treinos foram realizados com o otimizador *Adam* e taxa de aprendizado de 0,01. Os tamanhos de lote foram de 250 para treinamento e 125 para validação e para testes.

A qualidade das previsões nos cenários factuais (sem intervenção no tratamento) foi avaliada utilizando o *Root Mean Squared Error (RMSE)* [Cheng et al. 2022], calculado entre os valores de referência e os preditos de Y ao longo do horizonte de previsão. Os resultados dos testes factuais realizados podem ser vistos na Tabela I.

Table I. RMSE da predição factual após 5 execuções para τ passos (valor médio ± desvio padrão).

Modelo	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
Regressão Linear	0,548	0,683	0,727	0,832	0,915	1,023	1,117	1,199	1,244	1,274
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	0,147 ± 0,006	0,145 ± 0,000	0,139 ± 0,002	0,128 ± 0,003
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**	**0,138 ± 0,001**	**0,144 ± 0,001**	0,137 ± 0,001	**0,125 ± 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	0,139 ± 0,001	0,146 ± 0,001	**0,136 ± 0,001**	0,126 ± 0,001

Para os cenários contrafactuais, utilizamos o *Precision in Estimation of Heterogeneous Effect* (PEHE) [Cheng et al. 2022], que avalia o erro na estimativa do efeito do tratamento em nível individual, comparando a diferença entre os desfechos factual e contrafactual de referência com a diferença estimada pelo modelo ao longo do horizonte de previsão. Formalmente, o PEHE é definido como:

$$\epsilon_{\text{PEHE}} = \frac{1}{N} \sum_{i=1}^{N} \Big((Y_i(1) - Y_i(0)) - (\hat{Y}_i(1) - \hat{Y}_i(0)) \Big)^2 \tag{15}$$

onde $Y_i(1)$ - $Y_i(0)$ são os desfechos de referência do indivíduo i nos cenários com e sem tratamento, respectivamente, enquanto $\hat{Y}_i(1)$ e $\hat{Y}_i(0)$ são as estimativas correspondentes preditas pelo modelo. Os resultados das simulações contrafactuais utilizando essa métrica estão apresentados na Tabela II.

Table II. PEHE das predições factual e contrafactual após 5 execuções para τ passos (média ± desvio padrão).

Modelo	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
Regressão Linear	1,048	1,257	1,339	1,314	1,224	1,102	0,968	0,835	0,710	0,598
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	0,017 ± 0,011	0,011 ± 0,007	0,008 ± 0,005	0,005 ± 0,004
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**	**0,003 ± 0,001**	**0,001 ± 0,001**	**0,001 ± 0,001**	**0,001 ± 0,001**
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	0,007 ± 0,003	0,004 ± 0,003	0,003 ± 0,002	0,003 ± 0,002

6. CONCLUSÃO E TRABALHOS FUTUROS

O método de geração de dados proposto neste trabalho permite a criação de séries temporais sintéticas 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 desfechos potenciais. O uso de covariáveis também é versátil, permitindo simular cenários com variáveis observáveis, ocultas ou com papel causal conhecido.

Por limitação de espaço, a prova de conceito proposta foi apenas para a estrutura causal Direta com relação linear e intervenção pontual. Avaliações com estruturas adicionais (Cadeia e Confundidor) e funções não-lineares podem ser incluídas em extensões deste trabalho.

Outro avanço relevante seria permitir que o usuário informe uma estrutura causal personalizada por meio de uma matriz de adjacência 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.

Na geração de cenários contrafactuais também há oportunidades de melhorias, como a possibilidade de geração de mais de um contrafactual para o mesmo paciente e uma maior flexibilização do tipo de intervenção realizada no tratamento.

A criação de *benchmarks* específicos para diferentes domínios a partir do gerador, como classificação de políticas, seleção de variáveis causais ou simulações com interferência entre unidades, também seria uma ampliação de uso promissora.

REFERENCES

Arkhangelsky, D. and Imbens, G. Causal models for longitudinal and panel data: a survey. *The Econometrics Journal* 27 (3): C1–C61, 06, 2024.

Balkus, S. and Hejazi, N. Causaltables.jl: Simulating and storing data for statistical causal inference in julia. *Journal of Open Source Software* vol. 10, pp. 7580, 02, 2025.

Box, G. E. and Jenkins, G. M. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1970.

Bun, M., Gaboardi, M., Neunhoeffer, M., and Zhang, W. Continual release of differentially private synthetic data from longitudinal data collections. *Proc. ACM Manag. Data* 2 (2), 2024.

Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K. S., and Liu, H. Evaluation methods and measures for causal algorithms. *IEEE Transactions on Artificial Intelligence* vol. 3, pp. 924–943, 2022.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. 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*, 2014.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK, 2002.

Elman, J. L. Finding structure in time. *Cognitive Science* 14 (2): 179–211, 1990.

Enders, W. *Applied Econometric Time Series*. John Wiley & Sons, Hoboken, New Jersey, 2010.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation* 9 (8): 1735–1780, 11, 1997.

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., and Silva, R. Causal machine learning: A survey and open problems, 2022.

Kühnel, L., Schneider, J., Perrar, I., Adams, T., Moazemi, S., Prasser, F., Nöthlings, U., Fröhlich, H., and Fluck, J. Synthetic data generation for a longitudinal cohort study–evaluation, method extension and reproduction of published data analysis results. *Scientific Reports* 14 (1): 14412, 2024.

Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2005.

Melnychuk, V., Frauen, D., and Feuerriegel, S. Causal transformer for estimating counterfactual outcomes. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Mendis, K., Wickramasinghe, M., and Marasinghe, P. Multivariate time series forecasting: A review. In *Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition*. pp. 1–9, 2024.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

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

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5): 688–701, 1974.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. MIT Press, Cambridge, MA, USA, pp. 3104–3112, 2014.

Wright, S. Correlation and causation. *Journal of Agricultural Research* 20 (7): 557–585, 1921.