Global-Deep Synthetic Control: A Heterogeneous Treatment Effect Estimator for Shared Market

Priscila Grecov, Klaus Ackermann, Christoph Bergmeir [‡]

January 31, 2025

Abstract

Estimating heterogeneous treatment effects in time series data is crucial across numerous real-world scenarios. Nonetheless, ensuring the no-interference assumption for control units can often be challenging when gauging treatment effects via a counterfactual and synthetic control approach in a shared market environment. To address spillover effects that can bias the results, we introduce "Global-Deep Synthetic Control" (Global-DeepSC), which uses globally trained deep learning forecasting models to estimate the treatment effects on multiple treated instances. Global-DeepSC addresses the violation of the non-interference assumption by forecasting counterfactual outcomes using only pre-intervention data for learning. Such an approach curtails the reliance on post-intervention disturbances, thereby mitigating the interference of spillovers and enhancing the accuracy of estimations. Moreover, our framework incorporates data-driven clustering and partitioning stages to purge the modeling of control units affected by spillovers and pinpoint variables responsible for heterogeneous impacts. Experiments on synthetic and real-world datasets show that our method outperforms well-known structural causal models regarding treatment effect estimation and statistical significance tests. An experimental study on the effect of sales promotions on supermarket sales found that younger, lower-income individuals should be the primary target audience for the marketing strategy, as estimated by conditional treatment effect analysis.

Causal inference, heterogeneous treatment effect, time series forecasting, global forecasting method, spillover effects

1 Introduction

Researchers across various fields, including the social sciences, policy development, and the technology sector, often aim to evaluate the effects of alterations to policies or business strategies. These evaluations commonly use time series data and the potential outcome framework, where some units are exposed to a particular intervention while others are not. An example could be a car ride business offering incentives to some drivers (treated group), which motivates them to pick up more passengers, indirectly affecting other drivers who did not receive incentives (control group) by reducing their potential passenger pool. Simply taking the difference between these groups, even under random assignment, will provide misleading results. Given the non-random nature of these interventions across units, numerous statistical models have been developed to predict unobserved counterfactuals for affected units and estimate policy or strategy effects based on observed data—known as the Rubin Causal Model (RCM) [1]. Cornerstone approaches for these comparative studies include difference-in-differences (DiD) methods [2, 3, 4] and the Synthetic Control (SC) approach [5], both of which are rooted in the RCM and

[*]Department of Data Science and Artificial Intelligence, Monash University, Clayton, Australia.
[†]Department of Econometrics and Business Statistics, Monash University, Clayton, Australia.
[‡]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain.

utilise observational time series data. In recent years, several significant variants of DiD and SC methods have emerged, including penalised regression models [4, 6, 7], Bayesian structural time series models [8], factor and interactive fixed effects (FE) models [9], matrix completion [10], and a combination of DiD and SC [11]. These methods collectively represent counterfactual and synthetic control (CSC) methods for causal effect estimation.

In shared markets, where spillover effects on control units are likely, maintaining an utterly unaffected group can be difficult. However, CSC methods, such as in Brodersen et al. [8], have to rely on a group of control units or external predictors that remain not impacted by the policy in the pre- and post-treatment periods as covariates. Some studies, such as Menchetti et al. [12], Cerqua et al. [13], propose counterfactual forecasting methods that use only pre-intervention data of external predictors as covariates without relying on control series. Although these methods address the unavailability of control units, they may lose valuable information when control series (with or without spillovers) are available. Conversely, Grecov et al. [14, 15] trained its models by relying only on prior intervention data while still considering both treated and control instances and external predictors into the model, avoiding the loss of cross-relation information. This research is the first to introduce global forecasting methods (GFMs) for the average treatment effect (ATE); however, without exploring the heterogeneous treatment effect (HTE) analysis or discussing the framework setting for circumstances when the standard stable unit treatment value assumption (SUTVA) cannot be ensured.

To address these shortcomings and provide a reliable HTE estimation, we propose a new method called **Global-DeepSC**. This method performs CSC-based causal effect analysis in intricate scenarios involving spillovers or situations where certain instances are affected more strongly than others. Global-DeepSC executes multivariate counterfactual modelling, leveraging the field of GFMs for time series forecasting in conjunction with autoregressive deep neural networks (DNNs) in their recurrent form. Given the autoregressive and *global* forecasting aspects of Global-DeepSC, the model primarily considers the pre-intervention time series behaviour of the response variable (i.e., the treated units), external predictors, and control units as inputs. Unlike SCMs [8], Global-DeepSC does not rely on post-treatment values to modify the counterfactual prediction in the post-intervention period, thereby circumventing biased causal effect estimation that may occur when spillovers are inevitable. As such, we do not demand that all controls or external predictors remain entirely unaffected by the treatment. Instead, we require a subset of control units to be free of potential spillovers. This condition serves two primary purposes: firstly, to ensure proper forecasting evaluation by the model following the intervention, and secondly, to confirm the absence of other structural changes beyond the intervention of interest. Our proposed methodology is particularly applicable in shared marketplaces, where the SUTVA is often violated due to the market's inherent interconnectivity [16, 17]. Therefore, the theoretical perspective and benefit of our method's performance are based on its applicability in scenarios where current methods fail due to SUTVA violation. We need a weaker version of the SUTVA assumption, where a subset of control units is free from treatment interference. We employ a time series clustering step on the controls to construct this subset of control units with negligible effects. This step, carried out after training for counterfactual estimation, aims to isolate those controls that display distinct behaviour in the post-intervention period.

Our framework may be considered unique in the literature on causal effect analysis within shared markets, where the stringent non-interference assumption is impractical. Even the most recent research on counterfactual and individual treatment effect estimation (ITE), considering dynamic historical data outcomes, assumes that spillover effects are either non-existent, negligible, or controllable. Building on these premises, complex methodologies have been developed, such as modifications to Transformers [18] and RNNs [19], both of which must adhere to *Sequential Ignorability* or *No Unobserved Confounding* assumptions for causal inference—either in their pure form or with certain adjustments. Our work, however, is pioneering in relaxing the SUTVA assumption and addressing post-intervention confounders without overlooking them.

Our innovation lies in a distinctive perspective on the issue, demonstrating a less restrictive and more practical way to tackle it. We advocate the modification of established methodologies (similar to the approach adopted in Yao et al. [20], albeit with the retention of assumptions such as SUTVA and *Ignorability*) to address the unresolved complexities in this domain.

HTEs occur when different sub-populations respond uniquely to an intervention. Our model can estimate the treatment effect for each time series, thereby offering flexibility to aggregate the effects for any desired sub-population. This capability facilitates the identification of the optimal dimension of impact heterogeneity. HTEs are manifested through conditional average treatment effects (CATEs), representing treatment effects conditional on individual characteristics. After estimating the per-series effects, we determine CATE by aggregating the individual time series effects across various sub-populations. We employ causal forest models [21, 22] to segregate our data into subgroups characterised by distinct treatment effect magnitudes. Causal forests find partitions in the data based on population features that maximise the difference in mean effects, thereby identifying the heterogeneity in these effects. To the best of our knowledge, our GFM-DNN-CSC-based model for estimating HTE is the first to allow for CATE calculations based on aggregating ITEs that can learn from pre-treatment data only, relax the SUTVA requirements and identify not only unbiased treatment effects but also spillover effects and their impact heterogeneity. The main contributions of this study are as follows:

- We propose a novel Global-DeepSC framework capable of jointly estimating ITE, ATE or HTE under the circumstances of SUTVA violation by introducing a weaker version of SUTVA, which becomes feasible once we employ the GFM.
- We incorporate a data-driven clustering step to identify the control units without spillover effect from those influenced by the treatment.
- We adapt and extend the application of causal forest models for HTE when considering time series data.
- In addition to estimating the effect on treated units, our framework can also estimate the spillover effects on controls.

The structure of the remaining paper is organised as follows. A review of previous work is provided in Section 2. The subsequent Section 3 presents the problem context and notation and introduces the assumptions required for treatment effect identification. The methodology and implementation of our Global-DeepSC framework are discussed in Section 4. The experimental results and analyses for the simulated and real-world scenarios are reported in Sections 5 and 6, respectively. Finally, Section 7 concludes this paper.

2 Related Work

Treatment effects for observational data in static settings

In the realm of observational data analysis, established methodologies for assessing treatment effects have predominantly centred on calculating the ATE through diverse propensity score methodology, including matching, weighting, stratification, and doubly robust techniques [23, 24, 25, 26]. While matching techniques depend on an appropriate proximity measure to identify a group of neighbours, weighting approaches are typically crafted for calculating average effects. They are not readily applicable to HTE estimation. Recently, supervised models based on trees and forests [21, 27] have been considered adaptive proximity measures and have found application in non-parametric causal inference for HTE. Illustrations include BART [28], which utilises Bayesian nonparametric additive regression trees for deducing potential outcomes to estimate ITEs using observational data, and Causal Forest [21], which harnesses random forest techniques to evaluate HTEs within subgroups. Additionally, significant attention has been

directed towards determining counterfactual outcomes in static environments [29, 30, 31, 32]. In static scenarios, input data are cross-sectional; therefore, time-dependent variables, interventions, and results are lacking. Nevertheless, in addition to neglecting temporal data, these methods are predicated on the strong ignorability assumption, which essentially overlooks the impact of hidden confounding variables. Moreover, a substantial body of research has employed deep learning techniques, alternatively referred to as causal inference grounded in representation learning, to address this objective [29, 30, 33]. However, these approaches again are predicated on rigid ignorability assumptions. Various studies have sought to mitigate these assumptions by accounting for hidden confounders, including the application of domain adversarial training [34, 35]. Nevertheless, these methodologies fail to exploit the dynamic evolution of the observed variables and interpersonal relationships inherent in time-dynamic contexts.

Treatment effects for time series data

Initially developed in epidemiology, techniques for assessing time-varying outcomes predominantly employ basic linear models. These methods, such as G-computation [36], structural nested models [37], and marginal structural models (MSMs) [38], aim to calculate the ATE of treatments that change over time rather than HTE. Recognising the constraints of linear models, scholars proposed several Bayesian non-parametric approaches [39, 40, 41]. These methods, however, are predicated on robust assumptions regarding data generation and prove inadequate for multi-dimensional outcomes or static covariates. Alternative methodologies grounded in recurrent neural networks (RNNs) [42, 43] necessitate stringent identifiability assumptions unsuitable for contexts where intervention interference happens. Contemporary state-of-the-art techniques, which integrate the potential outcomes framework with neural network models, encompass G-Net [44], counterfactual recurrent network (CRN) [45], and recurrent marginal structural networks (RMSNs) [46]. Nonetheless, these methods continue to rely on ignorability assumptions (Sequential Ignorability or No Unobserved Confounding assumption), albeit in a more relaxed form, and do not tackle the spillover phenomenon. To manage more complex data, Melnychuk et al. [18] employ Transformers models whilst still adhering to certain ignorability assumptions, whereas Fujii et al. [47] fully relax these assumptions when estimating ITE in complex multiagent scenarios; however, both studies do not address SUTVA violation.F

The DiD design is another widely employed technique for determining causal effects from observational time series data [48, 49]. Traditionally applied in scenarios involving two groups and two-time periods, these designs depend on assumptions of parallel trends, which postulate that the counterfactual outcome free of intervention follow an identical temporal trajectory for both groups. An alternative method for causal inference in this scenario is provided by synthetic controls [5]. This technique, originally developed for cases with a single treated unit, estimates the counterfactual outcome under control by utilising a weighted average of control units. The weights are determined using pre-treatment information. Researchers have extended this approach to handle multiple treated units with staggered adoption by independently applying the synthetic control method to each unit receiving treatment [50, 51]. However, this extension has been shown to be suboptimal for ATE. Athey et al. [10] placed synthetic controls and other panel data estimators into a more comprehensive framework of matrix completion techniques. They approached the matrix of potential outcomes over time as partially observed, using matrix techniques for inferring the values of missing potential outcomes. Furthermore, researchers have integrated synthetic controls with DiD estimators to produce doubly robust estimates. This approach requires only one of the two components—either the synthetic control weights or the fixed effects regression model—to be accurately specified for obtaining consistent treatment effect estimates [11]. Recent advancements in causal effect estimation have employed time series methodologies to forecast outcomes in the absence of intervention [6, 8, 52]. The seminal approach introduced by Brodersen et al. [8] was conceived for instances where a singular time series is observed across multiple time points preceding and following the treatment. A

Bayesian state-space model is constructed using data gathered before treatment begins. This model is subsequently employed to forecast the counterfactual trajectory under control conditions in the post intervention phase. Antonelli and Beck [53] advanced using Bayesian time series models for HTE, considering multiple time series with staggered adoption and ensuring robustness to unmeasured confounding bias. However, these methods have neither explored the performance of their frameworks under SUTVA violations or the complete absence of any variance in ignorability assumptions nor investigated the benefits of using global forecasting methods.

Treatment effects considering spillovers

The utilisation of network information to enhance causal effect estimation in shared markets and networked observational data has emerged as a critical area of research in causal inference. Although Rakesh et al. [54] employs a variational auto-encoder to estimate ITE under networked observational data, this approach struggles to differentiate varying degrees of spillover effects among neighbouring instances and may not be effective for networks of considerable size. Fatemi and Zheleva [55, 56] aims to minimise interference and infer unbiased causal effects in network experiments but primarily focuses on experimental design analysis. Qu et al. [57] proposes creating distinct clusters for units based on observables and suggests an estimator for assessing treatment and spillover effects in heterogeneous groups, assuming conditional exchangeability. Awan et al. [58] propose a matching-based causal estimator for ATE calculation which achieves precise unit matching using subgraph counts of each unit's neighbourhood graph to account for network interference. However, both methods neglect the varying levels of interference between pairs of nodes and the presence of confounding bias. The framework proposed by Forastiere et al. [59] includes a mechanism that combines the treatment allocations of adjacent units to create a collective intervention alongside each individual's treatment. Nonetheless, Huang et al. [60] highlight the ineffectiveness of Forastiere et al. [59] approach when dealing with high-dimensional covariates because it relied on a one-dimensional scalar variable for interference determination and failure to address hidden confounder bias. In response, Huang et al. [60] introduces the SPNet model, which assumes first-order full interference and measures the interference intensity between pairs of nodes. This model employs a graph neural network, utilising node covariates and supplementary graph structure data to identify hidden confounding factors and generate unbiased estimates of causal effects. Nevertheless, their framework is unsuitable for dynamic settings involving time series data.

Our proposed framework distinguishes itself from existing works by addressing the gap and scarcity of previous research concerning HTE for time series data within scenarios that simultaneously violate both the SUTVA and Ignorability/Unconfoundness assumptions. Applying a GFM-RNN structure enables this approach (once it facilitates modelling counterfactual predictions using only pre-intervention data). Furthermore, the advancements implemented through novel data-driven components added to previous works in Grecov et al. [14, 15], Prasanna et al. [61], such as the clustering of control time series and causal forest partitioning, render it feasible to explicit the spillover effects on control units and identify the variables responsible for heterogeneous impacts, respectively.

3 Problem Setting, Notations and Assumptions

This section introduces the problem structure, notations, definitions of the causal effect estimates, and assumptions for identifying causal effects.

3.1 Problem Setting and Notations

We frame the counterfactual forecasting problem as follows. Consider $Y_{j,t}$ as a particular outcome of an instance $j \in \{1, \ldots, J\}$ at time $t \in \{1, \ldots, T\}$. This collection of J instances, that is, J time series, might represent administrative units, such as countries and districts, or different stores, firms, and products, provided they originate from similar data generation processes (DGPs). We set an intervention at time T_0, where $T_0 < T$, affecting only a subset I from all J instances, such that $1 < I < J$. We refer to these I instances, indexed by $i \in \{1, \ldots, I\}$, as treated units, $Y_{i,t}$, while the control group are the residual $C = J - I$ instances, indexed by $c \in \{1, \ldots, C\}$, not exposed to the intervention for all t, the $Y_{c,t}$. Consider $D_t \in \{0, 1\}$ as a binary indicator representing the intervention status of a given unit i at a specific time point $t \in \{T_0+1, \ldots, T\}$, where 1 signifies the occurrence of the treatment and 0 indicates the absence of this treatment. Following the potential outcome notation, for $t > T_0$, we represent $Y_{i,t}$ such that $Y_{i,t} = D_t Y_{i,t}^{(1)} + (1 - D_t) Y_{i,t}^{(0)}$, where $Y_{i,t}^{(1)}$ is the observed potential outcome series for some treated unit i, while $Y_{i,t}^{(0)}$ is the unobserved or counterfactual potential outcome time series that needs to be estimated. Hence, the counterfactual forecasting $Y_{i,t}^{(0)}$ can be expressed as follows, for $t = 1, ..., T_0, ..., T$:

$$\begin{cases} Y_{i,t>T_0}^{(0)} = P_{i,t>T_0}^{(0)} + u_t, \\ Y_{i,t>T_0}^{(1)} = P_{i,t>T_0}^{(0)} + \delta_{i,t} + u_t, \end{cases} \tag{1}$$

with $E(u_t) = 0$ and where $P_{i,t}^{(0)}$ serves as a mean-unbiased proxy for the counterfactual of treated units $Y_{i,t}^{(0)}$ in the absence of the treatment, generated by a model of choice. The term u_t denotes a stationary stochastic process, while the relationship between observed and potential outcomes is expressed by $Y_{i,t} = Y_{i,t}^{(0)} + D_t(Y_{i,t}^{(1)} - Y_{i,t}^{(0)})$. This effect estimation postulates the presence of unbiased surrogates for the counterfactual prediction $P_{i,t}^{(0)}$. Although it does not restrict their association with u_t, it necessitates that u_t be independently, identically distributed (iid) and unaffected by the intervention. Consequently, the intervention's timing should not be influenced by factors that alter the distribution of u_t. For each treated $Y_{i,t}$ and control $Y_{c,t}$ outcomes, we may observe a set of external predictors $X_{j,t} = (x_{j,t}^1, ..., x_{j,t}^k) \in \mathbb{R}^k,\ k > 0$. Our modelling strategy for estimating the counterfactual proxies $P_{i,t}^{(0)}$ considers only the pre-treatment data $Y_{i,t \leq T_0}$, $Y_{c,t \leq T_0}$, and $X_{j,t \leq T_0}$ as input variables for training the model. Hence, we can assume, without loss of generality, that $Y_{j,t}, X_{j,t} \in \mathcal{Z} \subseteq \mathbb{R}^d,\ d > 0$, and $Y_{i,t}^{(0)} \in \mathcal{Y} \subseteq \mathbb{R}^q,\ q > 0$

$$Y_{i,t>T_0}^{(0)} = \mathcal{M}(Y_{:j,t \leq T_0}, X_{:j,t \leq T_0}, \theta) + u_t \tag{2}$$

where $\mathcal{M} : \mathcal{Z} \times \Theta \to \mathcal{Y}$ represents a measurable function for every $\theta \in \Theta$, which is a parametric space of finite dimensions.

3.2 Causal Estimands

The counterfactual and synthetic control (CSC) methodology adopts the potential outcome framework described by Rubin [62], Neyman [63]. Within this framework, the effect of a policy intervention on each treated unit i at time $t > T_0$ is characterised as the discrepancy between the potential outcomes influenced by the policy $Y_{i,t}^{(1)}$ and those estimated counterfactual $Y_{i,t}^{(0)}$, expressed as $\delta_{i,t} = Y_{i,t}^{(1)} - Y_{i,t}^{(0)}$. Conversely, for control units, the observed outcomes $Y_{c,t}$ coincide with their outcomes in scenarios with no intervention, denoted by $Y_{c,t} = Y_{c,t}^{(0)}$ over the interval $1 \leq t \leq T$.

Given the counterfactual forecasting model trained under the setting mentioned above, our goal is to estimate the ITE, ATE, and CATE for both treated (ITT, ATT, and CATT)[1] and

[1]ITT, ATT and CATT stand for Individual Treatment Effect on Treated, Average Treatment Effect on Treated, and Conditional Average Treatment Effect on Treated, respectively.

non-treated (ITNT, ATNT, and CATNT)[2] instances, that is, for all J units. Assume $S_{j,t} \subset X_{j,t}$ is a subset of personal characteristics and N_s is the quantity of instances in the entire population sharing $S_{j,t} = s$, such that $\hat{\delta}_{j,t>T_0}(s) = E[Y^{(1)}_{j,t>T_0} - Y^{(0)}_{j,t>T_0} | X_j = s]$. Hence, the ITE, ATE and CATE are defined as

$$\begin{array}{lll} ITE_j & := \hat{\delta}_{j,t>T_0} = Y^{(1)}_{j,t>T_0} - Y^{(0)}_{j,t>T_0} & \\ ATE & := \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \frac{1}{J} \sum_{j=1}^{J} \hat{\delta}_{j,t}, & \\ CATE_s & := \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \frac{1}{J_s} \sum_{j:,S_{j,t}=s} \hat{\delta}_{j,t}(s), & \end{array} \tag{3}$$

where when $J = I$ and $j = i$ means we calculate the ITT, ATT and CATT, but if $J = C$ and $j = c$, the estimates are the ITNT, ATNT and CATNT. The central challenge in calculating these estimates lies in forecasting $Y^{(0)}_{j,t>T_0}$, which essentially entails predicting $P^{(0)}_{j,t>T_0}$ given that $Y^{(1)}_{j,t>T_0}$ is known.

3.3 Assumptions

Given our application of the CSC approach, we should ensure the following five common underlying assumptions for the identifiability of the treatment effect estimation when considering longitudinal observational data:

Assumption 1 (Null effect on controls)**.** *$Y_{c,t}$ is independent of D_t for all t; that is, $Y^{(1)}_{c,t} = Y^{(0)}_{c,t}$ for $1 \leq c \leq C$.*

Assumption 2 (Single Persistent Intervention)**.** *Unit i receives a single persistent intervention when $\exists\ T_0$ such that $D_t = 0\ \forall\ t \in \{1, \ldots, T_0\}$, but $\forall\ t \in \{T_0 + 1, \ldots, T\}$ the unit is consistently either subjected to treatment, $D_t = 1$, or assigned to the control group, $D_t = 0$.*

Assumption 3 (Distribution-Consistency)**.** *If $D_t = 1$ for a given unit i, then its potential outcome is equivalent in distribution to its observed (factual) outcome, i.e., $Y^{(1)}_{i,T_0+n} \overset{d}{=} Y^{obs}_{i,T_0+n}$, for $n \geq 1$.*

Assumption 4 (Positivity)**.** *Throughout all historical time periods, there exists a non-zero likelihood of either obtaining or not obtaining treatment, which means that $0 < \mathbb{P}(D_t = d_t | H_t = h_t) < 1$, if $\mathbb{P}(H_t = h_t) > 0$, where $d_t \in \{0, 1\}$ and h_t denotes some realisation of a unit's history.*

Assumption 5 (Conditional Stationarity for Identification)**.** *Define $\mathcal{I}_t = (\mathcal{I}_0, Y_{j,1}, ..., Y_{j,t}, X_{j,1}, ..., X_{j,t})$, where $\mathcal{I}_0$ is the information setup to time $t = 0$. Let $f_{Y^{(0)}_{j,t+1}|\mathcal{I}_t}$ indicate the conditional density of the outcome from unit j at time $t+1$ given the past information such that $f_{Y^{(0)}_{j,t+1}|\mathcal{I}_t}(y_{j,t+1}|\mathcal{I}_t)$ is the conditional density in the absence of treatment. Then, for any integer $t \geq 1$, $f_{Y^{(0)}_{j,t+1}|\mathcal{I}_t}\big(y_{j,t+1}\big|\mathcal{I}_t\big) = f_{Y^{(0)}_{j,1}|\mathcal{I}_0}\big(y_{j,t+1}\big|\mathcal{I}_t\big)$.*

Assumption 1 ensures that the control units remain unaffected by intervention. This assumption is fundamental to CSC methods and is essential for discerning the effects of the intervention. If $Y_{c,t}$ is independent of D_t, the treatment effect on treated units becomes identifiable given the deterministic and known timing of the intervention. Absent this independence, CSC-based counterfactual estimations would be compromised. We will see in Section 4 that for our proposed framework, this assumption's role diverges slightly from its traditional application in standard CSC methods. Given the independence between $Y_{c,t}$ and D_t and the employment of the GFM (which shares learned parameters over all series j for prediction), the $Y_{c,t}$ forecasting acts as a validation metric, confirming the model's forecasting accuracy for times $t > T_0$. In the

[2]ITNT, ATNT and CATNT stand for Individual Treatment Effect on Non-Treated, Average Treatment Effect on Non-Treated, and Conditional Average Treatment Effect on Non-Treated, respectively.

GFM framework, global parameters estimated during the pre-treatment phase are consistently utilised for post-intervention predictions for both control and treated time series. Consequently, we evaluate the model parameters' efficacy post-treatment by testing for a null effect on controls. Therefore, in an extreme scenario where all units would be treated, we could still estimate individual treatment effects, albeit with the caveat that we cannot use statistical testing to verify the estimation accuracy. Then, in theory, the framework we propose could also be used in the no-control-unit case.

Assumption 2 stipulates that the intervention is universally recognised, ensuring that all treated units are uniformly and continuously exposed to it for all $t > T_0$. This assumption also precludes any anticipatory effects related to the intervention. The consistency assumption (Assumption 3) stipulates that an individual's actual outcome matches the outcome that would be observed for them, considering their observed exposure history. Meanwhile, the positivity assumption (Assumption 4) establishes that there is a non-null probability of the treatment being assigned for each time step.

Finally, the causal estimates from Eq. (3) become identifiable if we can accurately forecast the counterfactuals for all treated and control units. Therefore, the identification of the causal effect for time series setups necessitates the conditional stationarity assumption (Assumption 5), consistent with findings from existing literature [12, 13, 52, 53]. Given pre-intervention data, Assumption 5 states that the outcome's conditional distribution without treatment effect is time-translation invariant, that is, understanding the conditional distribution of $Y_{j,t}^{(0)}$ before intervention equips us with the knowledge of its conditional distribution afterwards. Testing conditional stationarity involves checking the post-treatment forecast accuracy for unaffected control units, a process commonly referred to as a placebo test, which will be detailed in Section 4. Should the performance of this test prove satisfactory, we can conclude that the model has effectively learned the pre-treatment distributions. As a result, this assumption allows us to use the post-treatment forecast as an estimation of the counterfactual outcome $Y_{i,t}^{(0)}$ for treated units. This supports treatment effect identification from observational data, as defined in Eq. (2).

Furthermore, within the field of treatment effect estimation, it is common practice to incorporate two supplementary assumptions: the *Sequential Strong Ignorability* and the *Stable Unit Treatment Value Assumption – SUTVA* [23, 64]; however, they are not entirely assumed by our proposed framework. The Ignorability assumption states that $Y_{t+1}^{D_t=d_t} \perp\!\!\!\perp D_t \mid X_t$, with $Y^{D_t=d_t}$ denoting the potential outcome of Y given treatment status $D_t = d_t \in \{0, 1\}$. This is a primary identification assumption within the Double Machine Learning (DML) framework [65], with varying degrees of strictness observed among studies in this regard [18, 19, 45]. It assumes that the treatment status is conditionally unrelated to potential outcomes in a multidimensional vector of control variables. This suggests the presumption that no hidden confounding variables are present, which means that all covariates that have an impact on both the treatment allocation and the outcomes are included within the observational dataset. This assumption may be rapidly invalidated if X_t comprises variables that are not entirely exogenous. In data-rich environments, such as Facebook or Amazon, it is assumed that the data has sufficient information to account for any potential confounders. This assumption **cannot be tested** in the real world [38, 66]. Recently, Hünermund et al. [65] pointed out that including a single "bad control", even in a large covariate space, can substantially alter the conclusion about the treatment effect. Our framework **does not rely** on the untestability of this assumption, and offers alternative conditions that can be tested.

Conversely, SUTVA establishes that potential outcomes for any unit are not influenced by the treatments assigned to other units. Each unit has only one form or version of each treatment level, which leads to a specific set of outcomes [67, 23]. There are two distinct parts of SUTVA. First, it states the **no interference between units**. This part assumes the absence of spillover effects or interactions between units. Second, it assumes **no hidden**

variations in treatments, meaning that a unique version of each treatment level is consistently applied across all units. This implies that the treatment is well-defined and does not vary in ways not captured by the study design. Although SUTVA is a core assumption in many causal inference frameworks, it is often criticised for being unrealistic, where interactions between units can lead to spillover effects, thereby violating this assumption and risking inaccurate causal conclusions. Nevertheless, we do not need to comply strictly with SUTVA's *no interference* condition. If the treatment influences the control units or external covariates, then models that use post-treatment data (such as SCMs) will bias the causal estimates. By leveraging the GFM mechanism, we can circumvent these pitfalls by confining our analysis to the pre-treatment data of any unit and external predictors (Eq. (2)). Based on this, we propose a weaker version of SUTVA that permits causal effect estimation under potential spillover effects as follows:

Assumption 6 (Weaker SUTVA – interference assumption)**.** *Let $Y_{1:C,1:T}$ and $X_{1:J,1:T}$ be the entire panel of outcomes for the control units and the external covariates of all units, respectively. Let $Y_{1:C^*,t>T_0} \subset Y_{1:C,1:T}$ and $X_{1:J^*,t>T_0} \subset X_{1:J,1:T}$ be subsets of control units and external covariates. In panel settings where a single intervention happens at time T_0 and there are no treatment hidden forms, we assume the potential outcomes of the c^*-th control unit $Y_{c^*,t>T_0}$ as the external covariate vector $X_{j^*,t>T_0}$ (if post-treatment data is used to predict the outcome of a unit) such subsets of external covariates and control units are independent of treatment interference, i.e., $X^{(1)}_{j^*,t} = X^{(0)}_{j^*,t}$ and $Y^{(1)}_{c^*,t} = Y^{(0)}_{c^*,t}$ for all $t \in \{T_0 + 1, \ldots, T\}$.*

Assumption 6 constitutes a significant contribution of the proposed method as it adapts SUTVA to accommodate interference among units for shared market analysis. By modifying the no-interference rule within a global forecasting framework, we require only a subset of control units unaffected by the treatment for accurate point forecasts. Other controls may have been influenced; however, their pre-treatment information remains integral to our analysis. The data-driven clustering step incorporated as a component of our framework (details in Section 4.1.3) functions to identify this subset of control units without spillover.

4 Global-DeepSC

Global-DeepSC is a framework for estimating ITE, ATE, or CATE for time series data under the potential outcome (PO) and synthetic control (SC) approaches by globally forecasting counterfactuals for treated instances, relying solely on pre-intervention data. It comprises of three stages, as shown in Fig. 1. The first stage is the counterfactual forecasting process. The second is estimating ITE and aggregations (causal estimates). Finally, the third stage encompasses the inference and placebo testing process. This section commences with an introduction to the first stage, which includes the *global* time series forecasting method and, subsequently, the DNN architecture of our estimator, that is, the entire framework's engine consisting of data engineering steps, counterfactual forecasting, and a data-driven clustering process of the control unit time series. It then presents the data-driven partitioning process to identify the subpopulations for HTE and the estimation of causal estimates (the second stage). Lastly, the inference process and tests of the third stage are explained.

4.1 Stage 1: Counterfactual forecasting process

4.1.1 The autoregressive global forecasting method (GFM)

Let $Y = \{y_1, y_2, ..., y_t, ..., y_T\}$ be a time series, and $\{y_{T+1}, ..., y_{T+T^*}\} = F(y_1, y_2, ..., y_T) + \epsilon$ the autoregressive time series forecasting problem, where F is the function approximated by the model to predict the future series values, T^* the forecasting horizon, and ϵ the error associated with the approximation function F. Let $\mathbb{S}$ be a set of J univariate time series $\mathbb{S} = \{Z_j \in \mathbb{R}^{T_0},\ j = 1, ..., J\}$ where Z_j represents the observed time series of length T_0. The forecasting functions are

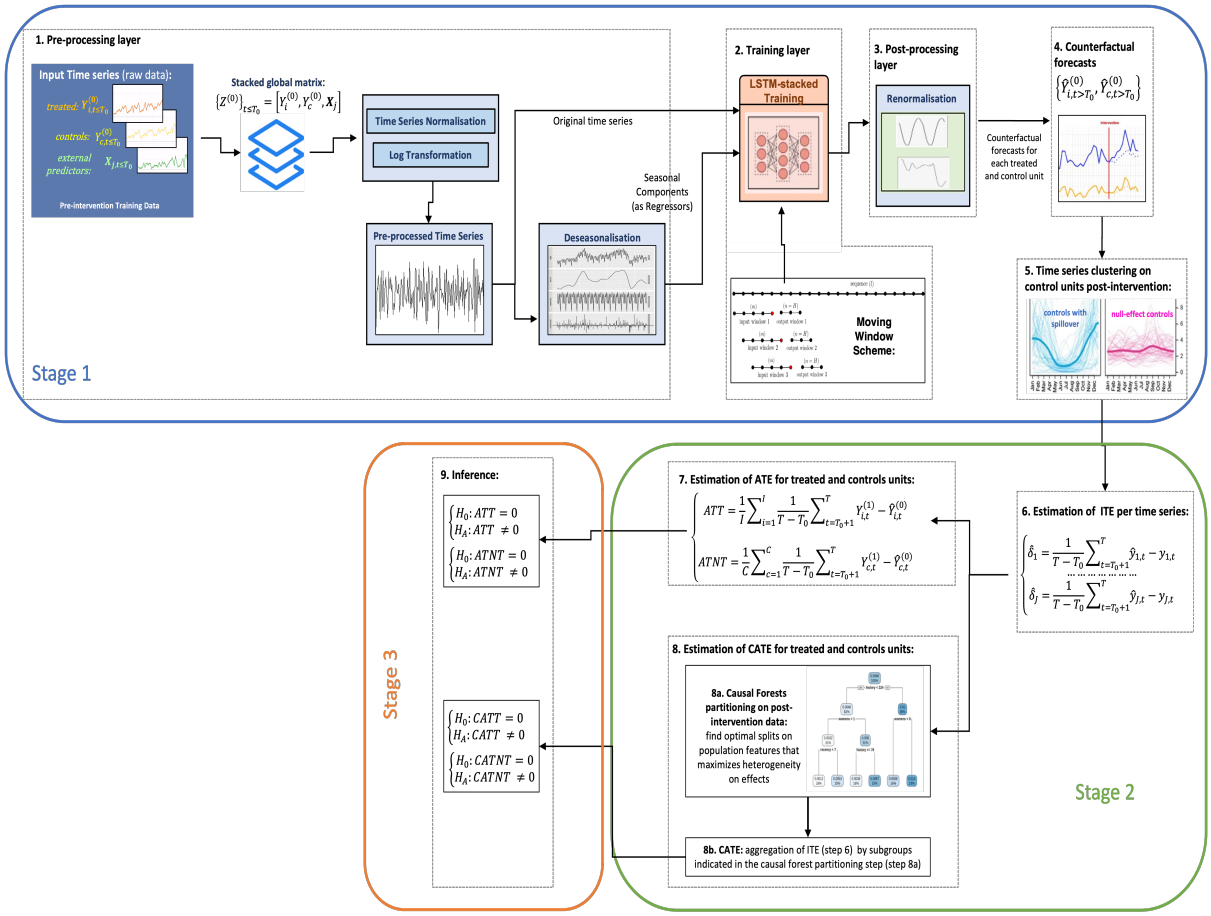


Figure 1: Global-DeepSC: Overall architecture in three stages.

then defined as $f : \mathbb{R}^{T_0} \to \mathbb{R}^{T_*}$. We might consider two methods to forecast a collection of time series: (1) the *local method* that fits a function for each time series, where the *local algorithm* A_L takes a time series Z_j and returns a forecasting function f_j, i.e., $A_L : \mathbb{R}^{T_0} \to (\mathbb{R}^{T_0} \to \mathbb{R}^{T_*})$ (which means an identical algorithm is deployed to each $Z_j \in \mathbb{S}$ generating distinct forecasting functions f_j for each Z_j); or (2) the *global method* that fits a unique function to all time series belonging to the collection, which means that the *global algorithm* A_G produces only one unique forecasting function g for the entire $\mathbb{S}$, $A_G : \mathbb{S} \to (\mathbb{R}^{T_0} \to \mathbb{R}^{T_*})$ [68], leading to shared errors and parameters for both $Y^{(1)}$ and $Y^{(0)}$ due to cross-learning.

Montero-Manso and Hyndman [68] demonstrate in their Proposition 1 that individual forecasts generated by distinct *local* functions (such as f_j coming from A_L algorithm) can be equivalently produced by an unique, comprehensive function g used by a global algorithm A_G. Following this, they prove that the complexity of *local* algorithms grows in proportion to the increasing sample set size, ultimately surpassing the relatively stable complexity of *global* algorithms. The complexity of the latter remains stable provided that the number of parameters within its singular forecasting function is maintained, as stated in their Proposition 2. Consequently, given an equivalent sample set size, *global* models can incorporate enhanced complexity – whether through the addition of features, extended memory, or other means – thereby yielding improved accuracy and superior generalisation compared with *local* models.

Our framework proposes $\mathcal{M}(\cdot)$ of Eq. (2) be this function g generated by the algorithm A_G method. The set of matrices $\mathbf{Y}_i = Y^{(0)}_{1:I,t\leq T_0}$, $\mathbf{Y}_c = Y^{(0)}_{1:C,t\leq T_0}$ and $\mathbf{X}_j = X^{(0)}_{1:J,t\leq T_0}$ are stacked together to create a larger matrix $\mathbf{Z}^{(0)} = \left[\mathbf{Y}_i\ \mathbf{Y}_c\ \mathbf{X}_j\right]'$ with $2 \times J$ time series, achieving a data pooling for use as input data. The motivation for choosing a *global* perspective over a *local*

one is that a *local* model precludes placebo tests while maintaining fixed parameters. Global structuring with shared parameters across time series is essential for testing forecast accuracy on unaffected controls, bolstering confidence in causal estimates. Moreover, owing to its *global* nature, Global-DeepSC can handle the issue of limited pre-intervention data by learning across all available series and pooling the data.

4.1.2 The autoregressive GFM joined RNN for counterfactual forecasting

Our general setup employs as $g(\cdot)$ an autoregressive recurrent neural network (RNN) model with the adoption of the long-short-term memory (LSTM) architecture such that

$$\{y^{(0)}_{j,T_0+1:T}\}_{j=1}^{J} = g(\mathbf{Z}^{(0)}_{1:T_0}, \Theta) + u_t, \tag{4}$$

where $\{u_t\} \sim iid\ (0, \sigma_u^2)$, Θ is the set of learnable parameters, and the architecture of the LSTM component of g follows the implementation described in Bandara et al. [69], Hewamalage et al. [70], Grecov et al. [14, 15]. Our modelling framework operates as follows: each time series $y_{j,t}$ from the pre-intervention periods of treated and control units, $\mathbf{Y}_i \cup \mathbf{Y}_c$, along with optional external covariates $\mathbf{x}_{j,t}$ from $\mathbf{X}_j$, are concatenated into a single matrix $\mathbf{Z}^{(0)}_{1:T_0}$ and fed into a *global* LSTM model in an autoregressive and recurrent manner, considering the values from the previous time step as well as the prior state of the network. Consequently, the network output $\mathbf{h}_{j,t}$ – which represents the post-intervention predictions for each series – of our autoregressive LSTM *global* model $g(\cdot)$, as characterised by its hidden states (h_t), can be mathematically formulated as:

$$\mathbf{h}^{LSTM}_{j,t} = g(\mathbf{h}_{j,t-1}, y_{:j,t-1}, \mathbf{x}_{:j,1:T_0}, \Theta), \tag{5}$$

where $g(\cdot)$ is a multilayer RNN with LSTM cells and rectified linear unit (ReLU) activation.

4.1.3 The forecasting engine of Global-DeepSC

The forecasting engine of Global-DeepSC comprises three distinct layers arranged sequentially as follows: 1) the pre-processing layer, 2) the Global-DeepSC training layer, and 3) the post-processing layer, as illustrated in steps 1 to 3 in Fig. 1. These pre-processing steps constitute the data engineering procedures executed to enhance the accuracy of the forecasting outcome. Initially, the pre-processing layer (step 1 in Fig. 1) commences by stacking all pre-treatment time series data of treated and control instances along with the external covariates time series to construct a unique matrix to be utilised as input for our DNN model. These time series are subsequently normalised via mean-scale transformation to mitigate variation and accentuate the intervention effect. This process addresses series-specific fixed effects and facilitates the estimation of within-effect policy intervention. Logarithmic transformation is then applied to stabilise the variance and ensure additive seasonality and trends, which is a conventional practice in forecasting. Subsequently, the seasonal components are extracted using the Seasonal and Trend Decomposition using Loess (STL) method [71], which are later integrated as exogenous variables in the Global-DeepSC model training.

In the training phase (step 2 in Fig. 1), an additional data engineering technique is implemented: the *moving-window* scheme. This approach involves dividing the time series data into fixed-size windows using the *moving-window* transformation strategy, generating multiple training instances to enhance model generalisation. The mean absolute error (MAE) for each window is calculated as the training loss function and minimised using the Adam optimiser [72]. The dataset is partitioned into three parts: training, validation, and test/forecast. The validation set is reserved for fine-tuning the LSTM hyper-parameters via the Sequential Model-Based Configuration (SMAC) algorithm [73], employing a Bayesian optimisation approach [70]. As proposed by Bandara et al. [74], we employ the Seasonal Exogenous (SE) training approach, which incorporates seasonal components as exogenous variables in training, providing an additional mechanism for Global-DeepSC to account for seasonality. TensorFlow$^{\text{TM}}$ version 2.0,

an open-source deep-learning toolkit, is utilised to implement the *global* LSTM employed for learning [75, 70].

In the final stage of the forecasting engine, the post-processing layer reverses the pre-processing transformations, restoring the model forecasting output to the original data scale, as illustrated in step 3 of Fig. 1. Post-processed time series forecasts constitute the counterfactual outcomes for each individual treated and control time series in the post-intervention period. The projected time series reconstructs the potential trajectory of each instance, has the treatment or intervention not been implemented, and is utilised to be contrasted against the observed time series to estimate the magnitude of the treatment effect.

4.1.4 The clustering process on control units

To construct this subset of control units with negligible effects, we introduce a time series clustering step on the controls, as illustrated in step 5 of Fig. 1. This step, carried out after training for counterfactual estimation, aims to isolate those controls that display distinct behaviour in the post-intervention period and is one of the innovations our framework presents. To this end, we employ a data-driven clustering technique to distinguish between the control units exhibiting null effects and those subject to certain spillovers, which utilises a dissimilarity measure combined with a hierarchical clustering algorithm to cluster the control time series.

To select an appropriate dissimilarity measure, we compared the accuracy of different model-free, model-based and complexity-based dissimilarity measures, resulting in 11 different dissimilarity measures tested throughout the simulated and real-world datasets employed in our experiments. All the measures were implemented using the `TSClust R` package [76]. After measuring the dissimilarity for these metrics, we perform the clustering using the hierarchical clustering (`hclust`) algorithm or the Partitioning Around Medoids (PAM) algorithm (more details in Montero and Vilar [76]). Each dissimilarity measure and clustering algorithm was also tested with either the original dataset or the original dataset's normalisation, PAA reduction, and SAX symbolisation (which we named the *tsersax* version). In our case, the ground truth of the clusters is known. This is evident for the simulated datasets. In the case of the real-world dataset, the geographical information encoded by the zip code allowed us to identify the control stores inside and outside the region containing the treated stores. Thus, we assess the clustering performance using classification accuracy. The best method for both datasets was the compression-based dissimilarity (CDM) measure [77, 78] combined with the `hclust` algorithm, which obtained accuracies of 71.4% and 70.0% for the synthetic and real-world experiments, respectively. The comprehensive findings can be found in the **Supplementary Material**. In summary, the data engineering and modelling procedures executed in this clustering step to identify which controls are unlikely to be influenced by spillovers are as follows: 1) collect the entire length of all control series; 2) apply normalisation, PAA reduction, and SAX symbolisation transformations to the time series from step 1; 3) use the CDM method to measure the dissimilarity between the transformed time series from step 2; and 4) apply hierarchical clustering to the dissimilarity values from step 3, utilising the parameter `cutree=2` to obtain two distinct clusters.

4.2 Stage 2: Estimation of treatment effect

In possession of the counterfactual forecasts produced in stage 1, we are ready to estimate the ITE_j for each treated and control unit as defined in Eq. (3) (step 6 in Fig. 1). Then, to obtain the ATE, we aggregate the ITE by averaging over all instances and the full period under treatment (step 7 of Fig. 1). Nevertheless, to estimate $CATE_s$ as defined in Eq. (3), we should find subsets of individual characteristics $S_{j,t} = s$ (the partitions) which maximise the dimension of impact heterogeneity; that is, where the same treatment yields heterogeneous effects. To

segregate our data into subgroups characterised by distinct treatment effect magnitudes, we employ causal forest strategies [21, 22] as described in sequence.

Causal forest partitioning (step 8 of Fig. 1) To identify these critical subgroups for HTE, users cannot rely solely on their expertise, particularly when working with high-dimensional data. Athey and Imbens [27] addressed this challenge by adapting random forest models to build trees that estimate treatment effects across data-driven groups. These groups are created by segmenting the population to maximise the treatment effect heterogeneity across the tree leaves. To minimise bias, causal forest models split the sample data into two non-overlapping subsets; this process is referred to as *honesty*. The first subset is the training subsample, which is used to build the tree by finding rules for partitioning a covariate space $\mathbf{X}$ and selecting the essential variables from $\mathbf{X}$ to predict the outcome $\hat{Y}_j$. The second subset, the estimation subsample, is used to estimate the within-leaf CATE using the fitted tree. Technical details regarding the formalism of the causal forest models can be found in **Supplementary Material**.

Our approach utilises only the tree structure output from this causal forest model and discards its CATE calculation, which is unsuitable for time series data. Hence, our approach treats the tree partition as if it was exogenously provided, as intended by the *honesty* process [79]. CATE is calculated by aggregating the ITE estimations ($\hat{\delta}_{j,t>T_0} = Y^{(1)}_{j,t>T_0} - Y^{(0)}_{j,t>T_0}$) produced by Global-DeepSC (Eq. (4)) by the subgroups of the terminal nodes of this tree produced by the causal forests, according to Eqs. (3). We implement the causal forests using the model `CausalForestDML` of Wager and Athey [21] and the `SingleTreeCateInterpreter()` function (used to plot the tree representation of the causal forest algorithm), both of which are available in the Python library [80]. The training input for learning the tree partitioning in our framework is the observed post-intervention data. To find the specific partitions for the treated and control groups, we run three distinct causal forest models: one for treated units (where input data is $Y^{(1)}_{i,t>T_0}$); one for controls units free from spillover (where input data is $Y^{(1)}_{c^*,t>T_0}$), and another for controls units with spillover (where the input data is $Y^{(1)}_{c-c^*,t>T_0}$).

4.3 Stage 3: Inference and placebo test

The residuals, defined as the difference between the forecasting outcomes (counterfactuals) and observed values, essentially represent the causal effect. For treated units, these residuals should be significantly different from zero, signifying the existence of a treatment effect. Meanwhile, the residuals for the control units should be substantially close to zero, implying a null effect of the intervention on controls, thus serving as a placebo test. Therefore, we conduct a double-tailed hypothesis testing procedure, wherein the null and alternative hypotheses are $H_0 : \widehat{ATE}, \widehat{CATE} = 0$ and $H_A : \widehat{ATE}, \widehat{CATE} \neq 0$ (step 9 in Fig. 1). Concerning the treated units ($\widehat{ATT}, \widehat{CATT}$), the rejection of the null hypothesis provides strong support for the existence of policy-induced effects. Conversely, for the subgroup of controls cleaned from spillover effects ($\widehat{ATNT}, \widehat{CATNT}$), as a placebo test, we expect to accept the null hypothesis to affirm the null treatment effect and confirm **Assumption 1**. However, the acceptance of null impacts is unlikely for the control units subject to spillovers. We follow the test statistic and *p-value* definitions from Chernozhukov et al. [7].

5 Experiments with Simulated Data

5.0.1 Synthetic Data

Given that counterfactuals are unobservable, assessing which estimates approximate the true treatment effect can be challenging. The standard practice for evaluating counterfactual forecasting uses synthetic datasets to compare estimates that could be closer to synthetic ground-

Table 1: Simulated Study: ATT/ATNT and Accuracy - scenarios without clustering (A) and clustering the controls (B)

	(A) MODEL	ATT/ATNT[a]	RMSE (for ATT/ATNT)	RMSE (for point forecasting)	WRMSPE[b]	p-values ($H_0 : ATT/ATNT = 0$)
Treated Units	(True Effect)	-1.890	-	-	-	-
	Global-DeepSC	-1.859	**0.0313**	**0.0788**	**0.00257**	1.12e-69
	Global-FFNN	-1.941	0.0513	0.0799	0.00261	2.39e-85
	DeepAR	-1.768	0.1218	0.1485	0.00485	5.17e-107
	CausalImpact	-1.510	0.3798	0.4220	0.01377	7.02e-100
	ArCo	-1.470	0.4196	0.4524	0.01477	7.41e-104
	DiD	-1.462	0.4283	-	-	6.03e-70
Control Units - spil	(Expected Effect)	0	-	-	-	-
	(True Effect/Spillover)	-0.382	-	-	-	-
	Global-DeepSC	-0.423	**0.0415**	**0.1095**	**0.00371**	3.32e-16
	Global-FFNN	-0.465	0.0836	0.1249	0.00423	7.57e-19
	DeepAR	0.039	0.4209	0.4454	0.01508	0.25200
	CausalImpact	-0.245	0.1368	0.2113	0.00715	0.0000159
	ArCo	-0.071	0.3108	0.3459	0.01171	0.0000196
	DiD	-0.538	0.1565	-	-	1.34e-10
Control Units	(True Effect)	0	-	-	-	-
	Global-DeepSC	0.225	**0.2251**	**0.2605**	**0.00888**	0.26200
	Global FFNN	0.249	0.2488	0.2867	0.00977	0.04320
	DeepAR	0.334	0.3339	0.3669	0.01251	6.60e-21
	CausalImpact	0.400	0.4004	0.4708	0.01605	0.00000331
	ArCo	0.353	0.3532	0.3921	0.01337	0.00000393
	DiD	0.538	0.5381	-	-	1.34e-10

	(B) MODEL	ATT/ATNT[a]	RMSE (for ATT/ATNT)	RMSE (for point forecasting)	WRMSPE[b]	p-values ($H_0 : ATT/ATNT = 0$)
Treated Units	(True Effect)	-1.890	-	-	-	-
	Global-DeepSC	-1.859	**0.0313**	**0.0788**	**0.00257**	1.12e-69
	Global-FFNN	-1.941	0.0513	0.0799	0.00261	2.39e-85
	DeepAR	-1.768	0.1218	0.1485	0.00485	5.17e-107
	CausalImpact	-1.281	0.6088	0.6877	0.02244	2.46e-49
	ArCo	-1.254	0.6357	0.6891	0.02249	4.68e-53
	DiD	-1.322	0.5677	-	-	2.96e-44
Control Units - spil	(Expected Effect)	0	-	-	-	-
	(True Effect/Spillover)	-0.261	-	-	-	-
	Global-DeepSC	-0.270	**0.0092**	0.0968	0.00322	2.11e-12
	Global-FFNN	-0.247	0.0137	**0.0804**	**0.00268**	1.33e-11
	DeepAR	0.112	0.3725	0.4139	0.01378	0.01470
	CausalImpact	-	-	-	-	-
	ArCo	-	-	-	-	-
	DiD	-	-	-	-	-
Control Units	(True Effect)	-0.092	-	-	-	-
	Global-DeepSC	0.143	0.2345	0.3005	0.01051	0.52900
	Global FFNN	0.009	0.1801	0.2854	0.00998	0.81100
	DeepAR	0.293	0.3843	0.4044	0.01414	1.88e-18
	CausalImpact	0.119	0.2104	0.3051	0.01067	0.62900
	ArCo	0.020	**0.1114**	**0.1530**	**0.00535**	0.20900
	DiD	0.238	0.3297	-	-	0.00476

[a] ATT is calculated for treated units whilst ATNT for control units. [b]Weighted Root Mean Squared Proportional Error.

truth treatment effects. Our simulation considers a multivariate scenario built by a DGP embracing two steps. First, we replicate a nonlinear Self-exciting Threshold Autoregressive (SETAR) model to create 100 time series for the intervention variable. Second, we employ a Monte Carlo simulation process of a VAR(3, 4) model, where the intervention time series from the first step is entered as a regressor. The resultant time series set with three exogenous variables iterating with our target variable (at the end, a total of 100 series for each variable, with a length of 224, 70 assigned as controls and 30 as treated) presents then a nonlinear relationship across the series (more details concerning this construction can be found in the **Supplementary Material**). The intervention takes place at $T_0 = 220$. To emulate a scenario with heterogeneous treatment effects, each treated instance is affected by different treatment levels. In addition, to build a scenario in which a spillover effect occurs, a portion of the control series is synthetically affected by a lower magnitude level of treatment. In this manner, we seek to emulate conditions similar to real-world scenarios.

5.0.2 The baselines and their implementation

We carry out the experiments to benchmark the Global-DeepSC performance against well-known structural CSC-based causal models such as the Artificial Counterfactual (ArCo) of Carvalho et al. [6] [81]; the CausalImpact of Brodersen et al. [8]; and a canonical DiD model version [7]. In contrast to Global-DeepSC, these models are local and linear-regression-based and, as structural causal models, use the post-treatment data to estimate the effects. ArCo model, which utilises constrained LASSO regression, is implemented via the `ArCo` R package [82]. In contrast, CausalImpact, available through the `CausalImpact` R package [83], employs a Bayesian structural time series model that is founded on diffusion-regression state-space models. We implement the canonical DiD model by performing linear regressions (using the `lm` function in R), where the DiD estimator is obtained from the resultant coefficient of the interaction between the dummy variables that indicate the treatment start and treated units. We also contrast the Global-DeepSC performance against two global models: i) a canonical FFNN implemented using version 2.0.0 of the Tensorflow package in Python as a global model, which, identically to Global-DeepSC, is fed only with pre-treatment data of treated controls, and external predictors; and ii) DeepAR from Salinas et al. [84] which is implemented using the `deepar` package from GluonTS [85]. As a GFM model, DeepAR utilises solely the pre-treatment data from both treated and control series. However, owing to its implementation, it incorporates post-treatment information from external predictors.

5.0.3 Results

Table 1 reports the simulated results for two scenarios: (A) when we use the true labels of the control units with and without spillovers, and (B) when we perform the clustering to distinguish them. First of all, when comparing the true effects of both control groups between the two scenarios, we can see that the clustering produced a slight divergence from the ground truth effects. Global-DeepSC outperformed all methods for ATE estimation and point forecasting for treated units in both scenarios and always obeyed the appropriate *p-values* for all situations (≤ 0.05 for treated units and controls with spillover, and > 0.05 for controls without spillover). Concerning controls with spillover, it is only outperformed by Global-FFNN in the point forecasting accuracy for scenario (B). For the SCMs, we verify that using the clustering method to select the controls without spillovers improves their placebo tests but not necessarily the estimation of ATT for treated units since less data is employed to train the models. Additional results regarding this simulated study (such as tests for significant differences in performance and variation in average performance) are found in the **Supplementary Material**.

6 Real-World Application

6.0.1 Dataset

Our empirical analysis is designed to estimate the impact of grocery store promotions on essential products, specifically milk, as substantiated by Mojir and Sudhir [86], on overall store sales. Given the high competitiveness in shared supermarket markets, we hypothesise that exclusive promotions on primary goods by a subset of stores may induce customers to switch their store choices and trigger spillover effects on neighbouring competitors that did not adopt the same marketing strategy. We utilised the proprietary NielsenIQ-Consumer Panel dataset, encompassing weekly grocery purchase records from households across retail stores in the USA from January 2016 to December 2018. After applying specific filters (detailed in **Supplementary Material**), we retained 148-time series corresponding to individual stores, composed of 102 treated stores and 46 control stores. The clustering step identified the possible control stores affected by spillover effects. Control stores were subsequently classified into two subgroups, those located within and outside the treated regions, based on their respective divergent or convergent post-intervention behaviour when juxtaposed with treated stores. The treatment period spanned from December 1st to 29th, 2018, yielding a training dataset of length 152. In addition to the sales time series (the target variable), we incorporated several exogenous variables, such as the USA holiday calendar, the store's and county's zip codes, and the parent store code.

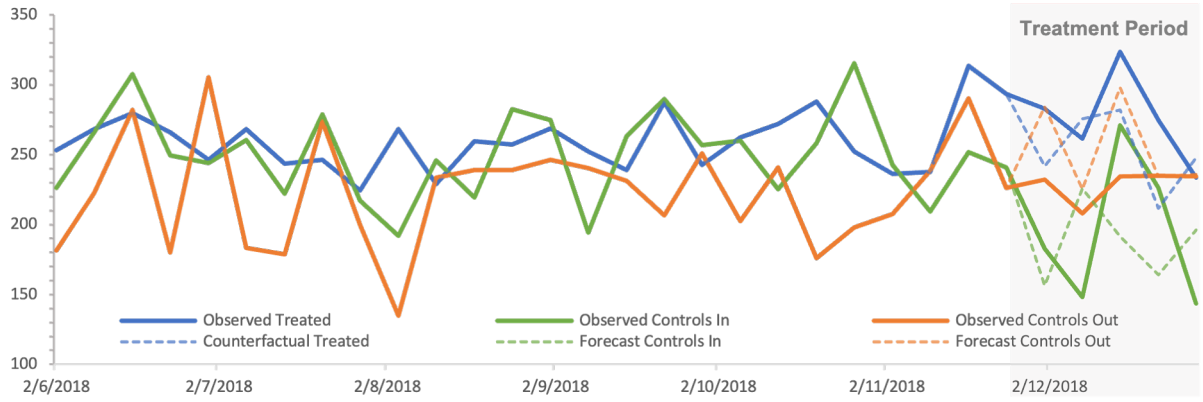


Figure 2: Counterfactual forecasting by Global-DeepSC.

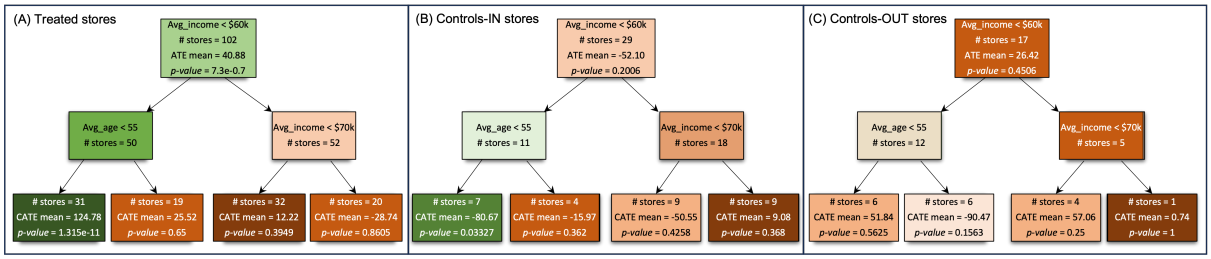


Figure 3: CATE estimated by Global-DeepSC using random forests to split the treatment effect by income and age subgroups. Green tones point to significant (higher) CATE, while orange/brown tones point to insignificant (lower) CATE.

Table 2: Empirical Study: Errors and ATE - clustering the controls

	(B) MODEL	RMSE (for 2W point forecasting)	WRMSPE[a]	Average Effect (ATT/ATE_{ctr})[b] Week 1	2 Weeks	*p-values* Week 1	2 Weeks
Treated Units	Global-DeepSC	-	-	40.88 (+14.5%)	13.37 (+4.9%)	7.3e-07	7.0e-04
	CausalImpact	-	-	-39.22 (-13.9%)	-33.09 (-12.9%)	0.195	0.0269
	ArCo	-	-	121.79 (+43.1%)	119.18 (+43.8%)	3.1e-07	2.3e-13
	DiD	-	-	36.30 (+12.8%)	42.39 (+43.8%)	0.3549	0.1831
Control Units - IN	Global-DeepSC	-	-	-52.10 (-22.4%)	-34.64 (-15.7%)	0.2006	0.1637
	CausalImpact	-	-	-66.22 (-28.5%)	-58.33 (-26.5%)	0.0179	0.0023
	ArCo	-	-	-42.10 (-18.1%)	-52.07 (-23.7%)	0.0086	4.4e-05
	DiD	-	-	-18.41 (-10.1%)	-24.54 (-14.8%)	0.6880	0.4847
Control Units - OUT	Global-DeepSC	**57.991**	**0.248**	26.42 (14.5%)	-25.61 (-15.5%)	0.4506	0.1213
	CausalImpact	58.363	0.250	-43.50 (-23.8%)	-56.82 (-34.5%)	0.3289	0.1408
	ArCo	72.671	0.311	-53.46 (-29.3%)	-70.62 (-42.7%)	0.0313	9.0e-06
	DiD	-	-	18.41 (+7.9%)	24.54 (+11.2%)	0.6880	0.4847

[a] Weighted Root Mean Squared Proportional Error. [b]ATT is calculated for treated units whilst ATE_{ctr} for control units.

6.0.2 Results

In the CSC approach, the true underlying impacts of the treated instances are not known in real-world scenarios. Therefore, to validate the model's forecasting performance, we focus on the errors only for stores that did not offer any milk deals during the post-treatment period and are clean from spillovers ("Control Units-OUT" in Table 2).

In Fig. 2, when contrasting the forecasting outcomes to the observed series, we see that the effect on treated stores is more significant in the first week. We conclude the same for the spillover effect in control stores located inside the treated area. These stores present a noticeable reduction in sales in the first two weeks (as a side effect), whereas, for controls outside the treated area, Global-DeepSC returns a more negligible effect. Thus, we concentrate on the impact analysis of offering deals for milk on total sales only in the first two weeks, as reported in Table 2.

The ATE results reported in Table 2 are only for the scenario where the control series were segregated using the cluster modelling, where we observe Global-DeepSC obtains the best performance for the "control-out" units. As expected, the ATE estimated by Global-DeepSC for treated stores pointed to a positive impact of 4.9% of growth in their sales and the desired inference results for treated and control units. On the other hand, only some SCMs, such as ArCo, benefit from the clustering process to identify and remove the controls with spillovers and produce, under this scenario, a correct direction of the impact for the treated series, added to lower *p-values*. However, the fact of employing fewer data during the training (by removing the "control-in" series from training) prejudices the inference result for the "control-out" units. Overall, Global-DeepSC is the only method that leads us to conclude (by all error and inference metrics together for either treated or control instances) that the strategy of offering deals for essential products promotes an increase in total sales and leads to some spillover effects on the control stores neighbouring the treated stores and null effects from controls outside the treated region.

Fig. 3 shows the CATE estimation results obtained by averaging the first week of treatment effects for different income and consumer age levels, where the optimal splits were determined by causal random forest models using 1,000 trees. We can observe (for treated stores) that the marketing strategy of offering deals for milk to induce the purchase of other products

is statistically significant for consumers with lower income and middle age (leftmost leaf of tree (A)). Moreover, the opposite side effect of decreasing sales in the "controls-in" stores is also statistically significant at similar income and age levels. In addition, the inference test confirmed a null effect for all leaves of the "controls-out" group tree. Hence, they can be employed as controls for the ATE estimation. These results lead us to conclude, first, that there are heterogeneous treatment effects in this type of scenario. Second, younger people from lower income levels are the most sensitive to deals because they might have tighter budgets; thus, they should be the target consumer group for this type of marketing strategy.

7 Conclusion

Discerning treatment effects in scenarios plagued by spillovers remains challenging, despite the plethora of impact evaluation tools. This issue often precludes the application of baseline CSC models, making it difficult to conduct impact analyses for relevant policy implementations. Our GFM-LSTM-based engine for HTE estimation using an SC approach demonstrates the effectiveness of learning a GFM from a nonlinearly related time series, being capable of mapping complex patterns, even under scenarios where we cannot ensure complete adherence to SUTVA. While recent research has leveraged DNN forecasting models for HTE estimation, the fusion of GFM models to perform policy evaluation remains largely unexplored, especially with the benefit of not requiring post-intervention data for unbiased HTE estimation amidst prevalent unit interferences. Hence, this paper presents a comprehensive experimental proposal for this field in such wider contexts. Our framework alleviates some stringent assumptions characteristic of traditional methods, enabling more reliable HTE assessments in real-world scenarios and thus facilitating decision-making. Empirical evidence underscores its superior performance compared to well-established SCMs. We incorporate a data-driven clustering step to isolate control instances susceptible to spillover effects, enhancing our ability to identify clean control units where manual differentiation is untenable. Evaluating CATE alongside ATE augments the interpretability of the analysis, enabling the identification of drivers that contribute to the differential impacts of the same intervention. Given the capacity of our method to estimate the effects for each time series individually, we can aggregate the individual effects across any chosen subgroup. We leverage a causal forest data-driven approach applied to longitudinal data to partition our population into optimal subgroups that unveil significant impact heterogeneity, thereby preventing the need for user-specified indicators for variables that might inform and guide our policies. Currently, we assume the same amount of a single treatment for all affected units. An avenue for future research would be to develop models that can quantify the impact of varying levels of treatment intensity, such as more minor versus more extensive promotions.

Acknowledgments

This research was supported by funding from Monash University Graduate Research, Australia.

References

[1] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[2] S. Athey and G. W. Imbens, "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, vol. 74, no. 2, pp. 431–497, 2006.

[3] T. G. Conley and C. R. Taber, "Inference with "difference in differences" with a small

number of policy changes," *The Review of Economics and Statistics*, vol. 93, no. 1, pp. 113–125, 2011.

[4] N. Doudchenko and G. W. Imbens, "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis," National Bureau of Economic Research, 1050 Massachusetts Avenue - Cambridge, MA 02138, Tech. Rep., 2016.

[5] A. Abadie, A. Diamond, and J. Hainmueller, "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American statistical Association*, vol. 105, no. 490, pp. 493–505, 2010.

[6] C. Carvalho, R. Masini, and M. C. Medeiros, "Arco: an artificial counterfactual approach for high-dimensional panel time-series data," *Journal of Econometrics*, vol. 207, no. 2, pp. 352–380, 2018.

[7] V. Chernozhukov, K. Wüthrich, and Y. Zhu, "An exact and robust conformal inference method for counterfactual and synthetic controls," *Journal of the American Statistical Association*, pp. 1–16, 2021.

[8] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott, "Inferring causal impact using Bayesian structural time-series models," *The Annals of Applied Statistics*, vol. 9, no. 1, pp. 247–274, 2015.

[9] L. Gobillon and T. Magnac, "Regional policy evaluation: Interactive fixed effects and synthetic controls," *Review of Economics and Statistics*, vol. 98, no. 3, pp. 535–551, 2016.

[10] S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi, "Matrix completion methods for causal panel data models," *Journal of the American Statistical Association*, pp. 1–15, 2021.

[11] D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager, "Synthetic difference-in-differences," *American Economic Review*, vol. 111, no. 12, pp. 4088–4118, December 2021.

[12] F. Menchetti, F. Cipollini, and F. Mealli, "Combining counterfactual outcomes and arima models for policy evaluation," *The Econometrics Journal*, vol. 26, no. 1, pp. 1–24, 2023.

[13] A. Cerqua, M. Letta, and F. Menchetti, *Losing Control (Group)? The Machine Learning Control Method for Counterfactual Forecasting.* SSRN, 2023.

[14] P. Grecov, K. Bandara, C. Bergmeir, K. Ackermann, S. C. Campbell, D. Scott, and D. I. Lubman, "Causal inference using global forecasting models for counterfactual prediction," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining 2021.* Springer, 2021, pp. 282–294.

[15] P. Grecov, A. N. Prasanna, K. Ackermann, S. Campbell, D. Scott, D. I. Lubman, and C. Bergmeir, "Probabilistic causal effect estimation with global neural network forecasting models," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[16] R. Johari, H. Li, I. Liskovich, and G. Y. Weintraub, "Experimental design in two-sided platforms: An analysis of bias," *Management Science*, 2022.

[17] M. Liu, J. Mao, and K. Kang, "Trustworthy and powerful online marketplace experimentation with budget-split design," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3319–3329.

[18] V. Melnychuk, D. Frauen, and S. Feuerriegel, "Causal transformer for estimating counterfactual outcomes," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 293–15 329.

[19] D. Cao, J. Enouen, Y. Wang, X. Song, C. Meng, H. Niu, and Y. Liu, "Estimating treatment effects from irregular time series observations with hidden confounders," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, pp. 6897–6905, Jun. 2023.

[20] L. Yao, S. Li, Y. Li, H. Xue, J. Gao, and A. Zhang, "On the estimation of treatment effect with text covariates," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 4106–4113.

[21] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[22] C. Tran and E. Zheleva, "Heterogeneous peer effects in the linear threshold model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 4175–4183.

[23] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[24] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[25] K. Kuang, P. Cui, H. Zou, B. Li, J. Tao, F. Wu, and S. Yang, "Data-driven variable decomposition for treatment effect estimation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 5, pp. 2120–2134, 2020.

[26] A. Wu, J. Yuan, K. Kuang, B. Li, R. Wu, Q. Zhu, Y. Zhuang, and F. Wu, "Learning decomposed representations for treatment effect estimation," *IEEE Transactions on Knowledge & Data Engineering*, vol. 35, no. 05, pp. 4989–5001, 2023.

[27] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016.

[28] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[29] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*. PMLR, 2016, pp. 3020–3029.

[30] J. Yoon, J. Jordon, and M. Van Der Schaar, "Ganite: Estimation of individualized treatment effects using generative adversarial nets," in *International conference on learning representations*, 2018.

[31] A. Curth and M. Van der Schaar, "Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1810–1818.

[32] Q. Li, Z. Wang, S. Liu, G. Li, and G. Xu, "Causal optimal transport for treatment effect estimation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 8, pp. 4083–4095, 2021.

[33] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International conference on machine learning*. PMLR, 2017, pp. 3076–3085.

[34] J. Berrevoets, J. Jordon, I. Bica, M. van der Schaar *et al.*, "Organite: Optimal transplant donor organ offering using an individual treatment effect," *Advances in neural information processing systems*, vol. 33, pp. 20 037–20 050, 2020.

[35] A. Curth and M. Van der Schaar, "On inductive biases for heterogeneous treatment effect estimation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 883–15 894, 2021.

[36] J. Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect," *Mathematical modelling*, vol. 7, no. 9-12, pp. 1393–1512, 1986.

[37] J. Robins and M. Hernan, "Estimation of the causal effects of time-varying exposures," *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pp. 553–599, 2008.

[38] J. M. Robins, M. A. Hernan, and B. Brumback, "Marginal structural models and causal inference in epidemiology," pp. 550–560, 2000.

[39] Y. Xu, Y. Xu, and S. Saria, "A bayesian nonparametric approach for estimating individualized treatment-response curves," in *Machine learning for healthcare conference*. PMLR, 2016, pp. 282–300.

[40] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," *Advances in neural information processing systems*, vol. 30, 2017.

[41] H. Soleimani, A. Subbaswamy, and S. Saria, "Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions," *arXiv preprint arXiv:1704.02038*, 2017.

[42] Z. Qian, Y. Zhang, I. Bica, A. Wood, and M. van der Schaar, "Synctwin: Treatment effect estimation with longitudinal outcomes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3178–3190, 2021.

[43] J. Berrevoets, A. Curth, I. Bica, E. McKinney, and M. van der Schaar, "Disentangled counterfactual recurrent networks for treatment effect inference over time," *arXiv preprint arXiv:2112.03811*, 2021.

[44] R. Li, Z. Shahn, J. Li, M. Lu, P. Chakraborty, D. Sow, M. Ghalwash, and L.-w. H. Lehman, "G-net: a deep learning approach to g-computation for counterfactual outcome prediction under dynamic treatment regimes," *arXiv preprint arXiv:2003.10551*, 2020.

[45] I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar, "Estimating counterfactual treatment outcomes over time through adversarially balanced representations," *arXiv preprint arXiv:2002.04083*, 2020.

[46] B. Lim, "Forecasting treatment responses over time using recurrent marginal structural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[47] K. Fujii, K. Takeuchi, A. Kuribayashi, N. Takeishi, Y. Kawahara, and K. Takeda, "Estimating counterfactual treatment outcomes over time in complex multiagent scenarios," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[48] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[49] O. C. Ashenfelter and D. Card, "Using the longitudinal structure of earnings to estimate the effect of training programs," *The Review of Economics and Statistics*, vol. 67, no. 4, pp. 648–660, 1985.

[50] A. Dube and B. Zipperer, "Pooling multiple case studies using synthetic controls: An application to minimum wage policies," IZA Discussion Papers, Tech. Rep., 2015.

[51] J. J. Donohue, A. Aneja, and K. D. Weber, "Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis," *Journal of Empirical Legal Studies*, vol. 16, no. 2, pp. 198–247, 2019.

[52] G. Papadogeorgou, F. Menchetti, C. Choirat, J. H. Wasfy, C. M. Zigler, and F. Mealli, "Evaluating federal policies using bayesian time series models: estimating the causal impact of the hospital readmissions reduction program," *Health Services and Outcomes Research Methodology*, vol. 23, no. 4, pp. 433–451, 2023.

[53] J. Antonelli and B. Beck, "Heterogeneous causal effects of neighbourhood policing in new york city with staggered adoption of the policy," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 186, no. 4, pp. 772–787, 2023.

[54] V. Rakesh, R. Guo, R. Moraffah, N. Agarwal, and H. Liu, "Linked causal variational autoencoder for inferring paired spillover effects," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1679–1682.

[55] Z. Fatemi and E. Zheleva, "Minimizing interference and selection bias in network experiment design," in *Proceedings of the international aaai conference on web and social media*, vol. 14, 2020, pp. 176–186.

[56] ——, "Network experiment design for estimating direct treatment effects," in *KDD International Workshop on Mining and Learning with Graphs (MLG 2020)*, 2020.

[57] Z. Qu, R. Xiong, J. Liu, and G. Imbens, "Efficient treatment effect estimation in observational studies under heterogeneous partial interference," *arXiv preprint arXiv:2107.12420*, 2021.

[58] U. Awan, M. Morucci, V. Orlandi, S. Roy, C. Rudin, and A. Volfovsky, "Almost-matching-exactly for treatment effect estimation under network interference," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 3252–3262.

[59] L. Forastiere, E. M. Airoldi, and F. Mealli, "Identification and estimation of treatment and interference effects in observational studies on networks," *Journal of the American Statistical Association*, vol. 116, no. 534, pp. 901–918, 2021.

[60] Q. Huang, J. Ma, J. Li, R. Guo, H. Sun, and Y. Chang, "Modeling interference for individual treatment effect estimation from networked observational data," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 3, pp. 1–21, 2023.

[61] A. N. Prasanna, P. Grecov, A. D. Weng, and C. Bergmeir, "Causal effect estimation with global probabilistic forecasting: A case study of the impact of covid-19 lockdowns on energy demand," *IEEE Transactions on Power Systems*, vol. 39, no. 2, pp. 3417–3430, 2024.

[62] D. B. Rubin, "Bayesian inference for causal effects: The role of randomization," *The Annals of statistics*, pp. 34–58, 1978.

[63] J. S. Neyman, "On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480)," *Annals of Agricultural Sciences*, vol. 10, pp. 1–51, 1923.

[64] Y. Zeng, R. Cai, F. Sun, L. Huang, and Z. Hao, "A survey on causal reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[65] P. Hünermund, B. Louw, and I. Caspi, "Double machine learning and automated confounder selection: A cautionary tale," *Journal of Causal Inference*, vol. 11, no. 1, p. 20220078, 2023.

[66] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.

[67] D. B. Rubin, "Randomization analysis of experimental data: The fisher randomization test comment," *Journal of the American statistical association*, vol. 75, no. 371, pp. 591–593, 1980.

[68] P. Montero-Manso and R. J. Hyndman, "Principles and algorithms for forecasting groups of time series: Locality and globality," *International Journal of Forecasting*, no. xxxx, 2021. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2021.03.004

[69] K. Bandara, C. Bergmeir, and H. Hewamalage, "LSTM-MSNet: Leveraging forecasts on sets of related time series with multiple seasonal patterns," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1586–1599, 2021.

[70] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *International Journal of Forecasting*, vol. 37, no. 1, pp. 388–427, 2021.

[71] R. Cleveland, W. Cleveland, J. McRae, and I. Terpenning, "STL: A Seasonal-Trend decomposition procedure based on loess," *J. Off. Stat.*, vol. 6, no. 1, pp. 3–33, 1990.

[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[73] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *International conference on learning and intelligent optimization*. Springer, 2011, pp. 507–523.

[74] K. Bandara, C. Bergmeir, S. Campbell, D. Scott, and D. Lubman, "Towards accurate predictions and causal 'what-if' analyses for planning and policy-making: A case study in emergency medical services demand," in *IJCNN*. IEEE, 2020, pp. 1–10.

[75] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[76] P. Montero and J. A. Vilar, "Tsclust: An r package for time series clustering," *Journal of Statistical Software*, vol. 62, pp. 1–43, 2015.

[77] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 206–215.

[78] E. Keogh, J. Lin, S.-H. Lee, and H. V. Herle, "Finding the most unusual time series subsequence: algorithms and applications," *Knowledge and Information Systems*, vol. 11, pp. 1–27, 2007.

[79] S. Athey and S. Wager, "Susan Athey and Stefan Wager: Estimating heterogeneous treatment effects in R," https://www.youtube.com/watch?v=YBbnCDRCcAI, 2021, accessed: 2024-06-01.

[80] EconML, "Estimation methods under unconfoundedness: Forest based estimators," https://econml.azurewebsites.net/spec/estimation/forest.html, 2023, accessed: 2023-06-01.

[81] R. Masini and M. C. Medeiros, "Counterfactual analysis with artificial controls: Inference, high dimensions, and nonstationarity," *Journal of the American Statistical Association*, vol. 116, no. 536, pp. 1773–1788, 2021.

[82] Y. R. Fonseca, R. Masini, M. C. Medeiros, and G. F. R. Vasconcelos, "Arco: Artificial counterfactual package," nov 2017, r package. [Online]. Available: https://CRAN.R-project.org/package=ArCo

[83] K. H. Brodersen and A. Hauser, "Causalimpact - an r package for causal inference using bayesian structural time-series models," 2014, r package. [Online]. Available: https://cran.r-project.org/web/packages/CausalImpact/vignettes/CausalImpact.html

[84] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *Int. J of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

[85] A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. Rangapuram, D. Salinas, J. Schulz, L. Stella, A. C. Türkmen, and Y. Wang, "GluonTS: Probabilistic time series models in python," *arXiv*, pp. 1–24, 2019.

[86] N. Mojir and K. Sudhir, "A model of multipass search: Price search across stores and time," *Management Science*, vol. 67, no. 4, pp. 2126–2150, 2021.