# Article Review: Inferring Causal Impact Using Bayesian Structural Time-Series Models

Filip Reierson[1]
Econometrics & Business Statistics
Monash University

October 17, 2023

**Abstract** In this report I summarise Brodersen et al. (2015) and reproduce simulations as described. I implement ways that these simulations can become reproduceable and describe how Brodersen et al. (2015) falls short in this respect. Additionally, I show that a frequentist version of the power curve and coverage give similar resulsts to that in Brodersen et al. (2015).

*Keywords*: synthetic control, simulation, visualisation

## Introduction

Organisations often need or benefit from assessing the contribution of various parts to a system as a whole. For example, a government agency may wish to understand which policies had the desired outcome, and a business may wish to understand what decision caused an increase in sales. This type of reasoning is known as causal inference.

In this report, I review Brodersen et al. (2015) which presents a practical approach to modelling causal impact in time series data using state space models. The paper was influential as measured by citations and people using the ac-

[1]This research is supported by an Australian Government Research Training Program (RTP) Scholarship Monash Graduate Excellence Scholarship and Monash Business School Graduate Research Scholarship.

companying R package. However, the paper's reproducibility is lacking as the dataset used as empirical justification is not available and the simulation section included most, but not all the required inputs. To alleviate these issues I have made this report fully reproducible so that the main insights can be verified. This report was prepared using R (R Core Team 2023), quarto for document preparation (Allaire et al. 2022), and R packages by Wickham et al. (2019), Pedersen (2023), Meschiari (2022), and Hester and Bryan (2022).

## Synthetic control

A researcher may want to investigate what the effect of a novel government policy has been. In causal language the question is, how different was the observed outcome to what would have been observed were the government policy not implemented? If there is an obvious control such as a similar neighbouring state that did not receive the treatment[2] for reasons unrelated to the outcome of interest, then the researcher may use that. However, if there are many candidate controls, none of which on their own are similar to the treated unit, then a different approach is called for. The case where a weighted average of candidate controls can be constructed to resemble the treated unit is the setting that is dealt with in Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010). This type of control is known as a synthetic control. Abadie's approach requires that a convex combination of the candidate controls can approximate the outcome for the treated series. The reasoning is that if this synthetic control predicts the treated series well before treatment and the controls were not affected by the treatment, then the synthetic control predicts what would have happened if the treated series had not been treated.

Abadie's approach requires that the researcher has a set of units which can through a weighted average approximate the treated series. Brodersen et al. (2015) propose that a synthetic control can also be constructed as long as there are series that predict the treated series, but may not be a similar kind of unit. For example, instead of advertisement clicks in different regions being used as controls, Google search trends could be used to retrospectively forecast a synthetic control. As with controls in general it is important that the predictors are themselves not affected by the intervention. Additionally, Brodersen et al. (2015) uses a spike and slab prior such that uncertainty in choosing predictors in the synthetic control is captured in credible intervals for the causal effect.

[2]Treatment, intervention, and campaign are used interchangeably to refer to some causal impact that is introduced.

## Causal inference

What would have happened in the absence of treatment is never observable in the treated series, but this theoretical series, known as a counterfactual, is still of scientific interest since it makes it possible to determine the causal impact of the treatment. Consider a random walk that is affected by a shock at some point in time, then the causal impact of the shock can be seen as the difference between what occurred and what would have occurred in the absence of the shock. The meaning of a difference can vary depending on the context and hypothesis. For example, when describing the causal effect on a flow quantity it can both make sense to talk about an increase of in a time period and the cumulative effect which is found by adding up multiple periods. To see why, consider Figure 1. On the other hand, for a stock quantity such as a population, it would not make sense to add up causal effects. In Brodersen et al. (2015) both the average effect across the treatment period and the cumulative effect is reported and these provide identical inference about relative effects. In Brodersen et al. (2015) a causal effect is considered to be present if the 95% credible interval for the average (or cumulative effect) excludes zero.

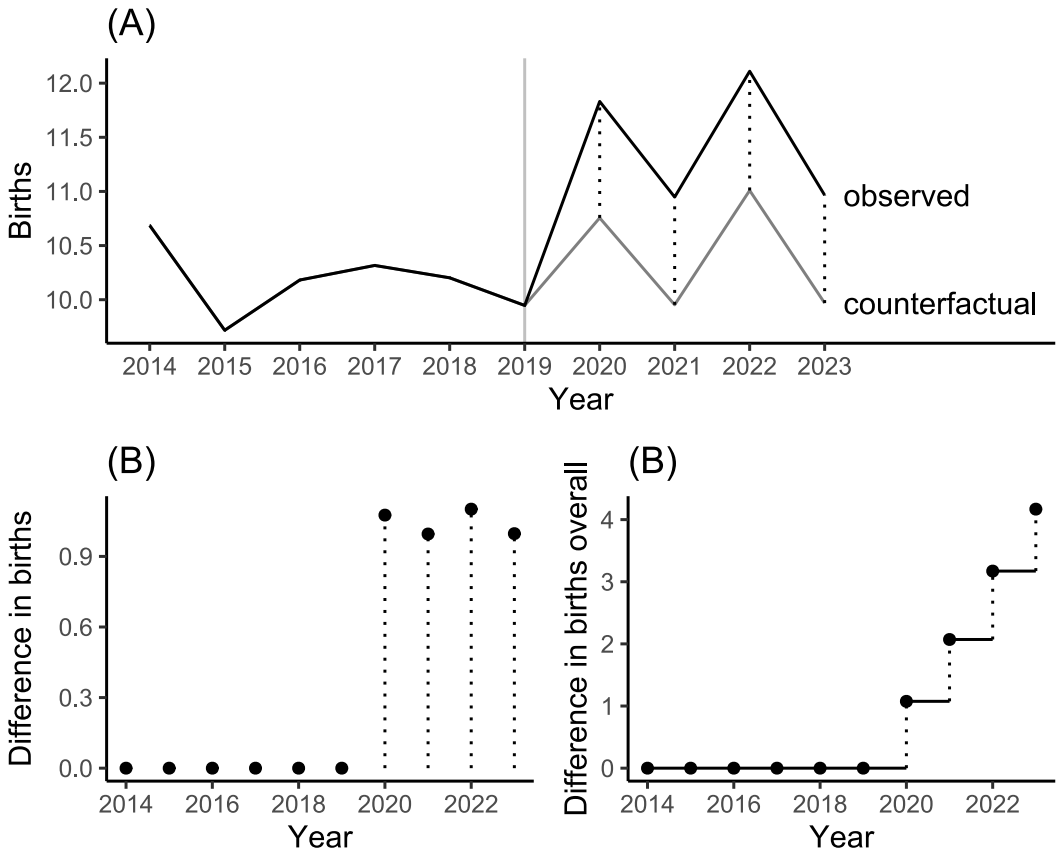


Figure 1: Taking sums of treatment effect for flow quantities such as births makes intuitive sense. (A) The vertical line indicates the time after which observations were impacted by the treatment. Dotted lines shows the treatment effect visually. (B) The treatment effect. (C) The cumulative treatment effect.

Brodersen et al. (2015) fails to describe what exactly their model is estimating in causal notation. A commonly used approach to causal inference is the Rubin causal model (RCM), the first ideas of which were introduced in Rubin (1974). The notation can be useful to specify exactly which causal effect is being estimated, even if the notation is not strictly necessary to perform the hypothesis testing as a practitioner.

Let $Y_t(1)$ be the potential cumulative (or average) outcome at time t, with treatment, and $Y_t(0)$ be the potential cumulative (or average) outcome at time t, without treatment. Then the treatment effect is $Y_t(1) - Y_t(0)$ at time t. Let D be an indicator random variable denoting treatment status, i.e., it is 1 if the unit is treated and 0 otherwise. The causal effect that is estimated by the synthetic control methods is the average treatment effect on the treated which is defined as,

$$\tau_{\text{att}} = \mathbb{E}[Y_t(1) - Y_t(0) \mid D = 1].$$

The potential outcome $Y_t(0)$ is a missing value if t is during the post-period which is why the approach by Brodersen is to retrospectively forecast the counterfactual to obtain an estimate of $\mathbb{E}[Y_t(0) \mid D = 1]$.

If the counterfactual expectation is correctly predicted then the estimate will be correct. In practice, untreated units or predictors must be used to estimate $E[Y_t(0) \mid D = 1]$ which requires the assumption that $Y_t(0) \perp D$ so that $E[Y_t(0) \mid D = 0] = E[Y_t(0) \mid D = 1] = E[Y_t(0)]$. If the assumption that $Y_t(1) \perp D$ also hold then it is possible to also estimate the average treatment effect,

$$\tau_{\text{ate}} = \mathbb{E}[Y_t(1) - Y_t(0)].$$

By using the causal notation above it is clear that the approach in Brodersen et al. (2015) gives a method for estimating different estimands depending on the assumptions permitted. Furthermore, it is worth noting that if the assumption of $Y_t(0) \perp D$ does not hold then even the ATT can not be determined by Brodersen et al. (2015). If there was an abundance of experimental units then it could be possible to correct for a violation of $Y_t(0) \perp D$ if $Y_t(0) \perp D, X$ for covariates $X$ that can predict propensity for treatment. This situation is documented by Dehejia and Wahba (2002) and has a rich literature which favours regression approaches. However, here as in Brodersen et al. (2015) I keep the stricter assumptions since there is typically just one or a few treated units in time series applications.

# The model

Brodersen et al. (2015) employs a version of synthetic control as well, but instead of using simple weights allows for a state space structure and estimates the treatment effect in a Bayesian paradigm. An advantage of this over previous methods is that if there is uncertainty about which series are most suitable as controls, then this uncertainty is captured in the credible intervals. The model can also accommodate temporal structure such as seasonality and ARMA models, which can be shown to have state space representations. Ignoring temporal dependence can lead to incorrect effect estimates and credible intervals.

With a flexible model overfitting can become an issue, but Brodersen et al. (2015) argues that including the set of controls using a Bayesian approach, e.g. spike and slab priors on coefficients, reduces the risk of overfitting since the model is not fully committed to one set of regressors. A downside of adopting this model is that some prior sensitivity is introduced particularly the level of noise in the level.

The methods in Brodersen et al. (2015) are implemented in the CausalImpact R package. The default model used in the R package is a model with a non-stationary trend, spike and slab prior on coefficients, and contemporaneous covariates. It is possible to allow the coefficients to vary according to a random walk as well, which is used for the simulation section in this report.

Any structural time-series model can be written as,

$$y_t = Z_t^\top \alpha_t + \epsilon_t$$
$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t,$$

where $y_t$ is the outcome and $\alpha_t$ is a state vector and the error terms $\epsilon_t$ and $\mu_t$ are Gaussian and independent of all other unknowns.

For the case of one predictor $x_t$, allowing for dynamic coefficient, and dynamic trend, this can be written in matrix form as, the observation equation,

$$y_t = \begin{bmatrix} 1 & x_t & 0 \end{bmatrix} \begin{bmatrix} \mu_{t+1} \\ \delta_{t+1} \\ \beta_{t+1} \end{bmatrix} + \epsilon_t,$$

and the state equation,

$$\begin{bmatrix} \mu_{t+1} \\ \delta_{t+1} \\ \beta_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_t \\ \beta_t \\ \delta_t \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \eta_{\mu,t} \\ \eta_{\delta,t} \\ \eta_{\beta,t} \end{bmatrix}.$$

Where, $\epsilon_t$ and $\eta_t = \begin{bmatrix} \eta_{\mu,t} & \eta_{\delta,t} & \eta_{\beta,t} \end{bmatrix}^\top$ are independent Gaussian and $\eta_t$ has a block diagonal variance. Brodersen et al. (2015) assumes a gamma prior for $1/\sigma_\mu^2$ and $1/\sigma_\delta^2$ the inverse of the variances for the level and trend components. The coefficients have a Bernoulli prior for the spike and conjugate normal-inverse Gamma distribution for the slab.

# Simulation

## The data generating process

In the simulation section of Brodersen the data is simulated from 1st of January 2013 to 30th of June 2014, with a causal effect at 1st of January 2014. Equivalently, data can be generated from a time series with $t \in T = \{1, 2, ..., 546\}$ and an intervention at t=366. In Brodersen et al. (2015) the simulations are intended to represent ad campaigns, although there appears to be no aspect of the simulation study that is specific to that application.

Brodersen uses the following data generating process,

$$
\begin{aligned}
y_t &= \beta_{t,1} x_{t,1} + \beta_{t,2} x_{t,2} + \mu_t + \epsilon_t \\
\beta_{t,i} &\sim \mathcal{N}\left(\beta_{t-1,i}, 0.01^2\right); \quad \beta_{0,i} = 0; \quad i \in \{1, 2\} \\
\mu_t &\sim \mathcal{N}\left(\mu_{t-1}, 0.1^2\right); \quad \mu_0 = 20 \\
\epsilon_t &\sim \mathcal{N}\left(0, 0.1^2\right).
\end{aligned}
$$

Based on visual inspection of Figure 3 (a) in Brodersen et al. (2015) it appears that $\mu_0 \approx 20$, not $\mu_0 = 0$ as stated in the text. A positive series is necessary for a multiplicative effect to have a meaningful interpretation for ad campaign outcomes such as clicks or sales, therefore I use $\mu_0 = 20$, which appears to be the intended value. Another point of ambiguity is that in Brodersen et al. (2015) the particular kind of sinusoid covariate that is used for $x_1$ and $x_2$ is not specified, only their period. This means that the data generating process is technically not specified in Brodersen et al. (2015) and so any reproduction can't expect the same results.

Brodersen et al. (2015) applies a multiplicative factor to imitate a causal effect so that the final observations are given by $y_t^* = y_t \mathbb{I}\{t < 366\} + y_t(1 + e)(1 - \mathbb{I}\{t < 366\})$, where $\mathbb{I}\{f(t)\}$ is the indicator function that evaluates to 1 when $t \in \{w : f(w)\}$ and 0 otherwise. It is easy to see that the simulation would be affected by the magnitude of $y_t$ since this imposed shock is relative. In reality many changes are gradual so the approach used by Brodersen et al. (2015) in simulations may be unrealistic. This intuition is visually supported by Figure 2, which shows how a simulated series would be shifted based on different effect sizes.

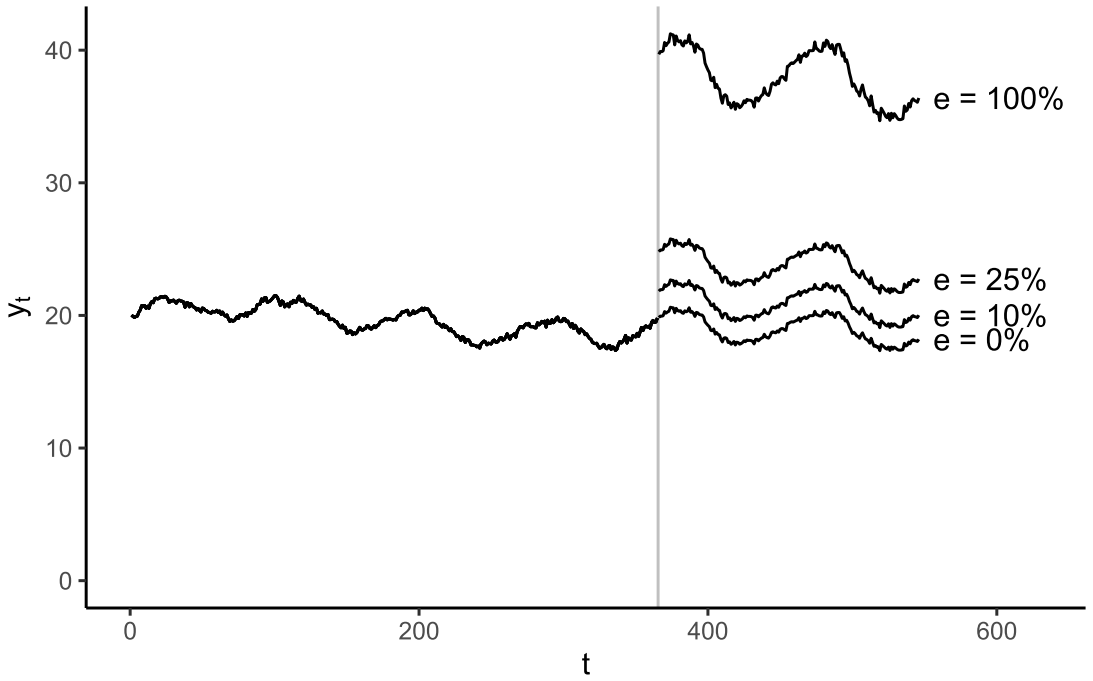


Figure 2: Examples of how Brodersen et al. (2015) applies a multiplicative factor in simulations to imitate the effect of a sustained intervention effect.

In Brodersen et al. (2015), the time series was simulated in 256 times for each effect size $e \in \{0, 0.001, 0.01, 0.1, 1\}$. Although, Figure 3 (b) actually shows effect sizes 25% and 50% which were not mentioned in the text. I opted to simulate $e \in \{0, 0.001, 0.01, 0.05, 0.1, 0.25, 1\}$. Additionally for $e = 0.1$ the time series was simulated 256 times for different campaign durations $\max T - 366 \in \{30, 60, 90, 120, 150, 180\}$ to study the coverage properties of changing the campaign duration.

## Model fitting

For illustrative purposes, a simulation realisation is visualised in Figure 3 where the true effect size was 0.1 and the ad campaign was 180 days. The same realisation was used to fit the CausalImpact model proposed by Brodersen which is visualised in Figure 4. Up until the campaign the model fits very well so it is not unreasonable to expect that the series $X_1$ and $X_2$ could predict $Y_t(0)$ once the intervention period has started. As is the case for regular forecasting, the counterfactual forecasts also get increasingly uncertain further past the intervention date.

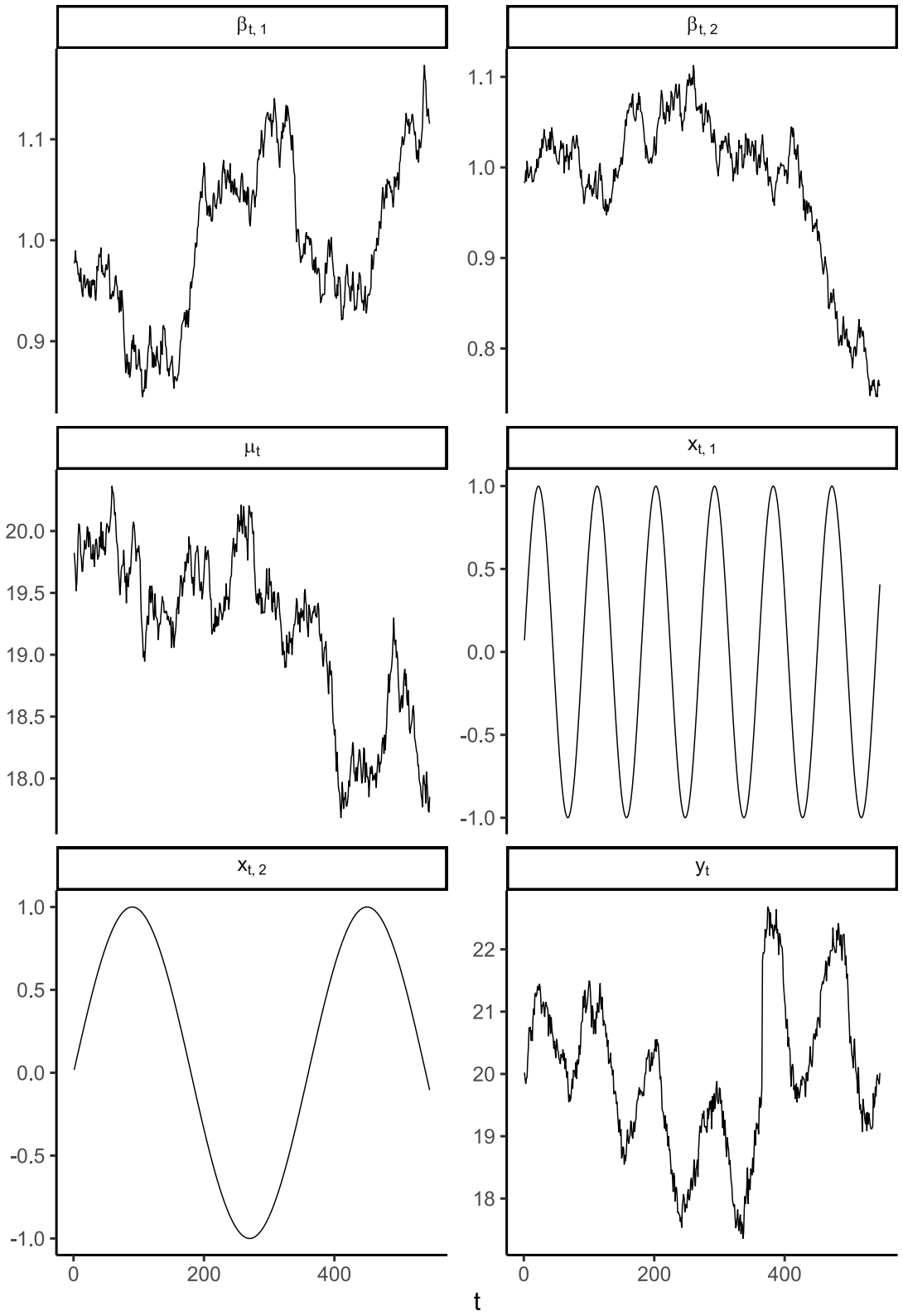


Figure 3: An example simulation realisation with effect size 0.1 and intervention, e.g. ad campaign, lasting 180 days.

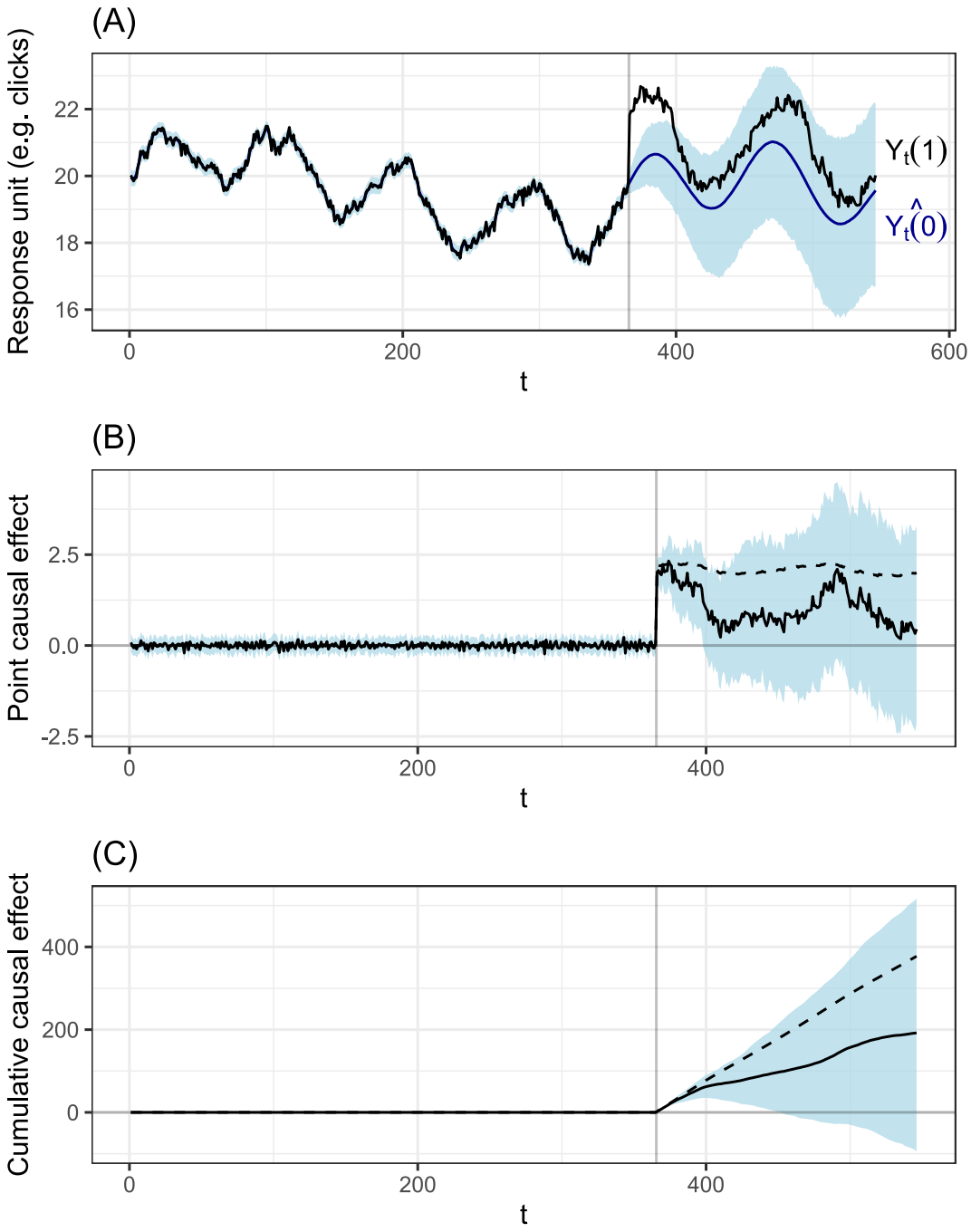


Figure 4: (A) Using the same simulation realisation as in Figure 3, the response is compared to a retrospectively forecasted counterfactual, $\hat{Y}_t(0)$, along with a 95% credibility interval which gets wider as the time since the start of the campaign increases. (B) The point estimate causal effect based on the observed response less the forecasted counterfactual. Dashed line shows ground truth (by construction). (C) The estimated cumulative causal effect defined as the sum of point causal effects.

## Results

One of the seemingly key plots in the Brodersen paper, Figure 3 (b), is reproduced in Figure 5 (A), using inputs ascertained from the simulation section of Brodersen et al. (2015). The plot shows the proportion of simulations in each of the effect size settings that rejected the null hypothesis of no causal impact in the positive direction. The null was rejected when the lower bound of the 95% credible interval for the average effect was above zero. As noted in Brodersen et al. (2015), this provides an estimate of sensitivity . The results are different to those of Brodersen, although this is not surprising given the lack of specific information around covariates used by Brodersen et al. (2015). Higher effect sizes are more likely to picked up, but the power will depend on the particular data generating process, in particular the amount of noise.

Figure 5 (B) confirms that for the most part the coverage probability is reasonable for different campaign durations. It also shows how the Bayesian approach leads to a shrinkage estimator, which on the whole provides a similar result qualitatively.

Figure 6 shows the impact on estimation accuracy of a change to standard deviation of dynamic coefficients 90 days into the treatment period. The situation considered is a 180 day treatment period and an effect size e=10%. The simulation was run 64 times for each setting. Stated mathematically, for some new standard deviation $c$, the coefficients evolve according to,

$$\beta_{t,i} \sim \begin{cases} \mathcal{N}\left(\beta_{t-1,i}, 0.01^2\right) & \text{if } t < 366 + 90 \\ \mathcal{N}\left(\beta_{t-1,i}, c^2\right) & \text{otherwise .} \end{cases}$$

The figure shows that for a small change to $c$ there is not a meaningful deterioration of estimation accuracy. This differs from Figure 4 (a) in Brodersen et al. (2015). The discrepancy can be attributed to this reproduction using unit sinusoids for the covariates while Brodersen et al. (2015) appear to be using something else, otherwise the estimation error would periodically converge with the no structural change case when covariates are zero. However, instead of guessing what Brodersen et al. (2015) may have meant, I use the most reasonable interpretation of the text. The structural change with a new standard deviation of 0.3 does however show a similar increase in relative absolute error to that in Brodersen et al. (2015), but clearly with much more influence from the sinusoidal nature of the covariate.

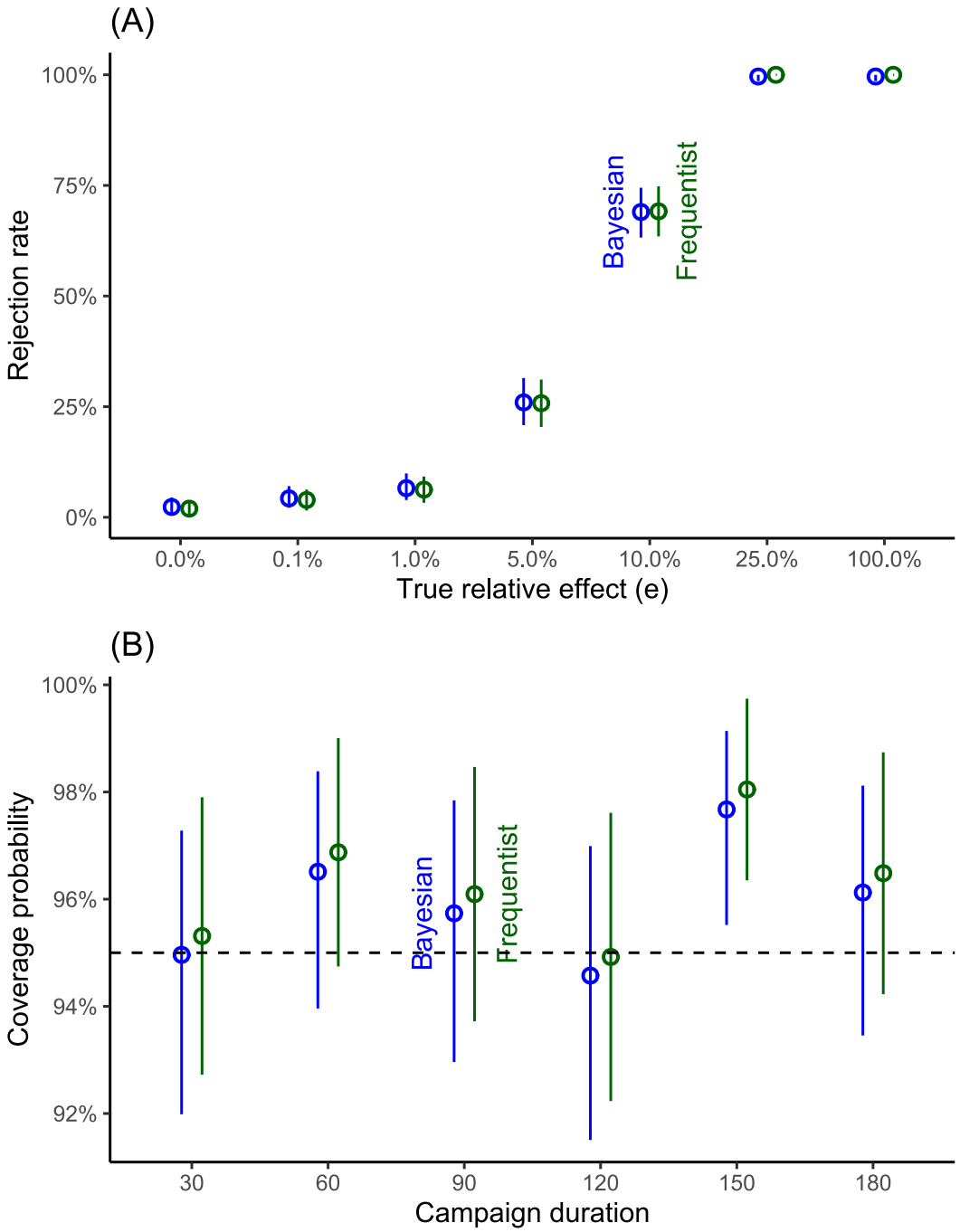


Figure 5: (A) The empirical prevalence of credibility intervals that exclude zero in the positive direction, i.e., a power curve, for the simulation setting with a 180 day intervention period. (B) The empirical coverage of causal effect credibility interval for simulations of different campaign durations. Frequentist 95% confidence interval estimated by $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. Bayesian 95% credibility interval determined with a uniform(0,1) prior. The point estimate in the frequentist case is the MLE $\hat{p}$ and in the Bayesian case the posterior expectation.

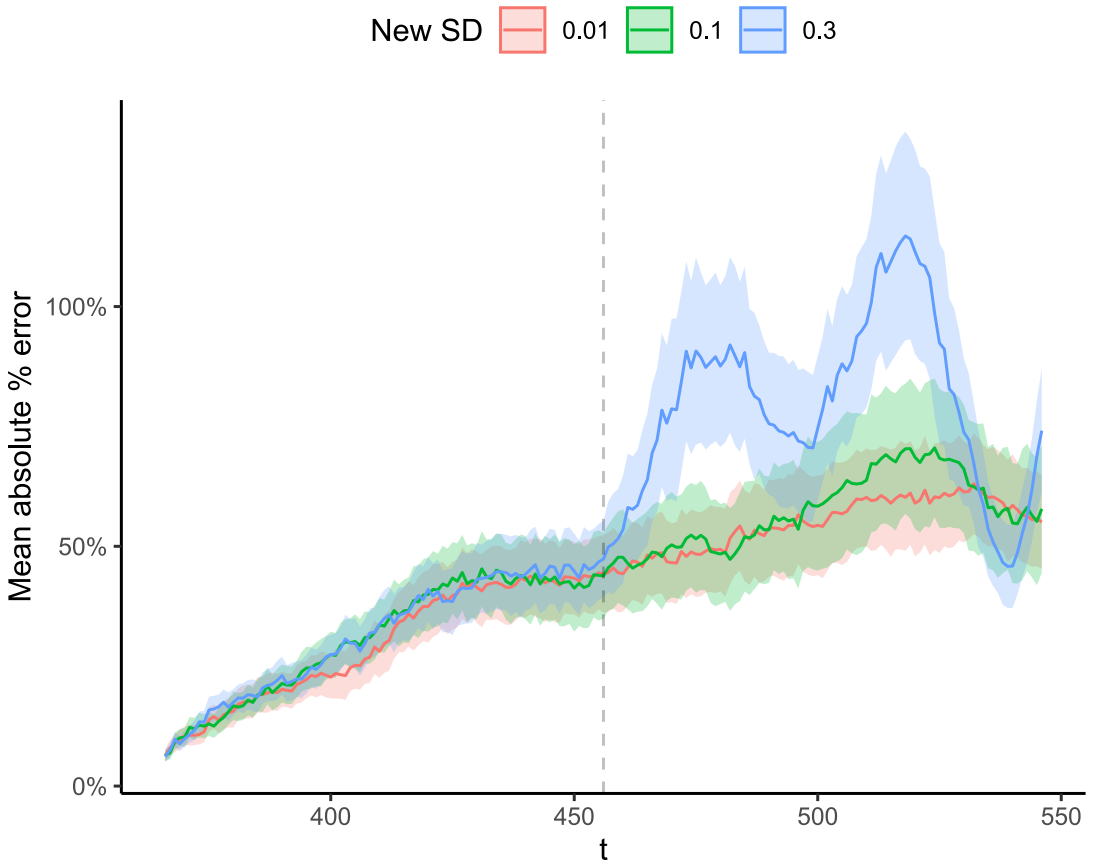


Figure 6: The absolute percentage error for pointwise treatment effect compared to the constructed true effect for a structural change in the coefficient dynamics. Vertical line indicates structural change. Shaded 95% confidence intervals estimated by $\hat{\mu} \pm 1.96\hat{\sigma}/\sqrt{64}$ at each time point, where $\hat{\mu}$ is the sample mean absolute % error, and $\hat{\sigma}$ is the sample standard deviation of the mean absolute % error. Note that a new SD of 0.01 corresponds to no structural change.

# Conclusion

Brodersen et al. (2015) provides a reasonable way for constructing synthetic controls in a largely automated and easy to implement way. However, the paper falls short in terms of describing causality in clear notation and the data generating process in simulation section falls short of best practices in reproducible research. In this report I have made the causal notation and language more explicit and repeated simulations in a way that is reproducible.

# References

Abadie, Alberto, Alexis Diamond, and And Jens Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Cali-

fornia's Tobacco Control Program." Article. *Journal of the American Statistical Association* 105 (490): 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." Article. *American Economic Review* 93 (1): 113–32. https://doi.org/10.1257/000282803321455188.

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2022. "Quarto." https://doi.org/10.5281/zenodo.5960048.

Brodersen, Kay H, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. 2015. "Inferring Causal Impact Using Bayesian Structural Time-Series Models." *The Annals of Applied Statistics*, 247–74.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." Review. *Review of Economics and Statistics* 84 (1): 151–61. https://doi.org/10.1162/003465302317331982.

Hester, Jim, and Jennifer Bryan. 2022. "Glue: Interpreted String Literals." https://CRAN.R-project.org/package=glue.

Meschiari, Stefano. 2022. "Latex2exp: Use LaTeX Expressions in Plots." https://CRAN.R-project.org/package=latex2exp.

Pedersen, Thomas Lin. 2023. "Patchwork: The Composer of Plots." https://CRAN.R-project.org/package=patchwork.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." Article. *Journal of Educational Psychology* 66 (5): 688–701. https://doi.org/10.1037/h0037350.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software*. https://doi.org/10.21105/joss.01686.