*The Annals of Applied Statistics*
2015, Vol. 9, No. 1, 247–274
DOI: 10.1214/14-AOAS788


# INFERRING CAUSAL IMPACT USING BAYESIAN STRUCTURAL TIME-SERIES MODELS

BY KAY H. BRODERSEN, FABIAN GALLUSSER, JIM KOEHLER, NICOLAS REMY AND STEVEN L. SCOTT

*Google, Inc.*

An important problem in econometrics and marketing is to infer the causal impact that a designed market intervention has exerted on an outcome metric over time. This paper proposes to infer causal impact on the basis of a diffusion-regression state-space model that predicts the counterfactual market response in a synthetic control that would have occurred had no intervention taken place. In contrast to classical difference-in-differences schemes, state-space models make it possible to (i) infer the temporal evolution of attributable impact, (ii) incorporate empirical priors on the parameters in a fully Bayesian treatment, and (iii) flexibly accommodate multiple sources of variation, including local trends, seasonality and the time-varying influence of contemporaneous covariates. Using a Markov chain Monte Carlo algorithm for posterior inference, we illustrate the statistical properties of our approach on simulated data. We then demonstrate its practical utility by estimating the causal effect of an online advertising campaign on search-related site visits. We discuss the strengths and limitations of state-space models in enabling causal attribution in those settings where a randomised experiment is unavailable. The CausalImpact R package provides an implementation of our approach.

**1. Introduction.** This article proposes an approach to inferring the causal impact of a market intervention, such as a new product launch or the onset of an advertising campaign. Our method generalises the widely used difference-in-differences approach to the time-series setting by explicitly modelling the counterfactual of a time series observed both before and after the intervention. It improves on existing methods in two respects: it provides a fully Bayesian time-series estimate for the effect; and it uses model averaging to construct the most appropriate synthetic control for modelling the counterfactual. The `CausalImpact` R package provides an implementation of our approach (http://google.github.io/CausalImpact/).

Inferring the impact of market interventions is an important and timely problem. Partly because of recent interest in big data, many firms have begun to understand that a competitive advantage can be had by systematically using impact measures to inform strategic decision making. An example is the use of "A/B experiments"

Received November 2013; revised September 2014.
*Key words and phrases.* Causal inference, counterfactual, synthetic control, observational, difference in differences, econometrics, advertising, market research.

to identify the most effective market treatments for the purpose of allocating resources [Danaher and Rust (1996), Leeflang et al. (2009), Seggie, Cavusgil and Phelan (2007), Stewart (2009)].

Here, we focus on measuring the impact of a discrete marketing event, such as the release of a new product, the introduction of a new feature, or the beginning or end of an advertising campaign, with the aim of measuring the event's impact on a response metric of interest (e.g., sales). The causal impact of a treatment is the difference between the observed value of the response and the (unobserved) value that would have been obtained under the alternative treatment, that is, the effect of treatment on the treated [Antonakis et al. (2010), Claveau (2012), Cox and Wermuth (2001), Heckman and Vytlacil (2007), Hitchcock (2004), Hoover (2012), Kleinberg and Hripcsak (2011), Morgan and Winship (2007), Rubin (1974, 2008)]. In the present setting the response variable is a time series, so the causal effect of interest is the difference between the observed series and the series that would have been observed had the intervention not taken place.

A powerful approach to constructing the counterfactual is based on the idea of combining a set of candidate predictor variables into a single "synthetic control" [Abadie, Diamond and Hainmueller (2010), Abadie and Gardeazabal (2003)]. Broadly speaking, there are three sources of information available for constructing an adequate synthetic control. The first is the time-series behaviour of the response itself, prior to the intervention. The second is the behaviour of other time series that were predictive of the target series prior to the intervention. Such control series can be based, for example, on the same product in a different region that did not receive the intervention or on a metric that reflects activity in the industry as a whole. In practice, there are often many such series available, and the challenge is to pick the relevant subset to use as contemporaneous controls. This selection is done on the *pre-treatment* portion of potential controls; but their value for predicting the counterfactual lies in their *post-treatment* behaviour. As long as the control series received no intervention themselves, it is often reasonable to assume the relationship between the treatment and the control series that existed prior to the intervention to continue afterwards. Thus, a plausible estimate of the counterfactual time series can be computed up to the point in time where the relationship between treatment and controls can no longer be assumed to be stationary, for example, because one of the controls received treatment itself. In a Bayesian framework, a third source of information for inferring the counterfactual is the available prior knowledge about the model parameters, as elicited, for example, by previous studies.

We combine the three preceding sources of information using a state-space time-series model, where one component of state is a linear regression on the contemporaneous predictors. The framework of our model allows us to choose from among a large set of potential controls by placing a spike-and-slab prior on the set of regression coefficients and by allowing the model to average over the set of controls [George and McCulloch (1997)]. We then compute the posterior distribution of the counterfactual time series given the value of the target series

Compute the posterior distribution of the counterfactual time series given
(1) the value of the target series in the pre-intervention period, and
(2) the values of the controls in the post-intervention period.

in the pre-intervention period, along with the values of the controls in the post-intervention period. Subtracting the predicted from the observed response during the post-intervention period gives a semiparametric Bayesian posterior distribution for the causal effect (Figure 1).

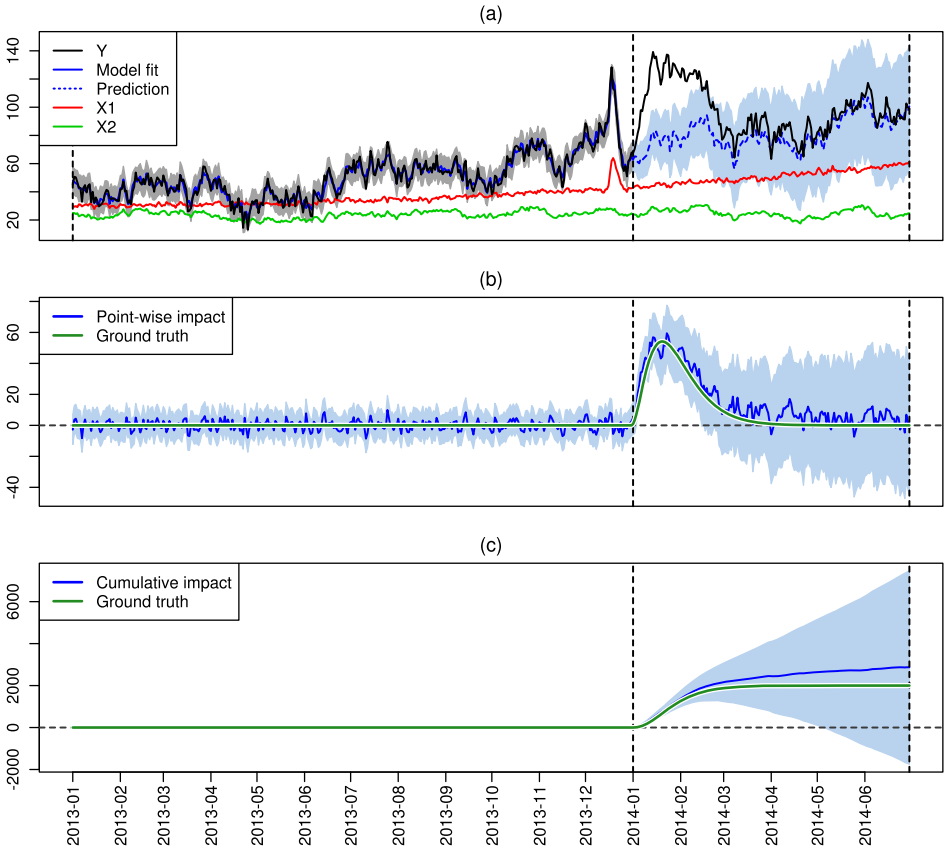


FIG. 1. *Inferring causal impact through counterfactual predictions.* (a) *Simulated trajectory of a treated market* ($Y$) *with an intervention beginning in January* 2014. *Two other markets* ($X_1$, $X_2$) *were not subject to the intervention and allow us to construct a synthetic control* [*cf. Abadie, Diamond and Hainmueller* (2010), *Abadie and Gardeazabal* (2003)]. *Inverting the state-space model described in the main text yields a prediction of what would have happened in* $Y$ *had the intervention not taken place* (*posterior predictive expectation of the counterfactual with pointwise* 95% *posterior probability intervals*). (b) *The difference between observed data and counterfactual predictions is the inferred causal impact of the intervention. Here, predictions accurately reflect the true* (*Gamma-shaped*) *impact. A key characteristic of the inferred impact series is the progressive widening of the posterior intervals* (*shaded area*). *This effect emerges naturally from the model structure and agrees with the intuition that predictions should become increasingly uncertain as we look further and further into the* (*retrospective*) *future.* (c) *Another way of visualizing posterior inferences is by means of a cumulative impact plot. It shows, for each day, the summed effect up to that day. Here, the* 95% *credible interval of the cumulative impact crosses the zero-line about five months after the intervention, at which point we would no longer declare a significant overall effect.*

*Related work.* As with other domains, causal inference in marketing requires subtlety. Marketing data are often observational and rarely follow the ideal of a randomised design. They typically exhibit a low signal-to-noise ratio. They are subject to multiple seasonal variations, and they are often confounded by the effects of unobserved variables and their interactions [for recent examples, see Chan et al. (2010), Leeflang et al. (2009), Lewis, Rao and Reiley (2011), Lewis and Reiley (2011), Seggie, Cavusgil and Phelan (2007), Stewart (2009), Takada and Bass (1998), Vaver and Koehler (2011, 2012)].

Rigorous causal inferences can be obtained through randomised experiments, which are often implemented in the form of geo experiments [Vaver and Koehler (2011, 2012)]. Many market interventions, however, fail to satisfy the requirements of such approaches. For instance, advertising campaigns are frequently launched across multiple channels, online and offline, which precludes measurement of individual exposure. Campaigns are often targeted at an entire country, and one country only, which prohibits the use of geographic controls within that country. Likewise, a campaign might be launched in several countries but at different points in time. Thus, while a large control group may be available, the treatment group often consists of no more than one region or a few regions with considerable heterogeneity among them.

A standard approach to causal inference in such settings is based on a linear model of the observed outcomes in the treatment and control group before and after the intervention. One can then estimate the difference between (i) the pre-post difference in the treatment group and (ii) the pre-post difference in the control group. The assumption underlying such *difference-in-differences* (DD) designs is that the level of the control group provides an adequate proxy for the level that would have been observed in the treatment group in the absence of treatment [see Abadie (2005), Angrist and Krueger (1999), Angrist and Pischke (2008), Antonakis et al. (2010), Ashenfelter and Card (1985), Athey and Imbens (2002), Campbell, Stanley and Gage (1963), Card and Krueger (1993), Donald and Lang (2007), Lester (1946), Meyer (1995), Robinson, McNulty and Krasno (2009), Shadish, Cook and Campbell (2002)].

DD designs have been limited in three ways. First, DD is traditionally based on a static regression model that assumes i.i.d. data despite the fact that the design has a temporal component. When fit to serially correlated data, static models yield overoptimistic inferences with too narrow uncertainty intervals [see also Bertrand, Duflo and Mullainathan (2002), Hansen (2007a, 2007b), Solon (1984)]. Second, most DD analyses only consider two time points: before and after the intervention. In practice, the manner in which an effect evolves over time, especially its onset and decay structure, is often a key question.

Third, when DD analyses *are* based on time series, previous studies have imposed restrictions on the way in which a synthetic control is constructed from a set of predictor variables, which is something we wish to avoid. For example, one strategy [Abadie, Diamond and Hainmueller (2010), Abadie and Gardeazabal

(2003)] has been to choose a convex combination $(w_1, \ldots, w_J)$, $w_j \geq 0, \sum w_j = 1$ of $J$ predictor time series in such a way that a vector of pre-treatment variables (not time series) $X_1$ characterising the treated unit before the intervention is matched most closely by the combination of pre-treatment variables $X_0$ of the control units w.r.t. a vector of importance weights $(v_1, \ldots, v_J)$. These weights are themselves determined in such a way that the combination of pre-treatment outcome time series of the control units most closely matches the pre-treatment outcome time series of the treated unit. Such a scheme relies on the availability of interpretable characteristics (e.g., growth predictors), and it precludes nonconvex combinations of controls when constructing the weight vector $W$. We prefer to select a combination of control series without reference to external characteristics and purely in terms of how well they explain the pre-treatment outcome time series of the treated unit (while automatically balancing goodness of fit and model complexity through the use of regularizing priors). Another idea [Belloni et al. (2013)] has been to use classical variable-selection methods (such as the Lasso) to find a sparse set of predictors. This approach, however, ignores posterior uncertainty about both which predictors to use and their coefficients.

here is how synthetic control works

why not including external characteristics?

should we predict using splitting samples to mitigate regularization bias?

The limitations of DD schemes can be addressed by using state-space models, coupled with highly flexible regression components, to explain the temporal evolution of an observed outcome. State-space models distinguish between a state equation that describes the transition of a set of latent variables from one time point to the next and an observation equation that specifies how a given system state translates into measurements. This distinction makes them extremely flexible and powerful [see Leeflang et al. (2009) for a discussion in the context of marketing research].

distinguished features of state-space model [?]

The approach described in this paper inherits three main characteristics from the state-space paradigm. First, it allows us to flexibly accommodate different kinds of assumptions about the latent state and emission processes underlying the observed data, including local trends and seasonality. Second, we use a fully Bayesian approach to inferring the temporal evolution of counterfactual activity and incremental impact. One advantage of this is the flexibility with which posterior inferences can be summarised. Third, we use a regression component that precludes a rigid commitment to a particular set of controls by integrating out our posterior uncertainty about the influence of each predictor as well as our uncertainty about which predictors to include in the first place, which avoids overfitting.

uncertainty about influence (i.e., coefficient) and which predictors to include (i.e., the set of the controls)

The remainder of this paper is organised as follows. Section 2 describes the proposed model, its design variations, the choice of diffuse empirical priors on hyperparameters, and a stochastic algorithm for posterior inference based on Markov chain Monte Carlo (MCMC). Section 3 demonstrates important features of the model using simulated data, followed by an application in Section 4 to an advertising campaign run by one of Google's advertisers. Section 5 puts our approach into context and discusses its scope of application.

**2. Bayesian structural time-series models.** Structural time-series models are state-space models for time-series data. They can be defined in terms of a pair of equations

$$y_t = Z_t^{\mathrm{T}} \alpha_t + \varepsilon_t, \tag{2.1}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \tag{2.2}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ and $\eta_t \sim \mathcal{N}(0, Q_t)$ are independent of all other unknowns. Equation (2.1) is the *observation equation*; it links the observed data $y_t$ to a latent $d$-dimensional state vector $\alpha_t$. Equation (2.2) is the *state equation*; it governs the evolution of the state vector $\alpha_t$ through time. In the present paper, $y_t$ is a scalar observation, $Z_t$ is a $d$-dimensional output vector, $T_t$ is a $d \times d$ transition matrix, $R_t$ is a $d \times q$ control matrix, $\varepsilon_t$ is a scalar observation error with noise variance $\sigma_t$, and $\eta_t$ is a $q$-dimensional system error with a $q \times q$ state-diffusion matrix $Q_t$, where $q \leq d$. Writing the error structure of equation (2.2) as $R_t \eta_t$ allows us to incorporate state components of less than full rank; a model for seasonality will be the most important example.

Structural time-series models are useful in practice because they are flexible and modular. They are flexible in the sense that a very large class of models, including all ARIMA models, can be written in the state-space form given by (2.1) and (2.2). They are modular in the sense that the latent state as well as the associated model matrices $Z_t$, $T_t$, $R_t$, and $Q_t$ can be assembled from a library of component sub-models to capture important features of the data. There are several widely used state-component models for capturing the trend, seasonality or effects of holidays.

A common approach is to assume the errors of different state-component models to be independent (i.e., $Q_t$ is block-diagonal). The vector $\alpha_t$ can then be formed by concatenating the individual state components, while $T_t$ and $R_t$ become block-diagonal matrices.

The most important state component for the applications considered in this paper is a regression component that allows us to obtain counterfactual predictions by constructing a synthetic control based on a combination of markets that were not treated. Observed responses from such markets are important because they allow us to explain variance components in the treated market that are not readily captured by more generic seasonal sub-models.

This approach assumes that covariates are unaffected by the effects of treatment. For example, an advertising campaign run in the United States might spill over to Canada or the United Kingdom. When assuming the absence of spill-over effects, the use of such indirectly affected markets as controls would lead to pessimistic inferences, that is, the effect of the campaign would be underestimated [cf. Meyer (1995)].

regression component includes observed responses from control markets, explaining variance not captured by generic structure modules (e.g., seasonal factors)

This is a simplified synthetic control: y(t)_treat ~ \sum_i{ w_i \times y(t)_{ctrl_i}}; why we prefer to a comb of control series without external characteristics? [also see my question on page 251]

2.1. *Components of state.*

*Local linear trend.* The first component of our model is a local linear trend, defined by the pair of equations

$$\begin{aligned} \mu_{t+1} &= \mu_t + \delta_t + \eta_{\mu,t}, \\ \delta_{t+1} &= \delta_t + \eta_{\delta,t}, \end{aligned} \tag{2.3}$$

where $\eta_{\mu,t} \sim \mathcal{N}(0, \sigma_\mu^2)$ and $\eta_{\delta,t} \sim \mathcal{N}(0, \sigma_\delta^2)$. The $\mu_t$ component is the value of the trend at time $t$. The $\delta_t$ component is the expected increase in $\mu$ between times $t$ and $t+1$, so it can be thought of as the *slope* at time $t$.

The local linear trend model is a popular choice for modelling trends because it quickly adapts to local variation, which is desirable when making short-term predictions. This degree of flexibility may not be desired when making longer-term predictions, as such predictions often come with implausibly wide uncertainty intervals.

There is a generalisation of the local linear trend model where the slope exhibits stationarity instead of obeying a random walk. This model can be written as

$$\begin{aligned} \mu_{t+1} &= \mu_t + \delta_t + \eta_{\mu,t}, \\ \delta_{t+1} &= D + \rho(\delta_t - D) + \eta_{\delta,t}, \end{aligned} \tag{2.4}$$

where the two components of $\eta$ are independent. In this model, the slope of the time trend exhibits AR(1) variation around a long-term slope of $D$. The parameter $|\rho| < 1$ represents the learning rate at which the local trend is updated. Thus, the model balances short-term information with information from the distant past.

*Seasonality.* There are several commonly used state-component models to capture seasonality. The most frequently used model in the time domain is

$$\gamma_{t+1} = -\sum_{s=0}^{S-2} \gamma_{t-s} + \eta_{\gamma,t}, \tag{2.5}$$

seasonality provides a cyclic pattern rather than an increase or a decrease in level; this definition is more restrictive than the generic definition of a set of dummies with one held out for rank condition.

where $S$ represents the number of seasons and $\gamma_t$ denotes their joint contribution to the observed response $y_t$. The state in this model consists of the $S-1$ most recent seasonal effects, but the error term is a scalar, so the evolution equation for this state model is less than full rank. The mean of $\gamma_{t+1}$ is such that the total seasonal effect is zero when summed over $S$ seasons. For example, if we set $S = 4$ to capture four seasons per year, the mean of the *winter* coefficient will be $-1 \times (\textit{spring} + \textit{summer} + \textit{autumn})$. The part of the transition matrix $T_t$ representing the seasonal model is an $S-1 \times S-1$ matrix with $-1$'s along the top row, 1's along the subdiagonal and 0's elsewhere.

[[-1,-1,-1],
[ 1, 0, 0],
[ 0, 1, 0]]

The preceding seasonal model can be generalised to allow for multiple seasonal components with different periods. When modelling daily data, for example, we

might wish to allow for an $S = 7$ day-of-week effect, as well as an $S = 52$ weekly annual cycle. The latter can be handled by setting $T_t = I_{S-1}$, with zero variance on the error term, when $t$ is not the start of a new week, and setting $T_t$ to the usual seasonal transition matrix, with nonzero error variance, when $t$ is the start of a new week.

*Contemporaneous covariates with static coefficients.* Control time series that received no treatment are critical to our method for obtaining accurate counterfactual predictions since they account for variance components that are shared by the series, including, in particular, the effects of other unobserved causes otherwise unaccounted for by the model. A natural way of including control series in the model is through a linear regression. Its coefficients can be static or time-varying.

A *static* regression can be written in state-space form by setting $Z_t = \beta^{\mathrm{T}}\mathbf{x}_t$ and $\alpha_t = 1$. One advantage of working in a fully Bayesian treatment is that we do not need to commit to a fixed set of covariates. The spike-and-slab prior described in Section 2.2 allows us to integrate out our posterior uncertainty about which covariates to include and how strongly they should influence our predictions, which avoids overfitting.

All covariates are assumed to be contemporaneous; the present model does not infer on a potential lag between treated and untreated time series. A known lag, however, can be easily incorporated by shifting the corresponding regressor in time. Covariates are assumed to be **contemporaneous**.

*Contemporaneous covariates with dynamic coefficients.* An alternative to the above is a regression component with *dynamic* regression coefficients to account for time-varying relationships [e.g., Banerjee, Kauffman and Wang (2007), West and Harrison (1997)]. Given covariates $j = 1, \ldots, J$, this introduces the dynamic regression component

How can we identify \beta_{j,t}?

feel like Bayesian statisticians do not care about identification; instead, they are satisfied if one of multiple possible solutions can be found?

$$\mathbf{x}_t^{\mathrm{T}}\beta_t = \sum_{j=1}^{J} x_{j,t}\beta_{j,t}, \qquad \beta_{j,t+1} = \beta_{j,t} + \eta_{\beta,j,t}, \tag{2.6}$$

**[YH] The motion function of beta in (2.6) is what we need when we allow the similarity to change across time periods!**

where $\eta_{\beta,j,t} \sim \mathcal{N}(0, \sigma^2_{\beta_j})$. Here, $\beta_{j,t}$ is the coefficient for the $j$th control series and $\sigma_{\beta_j}$ is the standard deviation of its associated random walk. We can write the dynamic regression component in state-space form by setting $Z_t = \mathbf{x}_t$ and $\alpha_t = \beta_t$ and by setting the corresponding part of the transition matrix to $T_t = I_{J\times J}$, with $Q_t = \mathrm{diag}(\sigma^2_{\beta_j})$.

*Assembling the state-space model.* Structural time-series models allow us to examine the time series at hand and flexibly choose appropriate components for trend, seasonality, and either static or dynamic regression for the controls. The

presence or absence of seasonality, for example, will usually be obvious by inspection. A more subtle question is whether to choose static or dynamic regression coefficients.

When the relationship between controls and treated unit has been stable in the past, static coefficients are an attractive option. This is because a spike-and-slab prior can be implemented efficiently within a forward-filtering, backward-sampling framework. This makes it possible to quickly identify a sparse set of covariates even from tens or hundreds of potential variables [Scott and Varian (2014)]. Local variability in the treated time series is captured by the dynamic local level or dynamic linear trend component. Covariate stability is typically high when the available covariates are close in nature to the treated metric. The empirical analyses presented in this paper, for example, will be based on a static regression component (Section 4). This choice provides a reasonable compromise between capturing local behaviour and accounting for regression effects.

An alternative would be to use dynamic regression coefficients, as we do, for instance, in our analyses of simulated data (Section 3). Dynamic coefficients are useful when the linear relationship between treated metrics and controls is believed to change over time. There are a number of ways of reducing the computational burden of dealing with a potentially large number of dynamic coefficients. One option is to resort to dynamic latent factors, where one uses $\mathbf{x}_t = B\mathbf{u}_t + \boldsymbol{\nu}_t$ with $\dim(\mathbf{u}_t) \ll J$ and uses $\mathbf{u}_t$ instead of $\mathbf{x}_t$ as part of $Z_t$ in (2.1), coupled with an AR-type model for $\mathbf{u}_t$ itself. Another option is latent thresholding regression, where one uses a dynamic version of the spike-and-slab prior as in Nakajima and West (2013). two alternatives: (1) specify AR-type model for u_t, or (2) use dynamic version of spike-and-slab priors

The state-component models are assembled independently, with each component providing an additive contribution to $y_t$. Figure 2 illustrates this process assuming a local linear trend paired with a static regression component.

2.2. *Prior distributions and prior elicitation.* Let $\theta$ generically denote the set of all model parameters and let $\alpha = (\alpha_1, \ldots, \alpha_m)$ denote the full state sequence. We adopt a Bayesian approach to inference by specifying a prior distribution $p(\theta)$ on the model parameters as well as a distribution $p(\alpha_0|\theta)$ on the initial state values. We may then sample from $p(\alpha, \theta|\mathbf{y})$ using MCMC.

Most of the models in Section 2.1 depend solely on a small set of variance parameters that govern the diffusion of the individual state components. A typical prior distribution for such a variance is

$$\frac{1}{\sigma^2} \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{s}{2}\right), \tag{2.7}$$

where $\mathcal{G}(a, b)$ is the Gamma distribution with expectation $a/b$. The prior parameters can be interpreted as a prior sum of squares $s$, so that $s/\nu$ is a prior estimate of $\sigma^2$, and $\nu$ is the weight, in units of prior sample size, assigned to the prior estimate.

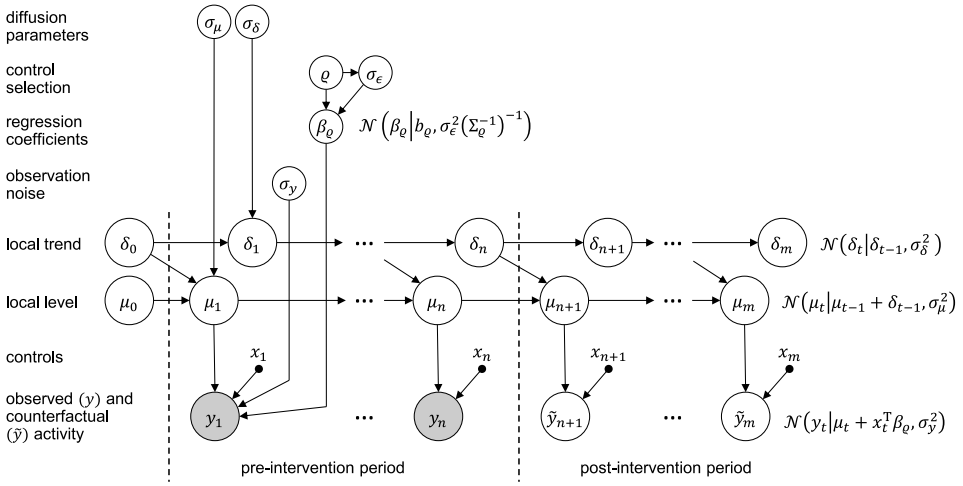


FIG. 2. *Graphical model for the static-regression variant of the proposed state-space model. Observed market activity* $\mathbf{y}_{1:n} = (y_1, \ldots, y_n)$ *is modelled as the result of a latent state plus Gaussian observation noise with error standard deviation* $\sigma_y$. *The state* $\alpha_t$ *includes a local level* $\mu_t$, *a local linear trend* $\delta_t$, *and a set of contemporaneous covariates* $\mathbf{x}_t$, *scaled by regression coefficients* $\beta_\varrho$. *State components are assumed to evolve according to independent Gaussian random walks with fixed standard deviations* $\sigma_\mu$ *and* $\sigma_\delta$ (*conditional-dependence arrows shown for the first time point only*). *The model includes empirical priors on these parameters and the initial states. In an alternative formulation, the regression coefficients* $\beta$ *are themselves subject to random-walk diffusion* (*see main text*). *Of principal interest is the posterior predictive density over the unobserved counterfactual responses* $\tilde{y}_{n+1}, \ldots, \tilde{y}_m$. *Subtracting these from the actual observed data* $y_{n+1}, \ldots, y_m$ *yields a probability density over the temporal evolution of causal impact.*

empirical priors?

We often have a weak default prior belief that the incremental errors in the state process are small, which we can formalise by choosing small values of $\nu$ (e.g., 1) and small values of $s/\nu$. The notion of "small" means different things in different models; for the seasonal and local linear trend models our default priors are $1/\sigma^2 \sim \mathcal{G}(10^{-2}, 10^{-2} s_y^2)$, where $s_y^2 = \sum_t (y_t - \bar{y})^2/(n-1)$ is the sample variance of the target series. Scaling by the sample variance is a minor violation of the Bayesian paradigm, but it is an effective means of choosing a reasonable scale for the prior. It is similar to the popular technique of scaling the data prior to analysis, but we prefer to do the scaling in the prior so we can model the data on its original scale. scale priors rather than observed data so that we can keep observed data as is.

violate?

When faced with many potential controls, we prefer letting the model choose an appropriate set. This can be achieved by placing a spike-and-slab prior over coefficients [George and McCulloch (1993, 1997), Polson and Scott (2011), Scott and Varian (2014)]. A spike-and-slab prior combines point mass at zero (the "spike"), for an unknown subset of zero coefficients, with a weakly informative distribution on the complementary set of nonzero coefficients (the "slab"). Contrary to what its name might suggest, the "slab" is usually not completely flat, but rather a Gaussian

with a large variance. Let $\varrho = (\varrho_1, \ldots, \varrho_J)$, where $\varrho_j = 1$ if $\beta_j \neq 0$ and $\varrho_j = 0$ otherwise. Let $\beta_\varrho$ denote the nonzero elements of the vector $\beta$ and let $\Sigma_\varrho^{-1}$ denote the rows and columns of $\Sigma^{-1}$ corresponding to nonzero entries in $\varrho$. We can then factorise the spike-and-slab prior as

$$p(\varrho, \beta, 1/\sigma_\varepsilon^2) = p(\varrho)p(\sigma_\varepsilon^2|\varrho)p(\beta_\varrho|\varrho, \sigma_\varepsilon^2). \tag{2.8}$$

The spike portion of (2.8) can be an arbitrary distribution over $\{0, 1\}^J$ in principle; the most common choice in practice is a product of independent Bernoulli distributions,

$$p(\varrho) = \prod_{j=1}^{J} \pi_j^{\varrho_j}(1 - \pi_j)^{1-\varrho_j}, \tag{2.9}$$

how to determine \rho_j?

where $\pi_j$ is the prior probability of regressor $j$ being included in the model.

Values for $\pi_j$ can be elicited by asking about the *expected model size* $M$, and then setting all $\pi_j = M/J$. An alternative is to use a more specific set of values $\pi_j$. In particular, one might choose to set certain $\pi_j$ to either 1 or 0 to force the corresponding variables into or out of the model. Generally, framing the prior in terms of expected model size has the advantage that the model can adapt to growing numbers of predictor variables without having to switch to a hierarchical prior [Scott and Berger (2010)].

For the "slab" portion of the prior we use a conjugate normal-inverse Gamma distribution,

$$\beta_\varrho|\sigma_\varepsilon^2 \sim \mathcal{N}(\mathbf{b}_\varrho, \sigma_\varepsilon^2(\Sigma_\varrho^{-1})^{-1}), \tag{2.10}$$

$$\frac{1}{\sigma_\varepsilon^2} \sim \mathcal{G}\left(\frac{\nu_\varepsilon}{2}, \frac{s_\varepsilon}{2}\right). \tag{2.11}$$

The vector $\mathbf{b}$ in equation (2.10) encodes our prior expectation about the value of each element of $\beta$. In practice, we usually set $\mathbf{b} = 0$. The prior parameters in equation (2.11) can be elicited by asking about the expected $R^2 \in [0, 1]$ as well as the number of observations worth of weight $\nu_\varepsilon$ the prior estimate should be given. Then $s_\varepsilon = \nu_\varepsilon(1 - R^2)s_y^2$.

is this R-sq in OLS?

The final prior parameter in (2.10) is $\Sigma^{-1}$, which, up to a scaling factor, is the prior precision over $\beta$ in the full model, with all variables included. The total information in the covariates is $X^{\mathrm{T}}X$, and so $\frac{1}{n}X^{\mathrm{T}}X$ is the average information in a single observation. Zellner's $g$-prior [Chipman, George and McCulloch (2001), Liang et al. (2008), Zellner (1986)] sets $\Sigma^{-1} = \frac{g}{n}X^{\mathrm{T}}X$, so that $g$ can be interpreted as $g$ observations worth of information. Zellner's prior becomes improper when $X^{\mathrm{T}}X$ is not positive definite; we therefore ensure propriety by averaging $X^{\mathrm{T}}X$ with its diagonal,

$$\Sigma^{-1} = \frac{g}{n}\{wX^{\mathrm{T}}X + (1 - w)\operatorname{diag}(X^{\mathrm{T}}X)\} \tag{2.12}$$

with default values of $g = 1$ and $w = 1/2$. Overall, this prior specification provides a broadly useful default while providing considerable flexibility in those cases where more specific prior information is available.

2.3. *Inference.* Posterior inference in our model can be broken down into three pieces. First, we simulate draws of the model parameters $\theta$ and the state vector $\alpha$ given the observed data $\mathbf{y}_{1:n}$ in the training period. Second, we use the posterior simulations to simulate from the posterior predictive distribution $p(\tilde{\mathbf{y}}_{n+1:m}|\mathbf{y}_{1:n})$ over the counterfactual time series $\tilde{\mathbf{y}}_{n+1:m}$ given the observed pre-intervention activity $\mathbf{y}_{1:n}$. Third, we use the posterior predictive samples to compute the posterior distribution of the pointwise impact $y_t - \tilde{y}_t$ for each $t = 1, \ldots, m$. We use the same samples to obtain the posterior distribution of cumulative impact.

*Posterior simulation.* We use a Gibbs sampler to simulate a sequence $(\theta, \alpha)^{(1)}$, $(\theta, \alpha)^{(2)}, \ldots$ from a Markov chain whose stationary distribution is $p(\theta, \alpha|\mathbf{y}_{1:n})$. The sampler alternates between a *data-augmentation* step that simulates from $p(\alpha|\mathbf{y}_{1:n}, \theta)$ and a *parameter-simulation* step that simulates from $p(\theta|\mathbf{y}_{1:n}, \alpha)$.

The data-augmentation step uses the posterior simulation algorithm from Durbin and Koopman (2002), providing an improvement over the earlier forward-filtering, backward-sampling algorithms by Carter and Kohn (1994), Frühwirth-Schnatter (1994), and de Jong and Shephard (1995). In brief, because $p(\mathbf{y}_{1:n}, \alpha|\theta)$ is jointly multivariate normal, the variance of $p(\alpha|\mathbf{y}_{1:n}, \theta)$ does not depend on $\mathbf{y}_{1:n}$. We can therefore simulate $(\mathbf{y}^*_{1:n}, \alpha^*) \sim p(\mathbf{y}_{1:n}, \alpha|\theta)$ and subtract $\mathrm{E}(\alpha^*|\mathbf{y}^*_{1:n}, \theta)$ to obtain zero-mean noise with the correct variance. Adding $\mathrm{E}(\alpha|\mathbf{y}_{1:n}, \theta)$ restores the correct mean, which completes the draw. The required expectations can be computed using the Kalman filter and a *fast mean smoother* described in detail by Durbin and Koopman (2002). The result is a direct simulation from $p(\alpha|\mathbf{y}_{1:n}, \theta)$ in an algorithm that is linear in the total (pre- and post-intervention) number of time points ($m$) and quadratic in the dimension of the state space ($d$). time requirement?

Given the draw of the state, the parameter draw is straightforward for all state components other than the static regression coefficients $\beta$. All state components that exclusively depend on variance parameters can translate their draws back to error terms $\eta_t$ and accumulate sums of squares of $\eta$, and, because of conjugacy with equation (2.7), the posterior distribution will remain Gamma distributed.

The draw of the static regression coefficients $\beta$ proceeds as follows. For each $t = 1, \ldots, n$ in the pre-intervention period, let $\dot{y}_t$ denote $y_t$ with the contributions from the other state components subtracted away, and let $\dot{\mathbf{y}}_{1:n} = (\dot{y}_1, \ldots, \dot{y}_n)$. The challenge is to simulate from $p(\varrho, \beta, \sigma_\varepsilon^2|\dot{\mathbf{y}}_{1:n})$, which we can factor into $p(\varrho|\mathbf{y}_{1:n})p(1/\sigma_\varepsilon^2|\varrho, \dot{\mathbf{y}}_{1:n})p(\beta|\varrho, \sigma_\varepsilon, \dot{\mathbf{y}}_{1:n})$. Because of conjugacy, we can integrate out $\beta$ and $1/\sigma_\varepsilon^2$ and be left with

$$\varrho|\dot{\mathbf{y}}_{1:n} \sim C(\dot{\mathbf{y}}_{1:n}) \frac{|\Sigma_\varrho^{-1}|^{1/2}}{|V_\varrho^{-1}|^{1/2}} \frac{p(\varrho)}{S_\varrho^{(N/2)-1}}, \tag{2.13}$$

where $C(\dot{\mathbf{y}}_{1:n})$ is an unknown normalizing constant. The sufficient statistics in equation (2.13) are

$$V_\varrho^{-1} = (X^{\mathrm{T}}X)_\varrho + \Sigma_\varrho^{-1}, \qquad \tilde{\beta}_\varrho = (V_\varrho^{-1})^{-1}(\mathbf{X}_\varrho^{\mathrm{T}}\dot{\mathbf{y}}_{1:n} + \Sigma_\varrho^{-1}b_\varrho),$$
$$N = \nu_\varepsilon + n, \qquad S_\varrho = s_\varepsilon + \dot{\mathbf{y}}_{1:n}^{\mathrm{T}}\dot{\mathbf{y}}_{1:n} + b_\varrho^{\mathrm{T}}\Sigma_\varrho^{-1}b_\varrho - \tilde{\beta}_\varrho^{\mathrm{T}}V_\varrho^{-1}\tilde{\beta}_\varrho.$$

To sample from (2.13), we use a Gibbs sampler that draws each $\varrho_j$ given all other $\varrho_{-j}$. Each full-conditional is easy to evaluate because $\varrho_j$ can only assume two possible values. It should be noted that the dimension of all matrices in (2.13) is $\sum_j \varrho_j$, which is small if the model is truly sparse. There are many matrices to manipulate, but because each is small, the overall algorithm is fast. Once the draw of $\varrho$ is complete, we sample directly from $p(\beta, 1/\sigma_\varepsilon^2|\varrho, \dot{\mathbf{y}}_{1:n})$ using standard conjugate formulae. For an alternative that may be even more computationally efficient, see Ghosh and Clyde (2011).

*Posterior predictive simulation.* While the posterior over model parameters and states $p(\theta, \boldsymbol{\alpha}|\mathbf{y}_{1:n})$ can be of interest in its own right, causal impact analyses are primarily concerned with the posterior incremental effect,

$$p(\tilde{\mathbf{y}}_{n+1:m}|\mathbf{y}_{1:n}, \mathbf{x}_{1:m}). \tag{2.14}$$

As shown by its indices, the density in equation (2.14) is defined precisely for that portion of the time series which is unobserved: the counterfactual market response $\tilde{y}_{n+1}, \ldots, \tilde{y}_m$ that would have been observed in the treated market, after the intervention, in the absence of treatment.

It is also worth emphasising that the density is conditional on the observed data (as well as the priors) and only on these, that is, on activity in the treatment market before the beginning of the intervention as well as activity in all control markets both before and during the intervention. The density is *not* conditioned on parameter estimates or the inclusion or exclusion of covariates with static regression coefficients, all of which have been integrated out. Thus, through Bayesian model averaging, we commit neither to any particular set of covariates, which helps avoid an arbitrary selection, nor to point estimates of their coefficients, which prevents overfitting.

The posterior predictive density in (2.14) is defined as a coherent (joint) distribution over all counterfactual data points, rather than as a collection of pointwise univariate distributions. This ensures that we correctly propagate the serial structure determined on pre-intervention data to the trajectory of counterfactuals. This is crucial, in particular, when forming summary statistics, such as the cumulative effect of the intervention on the treatment market.

Posterior inference was implemented in C++ with an R interface. Given a typically-sized data set with $m = 500$ time points, $J = 10$ covariates, and 10,000 iterations (see Section 4 for an example), this implementation takes less than 30 seconds to complete on a standard computer, enabling near-interactive analyses.

2.4. *Evaluating impact.* Samples from the posterior predictive distribution over counterfactual activity can be readily used to obtain samples from the posterior causal effect, that is, the quantity we are typically interested in. For each draw $\tau$ and for each time point $t = n+1, \ldots, m$, we set

$$\phi_t^{(\tau)} := y_t - \tilde{y}_t^{(\tau)}, \tag{2.15}$$

note: time index ranges from n+1 to m

yielding samples from the approximate posterior predictive density of the effect attributed to the intervention.

In addition to its pointwise impact, we often wish to understand the cumulative effect of an intervention over time. One of the main advantages of a sampling approach to posterior inference is the flexibility and ease with which such derived inferences can be obtained. Reusing the impact samples obtained in (2.15), we compute for each draw $\tau$

$$\sum_{t'=n+1}^{t} \phi_{t'}^{(\tau)} \qquad \forall t = n+1, \ldots, m. \tag{2.16}$$

The preceding *cumulative sum* of causal increments is a useful quantity when $y$ represents a *flow* quantity, measured over an interval of time (e.g., a day), such as the number of searches, sign-ups, sales, additional installs or new users. It becomes uninterpretable when $y$ represents a *stock* quantity, usefully defined only for a point in time, such as the total number of clients, users or subscribers. In this case we might instead choose, for each $\tau$, to draw a sample of the posterior *running average* effect following the intervention,

$$\frac{1}{t-n} \sum_{t'=n+1}^{t} \phi_{t'}^{(\tau)} \qquad \forall t = n+1, \ldots, m. \tag{2.17}$$

Unlike the cumulative effect in (2.16), the running average is always interpretable, regardless of whether it refers to a flow or a stock. However, it is more context-dependent on the length of the post-intervention period under consideration. In particular, under the assumption of a true impact that grows quickly at first and then declines to zero, the cumulative impact approaches its true total value (in expectation) as we increase the counterfactual forecasting period, whereas the average impact will eventually approach zero (while, in contrast, the probability intervals diverge in both cases, leading to more and more uncertain inferences as the forecasting period increases).

**3. Application to simulated data.** To study the characteristics of our approach, we analysed simulated (i.e., computer-generated) data across a series of independent simulations. Generated time series started on 1 January 2013 and ended on 30 June 2014, with a perturbation beginning on 1 January 2014. The data were simulated using a dynamic regression component with two covariates whose coefficients evolved according to independent random walks, $\beta_t \sim \mathcal{N}(\beta_{t-1}, 0.01^2)$,

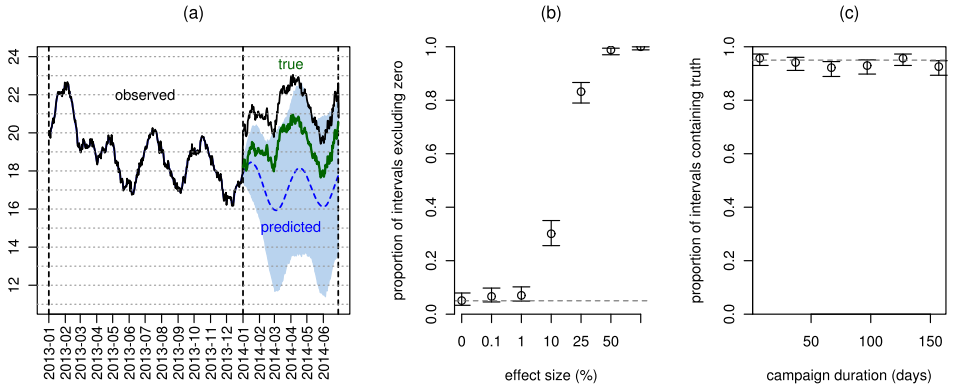


FIG. 3. *Adequacy of posterior uncertainty.* (a) *Example of one of the* 256 *data sets created to assess estimation accuracy. Simulated observations* (*black*) *are based on two contemporaneous covariates, scaled by time-varying coefficients plus a time-varying local level* (*not shown*). *During the campaign period, where the data are lifted by an effect size of* 10%, *the plot shows the posterior expectation of counterfactual activity* (*blue*), *along with its pointwise central* 95% *credible intervals* (*blue shaded area*), *and, for comparison, the true counterfactual* (*green*). (b) *Power curve. Following repeated application of the model to simulated data, the plot shows the empirical frequency of concluding that a causal effect was present, as a function of true effect size, given a post-intervention period of* 6 *months. The curve represents sensitivity in those parts of the graph where the true effect size is positive, and* 1*—specificity where the true effect size is zero. Error bars represent* 95% *credible intervals for the true sensitivity, using a uniform* Beta(1, 1) *prior.* (c) *Interval coverage. Using an effect size of* 10%, *the plot shows the proportion of simulations in which the pointwise central* 95% *credible interval contained the true impact, as a function of campaign duration. Intervals should contain ground truth in* 95% *of simulations, however much uncertainty its predictions may be associated with. Error bars represent* 95% *credible intervals.*

initialised at $\beta_0 = 1$. The covariates themselves were simple sinusoids with wavelengths of 90 days and 360 days, respectively. The latent state underlying the observed data was generated using a local level that evolved according to a random walk, $\mu_t \sim \mathcal{N}(\mu_{t-1}, 0.1^2)$, initialised at $\mu_0 = 0$. Independent observation noise was sampled using $\varepsilon_t \sim \mathcal{N}(0, 0.1^2)$. In summary, observations $y_t$ were generated using

$$y_t = \beta_{t,1} z_{t,1} + \beta_{t,2} z_{t,2} + \mu_t + \varepsilon_t .$$

To simulate the effect of advertising, the post-intervention portion of the preceding series was multiplied by $1 + e$, where $e$ (not to be confused with $\varepsilon$) represented the true effect size specifying the (uniform) relative lift during the campaign period. An example is shown in Figure 3(a).

*Sensitivity and specificity.* To study the properties of our model, we began by considering under what circumstances we successfully detected a causal effect, that is, the statistical power or sensitivity of our approach. A related property is the probability of *not* detecting an absent impact, that is, specificity. We repeatedly

sensitivity/recall: True Positive Rate (TPR) = TP / (TP + FN)
specificity: True Negative Rate (TNR) = TN / (TN + FP)

generated data, as described above, under different true effect sizes. We then computed the posterior predictive distribution over the counterfactuals and recorded whether or not we would have concluded a causal effect.

For each of the effect sizes 0%, 0.1%, 1%, 10% and 100%, a total of $2^8 = 256$ simulations were run. This number was chosen simply on the grounds that it provided reasonably tight intervals around the reported summary statistics without requiring excessive amounts of computation. In each simulation, we concluded that a causal effect was present if and only if the central 95% posterior probability interval of the cumulative effect excluded zero.

The model used throughout this section comprised two structural blocks. The first one was a local level component. We placed an inverse-Gamma prior on its diffusion variance with a prior estimate of $s/\nu = 0.1\sigma_y$ and a prior sample size $\nu = 32$. The second structural block was a dynamic regression component. We placed a Gamma prior with prior expectation $0.1\sigma_y$ on the diffusion variance of both regression coefficients. By construction, the outcome variable did not exhibit any local trends or seasonality other than the variation conveyed through the covariates. This obviated the need to include an explicit local linear trend or seasonality component in the model.

In a first analysis, we considered the empirical proportion of simulations in which a causal effect had been detected. When taking into account only those simulations where the true effect size was greater than zero, these empirical proportions provide estimates of the sensitivity of the model w.r.t. the process by which the data were generated. Conversely, those simulations where the campaign had no effect yield an estimate of the model's specificity. In this way, we obtained the power curve shown in Figure 3(b). The curve shows that, in data such as these, a market perturbation leading to a lift no larger than 1% is missed in about 90% of cases. By contrast, a perturbation that lifts market activity by 25% is correctly detected as such in most cases.

In a second analysis, we assessed the coverage properties of the posterior probability intervals obtained through our model. It is desirable to use a diffuse prior on the local level component such that central 95% intervals contain ground truth in about 95% of the simulations. This coverage frequency should hold regardless of the length of the campaign period. In other words, a longer campaign should lead to posterior intervals that are appropriately widened to retain the same coverage probability as the narrower intervals obtained for shorter campaigns. This was approximately the case throughout the simulated campaign [Figure 3(c)].

*Estimation accuracy.* To study the accuracy of the point estimates supported by our approach, we repeated the preceding simulations with a fixed effect size of 10% while varying the length of the campaign. When given a quadratic loss function, the loss-minimizing point estimate is the posterior expectation of the predictive density over counterfactuals. Thus, for each generated data set $i$, we

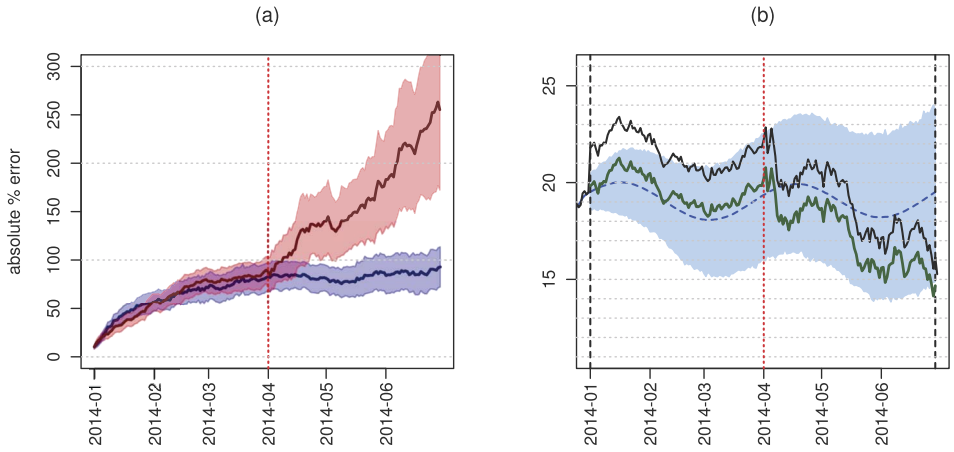


FIG. 4. *Estimation accuracy.* (a) *Time series of absolute percentage discrepancy between inferred effect and true effect. The plot shows the rate* (*mean* $\pm$ 2 *s.e.m.*) *at which predictions become less accurate as the length of the counterfactual forecasting period increases* (*blue*). *The well-behaved decrease in estimation accuracy breaks down when the data are subject to a sudden structural change* (*red*), *as simulated for* 1 *April* 2014. (b) *Illustration of a structural break. The plot shows one example of the time series underlying the red curve in* (a). *On* 1 *April* 2014, *the standard deviation of the generating random walk of the local level was tripled, causing the rapid decline in estimation accuracy seen in the red curve in* (a). blue: prediction; green: true counterfactual;
black: observed (= counterfactual + treatment effect)
after the vertical red line, predictions are deviating from the true counterfactual

computed the expected causal effect for each time point,

$$\hat{\phi}_{i,t} := \langle \phi_t | y_1, \ldots, y_m, x_1, \ldots, x_m \rangle \qquad \forall t = n+1, \ldots, m; i = 1, \ldots, 256. \tag{3.1}$$

To quantify the discrepancy between estimated and true impact, we calculated the absolute percentage estimation error,

$$a_{i,t} := \frac{|\hat{\phi}_{i,t} - \phi_t|}{\phi_t}. \tag{3.2}$$

This yielded an empirical distribution of absolute percentage estimation errors [Figure 4(a), blue], showing that impact estimates become less and less accurate as the forecasting period increases. This is because, under the local linear trend model in (2.3), the true counterfactual activity becomes more and more likely to deviate from its expected trajectory.

It is worth emphasising that all preceding results are based on the assumption that the model structure remains intact throughout the modelling period. In other words, even though the model is built around the idea of multiple (nonstationary) components (i.e., a time-varying local trend and, potentially, time-varying regression coefficients), this structure itself remains unchanged. If the model structure does change, estimation accuracy may suffer.

We studied the impact of a changing model structure in a second simulation in which we repeated the procedure above in such a way that 90 days after the

beginning of the campaign the standard deviation of the random walk governing the evolution of the regression coefficient was tripled (now 0.03 instead of 0.01). As a result, the observed data began to diverge much more quickly than before. Accordingly, estimations became considerably less reliable [Figure 4(a), red]. An example of the underlying data is shown in Figure 4(b).

The preceding simulations highlight the importance of a model that is sufficiently flexible to account for phenomena typically encountered in seasonal empirical data. This rules out entirely static models in particular (such as multiple linear regression).

**4. Application to empirical data.** To illustrate the practical utility of our approach, we analysed an advertising campaign run by one of Google's advertisers in the United States. In particular, we inferred the campaign's causal effect on the number of times a user was directed to the advertiser's website from the Google search results page. We provide a brief overview of the underlying data below [see Vaver and Koehler (2011) for additional details].

The campaign analysed here was based on product-related ads to be displayed alongside Google's search results for specific keywords. Ads went live for a period of 6 consecutive weeks and were geo-targeted to a randomised set of 95 out of 190 designated market areas (DMAs). The most salient observable characteristic of DMAs is offline sales. To produce balance in this characteristic, DMAs were first rank-ordered by sales volume. Pairs of regions were then randomly assigned to treatment/control. DMAs provide units that can be easily supplied with distinct offerings, although this fine-grained split was not a requirement for the model. In fact, we carried out the analysis as if only one treatment region had been available (formed by summing all treated DMAs). This allowed us to evaluate whether our approach would yield the same results as more conventional treatment-control comparisons would have done.

The outcome variable analysed here was search-related visits to the advertiser's website, consisting of organic clicks (i.e., clicks on a search result) and paid clicks (i.e., clicks on an ad next to the search results, for which the advertiser was charged). Since paid clicks were zero before the campaign, one might wonder why we could not simply count the number of paid clicks after the campaign had started. The reason is that paid clicks tend to cannibalise some organic clicks. Since we were interested in the net effect, we worked with the total number of clicks.

The first building block of the model used for the analyses in this section was a local level component. For the inverse-Gamma prior on its diffusion variance we used a prior estimate of $s/\nu = 0.1\sigma_y$ and a prior sample size $\nu = 32$. The second structural block was a static regression component. We used a spike-and-slab prior with an expected model size of $M = 3$, an expected explained variance of $R^2 = 0.8$ and 50 prior $df$. We deliberately kept the model as simple as this.

what is a prior sample size? how to understand it?

Since the covariates came from a randomised experiment, we expected them to already account for any additional local linear trends and seasonal variation in the response variable. If one suspects that a more complex model might be more appropriate, one could optimise model design through Bayesian model selection. Here, we focus instead on comparing different sets of covariates, which is critical in counterfactual analyses regardless of the particular model structure used. Model estimation was carried out using 10,000 MCMC samples.

*Analysis* 1: *Effect on the treated, using a randomised control.* We began by applying the above model to infer the causal effect of the campaign on the time series of clicks in the treated regions. Given that a set of unaffected regions was available in this analysis, the best possible set of controls was given by the untreated DMAs themselves (see below for a comparison with a purely observational alternative).

As shown in Figure 5(a), the model provided an excellent fit on the pre-campaign trajectory of clicks (including a spike in "week −2" and a dip at the end of "week −1"). Following the onset of the campaign, observations quickly began to diverge from counterfactual predictions: the actual number of clicks was consistently higher than what would have been expected in the absence of the campaign. The curves did not reconvene until one week after the end of the campaign. Subtracting observed from predicted data, as we did in Figure 5(b), resulted in a posterior estimate of the incremental lift caused by the campaign. It peaked after about three weeks into the campaign, and faded away after about one week after the end of the campaign. Thus, as shown in Figure 5(c), the campaign led to a sustained cumulative increase in total clicks (as opposed to a mere shift of future clicks into the present or a pure cannibalization of organic clicks by paid clicks). Specifically, the overall effect amounted to 88,400 additional clicks in the targeted regions (posterior expectation; rounded to three significant digits), that is, an increase of 22%, with a central 95% credible interval of [13%, 30%].

To validate this estimate, we returned to the original experimental data, on which a conventional treatment-control comparison had been carried out using a two-stage linear model [Vaver and Koehler (2011)]. This analysis had led to an estimated lift of 84,700 clicks, with a 95% confidence interval for the relative expected lift of [19%, 22%]. Thus, with a deviation of less than 5%, the counterfactual approach had led to almost precisely the same estimate as the randomised evaluation, except for its wider intervals. The latter is expected, given that our intervals represent prediction intervals, not confidence intervals. Moreover, in addition to an interval for the sum over all time points, our approach yields a full time series of pointwise intervals, which allows analysts to examine the characteristics of the temporal evolution of attributable impact.

The posterior predictive intervals in Figure 5(b) widen more slowly than in the illustrative example in Figure 1. This is because the large number of controls available in this data set offers a much higher pre-campaign predictive strength than in the simulated data in Figure 1. This is not unexpected, given that controls came

higher pre-campaign predictive strength can lead to a slowly increase in credible interval in the intervention period.

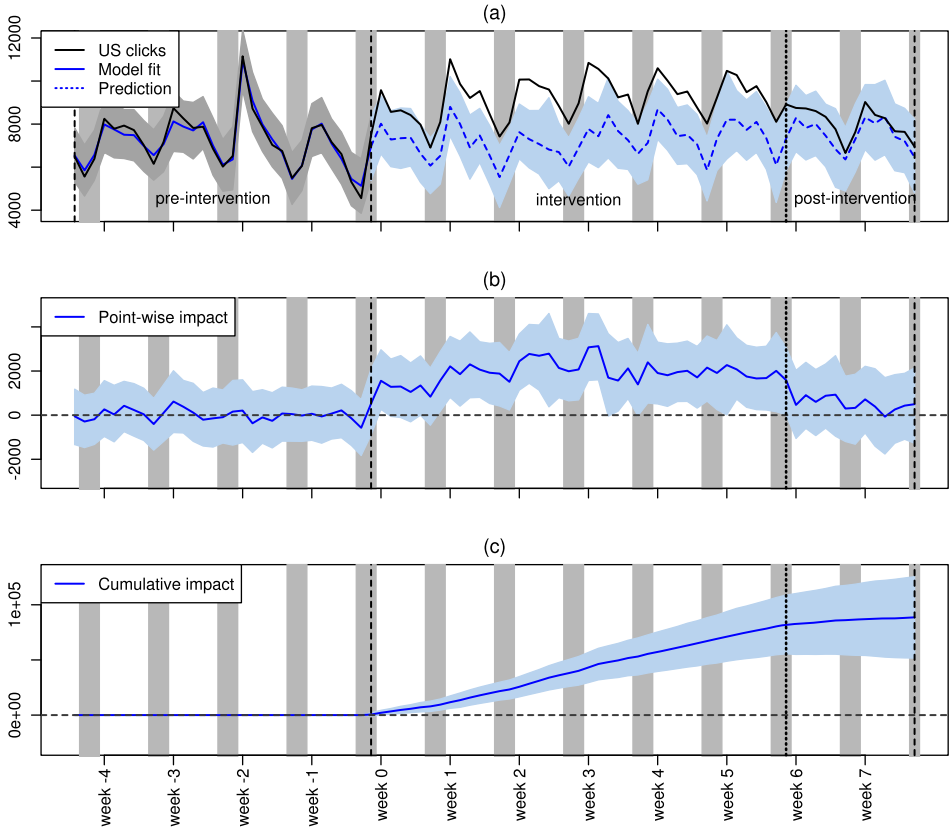


FIG. 5. *Causal effect of online advertising on clicks in treated regions.* (a) *Time series of search-related visits to the advertiser's website* (*including both organic and paid clicks*). (b) *Pointwise* (*daily*) *incremental impact of the campaign on clicks. Shaded vertical bars indicate weekends.* (c) *Cumulative impact of the campaign on clicks.*

from a randomised experiment, and we will see that this also holds for a subsequent analysis (see below) that is based on yet another data source for predictors. A consequence of this is that there is little variation left to be captured by the random-walk component of the model. A reassuring finding is that the estimated counterfactual time series in Figure 5(a) eventually almost exactly rejoins the observed series, only a few days after the end of the intervention.

intervention period: used for causal inference; post-intervention period: robust check

*Analysis* 2: *Effect on the treated, using observational controls.* An important characteristic of counterfactual-forecasting approaches is that they do not require a setting in which a set of controls, selected at random, was exempt from the campaign. We therefore repeated the preceding analysis in the following way: we discarded the data from all control regions and, instead, used searches for keywords related to the advertiser's industry, grouped into a handful of verticals, as

covariates. In the absence of a dedicated set of control regions, such industry-related time series can be very powerful controls, as they capture not only seasonal variations but also market-specific trends and events (though not necessarily advertiser-specific trends). A major strength of the controls chosen here is that time series on web searches are publicly available through Google Trends (http://www.google.com/trends/). This makes the approach applicable to virtually any kind of intervention. At the same time, the industry as a whole is unlikely to be moved by a single actor's activities. This precludes a positive bias in estimating the effect of the campaign that would arise if a covariate was negatively affected by the campaign.

As shown in Figure 6, we found a cumulative lift of 85,900 clicks (posterior expectation), or 21%, with a $[12\%, 30\%]$ interval. In other words, the analysis replicated almost perfectly the original analysis that had access to a randomised set of controls. One feature in the response variable which this second analysis failed to account for was a spike in clicks in the second week before the campaign onset; this spike appeared both in treated and untreated regions and appears to be specific to this advertiser. In addition, the series of point-wise impact [Figure 6(b)] is slightly more volatile than in the original analysis (Figure 5). On the other hand, the overall point estimate of 85,900, in this case, was even closer to the randomised-design baseline (84,700; deviation ca. 1%) than in our first analysis (88,400; deviation ca. 4%). In summary, the counterfactual approach effectively obviated the need for the original randomised experiment. Using purely observational variables led to the same substantive conclusions.

*Analysis* 3: *Absence of an effect on the controls*. To go one step further still, we analysed clicks in those regions that had been exempt from the advertising campaign. If the effect of the campaign was truly specific to treated regions, there should be no effect in the controls. To test this, we inferred the causal effect of the campaign on *unaffected* regions, which should *not* lead to a significant finding. In analogy with our second analysis, we discarded clicks in the treated regions and used searches for keywords related to the advertiser's industry as controls.

As summarised in Figure 7, no significant effect was found in unaffected regions, as expected. Specifically, we obtained an overall nonsignificant lift of 2% in clicks with a central 95% credible interval of $[-6\%, 10\%]$.

In summary, the empirical data considered in this section showed: (i) a clear effect of advertising on treated regions when using randomised control regions to form the regression component, replicating previous treatment-control comparisons (Figure 5); (ii) notably, an equivalent finding when discarding control regions and instead using observational searches for keywords related to the advertiser's industry as covariates (Figure 6); (iii) reassuringly, the absence of an effect of advertising on regions that were not targeted (Figure 7).

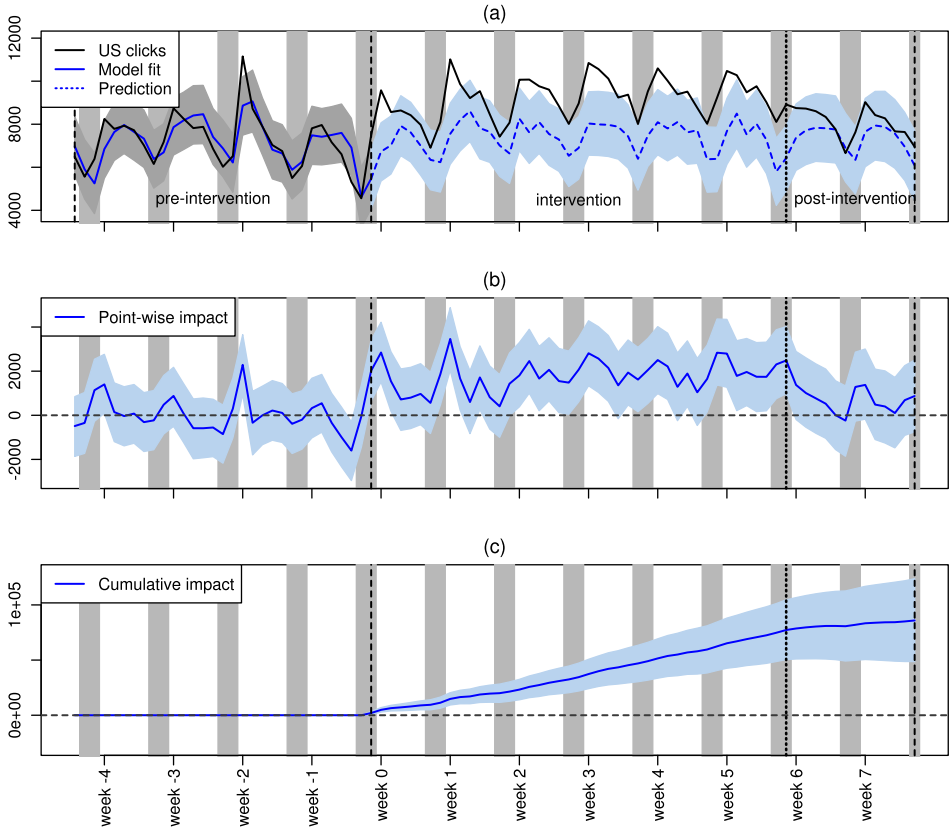


FIG. 6. *Causal effect of online advertising on clicks, using only* searches *for keywords related to the advertiser's industry as controls, discarding the original control regions as would be the case in studies where a randomised experiment was not carried out.* (a) *Time series of clicks on to the advertiser's website.* (b) *Pointwise (daily) incremental impact of the campaign on clicks.* (c) *Cumulative impact of the campaign on clicks. The plots show that this analysis, which was based on observational covariates only, provided almost exactly the same inferences as the first analysis (Figure 5) that had been based on a randomised design.*

**5. Discussion.** The increasing interest in evaluating the incremental impact of market interventions has been reflected by a growing literature on applied causal inference. With the present paper we are hoping to contribute to this literature by proposing a Bayesian state-space model for obtaining a counterfactual prediction of market activity. We discuss the main features of this model below.

In contrast to most previous schemes, the approach described here is fully Bayesian, with regularizing or empirical priors for all hyperparameters. Posterior inference gives rise to complete-data (smoothing) predictions that are only conditioned on past data in the treatment market and both past and present data in the control markets. Thus, our model embraces a dynamic evolution of states and,

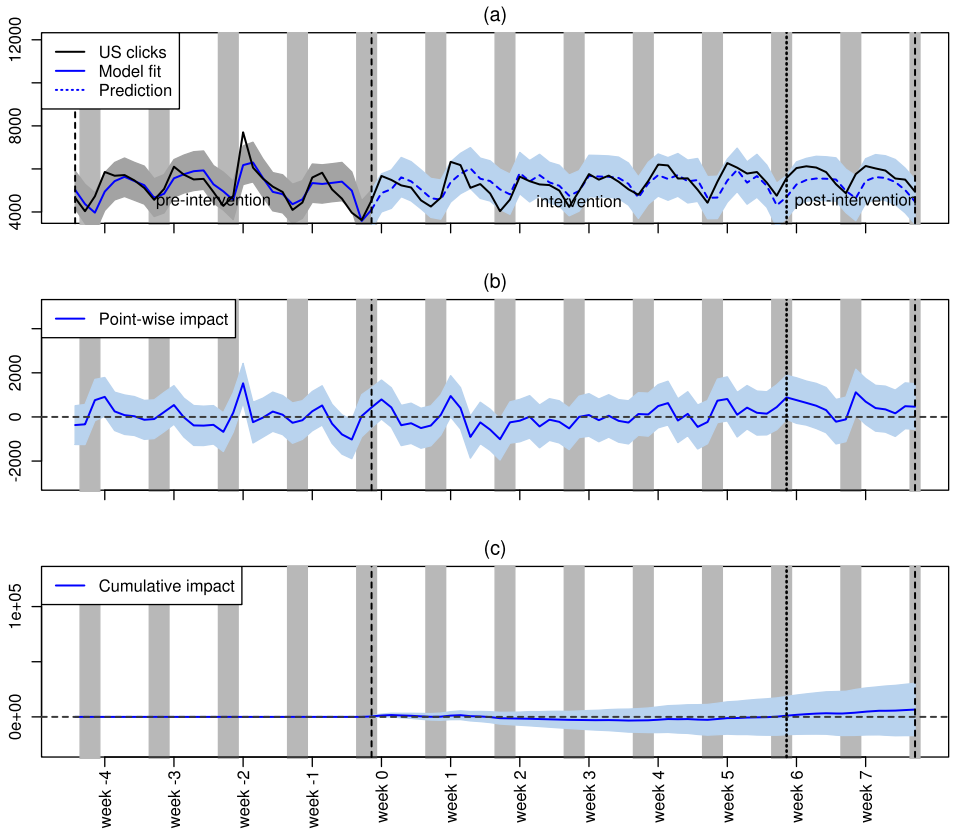


FIG. 7. *Causal effect of online advertising on clicks in* nontreated *regions, which should* not *show an effect. Searches for keywords related to the advertiser's industry are used as controls. Plots show inferences in analogy with Figure* 5. (a) *Time series of clicks to the advertiser's website.* (b) *Pointwise* (*daily*) *incremental impact of the campaign on clicks.* (c) *Cumulative impact of the campaign on clicks.*

optionally, coefficients (departing from classical linear regression models with a fixed number of static regressors), and enables us to flexibly summarise posterior inferences.

Because closed-form posteriors for our model do not exist, we suggest a stochastic approximation to inference using MCMC. One convenient consequence of this is that we can reuse the samples from the posterior to obtain credible intervals for all summary statistics of interest. Such statistics include, for example, the average absolute and relative effect caused by the intervention as well as its cumulative effect.

Posterior inference was implemented in C++ and R and, for all empirical data sets presented in Section 4, took less than 30 seconds on a standard Linux machine. If the computational burden of sampling-based inference ever became prohibitive,

one option would be to replace it by a variational Bayesian approximation [see, e.g., Brodersen et al. (2013), Mathys et al. (2011)].

Another way of using the proposed model is for power analyses. In particular, given past time series of market activity, we can define a point in the past to represent a hypothetical intervention and apply the model in the usual fashion. As a result, we obtain a measure of uncertainty about the response in the treated market after the beginning of the hypothetical intervention. This provides an estimate of what incremental effect would have been required to be outside of the 95% central interval of what would have happened in the absence of treatment.

The model presented here subsumes several simpler models which, in consequence, lack important characteristics, but which may serve as alternatives should the full model appear too complex for the data at hand. One example is classical multiple linear regression. In principle, classical regression models go beyond difference-in-differences schemes in that they account for the full counterfactual trajectory. However, they are not suited for predicting stochastic processes beyond a few steps. This is because ordinary least-squares estimators disregard serial autocorrelation; the static model structure does not allow for temporal variation in the coefficients; and predictions ignore our posterior uncertainty about the parameters. Put differently: classical multiple linear regression is a special case of the state-space model described here in which (i) the Gaussian random walk of the local level has zero variance; (ii) there is no local linear trend; (iii) regression coefficients are static rather than time-varying; (iv) ordinary least squares estimators are used which disregard posterior uncertainty about the parameters and may easily overfit the data.

Another special case of the counterfactual approach discussed in this paper is given by synthetic control estimators that are restricted to the class of convex combinations of predictor variables and do not include time-series effects such as trends and seasonality [Abadie (2005), Abadie, Diamond and Hainmueller (2010)]. Relaxing this restriction means we can utilise predictors regardless of their scale, even if they are negatively correlated with the outcome series of the treated unit.

Other special cases include autoregressive (AR) and moving-average (MA) models. These models define autocorrelation among observations rather than latent states, thus precluding the ability to distinguish between state noise and observation noise [Ataman, Mela and Van Heerde (2008), Leeflang et al. (2009)].

In the scenarios we consider, advertising is a planned perturbation of the market. This generally makes it easier to obtain plausible causal inferences than in genuinely *observational* studies in which the experimenter had no control about treatment [see discussions in Antonakis et al. (2010), Berndt (1991), Brady (2002), Camillo and d'Attoma (2010), Hitchcock (2004), Kleinberg and Hripcsak (2011), Lewis, Rao and Reiley (2011), Lewis and Reiley (2011), Robinson, McNulty and

Krasno (2009), Vaver and Koehler (2011), Winship and Morgan (1999)]. The principal problem in observational studies is endogeneity: the possibility that the observed outcome might not be the result of the treatment but of other omitted, endogenous variables. In principle, propensity scores can be used to correct for the selection bias that arises when the treatment effect is correlated with the likelihood of being treated [Chan et al. (2010), Rubin and Waterman (2006)]. However, the propensity-score approach requires that exposure can be measured at the individual level, and it, too, does not guarantee valid inferences, for example, in the presence of a specific type of selection bias recently termed "activity bias" [Lewis, Rao and Reiley (2011)]. Counterfactual modelling approaches avoid these issues when it can be assumed that the treatment market was chosen at random.

Overall, we expect inferences on the causal impact of designed market interventions to play an increasingly prominent role in providing quantitative accounts of return on investment [Danaher and Rust (1996), Leeflang et al. (2009), Seggie, Cavusgil and Phelan (2007), Stewart (2009)]. This is because marketing resources, specifically, can only be allocated to whichever campaign elements jointly provide the greatest return on ad spend (ROAS) if we understand the causal effects of spend on sales, product adoption or user engagement. At the same time, our approach could be used for many other applications involving causal inference. Examples include problems found in economics, epidemiology, biology or the political and social sciences. With the release of the `CausalImpact` R package we hope to provide a simple framework serving all of these areas. Structural time-series models are being used in an increasing number of applications at Google, and we anticipate that they will prove equally useful in many analysis efforts elsewhere.

**Acknowledgment.** The authors wish to thank Jon Vaver for sharing the empirical data analysed in this paper.

## REFERENCES

ABADIE, A. (2005). Semiparametric difference-in-differences estimators. *Rev. Econom. Stud.* **72** 1–19. MR2116973

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* **105** 493–505. MR2759929

ABADIE, A. and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the basque country. *Amer. Econ. Rev.* **93** 113–132.

ANGRIST, J. D. and KRUEGER, A. B. (1999). Empirical strategies in labor economics. *Handbook of Labor Economics* **3** 1277–1366.

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

ANTONAKIS, J., BENDAHAN, S., JACQUART, P. and LALIVE, R. (2010). On making causal claims: A review and recommendations. *Leadersh. Q.* **21** 1086–1120.

ASHENFELTER, O. and CARD, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Rev. Econ. Stat.* **67** 648–660.

ATAMAN, M. B., MELA, C. F. and VAN HEERDE, H. J. (2008). Building brands. *Mark. Sci.* **27** 1036–1054.

ATHEY, S. and IMBENS, G. W. (2002). Identification and inference in nonlinear difference-in-differences models. Working Paper 280, National Bureau of Economic Research, Cambridge, MA.

BANERJEE, S., KAUFFMAN, R. J. and WANG, B. (2007). Modeling Internet firm survival using Bayesian dynamic models with time-varying coefficients. *Electron. Commer. Res. Appl.* **6** 332–342.

BELLONI, A., CHERNOZHUKOV, V., FERNANDEZ-VAL, I. and HANSEN, C. (2013). Program evaluation with high-dimensional data. CeMMAP Working Paper CWP77/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, London.

BERNDT, E. R. (1991). *The Practice of Econometrics*: *Classic and Contemporary*. Addison-Wesley, Reading, MA.

BERTRAND, M., DUFLO, E. and MULLAINATHAN, S. (2002). How much should we trust differences-in-differences estimates? Working Paper 8841, National Bureau of Economic Research, Cambridge, MA.

BRADY, H. E. (2002). Models of causal inference: Going beyond the Neyman-Rubin-Holland theory. In *Annual Meetings of the Political Methodology Group*, Boston, MA.

BRODERSEN, K. H., DAUNIZEAU, J., MATHYS, C., CHUMBLEY, J. R., BUHMANN, J. M. and STEPHAN, K. E. (2013). Variational Bayesian mixed-effects inference for classification studies. *Neuroimage* **76** 345–361.

CAMILLO, F. and D'ATTOMA, I. (2010). A new data mining approach to estimate causal effects of policy interventions. *Expert Syst. Appl.* **37** 171–181.

CAMPBELL, D. T., STANLEY, J. C. and GAGE, N. L. (1963). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin, Boston.

CARD, D. and KRUEGER, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. Technical report, National Bureau of Economic Research, Cambridge, MA.

CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553. MR1311096

CHAN, D., GE, R., GERSHONY, O., HESTERBERG, T. and LAMBERT, D. (2010). Evaluating online ad campaigns in a pipeline: Causal models at scale. In *Proceedings of the* 16*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 7–16. ACM, New York.

CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **38** 65–134. IMS, Beachwood, OH. MR2000752

CLAVEAU, F. (2012). The Russo–Williamson Theses in the social sciences: Causal inference drawing on two types of evidence. *Stud. Hist. Philos. Biol. Biomed. Sci.* **43** 806–813.

COX, D. and WERMUTH, N. (2001). Causal inference and statistical fallacies. In *International Encyclopedia of the Social & Behavioral Sciences* (Neil J. Smelser and P. B. Baltes, eds.) 1554–1561. Pergamon, Oxford.

DANAHER, P. J. and RUST, R. T. (1996). Determining the optimal return on investment for an advertising campaign. *European J. Oper. Res.* **95** 511–521.

DE JONG, P. and SHEPHARD, N. (1995). The simulation smoother for time series models. *Biometrika* **82** 339–350. MR1354233

DONALD, S. G. and LANG, K. (2007). Inference with difference-in-differences and other panel data. *Rev. Econ. Stat.* **89** 221–233.

DURBIN, J. and KOOPMAN, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89** 603–615. MR1929166

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal.* **15** 183–202. MR1263889

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

GHOSH, J. and CLYDE, M. A. (2011). Rao-Blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *J. Amer. Statist. Assoc.* **106** 1041–1052. MR2894762

HANSEN, C. B. (2007a). Asymptotic properties of a robust variance matrix estimator for panel data when $T$ is large. *J. Econometrics* **141** 597–620. MR2413481

HANSEN, C. B. (2007b). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *J. Econometrics* **140** 670–694. MR2408922

HECKMAN, J. J. and VYTLACIL, E. J. (2007). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In *Handbook of Econometrics* 6, *Part B* (J. J. Heckman and E. E. Leamer, eds.) 4779–4874. Elsevier, Amsterdam.

HITCHCOCK, C. (2004). Do all and only causes raise the probabilities of effects? In *Causation and Counterfactuals*. MIT Press, Cambridge.

HOOVER, K. D. (2012). Economic theory and causal inference. In *Philosophy of Economics* 13 (U. Mäki, ed.) 89–113. Elsevier, Amsterdam.

KLEINBERG, S. and HRIPCSAK, G. (2011). A review of causal inference for biomedical informatics. *J. Biomed. Inform.* **44** 1102–1112.

LEEFLANG, P. S., BIJMOLT, T. H., VAN DOORN, J., HANSSENS, D. M., VAN HEERDE, H. J., VERHOEF, P. C. and WIERINGA, J. E. (2009). Creating lift versus building the base: Current trends in marketing dynamics. *Int. J. Res. Mark.* **26** 13–20.

LESTER, R. A. (1946). Shortcomings of marginal analysis for wage-employment problems. *Amer. Econ. Rev.* **36** 63–82.

LEWIS, R. A., RAO, J. M. and REILEY, D. H. (2011). Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the* 20*th International Conference on World Wide Web*. *WWW'*11 157–166. ACM, New York.

LEWIS, R. A. and REILEY, D. H. (2011). Does retail advertising work? Technical report.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of $g$ priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

MATHYS, C., DAUNIZEAU, J., FRISTON, K. J. and STEPHAN, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Front. Human Neurosci.* **5** 39.

MEYER, B. D. (1995). Natural and quasi-experiments in economics. *J. Bus. Econom. Statist.* **13** 151.

MORGAN, S. L. and WINSHIP, C. (2007). *Counterfactuals and Causal Inference*: *Methods and Principles for Social Research*. Cambridge Univ. Press, Cambridge.

NAKAJIMA, J. and WEST, M. (2013). Bayesian analysis of latent threshold dynamic models. *J. Bus. Econom. Statist.* **31** 151–164. MR3055329

POLSON, N. G. and SCOTT, S. L. (2011). Data augmentation for support vector machines. *Bayesian Anal.* **6** 1–23. MR2781803

ROBINSON, G., MCNULTY, J. E. and KRASNO, J. S. (2009). Observing the counterfactual? The search for political experiments in nature. *Polit. Anal.* **17** 341–357.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (2008). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In *Epidemiology and Medical Statistics* (J. P. Miller, C. R. Rao and D. C. Rao, eds.). *Handbook of Statist.* **27** 28–63. Elsevier, Amsterdam. MR2500431

RUBIN, D. B. and WATERMAN, R. P. (2006). Estimating the causal effects of marketing interventions using propensity score methodology. *Statist. Sci.* **21** 206–222. MR2324079

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

SCOTT, S. L. and VARIAN, H. R. (2014). Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modeling and Optimization* **5** 4–23.

SEGGIE, S. H., CAVUSGIL, E. and PHELAN, S. E. (2007). Measurement of return on marketing investment: A conceptual framework and the future of marketing metrics. *Ind. Mark. Manage.* **36** 834–841.

SHADISH, W. R., COOK, T. D. and CAMPBELL, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Cengage Learning, Seattle, WA.

SOLON, G. (1984). *Estimating Autocorrelations in Fixed-Effects Models*. National Bureau of Economic Research, Cambridge, MA.

STEWART, D. W. (2009). Marketing accountability: Linking marketing actions to financial results. *J. Bus. Res.* **62** 636–643.

TAKADA, H. and BASS, F. M. (1998). Multiple time series analysis of competitive marketing behavior. *J. Bus. Res.* **43** 97–107.

VAVER, J. and KOEHLER, J. (2011). Measuring ad effectiveness using geo experiments. Technical report, Google Inc.

VAVER, J. and KOEHLER, J. (2012). Periodic measurement of advertising effectiveness using multiple-test-period geo experiments. Technical report, Google Inc.

WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. Springer, New York. MR1482232

WINSHIP, C. and MORGAN, S. L. (1999). The estimation of causal effects from observational data. *Annu. Rev. Sociol.* 659–706.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with $g$-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

GOOGLE, INC.
1600 AMPHITHEATRE PARKWAY
MOUNTAIN VIEW, CALIFORNIA 94043
USA
E-MAIL: kbrodersen@google.com
gallusser@google.com
jkoehler@google.com
nicolasremy@google.com
stevescott@google.com