UPPSALA
UNIVERSITET

Working Paper 2018:1

Department of Statistics

Using High Frequency Pre-treatment Outcomes To Identify Causal Effects In Non-experimental Data

Mårten Schultzberg

Working Paper 2018:1
January 2019; September 2019 (revised)
Department of Statistics
Uppsala University
Box 513
SE-751 20 UPPSALA
SWEDEN

Working papers can be downloaded from www.statistics.uu.se

Title: Using high frequency pre-treatment outcomes to identify causal effects in non-experimental data

Author: Mårten Schultzberg
E-mail: marten.schultzberg@statistik.uu.se

Using high frequency pre-treatment outcomes to identify causal effects in non-experimental data

Mårten Schultzberg[1]

PhD-student, Department of Statistics, Uppsala University, Sweden
Email: marten.schultzberg@statistics.uu.se

September 24, 2019

[1]Acknowledgement: I would like to thank Per Johansson for discussions and suggestions to this paper, Bengt Muthén and Mattias Nordin for feedback on earlier drafts, and the editor and one anonymous reviewer for constructive suggestions and improvements.

Abstract

In observational studies it is common to use matching strategies to consistently estimate the average treatment effect of the treated (ATET) under an unconfoundedness assumption. Matching is often based on a set of time-invariant covariates together with one or a few pre-treatment measurements of the outcome. This paper proposes strategies to consistently estimate the ATET using information derived from a large number of pre-treatment measurements of the outcome. The key to this strategy is to use two-level time-series model estimates to summarize the inter-unit heterogeneity in the sample. It is illustrated how this approach is in line with the conventional identifying assumptions. The theoretical results and estimation strategies are illustrated by a study of electricity consumption.

Keywords: DSEM, intensive longitudinal data, matching, treatment effects, two-level time-series

1 Introduction

The collection of high frequency repeated measurements data has been substantially simplified by the technological progress of personal electronic devices like smart phones, smart watches, fitness trackers, and the "Internet of things". This has lead to that intensive longitudinal data (ILD) obtained from, e.g., experience sampling methods, ecological momentary assessment, and ambulatory assessment, are becoming more prevalent (see Trull and Ebner-Priemer (2009) for an overview). Typically, these types of data contain very few covariates which suggest limited utility for causal analysis using traditional selection on observables estimators (e.g. matching and regression analysis). This paper proposes a framework for the identification and estimation of the average treatment effect of the treated (ATET) estimand with ILD on the pre-treatment outcomes. By using novel multilevel time-series models, the heterogeneity among a large number of long time-series can easily be characterized by a small set of random effect estimates. This paper investigates how this information can be used to *non-parametrically identify* the ATET. As a motivational example, data on electricity consumption are used. These data, typical for electricity-consumption studies, consist of repeated measurements of electricity consumption for all observed units, before and after a policy intervention, together with one or a few time-varying covariates.

The non-parametric identification of the ATET in the proposed framework is obtained from the traditional uncounfoundedness assumption (Rubin, 1973, 1974), together with a sequential ignorability assumption on the time varying covariates (see, e.g., Imai et al. 2010). By adjusting the traditional notation to incorporate ILD, we show how the standard identifying assumptions maps to information that can be derived from time-series data. The novelty of this strategy is that unobserved confounders can be replaced by information extracted from time-series pre-treatment outcome data in the form of unit-specific random coefficients, which govern the treatment assignment and outcome trajectory if not treated. The heterogeneity is estimated using two-level time-series models, with time on level-1 and units on level-2. Based on these random coefficient estimates and, possibly, a small set of covariates, the ATET is non-parametrically estimated by calculating the counterfactual outcome (potentially adjusting for sequentially ignorable time-varying covariates) for all post treatment time periods using non-treated with the 'same' unit-specific coefficient and covariates as the treated. The importance and utility of repeated measurements for drawing casual inference in observational studies have been discussed extensively in the causal literature; for recent

papers on this topic see, e.g., Chabé-Ferret (2015) and O'Neill et al. (2016). However, the main focus of the 'matching literature' has been on how to efficiently utilize, and/or select among a large set of time-invariant covariates and/or time-varying covariates observed right before treatment. As a consequence, the discussion about if and how pre-treatment measurements should be included in matching is usually based on the presumption that only one, or possibly a few, pre-treatment measurements are available. The case when the available set of covariates is small but the number of pre-treatment measurements of the outcome is large is seldom discussed. This is likely due to the fact that such data historically have been rare and difficult to obtain.

Just as we do in this paper, the synthetic control methods, first introduced in Abadie and Gardeazabal (2003), are also to a large extent utilizing pre-treatment measurements in the effect estimation. This literature is focused on case studies with one treated unit and group-level aggregated data. In Abadie et al. (2010), a weighted set of pre-treatment measurements of outcomes from a pool of non-treated are used to construct a synthetic control group which is used to construct the counterfactual trajectory of the outcome for the treated unit under no treatment. This is similar to the framework proposed in this paper in terms of using pre-treatment data, but Abadie et al. (2010) do not aim to identify an average treatment effect for a given population. While it is possible to encompass multiple treated units by aggregating the outcome of all treated (Abadie et al., 2010), substantial heterogeneity among treated and control units would lead to efficiency losses associated with aggregating the treated units before constructing the counterfactual trajectory using a synthetic control group. Based on the idea of synthetic controls, other strategies have been developed by, e.g., Brodersen et al. (2015); Robbins et al. (2017); Xu (2017). In Brodersen et al. (2015) the outcome of the synthetic control is modeled using a Bayesian time-series model. The counterfactual outcome of the treated is predicted based on the chosen model. This approach is similar to the present paper in the sense that parametric time-series models are fitted on pre-treatment measurements of the outcome. Robbins et al. (2017) extend the synthetic control framework to multiple treated units and to high dimensional micro data. Xu (2017) suggest combining standard fixed effects panel models with the synthetic control framework in a large N and T factor model for interactive effects as proposed by Bai (2009).

The main difference between the synthetic control frameworks and the framework proposed in this paper is that here the pre-treatment outcomes are used only as tools to non-parametrically identify the ATET, whereas in the synthetic control frameworks the modelling of the pre-treatment

outcomes is also a part in estimating the counterfactual trajectory of the outcome of the treated. Secondly, there are important differences in how the frameworks utilize time-varying covariates. For example, the proposed framework can utilize time-varying covariates that do not vary across units to identify the ATET. A third difference is that none of the synthetic control methods can handle missing data in the pre-treatment outcome time-series, whereas the proposed strategy can incorporate missing data in both the pre-treatment outcomes and the time-varying covariates.[1]

There is a rich pool of time-series clustering and matching strategies available see; e.g., Fu (2011) and Aghabozorgi et al. (2015) for overviews. The focus of this literature is mainly on data mining, and finding similar sequences for predictive purposes. However, to the best of our knowledge, this literature has not been utilizing time-series matching to identify causal effects.

The theoretical results derived in this paper can be combined with other time-series models than the model used in this paper. The two-level time-series model used is chosen for several reasons; it is an intuitive, easy to use, and, for most purposes, a sufficiently sophisticated method to achieve closeness in the outcome under no treatment between the treatment and control group. In addition, the model-based time-series strategy enables sensitivity analysis of key identifying assumptions such as the unconfoundedness and overlap assumptions, discussed in detail in the following sections.

In summary, this paper specifies the assumptions under which high frequency pre-treatment measurements of the outcome can be utilized, as a substitute for, or in addition to, observed covariates, to non-parametrically identify the ATET. It is shown that when the pre-treatment data are rich enough and these assumptions hold, the heterogeneity across observational units in unobserved covariates affecting the outcome under no treatment can be derived from the pre-treatment data.

The remainder of this paper is structured as follows. Section 2 presents the notation and gives the theoretical results. Section 3 presents estimation strategies utilizing the theoretical results. Section 4 presents a small Monte Carlo simulation study aimed to illustrate a situation where the proposed framework can be a useful alternative to the synthetic control framework. Section 5 presents an empirical example with electricity consumption data. Section 6 presents a discussion and concluding remarks.

[1] Brodersen et al. (2015) can handle missing data in the outcome of a single treated group but not in the pool of non-treated units.

2 Identification of the ATET

This section formalizes the assumptions under which the ATET can be identified from the pre-treatment outcome data. Consider an observational study setting where observational units, $i = 1, ..., N$, are measured repeatedly for T consecutive time periods. Time periods $t = 1, 2, ..., t_1 - 1$ and $t = t_1, t_1 + 1, ..., T$ refer to time periods pre and post assignment of treatment, respectively. Let D_{it} be an indicator function, taking the value 0 if not treated and 1 if treated, at time period t for unit i. Once treated, $D_{it} = 1$ throughout the study. Furthermore, let $Y_{it}(d)$ be the potential outcome (Splawa-Neyman et al., 1990 translating Neyman, Rubin, 1973) at time period t where d is set to 1 if the observational unit i is given treatment at time t and 0 if not given treatment at time t. The ATET for time period t is given by

$$\text{ATET}_t = E[Y_{it}(1) - Y_{it}(0)|D_{it} = 1], \ \ \forall \ t = t_1, ..., T. \tag{1}$$

There are three assumptions commonly used to identify the ATET from observational data. Throughout this paper, if nothing else is stated, it is assumed that two of these, the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980) and the assumption of overlap holds. The third assumption, known as the weak unconfoundedness assumption (UA), is that the distributions of the potential outcomes under non-treatment are the same for the treated and non-treated (controls) given a set, $\mathbf{Z}_i^{(*)}$, of time-invariant covariates and time-varying covariates observed at specific time periods although never later than at time period t, i.e., no future observations. For example, at time t, $\mathbf{Z}_i^{(*)}$ may contain some time-invariant covariate w_i, one lagged time-varying covariate X_{it-1}^1 and another time-varying covariate at the current time period X_{it}^2. Throughout this paper, superscripts with and without parenthesis are variable-subset and scalar-variable indices, respectively, and '$^{(*)}$' is used to indicate any subset of a set that contains information from that set, sufficient for the UA to hold conditioned on this subset.

In line with Dawid (1979), the UA is defined formally as

$$Y_{it}(0) \perp\!\!\!\perp D_{it_1}|\mathbf{Z}_i^{(*)}, \ \ \forall \ t = t_1, ..., T. \tag{2}$$

This UA assumption is similar to the assumption made by Angrist and Kuersteiner (2011), but where they restrict the analysis to a single time-series. If $\mathbf{Z}_i^{(*)}$ is known and observed, the ATET can be consistently estimated under the UA by controlling for $\mathbf{Z}_i^{(*)}$, however, in general $\mathbf{Z}_i^{(*)}$ is not

known. The question of interest is what part of $\mathbf{Z}_i^{(*)}$ can be derived from pre-treatment outcome data for identification and estimation of the ATET.

To make the identifying assumptions precise, some additional notation is required. Let $\mathbf{W}_i$ be a $1 \times P$ vector containing all possible time-invariant covariates, and $\mathbf{X}_i$ be a $T \times K$ matrix containing all possible time-varying covariates at all time-periods of the study for unit i. Denote by $\mathbf{W}_i^{(p)}$ a $1 \times P_p$ vector, a subset of $\mathbf{W}_i$ where $P_p \leq P$. Let $\mathbf{X}_i^{(k)}$ be a $T \times K_k$ matrix, a subset of $\mathbf{X}_i^{(k)}$ containing $K_k \leq K$ time-varying covariates. Moreover, let $\mathbf{X}_{it}(\mathbf{L})$ denote a subset of $\mathbf{X}_i$ in terms of time periods, including only the time period t and corresponding lags, where $\mathbf{L}$ is a $K \times 1$ vector containing the number of lags (including t itself) for each of the K time-varying covariates. For example, for K=10, $K_k = 3$ and $\mathbf{L_k} = (1, 3, 0)'$, it follows that $\mathbf{X}_{it}^{(k)}(\mathbf{L_k})$ extracts $\begin{bmatrix} X_{it}^1 & X_{it-1}^1 \end{bmatrix}'$, $\begin{bmatrix} X_{it}^3 & X_{it-1}^3 & X_{it-2}^3 & X_{it-3}^3 \end{bmatrix}'$, and $\begin{bmatrix} X_{it}^{10} \end{bmatrix}$, assuming that covariates 1,3, and 10 are the three covariates chosen from the 10 possible.

Let the data generating process of the outcome under no treatment at time period t be given by

$$Y_{it}(0) = f_1\left(\mathbf{X}_{it}^{(1)}(\mathbf{L}_1), \mathbf{W}_i^{(1)}, \nu_{it}\right) \ \forall\ t = t_1, ..., T, \tag{3}$$

where f_1 is some function, and ν_{it} is independently and identically distributed (iid) for all t for each observational unit i, independent of $\mathbf{X}_{it}^{(1)}(\mathbf{L}_1)$ and $\mathbf{W}_i^{(1)}$. For this strategy to work, the length of the lags of the covariates used for identification must be bounded by what is observed, i.e., $\max(\mathbf{L}_1) < t_1$, which implies that the memory of the process is shorter than the length of the pre-treatment period. Correspondingly, let the treatment assignment mechanism at time period t_1 be given by

$$D_{it_1} = g\left(\mathbf{X}_{it_1}^{(2)}(\mathbf{L}_2), \mathbf{W}_i^{(2)}, \xi_{it_1}\right), \tag{4}$$

where g is some function, and ξ_{it_1} is iid for all t for each observational unit i, and independent of $\mathbf{X}_{it_1}^{(2)}(\mathbf{L}_2)$, and $\mathbf{W}_i^{(2)}$.

To facilitate the understanding of this paper, Assumption 1 gives a re-expression of the UA (Equation 2) in terms of the data generating process and the treatment assignment mechanism.

Assumption 1 *The outcome at time period t under no treatment for unit i is given by*

$$Y_{it}(0) = f_1\left(\mathbf{X}_{it}^{(1)}(\mathbf{L}_1), \mathbf{W}_i^{(1)}, \nu_{it}\right), \ \forall\ t = 1, ..., T,$$

such that ν_{it} *is independent of* $\{\mathbf{X}_{it}^{(1)}(\mathbf{L}_1), \mathbf{W}_i^{(1)}\}$*, and* D_{it_1} *for all i and* $t = t_1, ..., T$*.*

Assumption 1 says that the outcome process under no treatment is the same during the full study period. Furthermore, the error term is assumed independent of D_{it_1} for all $t \geq t_1$. If all covariates affecting the outcome under no treatment are observed, the ATET is clearly identified by adjusting for all these covariates. The aim of the rest of this section is to relax Assumption 1 such that the ATET can be identified also when not all covariates affecting the outcome is observed. For clarity of exposition, this is done in several steps, with the final assumption presented in Section 2.2.

Define the sets of covariates affecting both the assignment and the outcome under no treatment, $\mathbf{W}^{(s)} = \mathbf{W}^{(1)} \cap \mathbf{W}^{(2)}$ and $\mathbf{X}_{it}^{(s)}(\mathbf{L}_s) = \mathbf{X}_{it}^{(1)}(\mathbf{L}_1) \cap \mathbf{X}_{it_1}^{(2)}(\mathbf{L}_2)$. Under the Assumption 1, i.e., that ν_{it} is independent of ξ_{it_1} for all $t = t_1, ..., T$, a sufficient set, $\mathbf{Z}_i^*$, for which the UA is satisfied, is given directly by $\mathbf{Z}_i^* = \left\{\mathbf{W}_i^{(s)}, \mathbf{X}_{it}^{(s)}(\mathbf{L}_s)\right\}$. This means that it is *sufficient* to condition on the variables, and measurements thereof, that are common to the outcome during time periods $t = t_1, ..., T$ and the treatment assignment at time t_1.

Assumption 1 has strong implications for the time-varying covariates affecting both the treatment assignment and the outcome. Since any time-varying covariates in $\mathbf{X}_{it}^{(s)}(\mathbf{L}_s)$ must be controlled for also after treatment to identify the ATET, it must hold that

$$Y_{it}(0) \perp\!\!\!\perp D_{it_1} \Big| \left\{\mathbf{W}_i^{(1)}, \mathbf{X}_{it_1}^{(s)}(\mathbf{L}_s), ..., \mathbf{X}_{iT}^{(s)}(\mathbf{L}_s)\right\} \ \forall\ t = t_1, ..., T, \tag{5}$$

which implies that the ATET is only identified for all $t = t_1, ..., T$ if $\mathbf{X}_{it}^{(s)}(\mathbf{L}_s)$ is exogenous. In this context, that $\mathbf{X}_{it}^{(s)}(\mathbf{L}_s)$ is exogenous means that $\mathbf{X}_{it}^{(s)}(\mathbf{L}_s)$ cannot be causally affected by the treatment. This is well known and studied in several fields, e.g, in path analysis it is called the sequential ignorability assumption (see, e.g, Imai et al. 2010). Variants of this assumption, in the context of identifying causal effects from a single time-series, are discussed in Angrist and Kuersteiner (2011); Angrist et al. (2017); Bojinov and Shephard (2019).

Without loss of generality, the function f_1 can be parameterized by the vector $\boldsymbol{\theta}_i = h(\mathbf{W}_i^{(1)})$, where h is a vector of functions all allowed to be non-injective. The non-injectiveness, i.e., $h^{-1}(\boldsymbol{\theta}_i) \neq \mathbf{W}_i^{(1)}$ imply that the functions in h are allowed to reduce the information in $\mathbf{W}_i^{(1)}$ arbitrarily. The meaning of this reparametrization is illustated in the next subsection. Under Assumption 1, it follows directly from $\boldsymbol{\theta}_i = h(\mathbf{W}_i^{(1)})$ that

$$Y_{it_1}(0) \perp\!\!\!\perp D_{it_1} \Big| \left\{\boldsymbol{\theta_i}, \mathbf{X}_{it_1}^{(1)}(\mathbf{L}_1), ..., \mathbf{X}_{iT}^{(1)}(\mathbf{L}_1)\right\}. \tag{6}$$

Let $\boldsymbol{\theta}^{(s)}$ be the subset of $\boldsymbol{\theta}$ defined as $\boldsymbol{\theta}^{(s)} = h(\mathbf{W}^{(s)})$. It follows that

$$Y_{it_1}(0) \perp\!\!\!\perp D_{it_1} \Big| \left\{ \boldsymbol{\theta_i}^{(s)}, \mathbf{X}_{it_1}^{(s)}(\mathbf{L}_s), ..., \mathbf{X}_{iT}^{(s)}(\mathbf{L}_s) \right\}. \tag{7}$$

Finally, for the estimation it is helpful to define $\boldsymbol{\theta}_i^{(*)}$ and $\mathbf{X}_{it}^{(*)}(\mathbf{L}_*)$, two sets fulfilling $\boldsymbol{\theta}_i \supseteq \boldsymbol{\theta}_i^{(*)} \supseteq \boldsymbol{\theta}_i^{(s)}$ and $\mathbf{X}_{it}^{(1)}(\mathbf{L}_1) \supseteq \mathbf{X}_{it}^{(*)}(\mathbf{L}_*) \supseteq \mathbf{X}_{it}^{(s)}(\mathbf{L}_s)$, respectively. Since any such $\boldsymbol{\theta}_i^{(*)}$ and $\mathbf{X}_{it}^{(*)}(\mathbf{L}_*)$ are sufficient for the UA, the data generating process of the outcome under no treatment at time period t under for unit i can, without loss of generality, be rewritten as

$$Y_{it}(0) = f_* \left(\mathbf{X}_{it}^{(*)}(\mathbf{L}_*), \boldsymbol{\theta}_i^{(*)}, \nu_{it}^* \right), \ \forall \ t = 1, ..., T,$$

for any f_* such that ν_{it}^* is independent of $\{\mathbf{X}_{it}^{(*)}(\mathbf{L}_*), \boldsymbol{\theta}_i^{(*)}\}$, and D_{it_1} for all i and $t = t_1, ..., T$.

In summary, to identify the ATET for time periods $t = t_1, ..., T$, under Assumption 1, it is sufficient to condition on the parameters $\boldsymbol{\theta}_i^{(*)}$ describing the outcome time-series under no treatment and its relation to the time-varying covariates in $\mathbf{X}^{(*)}$, where the covariates in $\mathbf{X}^{(*)} \cap \mathbf{X}^{(s)}$ have to be exogenous.

2.1 Illustration of Assumption 1 and its implications

The purpose of this section is to give a better understanding of the notation and the intuition of Assumption 1. In empirical applications, the functional form of treatment assignment process and the outcome process under no treatment and their parametrizations $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}_i$ are generally not observed. Here, we give explicit functions and parametrizations to illustrate Assumption 1 in detail. In addition, to illustrate that this strategy is robust against the functional forms of h, arbitrary and purposely complex relations are chosen. Again, denote by X_{it}^q covariate q of unit i at time t. Let the outcome process under no treatment be given by

$$Y_{it}(0) = \theta_{1i} + \theta_{2i} X_{it-2}^1 + \theta_{3i} X_{it}^3 + \theta_{4i} X_{it-1}^3 + \nu_{it}, \tag{8}$$

where ν_{it} is iid, with $Var(\nu_{it}) = \theta_{5i}$, for $t = 1, ..., T$, where the parameters are given by

$$\begin{aligned}
\theta_{1i} &= \beta_{11} + \beta_{12} W_i^1 W_i^2 \\
\theta_{2i} &= \beta_{21} + \beta_{22} (W_i^2)^2 \\
\theta_{3i} &= c \\
\theta_{4i} &= \beta_{42} |W_i^3| \\
\log(\theta_{5i}) &= \beta_{51} + \beta_{52} (W_i^3)^2
\end{aligned}$$

where W_i^j's are elements of $\mathbf{W}_i$. Further assume that the treatment assignment process is given by

$$D_{t_1} = g(\gamma_{1i} + \gamma_{2i} X^1_{it_1-2} + \xi_{it_1}),$$

where ξ_{it} is iid, independent of ν_{it}, with $Var(\xi_{it}) = \gamma_{3i}$, g is some suitable link function, and the parameters are given by

$$\begin{aligned} \gamma_{1i} &= \zeta_{1,1} + \zeta_{12}(W_i^2)^2 + \zeta_{13}(W_i^4)^2 \\ \gamma_{2i} &= \zeta_{21} + \zeta_{22} e^{W_i^4} \\ \log(\gamma_{3i}) &= \zeta_{32} W_i^1. \end{aligned}$$

Note that some functions of W's, e.g. $(W_i^2)^2$, are non-injective as discussed above. That non-injectiveness is unproblematic follows directly from Equation 6, in which elements of $\mathbf{W}_i^{(1)}$ play no part. In other words, $\boldsymbol{\theta}_i$ can replace $\mathbf{W}_i^{(1)}$ regardless of the information in $\mathbf{W}_i^{(1)}$ it contains. For this example, $\mathbf{W}_i^{(1)} = \{W_i^1, W_i^2, W_i^3\}$, $\mathbf{X}_{it_1}^{(1)}(\mathbf{L}_1) = \{X_{it}^1, X_{it-1}^1, X_{it-2}^1, X_{it}^3, X_{it-1}^3\}$ with $\mathbf{L}_1 = (2,1)'$, $\boldsymbol{\theta}_i = \{\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}, \theta_{5i}\}$. Moreover, $\mathbf{W}_i^{(2)} = \{W_i^1, W_i^2, W_i^4\}$, $\mathbf{X}_{it_1}^{(2)}(\mathbf{L}_2) = \{X_{it_1}^1, X_{it_1-1}^1, X_{it_1-2}^1\}$ with $\mathbf{L}_2 = 2$, and $\boldsymbol{\gamma}_i = \{\gamma_{1i}, \gamma_{2i}, \gamma_{3i}\}$. Hence, $\mathbf{W}_i^{(s)} = \{W_i^1, W_i^2\}$, $\mathbf{X}_{it_1}^{(s)}(\mathbf{L}_s) = \{X_{it_1-2}^1\}$, and $\boldsymbol{\theta}_i^{(s)} = \{\theta_{1i}, \theta_{2i},\}$, and the minimal sufficient set[2] for the UA to hold is given by $\mathbf{Z}_i^* = \{W_i^1, W_i^2, X_{it_1-2}^1\}$, or equivalently, $\mathbf{Z}_i^* = \{\theta_{1i}, \theta_{2i}, X_{it_1-2}^1\}$.

By the results in the previous section, this implies that under Assumption 1, the following statements[3] hold for the first time period after treatment assignment, i.e., t_1

$$\begin{aligned} &Y_{it_1}(0) \perp\!\!\!\perp D_{i,t_1} | \mathbf{W}_i^{(1)}, \mathbf{X}_{it_1}^{(1)}(\mathbf{L}_1) \Leftrightarrow Y_{it_1}(0) \perp\!\!\!\perp D_{i,t_1} | W_i^1, W_i^2, W_i^3, X_{it}^1, X_{it-1}^1, X_{it-2}^1, X_{it}^3, X_{it-1}^3) \\ &Y_{it_1}(0) \perp\!\!\!\perp D_{i,t_1} | \mathbf{W}_i^{(s)}, \mathbf{X}_{it_1}^{(s)}(\mathbf{L}_s) \Leftrightarrow Y_{it_1}(0) \perp\!\!\!\perp D_{it_1} | W_i^1, W_i^2, X_{it_1-2}^1 \\ &Y_{it_1}(0) \perp\!\!\!\perp D_{i,t_1} | \boldsymbol{\theta}_i^{(s)}, \mathbf{X}_{it_1}^{(s)}(\mathbf{L}_s) \Leftrightarrow Y_{it_1}(0) \perp\!\!\!\perp D_{it_1} | \theta_{1i}, \theta_{2i}, X_{it_1-2}^1, \end{aligned} \tag{9}$$

and, e.g.,

$$Y_{it_1}(0) \perp\!\!\!\perp D_{i,t_1} | \boldsymbol{\theta}_i^{(*)}, \mathbf{X}_{it_1}^{(*)}(\mathbf{L}_*) \Leftrightarrow Y_{it_1}(0) \perp\!\!\!\perp D_{it_1} | \theta_{1i}, \theta_{2i}, \theta_{3i}, X_{it_1-2}^1, X_{it_1}^3.$$

In the following section we address the challenge of identifying the ATET when neither $\mathbf{W}^{(s)}$ or $\boldsymbol{\theta}_i^{(s)}$ are observed, by specifying the assumptions under which $\boldsymbol{\theta}_i^{(*)}$ and the ATET can be identified and estimated from pre-treatment data simultaneously.

[2]The minimal sufficient set for the UA to hold is known here because the data generating processes are known.

[3]The conditional independence will hold for any superset of the minimal sufficient set.

2.2 Deriving $\boldsymbol{\theta}_i^{(*)}$ from data for identification of the ATET

Let $\mathbf{Y}_{it}(L_y, 0) = (Y_{it-1}(0), Y_{it-2}(0), ..., Y_{it-L_y}(0))$ and $\mathbf{X}_i^{(obs)}$ be the set of observed time-varying covariates.

Assumption 2 *The distributions of the covariates in* $\mathbf{X}^{(obs)}$ *are the same under treatment and no treatment for each* $t = 1, ..., T$. *Furthermore, the outcome process under no treatment can be expressed as*

$$Y_{it}(0) = f_* \left(\mathbf{Y}_{it}(L_y, 0), \mathbf{X}_{it}^{(obs)}(\mathbf{L}_{obs}), \boldsymbol{\theta}_i^*, \epsilon_{it} \right) \; \forall \, t = 1, ..., T, i = 1, ..., N$$

such that

$$\epsilon_{it} \perp\!\!\!\perp D_{it_1} \forall \, t = 1, ..., T$$

and

$$Y_{it}(0) \perp\!\!\!\perp D_{it_1} \Big| \left\{ \boldsymbol{\theta}_i^*, \mathbf{X}_{it_1}^{(obs)}(\mathbf{L}_{obs}), \mathbf{X}_{it_1+1}^{(obs)}(\mathbf{L}_{obs}), ..., \mathbf{X}_{iT}^{(obs)}(\mathbf{L}_{obs}) \right\} \; \forall \; t = 1, ..., T.$$

Assumption 2 states that the outcome can be modeled as a time series process of lags and observed time-varying covariates with error terms independent of the treatment assignment at all t. Assumption 2 further implies that the lags of the outcome are sufficiently good proxies for unobserved time-varying covariates, $\mathbf{X}_{it}^{(unobs)}$, to identify $\boldsymbol{\theta}_i^*$, where $\mathbf{X}_{it}^{(unobs)} \cup \mathbf{X}_{it}^{(obs)} = \mathbf{X}_{it}^{(*)}$ and $\mathbf{X}_{it}^{(unobs)} \cap \mathbf{X}_{it}^{(obs)} = \emptyset$. This is a relaxation of Assumption 1 since this implies that potentially *all* time-invariant covariates in $\mathbf{W}^{(1)}$ and some time-varying covariates in $\mathbf{X}^{(1)}$ may be unobserved. However, there are two restrictions on the relaxation for the time-varying covariates. First, under Assumption 2, $\mathbf{X}^{(obs)} \supseteq \mathbf{X}^{(s)}$ must hold to enable necessary conditioning in the post period. Second, there are requirements for $\mathbf{Y}_{it}(L_y, 0)$ to be a relevant proxy for $\mathbf{X}_{it}^{(unobs)}$. The first requirement, that the lags, $\mathbf{Y}_{it}(L_y, 0)$ must be irrelevant for the treatment assignment conditioned on $\mathbf{X}_{it}^{(obs)}$, is implied by Assumption 2. However, it must also hold that

$$\mathbf{X}_{it}^{(unobs)} \not\perp\!\!\!\perp \mathbf{Y}_{it}(L_y, 0) | \mathbf{X}_{it}^{(obs)}, \boldsymbol{\theta}_i^* \Leftrightarrow \mathbf{X}_{it}^{(unobs)} \not\perp\!\!\!\perp \mathbf{Y}_{it}(L_y, 0) | \mathbf{X}_{it}^{(obs)}.$$

That is, $\boldsymbol{\theta}_i^*$ cannot predict $\mathbf{X}_{it}^{(unobs)}$ conditional on $\mathbf{X}_{it}^{(obs)}$ and $\mathbf{Y}_{it}(L_y, 0)$, and, furthermore the lagged outcomes need to be relevant for $\mathbf{X}_{it}^{(unobs)}$ conditional on $\mathbf{X}_{it}^{(obs)}$ (see, e.g., Wooldridge 2010; de Luna et al. 2017 for the definitions of proxy variables in the treatment effect literature), which essentially means that $\mathbf{X}_{it}^{(unobs)}$ must be dependent over time. Note that if there are no time-varying

covariates affecting the outcome and the treatment assignment then the sufficient set $\boldsymbol{\theta}_i^{(*)}$ can be identified from flexible time-series specification of f^* only. The identification strategy is closely connected to the indirect inference framework (see, e.g., Gourieroux et al. 1993; Smith 2016). The parameters, $\boldsymbol{\theta}^{(*)}$ of a possibly misspecified model are assumed to contain information about the parameters $\boldsymbol{\theta}^{(s)}$. Like in the indirect inference setting, the dimension of $\boldsymbol{\theta}_i^{(*)}$ may be larger than of the set $\boldsymbol{\theta}_i^{(s)}$.

The assumption $\epsilon_{it} \perp\!\!\!\perp D_{it_1} \forall\, t = 1, ..., T$ is a strong assumption, present in one form or another in all causal-effect identification from observational data. However, if the number of parameters estimated to derive $\boldsymbol{\theta}_i^*$ is less than the number of pre-treatment observations, model checks and sensitivity analysis for the plausibility of this assumption are possible. This is discussed in more detail in Section 3.3 and Appendix A.

The first sentence in Assumption 2 states that the distributions of the covariates in $\mathbf{X}^{(obs)}$ are the same under treatment and no treatment for each $t = 1, ..., T$. This is important for the identification of $\boldsymbol{\theta}_i^*$: If, e.g, the variation in the covariates in the pre-treatment period is much smaller than in the post-treatment period, or if the variation in the time-varying covariates differs among the treated and controls, it may not be possible to correctly identify the relation between unobserved parts of $\mathbf{W}^{(1)}$ and the post-treatment outcome under no treatment. This would in turn mean that the derived $\boldsymbol{\theta}_i^*$ is not sufficient for the identification of the ATET estimand. However, as the time-varying covariates are observed, the 'overlap' in variation of the time-varying covariates in the pre and post periods, as well as between the treated and non-treated, can be empirically evaluated.

3 Estimation strategies

Moving from identification to estimation, we suggest the following strategy: Fit a two-level time-series model to the pre-treatment data and match on $\hat{\boldsymbol{\theta}}_i^*$ and any observed time-invariant covariates, using a suitable matching strategy and a regression estimator to adjust for possible time-varying covariates. For inference, the standard errors of Abadie and Imbens (2006, 2011) can be applied.

As discussed above, to estimate ATET consistently, $\mathbf{X}^{(s)}$ must be conditioned on. If none of the covariates in $\mathbf{X}^{(1)}$ vary across units (i.e. $\mathbf{X}_{it}^{(1)} = \mathbf{X}_{i}^{(1)} \forall\, t$), it directly follows that $\mathbf{X}^{(s)} = \emptyset$ and the time-varying covariates can be excluded. However, in addition to the matching on $\boldsymbol{\theta}_i^*$, all time-

varying covariates in $\mathbf{X}^{(obs)}$ believed to be in $\mathbf{X}^{(s)}$ and varying across units should be controlled for explicitly in estimation. This can be done by assuming a functional form in the relationship between the outcome and the covariates guided by the estimates from the two-level model. Importantly, this implies that two units with similar $\hat{\boldsymbol{\theta}}_i^*$ may be matched even if they do not have the same X_{it}^q's for any time-periods, as long as they have 'overlap' in the variation in X_{it}^q, as discussed in the previous section. In Section 4, this important feature is studied using Monte Carlo simulations.

The identification strategy based on $\boldsymbol{\theta}_i^*$ does not impose a particular ATET estimator, and there are several ways of estimating the ATET consistently based on $\boldsymbol{\theta}_i^*$. There are many alternatives for matching, not further discussed here, see Rosenbaum (2019) for a recent overview. For general discussions on inference for matching estimators, see Bodory et al. (2018); Iacus et al. (2019). Alternatively, one can use a completely Bayesian inference approach, estimating a DSEM on the full data set, i.e., pre and post-treatment, explicitly modelling all the counterfactuals[4], or use a difference-in-differences estimator. Both these alternative identification and estimation strategies make assumptions of functional forms in different ways. For this reason we are inclined to instead use matching estimators, which makes it possible to non-parametrically identify the ATET under assumption that can be evaluated from pre-treatment data, without consulting post-treatment data.

3.1 Estimating $\boldsymbol{\theta}^*$

The key to the proposed strategy is to be able to estimate $\boldsymbol{\theta}_i^*$ from pre-treatment data. Several considerations have to be made in the estimation of $\boldsymbol{\theta}_i^*$. The most obvious challenge is to condense the information in the pre-treatment measurements. For example, the naïve estimator $\hat{\boldsymbol{\theta}}_i^* = \{Y_{i1}, ..., Y_{it_1-1}\}$ could be used but would require a unreasonably large pool of controls for close match for each time-period. Moreover, if the outcome is a function of exogenous time-varying covariates, similarity in the pre-treatment outcome is not guaranteed to imply similarity in the post-period under no treatment. This would be the case if, e.g., the relation between the exogenous variables of the units is not constant over time, a setting addressed in the Monte Carlo study in Section 4.

This paper instead suggests approximating the function f^* (Assumption 2) with a parametric time-series model for each observational unit. Once a time-series is fitted to an observational unit's pre-treatment data, the usual time-series tools to evaluate fit can be used to guide model

[4]This corresponds to estimating the Average Treatment Effect rather than the ATET.

adjustments. Even if the fit is poor for some observational units, the parameter estimates might give a sufficiently detailed description of the heterogeneity in the processes to achieve balance in the unobserved time-invariant covariates. A parametric time-series model can directly handle the challenge of time-varying covariates. The time-series model can be specified to include time-varying covariates observed at the observational unit level. However, also variables that do not vary across observational units but only over time can be included to improve the description of the heterogeneity across units. That is, although the time-varying covariate itself is common to all units, the unit-specific parameter moderating the relation between this covariate and the units' outcome may be an important part of $\boldsymbol{\theta}_i^*$. In the empirical electricity consumption example (Section 5), an example of such a variable is temperature which is common to observational units located in the same city. There, even though all units have the same temperature, the unit-specific temperature *dependency* is a potentially important part of $\boldsymbol{\theta}_i^*$. This means that any time-varying covariates believed to affect the outcome process can, and should, be included, to improve the estimate of $\boldsymbol{\theta}_i^*$. Including important time-varying covariates also makes Assumption 2 more realistic as less information is approximated by lags of Y. Appendix B gives a simple illustration of the potential gain of including parameters describing the dynamic aspects of the outcome processes in $\boldsymbol{\theta}_i^*$.

3.2 Multivariate two-level time-series approach

In this section a strategy for estimating $\boldsymbol{\theta}_i^*$ under Assumption 2 is proposed, using the novel Dynamic Structural Equation Modelling (DSEM) framework (Asparouhov et al., 2018) available in Mplus version 8 (Muthén and Muthén, 2017). The DSEM framework is a general multivariate two-level time-series modelling framework with time on level-1 and observational units on level-2. Bayesian MCMC estimation is employed to accommodate many random effects. Non-informative priors are used. For the purposes of this paper, some especially practical DSEM features are the ability to include time-varying covariates with measurement frequencies different from that of the outcome, being able to fit random effects of time-varying covariates that do not vary across observational units, and being able to allow for observational unit-specific auto-regressive coefficients and residual variances. Since zero estimates in unit-specific parameters are allowed for any observational unit, fitting one DSEM can encompass units with various different orders of VARMA models in one estimation since all restricted models nested in the fitted model are encompassed. Combined, this means that DSEM helps to reduce the amount of information in the pre-treatment period by

allowing for similarity and dissimilarity across units in a natural way. Appendix C.1 presents a simple example illustrating how a slightly over fitted two-level time-series model can distinguish between observational units with different orders of auto-regressive outcome processes. DSEM can also handle missing data and unbalanced panels of time-series, which implies that units can enter the study at different time periods. For details about missing data handling and sample size requirements in terms of N and T, see Appendix C.

One possible objection to the two-level time-series approach is that the parametric assumptions of the distributions of the random coefficients pre-supposes similarity that is imposed on the estimates. Although this objection can be valid, at least for short time-series, this problem will in such cases show up as lack of balance in the pre-treatment outcome. That is, if the time-series are so short that the distributional assumption of the random coefficients forces the estimates to be too similar, to a degree where the heterogeneity is underestimated, it will be apparent. The following section discuss how to do sensitivity analysis to detect such problems.

3.3 Model specification and sensitivity analysis

Several of the identifying assumptions can be checked to validate the plausibility of the identification strategy. SUTVA cannot be checked and must follow from the context or design of the study. The overlap assumption, i.e., that there exist units similar to the treated units in the pool of controls can be evaluated. Traditionally, this assumption is evaluated by checking overlap in $\mathbf{W}^{obs}$. Since we have substituted time-invariant covariates with $\boldsymbol{\theta}_i^*$ the corresponding evaluation is to check overlap in $\hat{\boldsymbol{\theta}}_i^*$. This can be thought of as a loose analogue to using the estimated propensity score (PS) to check overlap with a large number of covariates. In the standard setting the treatment assignment process is summarized by the estimated PS, here the time-invariant characteristics governing the pre-treatment outcome process are summarized by the estimated $\hat{\boldsymbol{\theta}}_i^*$. As mentioned in Section 2.2, the assumptions for the identification of $\boldsymbol{\theta}_i^*$ can also be checked by comparing the distribution of unit specific time-varying covariates. It is important to understand that if the assumptions for identification of $\boldsymbol{\theta}_i^*$ do not hold, checking the overlap in $\hat{\boldsymbol{\theta}}_i^*$ is futile, as overlap in $\hat{\boldsymbol{\theta}}_i^*$ does not imply overlap in $\boldsymbol{\theta}_i^*$ in that case.

Finally, under Assumption 2, sensitivity analyses of the unconfoundedness assumption can be performed. Given the final sample of treated and matched controls the distributions of the error terms, $e_{it} = Y_{it} - \hat{Y}_{it}$, for the treated and controls can be studied for all time periods $t = 1, .., t_1 - 1$,

where $\hat{Y}_{it}^0$ is the fitted value. If the pre-treatment period is long enough, the period can be split into a train and a test period: sensitivity analysis can be performed by calculating the residuals for the second period from the model estimated on the first period. However, even if the full pre-treatment period is used in the design, as long as the number of parameters is smaller than the number of observation, a less rigorous sensitivity analysis is still possible. We propose using a simplified sensitivity analysis by restricting the evaluation to the differences in means, i.e, study the balance of $\frac{1}{N_1}\sum_{i:D_i=1} e_{it}$ and $\frac{1}{N_0}\sum_{i:D_i=0} e_{it}$, where $N = N_1 + N_0$, for all time periods by comparing the point-wise standardized mean difference to the rule of thumb of being smaller than ± 0.25 (Imbens and Wooldridge, 2009), illustrated in the empirical example in Section 5. If the estimates of $\boldsymbol{\theta}_i^*$ do not contain at least as much information about the heterogeneity in the processes as $\boldsymbol{\theta}_i$, it is implausible that balance would be obtained for all pre-treatment time periods. There are many alternatives for model checking and sensitivity analysis; for additional suggestions and a simulated illustration, see Appendix A.

4 Monte Carlo Simulation - Comparison with existing strategies

In this section, the proposed identification strategy is compared to different versions of the synthetic control method. With time-series data on pre-treatment outcomes, synthetic control methods are standard tools for identifying causal effects, which makes it a reasonable comparison. The purpose is to give an example of a setting in which the proposed strategy is advantageous and illustrate and explain why. The focus of the simulations is on the ability to incorporate and control for time-varying covariates. As explained above, the proposed strategy allows for explicit matching on the relation between observed time-varying covariates (unit-specific or common) and the outcome under no treatment, which enables units with similar behaviour to be matched even if they have different values on such covariates at any given time-period. The identification strategy will therefore hold even if the relation between the time-varying covariates of the units change over time, as long as the relation between the time-varying covariate and the outcome remains constant for each unit, and the time-varying covariate is controlled for in the post-period.

The following data generating process is considered in the simulation. The outcome of interest

is given by

$$Y_{it} = \mu_i + \eta Y_{it-1} + \phi_i X_{it}^1 + \epsilon_{it}, \tag{10}$$

for $t = 1, ..., T$, where $\mu_i \sim N(0, \sigma_\mu^2)$, $\phi_i \sim N(0, \sigma_\phi^2)$, $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$, and

$$X_{it}^1 = \phi_2 X_{it-1}^1 + \mathbb{1}(t \geq t_1)\gamma_i + \zeta_{it}, \tag{11}$$

where $\mathbb{1}(t > t_1)$ is a indicator function taking the value one if t is larger than or equal to t_1, $\gamma_i \sim N(0, \sigma_\gamma^2)$, and $\zeta_{it} \sim N(0, \sigma_\zeta^2)$. Note that model is a special case of DSEM. In the DSEM framework, Equation 10 is the level-1 equation, also called within-unit level, and the distributions of μ_i and ϕ_i are modeled in the level-2 system of equations, also called the between-unit level (Asparouhov et al., 2018).

Here we consider the $T/2 = t_1$ first time periods to be pre-treatment periods, and the second half as post-treatment. Both the outcome and the time-varying covariate are observed during the pre-treatment period. The indicator function creates a change in the relation between the time-varying covariates of the observational units. Since the shift in X_{it}^1 is independent of everything else and has expected value zero, X remains exogenous wherefore Assumption 2 holds. For simplicity, we let the treatment effect be zero for all treated units, and measure the deviation from zero of the point-wise ATET estimate in the post-treatment period to evaluate how similar the treatment and control groups are.

It is an extreme case as all shifts in X_{it}^1 occur at the same time as treatment. However, this is chosen to illustrate the challenge that exogenous time-varying covariates may pose for identifying the ATET. The severity of the shift would be smaller if it happened at different times for different units, but the issue would remain unless the relations between the time-varying covariates are constant for the full study time.[5]

Two total sample sizes are considered: 120 and 400. In both cases the control group is three times larger than the treated (30,90 and 100, 300, respectively). Time-series of two different lengths are considered, 100 and 728. The 728 (2*364) is used to mimic the setting of the empirical example in Section 5. For each setting, 500 replications with independent samples are used.

[5] A simulation study with random time periods for the shifts of each units time-varying covariate was conducted to confirm this, these results are not included in the paper.

4.1 Strategies

Table 1 presents the factors and levels of the simulation. All strategies were implemented in R using default settings if nothing else is stated. For the DSEM estimation Mplus version 8.3 was called via the MplusAutomation R package (Hallquist and Wiley, 2018). The details of the strategies are the following: Two-level time-series matching (TLTM) is the strategy proposed in this paper where a DSEM model according to Equation 10 is fitted on the full sample, and caliper matching using the R-package MatchIt (King and Stuart, 2011) is used to find matching controls. The X_{it}^1 dependency is removed from the post-period outcomes using the unit-specific $\hat{\phi}_i$ parameters before the group difference is calculated. The original synthetic control (SCORG) is the strategy proposed by Abadie et al. (2010) applied to multiple treated units as proposed in Abadie et al. (2015), implemented using the R-package Synth (Abadie et al., 2015). That is, the outcome and X_{it}^1 of the treated units are aggregated and a synthetic control is then found for the whole treated group at once. The covariates, i.e, the aggregated X_{it}^1 for the treated and the unit-specific X_{it}^1 for the controls, are added as a time-varying covariate. Causal Impact (CI) it the strategy proposed by Brodersen et al. (2015) implemented using the R-package CausalImpact described in the same paper. This strategy was originally proposed for a single treated unit, however, as it is an interesting extension of the standard synthetic control it is added here using the same aggregated treatment-group setting as for SCORG. This package allows for time-varying covariates for the single treated unit, wherefore we used the temperature aggregated over the treated units as a time-varying covariate for the treated group. The package can not include time-varying covariates for the pool of controls. Robbins et al. (2017) proposed a framework for synthetic controls for microdata, implemented using the R-package MicroSynth (Robbins and Steven, 2019). This package creates synthetic controls for multiple treated units without aggregating. Due to convergence issues, only every 5th time period are used with this package. This package allows for including time-varying covariates in two ways, the full time-series (SC), and the time-invariant unit aggregate of the time-varying covariate (SCAG). The generalized synthetic control (GCS) is the strategy proposed by Xu (2017), implemented using the R-package gsynth (Xu and Liu, 2018). The GSC combines a standard fixed effects panel model with the synthetic control framework and a large N and T factor model for interactive effects as proposed by Bai (2009). A two-way fixed effect is fitted with X_{it}^1 as a regressor and the factor model is fitted with 0-5 factors using crossvalidation.

Table 1: Factors and levels of the Monte Carlo simulation study.

Factor	Levels
Strategy	TLTM, CI, SC, SCAG, SCORG, GSC
$N_1 + N_0$	30+90, 100+300
T	100, 728
σ_ϕ^2	0.5, 3

4.2 Results

The results are evaluated in terms of the MSE of the ATET estimate over the post-treatment period. Table 2 displays the mean and standard deviation of the MSE over all factors and levels. In some settings, some synthetic control strategies did not converge. The number of the 500 samples for which the strategies successfully found a design is reported in square brackets of each cell. The small number of convergent replications for some strategies it due to the setting not fitting the proposed strategy. They are included here to illustrate the limitations of the methods with large samples in terms of units and time-periods, and multiple treated units.

With large variation in ϕ_i over units, i.e., $\sigma_\phi^2 = 3$ (bottom half of Table 2), all strategies have substantial MSE's as compared to TLTM. As expected, for all strategies the mean MSE decreases with increasing N, as when N is larger the group mean of the X_{it}^1 shift is closer to zero by the law of large numbers. The strategies SC and SCAG have very large MSE in all settings with $\sigma_\phi^2 = 3$. When the heterogeneity in the dependency in X_{it}^1 is smaller ($\sigma_\phi^2 = 0.5$), the MSE decreases for all strategies. For TLTM, CI, GSC, and SCORG, all have MSE close to zero for all settings, except when both N and T are small, where only TLTM achieves close to zero MSE. SC and SCAG are again outperformed by all other strategies.

In settings where there is large variation in the units' relation between the outcome and the exogenous time-varying covariate, it is important to make the groups similar in the relation itself (ϕ_i) and then control for such covariates (X_{it}^1), rather than to make the groups similar in the raw unconditional outcome ($Y_{it}(0)$). That is, similarity in outcomes does not strictly imply similarity in $\boldsymbol{\theta}_i$. Similarity in raw pre-treatment outcome data can occur naturally in three ways here, (1) similarity in both μ_i and ϕ_i with similar X_{it} for all t, (2) dissimilarity in μ_i and ϕ_i with dissimilar X_{it} that balances the outcomes, and (3) similarity in X_{it} but dissimilarity in μ_i and ϕ_i such that the

Table 2: Mean, (standard deviation), and [number of succesful replications] of the average pointwise MSE for all strategies under different N, T and time-invariant covariate dependency. Empty cells indicate that none of the 500 replicates found a solution without errors.

			$\sigma^2_\phi = 0.5$				
N	T	TLTM	CI	GSC	SCORG	SC	SCAG
120	100	.04 (.03)[500]	1.8 (2.4)[500]	8.1 (11.9)[52]	1.3 (1.6)[500]	36.4 (51.8)[268]	29.4 (39.2)[500]
120	728	.12 (.06)[500]	.21 (.25)[500]	.31 (.37)[180]	.34 (.40)[500]	-(-)[0]	- (-)[0]
400	100	.04 (.02)[500]	.07 (.08)[500]	1.1 (.36)[3]	.1 (.12)[500]	1.6 (2.08)[255]	1.3 (1.67)[500]
400	728	.03 (.02)[500]	1.8 (2.5)[500]	2.6 (3.5)[189]	1.2 (1.5)[500]	52.9 (77.6)[100]	46.5 (64.77)[192]
			$\sigma^2_\phi = 3$				
120	100	.15 (.09)[500]	6.2 (8.3)[500]	15.6 (23.5)[176]	4.1 (5.0)[500]	174.2 (310)[252]	116.9 (180)[499]
120	728	.12 (.06)[500]	5.7 (8.1)[500]	7.7 (9.7)[263]	4.6 (5.6)[500]	- (-)[0]	- (-)[0]
400	100	.04 (.03)[500]	1.8 (2.4)[500]	8.1 (11.9)[52]	1.3 (1.6)[500]	36.4 (51.8)[268]	29.4 (39.16)[500]
400	728	.03 (.02)[500]	1.7 (2.5)[500]	2.6 (3.5)[189]	1.2 (1.5)[500]	52.9 (77.6)[100]	46.5 (64.77)[192]

Note: TLTM=Two Level Time-series Matching, CI=CausalImpact (Brodersen et al., 2015), GSC=Generalized Synthetic Control (Xu and Liu, 2018), SCORG = Synthetic control (Abadie et al., 2015), SC=Synthetic control non-aggregated time-varying covariates (Robbins and Steven, 2019), and SCAG=Syntentic Control aggregated time-varying covariates (Robbins and Steven, 2019).

outcomes are balanced. Only treatment and control groups that are similar in the first way would imply identification of the ATET, and even then, the time-varying covariates must be controlled for in the post-period, or be constant across units. To illustrate, under the data generating process in Equation 10, two units i and j with very similar Y under the same X^1, may be qualitatively very different from each other. For example, $X_{it} = X_{jt} = 2$, $\mu_i = 1$, $\phi_i = 4$, $\mu_j = 8$, and $\phi_j = 0.5$ would give the same outcome, i.e., $Y_{it} = 1 + 4 \times 2 = 8 + 0.5 \times 2 = Y_{jt}$. The proposed strategy aims to make the matched control group as similar to the treated group as possible in terms of time-invariant characteristics ($\boldsymbol{\theta}^*_i$) which includes relationships between time-varying covariates and outcome, which ultimately makes the groups similar in the outcome conditioned on the time-varying covariates.

5 Empirical example with electricity consumption data

In the motivating example for this paper, taken from Öhrlund et al. (2019), the focus is on the effects of a dynamic tariff in contrast to a flat-fee price tariff on electricity grid fee for firms. The aim of a dynamic price tariff instead of a flat fee is to lower the peaks in the grid. High peaks are associated with high costs for the company supplying the grid. This specific dynamic tariff has costs proportional to the customers' highest peak of consumption during each month, whereas the flat fee tariff is based on the total kWh usage each month. By reducing each firm's highest peaks, the grid-supplying company aims to lower the cumulative peaks in the grid.

This example uses data from a specific company supplying electricity grid to firms in the cities of Sandviken and Sundsvall in Sweden. All the firms had the flat fee tariff in 2014. The dynamic tariff was introduced to all firms in Sandviken in 2015 but a flat fee remained in Sundsvall. There are 212 firms in Sandviken and 1055 in Sundsvall, which means that 16.7% of the sample is in the treated group. 157 firms in Sandviken and Sundsvall were excluded from the sample due to lack of variation or too large amount of pre-treatment periods with zero consumption. In total, 1110 firms from both cities are included in the matching. In the final effect estimation, 184 of the firms in Sandviken were included and the matched control group consisted of 140 firms from Sundsvall. The daily electricity consumption is observed for all firms from 2014 up to and including 2016, three years in total. All the details of this study can be found in Öhrlund et al. (2019).

There are three important aspects to account for when comparing electricity consumption of firms. The first, most obvious aspect is outdoor temperature. It is important since heating is one of the largest sources of electricity consumption. The second aspect is the price of electricity. The treatment in this case concerns the price of the grid, but the price of the watt hours is unregulated by the treatment. Finally, the third aspect that must be accounted for is changing market conditions, including recessions and changes in demand etc. The city of Sundsvall was chosen as a control group to Sandviken with these aspects in mind. The temperature difference is small since the cities are closely located (200km)[6]. Regarding the market, the population size of Sundvall is around two times as large as Sandviken, however, the two cities have similar industry structure, are part of the same region, and have access to the same electricity market.

[6]The small remaining difference is addressed in the estimation section below.

5.1 Identification and Estimation

The parameter of interest is the ATET. The SUTVA should hold, because there is only one form of the treatment, and, given that firms are cost minimizing, it is not likely that the firms' electricity consumption is affected by other firms' treatment status. With regards to overlap, this evaluation is a special case as the assignment is at the city level with no overlap in location. This means that it has to be assumed that a firm located in Sandviken could have been located in Sundsvall, given all possible values of all potential confounders.

The fact that the firms cannot choose to be treated makes Assumption 2 more plausible. However, since the treatment is at the location level, any time-varying covariates that vary between these two locations should be controlled for. The cities are chosen as they are close to each other and share most of the environmental factors such as electricity price and recessions. One thing that might be slightly different is the outdoor temperature in the two cities, as temperature is not independent of treatment-assignment (location) and temperature affects the outcome, $\mathbf{X}_{it}^{(1)} \cap \mathbf{X}_{it}^{(2)} = Temp_{jt}$ where j =(Sandviken, Sundsvall). Because temperature cannot be causally affected by the treatment assignment it follows that $Temp_{jt}(0) = Temp_{jt}(1)$ for all j and t. This means that under Assumption 2, it is sufficient to condition on $\mathbf{W}^{(1)} \cap \mathbf{W}^{(2)}$, which likely contains information about heating systems, type of firm, insulation, facilities, etc., and the temperature lags to identify the ATET for all post-treatment time periods. In line with Equation 6, it is also sufficient to condition on $\boldsymbol{\theta}_i^*$, discussed in detail below, and the temperature lags. That is, Assumption 2 may hold in this setting even if $\mathbf{W}^{(s)}$ is unknown as long as sufficient information about the time-invariant characteristics that govern the outcome process under no treatment, $\boldsymbol{\theta}_i^*$, can be estimated from the pre-treatment outcome and temperature data.

To consistently estimate the ATET under Assumption 2 and assert the overlap assumption, a propensity-based caliper matching estimator is used. Firms in Sundsvall that are similar to firms in Sandviken are selected to construct a matched control group, and the electricity consumption of this matched control group and Sandviken are compared. The observed firm characteristics are few: two variables measuring the ampere dimension of the power subscription (Amp50 and Amp60) and one variable for the estimated difference in cost if the firm does not change their behaviour under

the dynamic tariff given by

$$\Delta_{cost,i} = \frac{\text{Yearly distribution tariff cost during the pre-treatment period|Demand-based tariff(new)}}{\text{Yearly distribution tariff cost during the pre-treatment period|Energy-based tariff(old)}}. \tag{12}$$

This is clearly a small set of characteristics: It lacks important factors for electricity consumption such as machine park, size of facilities, number of employees, heating system, etc.. However, since the daily electricity consumption and temperature data for the full pre-treatment year of the study are available for all firms in the study, it should be possible to estimate $\boldsymbol{\theta}_i^*$ with high accuracy. The estimation of $\boldsymbol{\theta}_i^*$ is done in several steps: First, a rich DSEM with a large set of substantively motivated random coefficients is estimated. If, after matching on the unit-specific parameter estimates from this model, Assumption 2 is indicated to be fulfilled in the sensitivity analysis, then this indicates that the set of random coefficients makes up a sufficient distillation of the characteristics. If not, a different DSEM can be fitted where random coefficients are added or removed according to expertise or model-fit measures.

The parameter vector $\boldsymbol{\theta}_i^*$ is estimated using the two-level time-series approach discussed in Section 3.2. After fitting and refitting, balance in outcome was achieved in all pre-treatment time periods. The final within-firm time-series model used for the electricity consumption under no treatment was given by

$$Y_{it} = \mu_i + \phi_{kWh,i} Y_{it-1} + \phi_{Temp,i} Temp_{j,t} + \phi_{Temp-1} Temp_{j,t-1} + \boldsymbol{\gamma}'_{i,month} \boldsymbol{M}_t + \epsilon_{it}, \tag{13}$$

for $t = 1, ...t_1 - 1$, where $\epsilon_{it} \sim N(0, \sigma_i^2)$, Y is the observed daily kWh consumption of electricity under the flat free, $\mathbf{M_t}$ is a vector of time-varying dummies for month, and $Temp_t$ is the outdoor temperature at day t in city j = (Sandviken, Sundsvall). The vector $\left(\mu_i, \phi_{kWh,i}, \phi_{Temp,i}, \phi_{Temp-1}, \boldsymbol{\gamma}'_{i,month}, \sigma_i^2\right)$ contains all random, i.e., firm-specific coefficients which are modeled on level-2. The final level-2

model is given by

$$
\begin{aligned}
\mu_i &\sim N(\mu_\mu, \sigma^2_\mu) \\
\phi_{kWh,i} &\sim N(\mu_{\phi_{kWh}}, \sigma^2_{\phi_{kWh}}) \\
\phi_{Temp,i} &\sim N(\mu_{\phi_{Te}}, \sigma^2_{\phi_{Te}}) \\
\phi_{Temp-1,i} &\sim N(\mu_{\phi_{Te-1}}, \sigma^2_{\phi_{Te-1}}) \\
\log(\sigma^2_i) &\sim N(\mu_{\log\sigma^2}, \sigma^2_{\log\sigma^2}) \\
\boldsymbol{\gamma_{ik}} &\sim N(\mu_{\gamma_k}, \sigma^2_{\gamma_k}), k = 1, ..., 11 \\
\text{Probit}(D_i) &= \beta_0 + \beta_1\mu_i + \beta_2\phi_{kWh,i} + \beta_3\phi_{Temp,i} + \\
&\quad \beta_4 \log(\sigma^2_i) + \boldsymbol{\gamma}'_i\boldsymbol{\beta} + \beta_{16}\Delta_{\text{cost},i} + \beta_{17}\text{Amp50}_{\ i} + \beta_{18}\text{Amp60}_i,
\end{aligned}
\tag{14}
$$

where $\boldsymbol{\gamma}_i$ is the 1×11 vector of the month-dummy estimates and $\boldsymbol{\beta} = (\beta_5, ..., \beta_{15})^T$. Again, to be explicit, $\boldsymbol{\theta}^*_i = (\mu_i, \phi_{kWh,i}, \phi_{Temp,i}, \phi_{Temp-1,i}, \log(\sigma^2_i), \gamma_{i1}, \gamma_{i2}, ..., \gamma_{i11}, Amp50_i, Amp60_i, \Delta_{\text{cost},i})$[7], and the corresponding estimates are used as a substitute for the, in large part unobserved, sufficient set of covariates $\mathbf{W}^{(1)} \cap \mathbf{W}^{(2)}$. Here, since the propensity score (PS) based caliper matching is used, the PS is estimated directly on level-2 using the probit link for the treatment assignment, Probit(D_i). Electricity consumption at day t is modeled as a function of the consumption the previous day, the temperature that day, and the month. The large number of parameters capturing potential heterogeneity in the temperature dependency is motivated by the often large effect outdoor temperature has on electricity consumption. For Mplus-estimation output and interpretations of key parameters, see Appendix C.

5.2 Sensitivity analysis

As discussed in Section 3.3, we make sure that the assumptions for the identification of $\boldsymbol{\theta}^*_i$ are fulfilled before checking overlap in the covariates or the PS. We do so by checking that the temperatures in the cities have similar range and variability, which they do. Figure 1 displays the overlap in the propensity of being located in Sandviken. Clearly there are firms in Sundsvall with similar propensity to those in Sandviken. However, the opposite is not true which is not a problem since the estimand of interest is the ATET. By using caliper matching based on the estimated propensity

[7]Note that, e.g, $Amp50_i$ is allowed in $\boldsymbol{\theta}^*_i$ as the vector of functions h, discussed in Section 2, is allowed to contain the identity function.

score and $\hat{\boldsymbol{\theta}}_i^*$, the overlap will be fulfilled as firms in Sundsvall with non-overlapping propensities will be excluded.

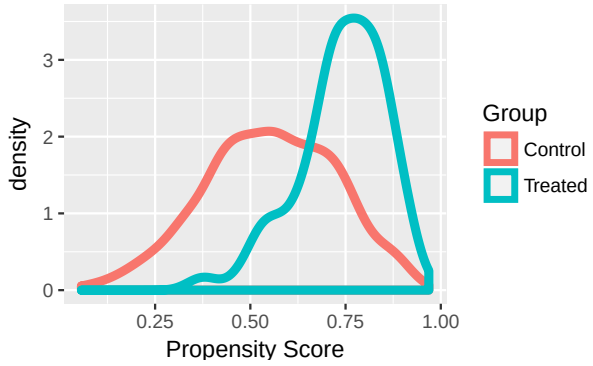


Figure 1: Overlap in firms' propensity of being located in Sandviken (being in the treated group).

The final model displayed in Equations 13 and 14 was obtained by studying balance in the pre-treatment outcome using the balance measure, Mean Difference standardized (MDS), suggested in Imbens and Rubin (2015) for continuous outcomes, given by

$$MDS_t = \frac{\bar{y}_{t,d=1} - \bar{y}_{t,d=0}}{\sqrt{\frac{s^2_{t,d=0} + s^2_{t,d=1}}{2}}}. \tag{15}$$

Figure 2 displays the pre-treatment balance for the final model. MDS_t is clearly smaller than the rule of thumb of 0.25 (Imbens and Wooldridge, 2009) for all pre-treatment t, indicating balance. Clearly, balance is obtained in all the outcomes of all pre-treatment periods. This indicates that the model is a sufficiently good description of the outcome processes to capture the important heterogeneities.

Figure 3a displays the average electricity consumption of the supplied firms in the two cities the year before treatment. It is clear that the seasonality pattern of electricity consumption is similar for the two cities but also indicates that the firms in Sundsvall consume more electricity on

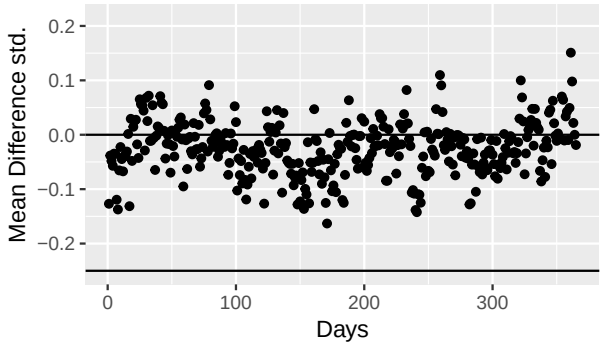
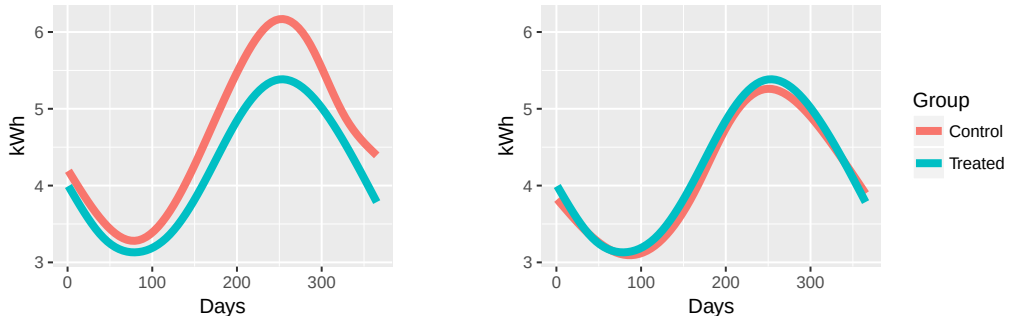


Figure 2: MDS_t for all pre-treatment time periods.

average. Therefore a naïve comparison of the electricity consumption during 2015 and 2016 would most likely provide a biased estimate of the effects from the dynamic tariff. Figure 3b displays the smoothed average after matching based on the final model, i.e. the control group has changed from all of Sundsvall to the matched control group, and the curve of the treated group is identical to that in Figure 3a. In this particular case, successful matching implies that matched firms have similar response to temperature shifts, and similar level of consumption, although there are small differences in temperature between the cities. $\boldsymbol{\theta}_i^*$ is clearly sufficiently well estimated to capture the non-constant temperature difference between the cities over the pre-treatment year.

After the matching was deemed successful, the post-treatment data were consulted to estimate the ATET. For all details of the estimated effect, see Öhrlund et al. (2019). The point estimate of the ATET, estimated with a panel regression model controlling for temperature, was, using the standard errors of Abadie and Imbens (2011), significantly different from zero and found to be -0.32 kWh averaged over the two post-treatment years, which is about 7.4% of the average pre-treatment consumption.



(a) Before matching

(b) After matching

Figure 3: Grouped smoothed pre-treatment electricity consumption before and after matching. In the left figure all firms in Sundsvall are included in the control-group, whereas in the right figure only the matched control firms are included. The time-series starts May 1 year 2014 and ends April 30 the following year.

6 Discussion

This paper has proposed a method for identifying causal effects from non-experimental data by utilizing time-series measurements of the pre-treatment outcome to substitute for, or add to, time-invariant covariates. The results in this paper show under which assumptions, and how, pre-treatment measurements of the outcome can be utilized in the identification and the estimation of the Average Treatment Effect of the Treated (ATET). The theoretical results and estimation methods can be used in addition to non-complete sets of observed confounders, and in some situations completely substitute for time-invariant covariates.

The strategies suggested in this paper build on being able to characterize a high frequency pre-treatment outcome process of an observational unit by a small set of informative statistics. In many situations, these process characteristics are directly related to unobserved characteristics of the subjects under observation. One might question why observational units are not simply matched on a simple summary statistic, e.g., within-unit pre-treatment means. However, the results in this study show that the within-unit short time dynamics can hold important information about heterogeneity across units. In other words, units with similar average levels may have different processes around their average level implying that there might be time periods for which the groups are not balanced even if they are balanced on average over time. By extracting these characteristics from the pre-treatment outcome data in a more fine tuned manner, unobserved variables that are by construction important as they are possible confounders, can be captured by parameter estimates. These estimates can then be used to identify causal effects by e.g. adjustment or matching.

Given that there are many time-series characterization and clustering/matching strategies and strategies for modelling the counterfactuals using time-series models, some important pros and cons of the suggested strategy are discussed in this paper. The matching estimation strategy suggested in this paper uses a parametric two-level time-series model to characterize each units outcome process. However, the matching strategy, evaluated on the pre-treatment outcome, can be used in combination with any type of ATET estimator and inference, which means that even though complex Bayesian time-series analysis is used to achieved balance, any standard effect estimator can be utilized in the final step of the effect evaluation. If there are no necessary time-varying covariates, it is even possible to non-parametrically estimate the ATET.

Comparing the proposed two-level time-series characterizing strategy to non-model based strate-

gies, the greatest difference is the ability in the proposed strategy to include time-varying covariates that do not vary across observational units. More specifically, the relation between such covariates and the outcome of each unit can in a natural way be utilized in the estimation. This is the main contribution of the proposed strategy as the relation between time-varying covariates that do not vary across observational units and the outcome of each unit might reveal important differences in outcome processes, as was illustrated by the inclusion of temperature and season in the empirical example. In addition, the proposed strategy has, in our opinion, one great advantage in a causal inference setting, namely that it provides the possibility to check the overlap assumption. The two-level time-series model-based approach quantifies the heterogeneity in terms of parameter estimates, estimates that can in turn be used as independent variables in a propensity score estimation used to check and make sure that the essential overlap assumption holds.

In conclusion, this paper has extended to a time-series setting the notation and assumptions used in cross-sectional observational studies to identify the ATET. The time-series setting opens up the possibility to alter the assumptions utilizing the pre-treatment data more extensively in the identification and estimation, replacing unobserved covariates with information derived from the pre-treatment outcome data. This development should enable more studies to be able to identify the ATET, even in fields where covariates are difficult or expensive to collect.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Synth : An R Package for Synthetic Control Methods in Comparative Case Studies . *Journal of Statistical Software*, 42(13).

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict : A Case Study of the Basque Country The Economic Costs of Conflict : A Case Study of the Basque Country. *The American Economic Review*, 93(1):113–132.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, 29(1):1–11.

Aghabozorgi, S., Seyed Shirkhorshidi, A., and Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, 53:16–38.

Angrist, J. D., Jordà, Ò., and Kuersteiner, G. M. (2017). Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited. *Journal of Business and Economic Statistics*, 36(3):371–387.

Angrist, J. D. and Kuersteiner, G. M. (2011). Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score. *Review of Economics and Statistics*, 93(3):725–747.

Asparouhov, T., Hamaker, E. L., and Muthén, B. (2018). Dynamic Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3):1–30.

Bai, J. (2009). Panel Data Models With Interactive Fixed Effects. *Econometrica*, 77(4):1229–1279.

Bodory, H., Camponovo, L., Huber, M., and Lechner, M. (2018). The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators. *Journal of Business and Economic Statistics*, 0015(May).

Bojinov, I. and Shephard, N. (2019). Time Series Experiments and Causal Estimands: Exact Randomization Tests and Trading. *Journal of the American Statistical Association*, 0(0):1–36.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9(1):247–274.

Chabé-Ferret, S. (2015). Analysis of the bias of Matching and Difference-in-Difference under alternative earnings and selection processes. *Journal of Econometrics*, 185(1):110–123.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31.

de Luna, X., Fowler, P., and Johansson, P. (2017). Proxy variables and nonparametric identification of causal effects. *Economics Letters*, 150:152–154.

Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118.

Hallquist, M. N. and Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural equation modeling : a multidisciplinary journal*, 25(4):621–638.

Iacus, S. M., King, G., and Porro, G. (2019). A Theory of Statistical Inference for Matching Methods in Causal Research. *Political Analysis*, 27:46–68.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction.* Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1):5–86.

King, G. and Stuart, E. A. (2011). MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *Journal Of Statistical Software*, 42(8):1–28.

Muthén, L. K. and Muthén, B. O. (2017). *Mplus User ' s Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Neyman, J. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Translated in Statistical Science*, 5(4):465–480.

Öhrlund, I., Schultzberg, M., and Bartusch, C. (2019). Identifying and estimating the effects of a mandatory billing demand charge. *Applied Energy*, 237.

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., and Sekhon, J. S. (2016). Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology*, 16(1-2):1–21.

Robbins, M. and Steven, D. (2019). microsynth: Synthetic Control Methods with Micro- And Meso-Level Data.

Robbins, M. W., Saunders, J., and Kilmer, B. (2017). A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention. *Journal of the American Statistical Association*, 112(517):109–126.

Rosenbaum, P. R. (2019). Modern Algorithms for Matching in Observational Studies. *Annual Review of Statistics and Its Application*.

Rubin, D. B. (1973). Matching to Remove Bias in Observational Studies. *Biometrics*, 29(1):159–183.

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

Rubin, D. B. (1980). Bias Reduction Using Mahalanobis-Metric Matching. *Biometrics*, 36(2):293–298.

Schultzberg, M. and Muthén, B. (2018). Number of Subjects and Time Points Needed for Multilevel Time-Series Analysis: A Simulation Study of Dynamic Structural Equation Modeling. *Structural Equation Modeling*, 25(4).

Smith, A. A. (2016). Indirect Inference BT - The New Palgrave Dictionary of Economics. pages 1–6. Palgrave Macmillan UK, London.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472.

Trull, T. J. and Ebner-Priemer, U. W. (2009). Using Experience Sampling Methods/Ecological Momentary Assessment (ESM/EMA) in Clinical Assessment and Clinical Research: Introduction to the Special Section. *Psychological Assessment*, 21(4):457–462.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press. MIT Press.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Xu, Y. and Liu, L. (2018). gsynth: Generalized Synthetic Control Method.

A Sensitivity Analysis

Substantial differences in the distributions at any pre-treatment time period period falsifies Assumption 2. To illustrate how the distribution of residuals may be used, consider the following simplified example. Let $T = 100$ and $N = 200$ $(50 + 150)$ One data sets are generated from each of the following two data generating processes

$$Y_{it}(0) = \mu_i + \phi_i X_{it} + \epsilon_{it} \qquad \text{(DGP 1)}$$

$$Y_{it}(0) = \mathbb{1}(D_i = 1)(\mu_{2i} + \phi_{2i} X_{it}) + (1 - \mathbb{1}(D_i = 1))(\mu_i + \phi_i X_{it}) + \epsilon_{it}, \qquad \text{(DGP 2)}$$

where $\mu_i \sim N(0,1), \phi_i \sim N(0,1), \epsilon_{it} \sim N(0,1), \mu_{2i} \sim N(2,1), \phi_{2i} \sim N(2,1)\ \forall\ i = 1, ..., N$. That is, in DGP 1 the treatment groups overlap completely in the parameters, in DGP 2 there is very little overlap. For DGP 1, successful matching should be possible whereas in DGP 2 this should be difficult. We fit a DSEM corresponding to DGP 1 to both samples and plot the distribution of t=90,91,...,99,100 to save space. Figure 4 displays the boxplots of the distributions of residuals. The top and bottom panels displays the residuals for DGP 1 and DGP 2, respectively. The failure of finding a comparable groups under DGP 2 is clear in the lack of overlap in the residual distributions. Note that other issues than overlap, e.g., severe model misspecifications are likely to be visible. This tool will indicate that the matching is not successful. It will not, however, indicate what the error is. A residual plot like the one in Figure 4 should alert the researcher that something is off. Inspection of the overlap in each specific parameter and modelfit in the DSEM can then guide the design.

The similarity in distribution could be formally tested by, for example, the distributions of residuals in the two groups that be tested using Kolmogorov-Smirnov tests at each time point from 1 to $t_1 - 1$ using some significance correction to account for the large number of tests.

A simplified sensitivity analysis is to restrict the evaluation to the differences in means, i.e, study the balance of $\overline{Y}_t^1$ and $\overline{Y}_t^0$ for all time periods as discussed in Section 3.3 and illustrated in the empirical example in Section 5. This caould be performed pointwise in t, or applying the frameworks proposed in Angrist and Kuersteiner (2011); Angrist et al. (2017); Bojinov and Shephard (2019). That is, test for an effect under a period, preferably the time periods leading up to the treatment assignment, where there should be no effect under Assumption 2. If the estimates of $\boldsymbol{\theta}_i^*$ do not contain at least as much information about the heterogeneity in the processes as $\boldsymbol{\theta}_i^{(s)}$, it is implausible that balance would be obtained for all pre-treatment time periods. In this paper, the

sensitivity analysis is simply done by comparing the point-wise standardized mean difference to the rule of thumb of being smaller than ± 0.25 (Imbens and Wooldridge, 2009).

The above mentioned sensitivity analysis strategies can also be used to help specify $\mathbf{L}_{obs}$. That is, $\mathbf{L}_{obs}$ can first be specified with larger lags than might be expected, lags with estimated zero effects can then be removed by iteratively fitting the two-level time-series model, evaluating the parameter estimates.

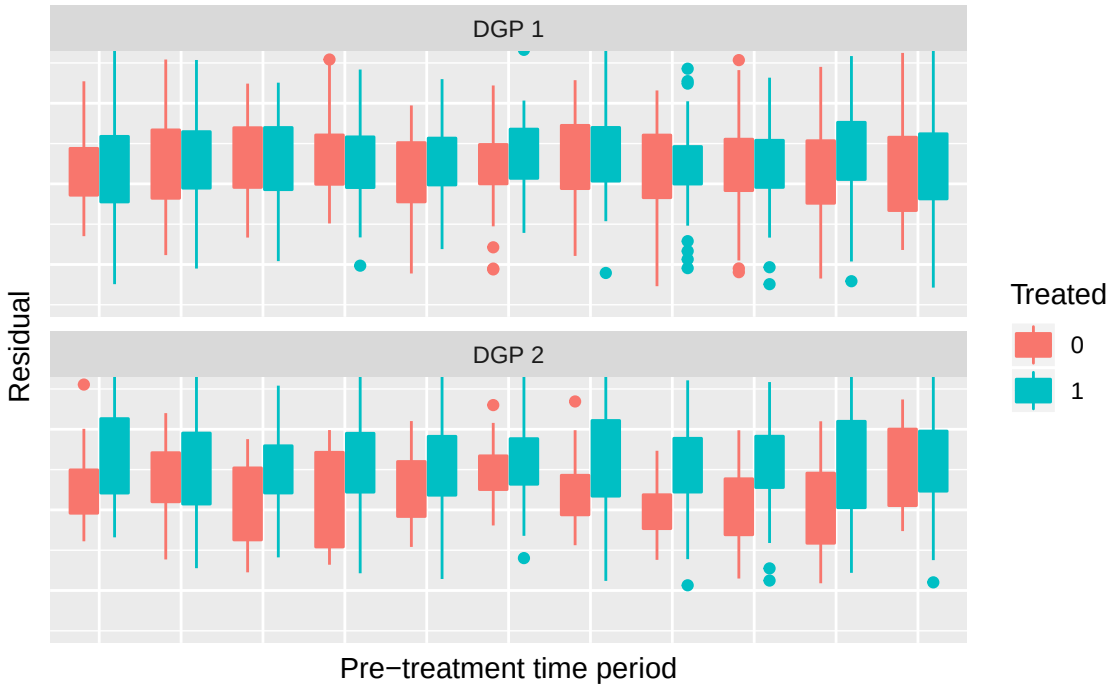


Figure 4: Distributions of residuals for the last 10 pre-treatment time periods. Two data generating processes are considered, with and without overlap in the parameter estimates.

B Possible gains of including dynamic parameters in matching

Here it is illustrated that balance evaluation based on pre-treatment within-observational unit means does not imply balance in the group means at each pre-treatment period, and, how this balance can be improved by evaluating the balance also in slopes of time-varying covariates. Data are generated as

$$Y_{it} = \mu_i + \phi_i X_t + \epsilon_{it} \tag{16}$$

where the error term is independently and identically distributed $\epsilon_{it} \sim N(0, 1)$ for all i and t. $\phi_i \sim n(0, 2)$. Here $X_t = \sin(0.05t)$ to mimic a covariate with clear seasons. The outcome Y could for example be electricity consumption within a household i, where the variation around the mean varies with temperature X that varies with season over time. One sample of 50 treatment and 150 controls is generated.

A simple matching algorithm is used: 50 of the controls are randomly sampled without replacement from the pool of controls repeatedly. A DSEM according to Equation 16 is fitted to the treated and and randomly sampled controls. This is repeated until (1) the absolute difference in the mean of the raw outcomes is smaller than 0.001, and (2) the absolute difference in the mean of the mean of the raw data and the mean of $\hat{\phi}$ are both smaller than 0.005. The distribution of coefficients, estimated one time-series at a time, in the sample before matching is given in Figure 5. The two groups clearly overlap in the distributions of both μ and ϕ.

To ensure that the groups are balanced at all post-treatment time points of interest, the goal of the matching is to make the groups as similar as possible in the pre-treatment period. If the groups are comparable for all pre-treatment time points it should increase the likelihood for the balance to continue after this period unless any treatment is added. Figure 6 displays the pre-treatment smoothed group means after matching based on only the within-unit mean and both the mean and $\hat{\phi}_i$. It is clear that the mean level over the full pre-treatment period is approximately balanced in both cases. However, at several time points the groups differ substantially when only the within-unit mean is balanced. It is clear that this increases the *point-wise similarity* in pre-treatment processes drastically. These results illustrate the potential importance of fully utilizing the pre-treatment data and its dynamic characteristics in the identification.

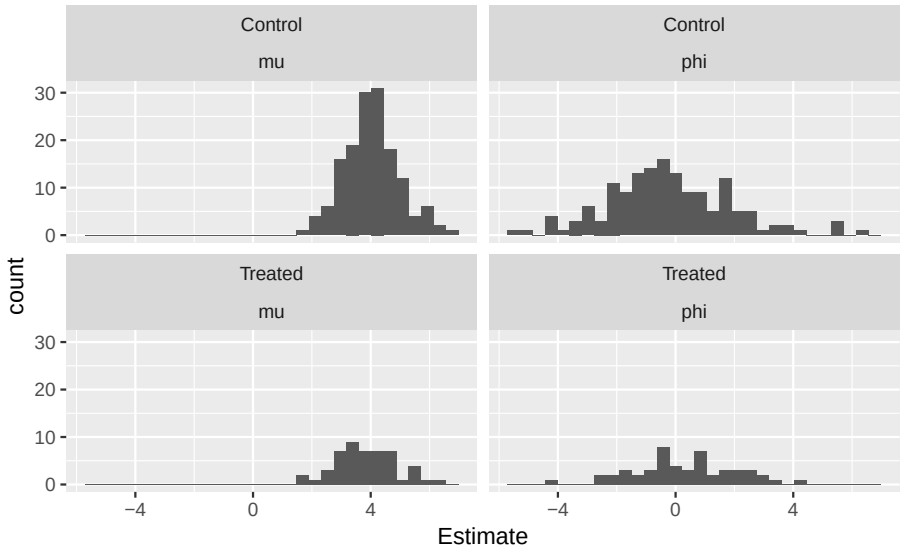


Figure 5: Distribution of random-coefficient estimates in the treatment group and the pool of controls.

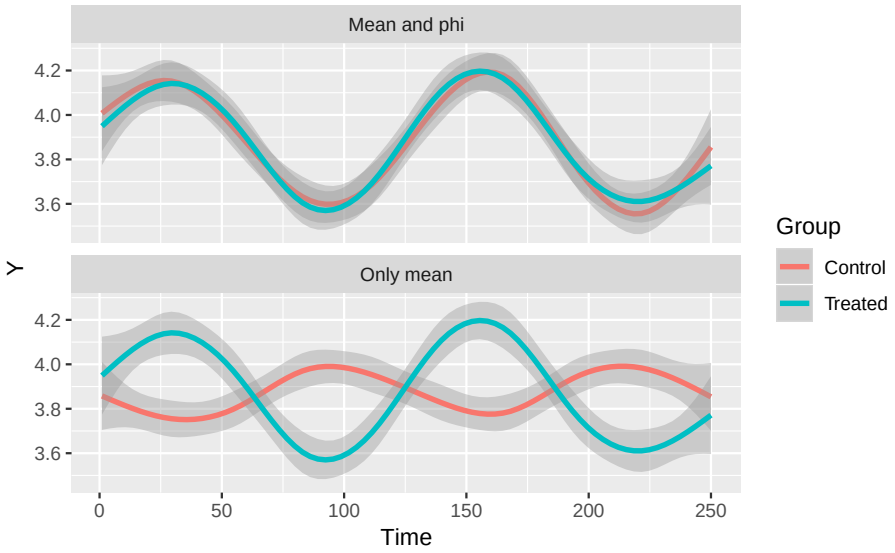


Figure 6: Balance checked against the pre-treatment within-observational unit mean only.

C DSEM

C.1 Illustration of the dimension reduction using two-level time-series analysis

In this section a small simulation study illustrates how a general multi-level time-series models, e.g., a DSEM, can be used to conveniently characterize a large number of long time-series. To illustrate the utility of the models, 4 groups with outcome processes following different AR-processes are constructed. The data are generated as

$$Y_{jit} = \mu_j + \phi_{1ij} Y_{it-1} + \phi_{2ij} Y_{it-2} + \epsilon_{ijt} \tag{17}$$

where $j = 1, 2, 3, 4.$ refers to the groups. Furthermore,

$$\begin{aligned}
\mu_{i1}, \mu_{i2} &\sim N(1, 0.2) \\
\mu_{i3}, \mu_{i4} &\sim N(3, 0.1) \\
\phi_{1i1}, \phi_{1i2} &\sim N(0.5, 0.3) \\
\phi_{1i3}, \phi_{1i4} &\sim N(-0.2, 0.3) \\
\phi_{2i1}, \phi_{2i4} &= 0 \\
\phi_{2i2}, \phi_{2i3} &\sim N(-0.4, 0.3) \\
\epsilon_{i1t} &\sim N(0, 1) \\
\epsilon_{i2t} &\sim N(0, 3) \\
\epsilon_{i3t} &\sim N(0, 0.5) \\
\epsilon_{i4t} &\sim N(0, 0.1),
\end{aligned}$$

with $i = 1..., N$ and $t = 1, ..., T$, in this case N=60, T=300. Figure 7 displays the time-series of all 60 observational units for 300 time periods. From the left panel it is difficult to distinguish between the different groups. The following model is fitted to this data

$$Y_{jit} = \mu_j + \phi_{1ij} Y_{it-1} + \phi_{2ij} Y_{it-2} + \phi_{3ij} Y_{it-3} + \epsilon_{ijt}. \tag{18}$$

The parameter estimates divided by groups are displayed in Figure 8. By looking at several parameters these four groups can easily be identified. For example, group 1 and 2 cannot be distinguished

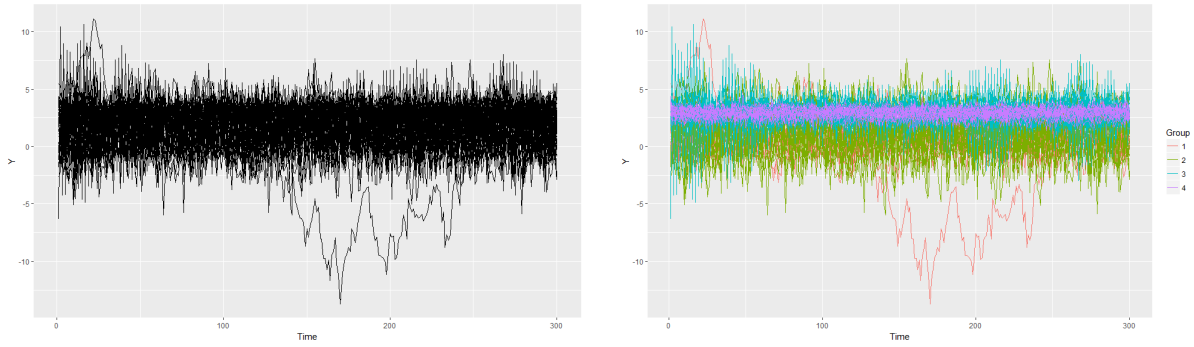


Figure 7: The time-series of the 60 observational units. In the right panel the time-series are coloured according to group.

from based on the mean only, however, using the mean and/or the variance the separation is clear. Although not all parameters have perfect estimates, the model successfully captures and distinguishes between the four different groups, including two nested AR processes, in one estimation. Figure 9 displays an example from each of the four groups. It is clear that the levels and variances are different, however, the difference in the order and strength of the autocorrelation is less obvious from the plots.

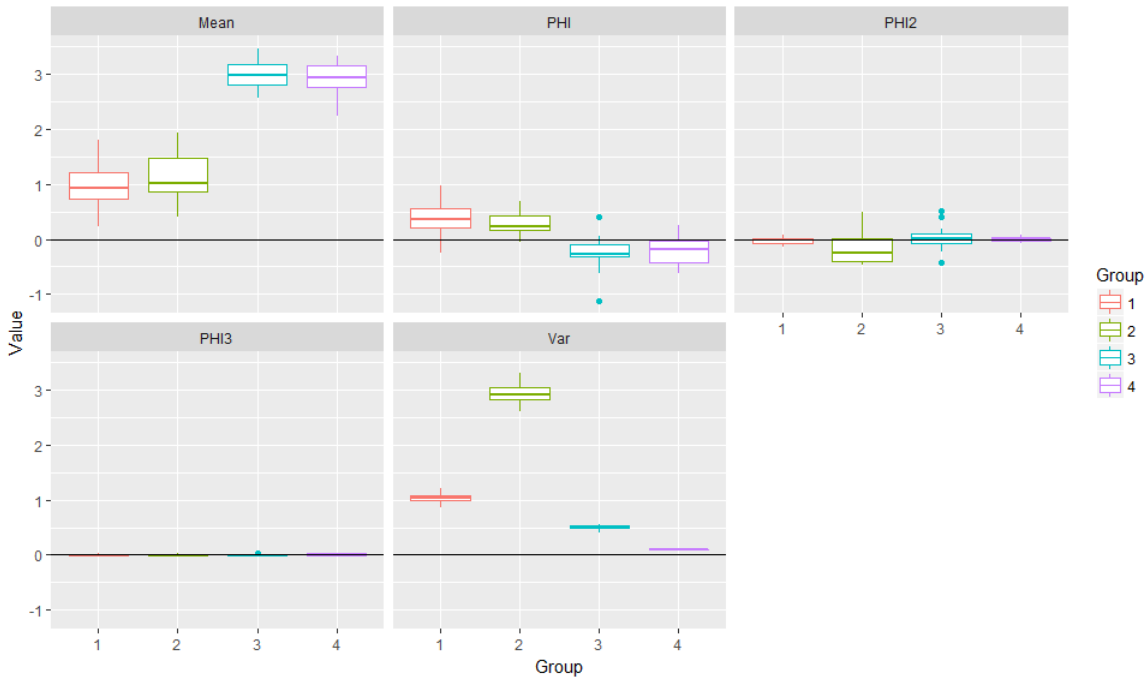


Figure 8: Grouped parameter estimates for all parameters of the model.

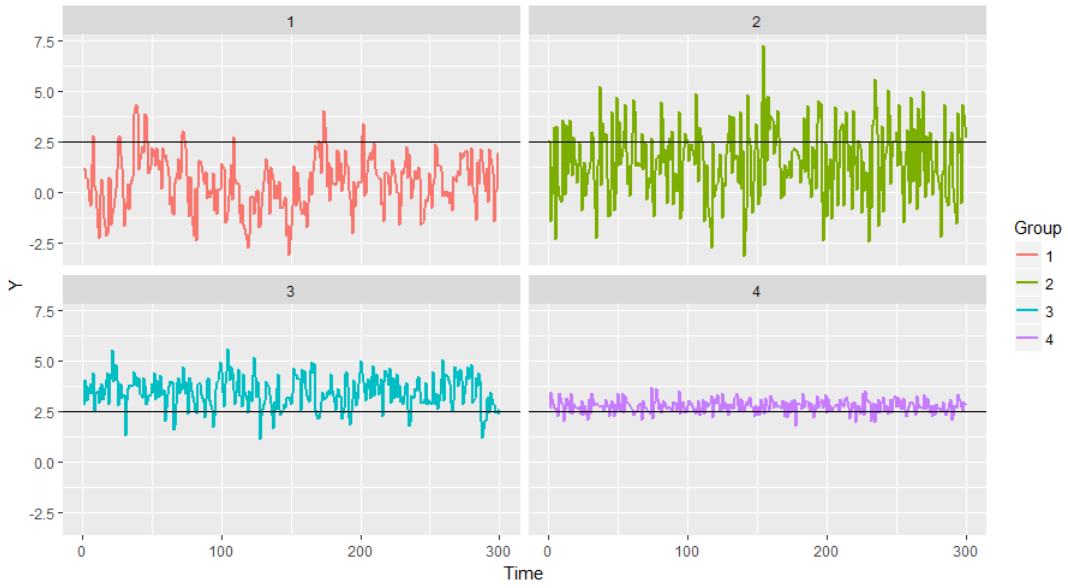


Figure 9: One example process from each of the four groups. The solid line is given as a level reference point.

C.2 Empirical example - Details

Table 3 shows some key estimates. The parameters correspond to the parameters in Equation 14. Starting from the bottom of the table, it is clear that there is, e.g., a large variation in the residual variance term and a number of other parameters including the mean level consumption, all indicating heterogeneity across firms in consumption processes. Some firms may have a low mean level but a large variance due to high autocorrelation and residual variance, whereas other might be more dependent on the temperature and the season. Indeed, several of these unit-specific parameters also have significant effects on the treatment assignment indicating the importance of including these in the balancing procedure.

Table 3: Estimates and 95% credibility intervals for key parameters of the Y and D regressions.

Parameter	Estimate	CI(low)	CI(high)
β_1	-0.008	-0.054	0.045
β_2	0.266	-0.2	0.713
β_3	1.878	-0.387	4.314
β_4	-0.072	-0.183	0.036
β_8	0.655	0.043	1.328
β_9	-0.958	-1.674	-0.308
β_{10}	-0.366	-1.18	0.381
β_{11}	2.049	0.888	3.126
β_{12}	-0.764	-1.561	0.025
β_{16}	0.38	-0.168	0.942
β_{17}	0.176	-0.1	0.432
β_{18}	0.333	0.033	0.647
μ_μ	4.597	4.413	4.787
$\mu_{\phi_{kWh}}$	0.447	0.433	0.460
$\mu_{\phi_{Te}}$	-0.042	-0.046	-0.039
μ_{γ_8}	0.405	0.383	0.426
μ_{γ_9}	0.368	0.345	0.391
$\mu_{\gamma_{10}}$	0.363	0.343	0.383
$\mu_{\gamma_{11}}$	0.176	0.158	0.193
$\mu_{\gamma_{12}}$	0.054	0.037	0.070
$\mu_{\log\sigma^2}$	-0.807	-0.888	-0.728
σ^2_μ	10.387	9.532	11.274
$\sigma^2_{\phi_{kWh}}$	0.057	0.052	0.062
$\sigma^2_{\phi_{Te}}$	0.003	0.003	0.004
$\sigma^2_{\gamma_8}$	0.085	0.075	0.096
$\sigma^2_{\gamma_9}$	0.087	0.076	0.1
$\sigma^2_{\gamma_{10}}$	0.059	0.051	0.068
$\sigma^2_{\gamma_{11}}$	0.043	0.038	0.05
$\sigma^2_{\gamma_{12}}$	0.046	0.04	0.053
$\sigma^2_{\log\sigma^2}$	1.813	1.669	1.976

C.3 Sample size recommendations

The sample size requirements in terms of N and T for using the DSEM strategy depends heavily on the complexity of the fitted model. With high-order autocorrelation structures and or many time-varying covariates, and thereby many random coefficients, the requirements increase. However, simulations in Schultzberg and Muthén (2018) suggests that models without complex level-2 models, i.e., where only the distribution of the random coefficients excluding their structural relations, are the least demanding. According to results in that paper, $T \geq 75$ and $N \geq 100$ should enable quite complex models. Due to the shared hyper parameters, a larger N, say 250 or more, can reduce the T requirements further in many situations.

Note that DSEM does allow for unbalanced time-series which implies that settings where pre-treatment observations and or post-treatment periods are different can be handled in a straightforward fashion. Of course, to find similar units, it is desirable to observe them under a sufficiently long period to ensure that similarities are not artefacts of seasonal effects.

C.4 Missing data in the pre-treatment period

As mentioned above, another important benefit of fitting a parametric Bayesian two-level times-series model such as the DSEM and match on the unit-specific parameters is that the model can easily handle missing data in the outcome and time-varying covariates under the missing at random assumption. None of the above mentioned strategies besides TLTM have built in strategies for handling missing data. The CI can handle missing data in the treated units outcome, but not the outcomes of the control. Figure 10 displays the MSE with and without 25% missing data in all units in the pre-treatment period for TLTM with N=120 and T=100. As shown in Schultzberg and Muthén (2018), the accuracy of the TLTM model improves in both N and T, wherefore the results for the smallest N and T applies for all larger values. Clearly, there is only a small increase in MSE due to the missing data and the TLTM approach is still preferrable to the other approches in this setting even when the other strategies have the full sample.

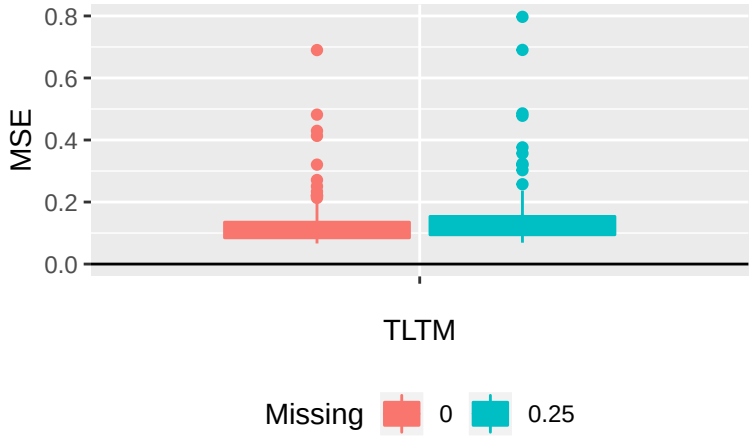


Figure 10: MSE of the DSEM strategy with N=120, T=100, with and without 25 percent missing data in the outcome in the pre-treatment period for all units.