DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

Tree-based Synthetic Control Methods: Consequences of moving the US Embassy

Nicolaj N. Mühlbach

CREATES Research Paper 2020-04

Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Tree-based Synthetic Control Methods: Consequences of moving the US Embassy*

Nicolaj Nørgaard Mühlbach†

This version: May 18, 2020

Abstract

We recast the synthetic controls for evaluating policies as a counterfactual *prediction* problem and replace its linear regression with a nonparametric model inspired by machine learning. The proposed method enables us to achieve more accurate counterfactual predictions. We apply our method to a highly-debated policy: the move of the US embassy to Jerusalem. In Israel and Palestine, we find that the average number of weekly conflicts has increased by roughly 103% over 48 weeks since the move was announced on December 6, 2017. Using conformal inference and placebo tests, we justify our model and find the increase to be statistically significant.

Word count: 11,858

JEL Classification: C14; C21; C54; D02; D74; F51

Key words: Treatment effects; Program evaluation; Synthetic control; Machine learning; US embassy move

---

*The author is grateful to Alberto Abadie, Guido Imbens, Stefan Wager, Christian B. Hansen, Christian Bjørnskov, and Bent Jesper Christensen for helpful comments and suggestions, and to the Center for Research in Econometric Analysis of Time Series (CREATES), the Dale T. Mortensen Center, Aarhus University, and the Danish Social Science Research Council for research support.

†Department of Economics and Business Economics, Aarhus University, muhlbach@econ.au.dk.

1 Introduction

In social science, we are often interested in the effects of policy interventions on aggregate entities to evaluate previous, understand current, or counsel future policies. The aggregate units may be firms, organizations, geographic areas, etc. Data often stem from observational studies. Estimating such effects has been heavily studied, and various methods apply to different data available (for reviews, see Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018)). One approach is to compare the treated unit to a control unit not exposed to the event. One of the first examples is Card (1990), who uses Southern US cities as a comparison group to estimate the effect of an unanticipated Cuban migratory influx in Miami. However, the design of a comparative case study faces certain challenges. First, it is not always transparent how specific control units are chosen, and the appropriate control may be chosen ex-post. Running several regressions may lead to publication bias (Franco et al., 2014). Second, many of the current methods to evaluate policies are based on regressions that try to maximize the pre-treatment fit, which may not generalize well out-of-sample. The situation illustrates the classical bias-variance trade-off, where methods are often chosen to minimize bias rather than balancing bias for variance. If one could build an econometric model that would accurately predict the outcome of the treated unit post-treatment in a counterfactual state absent the treatment, it may be helpful to evaluate the intervention. This is especially useful if pre-treatment inference is not a goal in itself (for a discussion on recasting economic problems as prediction problems, see Kleinberg et al. (2015)). Third, the standard approach to comparative case studies is to specify a linear functional form to capture the relationship between the treated unit and the control units. This may be restrictive if we are trying to answer questions for which no theoretical model exists. In addition, the standard approach does not take nonlinearities, especially interactions, into account except those explicitly modeled by the researcher. If the process that generates the outcomes for the treated unit in the pre-treatment period is nonlinear in the control outcomes, the resulting bias may be severe.

Building on ideas from Abadie and Gardeazabal (2003), Abadie et al. (2010) solve the first challenge. In the presence of a single treated unit and several control candidates, synthetic controls form a set of weights such that the weighted average of the control units approximately matches the treated unit in the pre-treatment period. The same weights are then channeled to the post-treatment period to estimate a synthetic control group that constitutes the counterfactual state of the world in which the treated unit was not exposed to the treatment. The issue of overfitting, however, remains unsolved.

Doudchenko and Imbens (2016) take on the second challenge by proposing a regularized version of the synthetic control method, namely the elastic net estimator. Relying on ideas from machine learning, the elastic net estimator shrinks the weights toward zero and sets some of them exactly to zero. Especially in moderately-high dimensions, this approach has shown promise in forecasting studies. Also, the selection property by zeroing out some weights has attractive interpretations as it allows researchers to pinpoint which control units have no explanatory power when forming the counterfactual control.

Both methods, however, specify a linear model that is not capable of automatically detecting nonlinearities among the control units. In particular, we expect many low-order interactions of the control outcomes to be informative in explaining the outcomes of the treated unit. For instance, consider the empirical application in Abadie et al. (2010) regarding cigarette sales in the US. While the sales in California may be modeled as a weighted average of the sales in New York and Florida given a common cigarette consumption pattern along the coasts, a decrease in sales in New York may be associated with an even bigger decrease in California given a low period of sales in Florida. This could happen if the people of California see themselves as trendsetters in regards to health; when people in both New York and Florida are decreasing their cigarette consumption, people of California want to reduce their consumption even further. Note that it is becoming natural for lasso-based estimators to include interactions and higher-order terms in contrast to synthetic controls. But important interactions and higher-order terms can be difficult to anticipate ex-ante. The kitchen sink approach would be to include all higher-order terms up to a pre-specified order, e.g. to third order. This approach, however, quickly faces its own problems. With 10 control units, all third-order terms would count $10 + {}^{(10+3-1)!}/_{(10-1)!3!} = 230$, which is infeasible to handle for most parametric estimators given finitely many observations. Thus, if nonlinearities are deemed important or the domain is unknown, we argue for flexible methods that can handle such nonlinearities in a data-driven manner.

We recast the problem of *estimating* a synthetic control as the problem of *predicting* one, similarly to both Doudchenko and Imbens (2016) and Athey et al. (2019) who advocate for powerful prediction methods. This way, we do not have to rely on linear, parametric models that potentially misspecify the true underlying model. This is beneficial, for instance, whenever the researcher does not have the domain knowledge required to specify a theoretical model. We choose a popular method from the machine learning literature that handles interactions and other nonlinearities automatically. For instance in our application on conflict levels in the Middle East, when we seek to understand which periods are similar in terms of the level of conflict, it is difficult to consider conflict levels in Iraq and Saudi Arabia separately without an interaction between them. Imagine some violent and frequent conflicts in the South of Iraq in a given period. The regime of Saudi Arabia may react by increasing the appearance of police forces in major cities, and as a result, the number of conflicts falls. If such interactions matter for the conflict level in Israel and Palestine, we would incur an omitted variable bias by leaving them out.

Nonparametric approaches to estimating treatment effects do exist in the econometric toolbox. Similarly to our method, Athey and Imbens (2016), Wager and Athey (2018), and Athey et al. (2019) also rely on ideas from machine learning to study heterogeneous treatment effects using nonparametric models. They propose various modifications to the random forests algorithm by Breiman (2001). Our method differs because we observe units over time with treatment happening at a certain point, whereas the other papers are based on a cross section of units. Moreover, their methods are most suitable when a large set of both observations and covariates is available as they focus on heterogeneous treatment effects, whereas we focus on average treatment effects. As another example, Hartford et al. (2017) use deep neural nets for counterfactual prediction. We find, however, that many applications in social science and ours included do not enjoy the luxury of having sufficiently large datasets available to apply (deep) neural nets.

We propose the tree-based synthetic control method as an alternative to the synthetic controls for applications where the researcher prefers accurate post-treatment predictions over the ability to do pre-treatment inference, and when the empirical question is not guided by any theoretical model that can justify specific assumptions on the empirical model. We adopt the design of synthetic controls that models the treated unit as a conditional expectation of the control units. We also consider all potential controls in the donor pool transparently. If any particular control units do not contribute to explaining the treated unit, the method is flexible enough to leave them out. Our method is inherently nonlinear when modeling the controls, and additionally, interactions and higher-order terms are included in a data-driven manner.

The proposed method uses the pre-treatment periods to estimate the relationship between the treated and all the control units and imposes this relationship onto the post-treatment period, similarly to Abadie et al. (2010) and Doudchenko and Imbens (2016). To model the conditional expectation, we apply the canonical random forests regression model. Random forests have proved successful in many applications (see for instance Montgomery and Olivella (2018) for a recent paper in political science, or Guha and Ng (2019) in IO). Further, variants of random forests have already been employed in the treatment effects literature either directly (Athey and Imbens, 2016; Athey et al., 2019; Wager and Athey, 2018) or indirectly (Chernozhukov et al., 2017a,1). Common to these papers is that they rely on the unconfoundedness assumption and assume there is a relationship between outcomes for a given unit over time (estimated by regressing control unit outcomes in treated periods on lagged outcomes) that is stable across units. In contrast, the synthetic control literature assumes there is a relationship between different units (estimated by regressing treated unit outcomes on control outcomes) that is stable over time. Our approach falls into the latter.

Intuitively, for each period where the treated unit is treated, our model locates a few corresponding pre-treatment periods based on the control units and uses the average of the pre-treatment outcomes of the treated unit as a counterfactual prediction in the post-treatment period. Stated differently, our model aggregates the pre-treatment periods into similar subgroups based on the control units. Then, it computes the average of the outcomes of the treated unit in each of the subgroups. In the post-treatment period, the model remembers how to group the periods and assigns the corresponding pre-treatment average to each of the periods. This gives an estimate of the potential outcome for the treated unit in the absence of the treatment. Having an estimate for all periods after the intervention, we compute the average of the differences between the estimate and the actual outcome, similarly to Chernozhukov et al. (2017).

We showcase the tree-based synthetic control method by estimating the effect of moving the US embassy from Tel Aviv to Jerusalem on the number of weekly conflicts in Israel and Palestine. It is beyond our interest to judge the particular political decision, rather we propose a method to estimate its impact. We use conflict data from December 28, 2015, to November 3, 2018, for Israel and Palestine as well as for 11 of the remaining countries in the Middle East as controls. The data are provided by the Armed Conflict Location & Event Data Project (Raleigh et al., 2010). Our results indicate that the weekly number of conflicts has increased by 26 incidents on average after the move was announced on December 6, 2017, until November 3, 2018. This corresponds to more than doubling the number of conflicts. We use the recently proposed conformal inference test by Chernozhukov et al. (2017b) to formally justify our results. The increase is statistically significant at a 1% level.

The rest of the paper is organized as follows. Section 2 reviews the related literature and introduces the tree-based synthetic control method. Section 3 considers the context of Israel and Palestine and presents the results alongside several robustness checks. Section 4 compares our method to state-of-the-art econometric methods. Section 5 concludes.

2 Synthetic Control Methods

2.1 *Setup*

We consider $N+1$ cross-sectional units observed in T periods and assume without loss of generality that only the first unit is exposed to the treatment, leaving N units as controls[1]. We index units by $i = 0, \ldots, N$ and time by $t = 1, \ldots, T_0, \ldots, T$ with the first T_0 periods before the treatment. Let $Y_{i,t}^N$ denote the potential outcome that would be observed for the ith unit at time t in absence of treatment, and similarly, let $Y_{i,t}^I$ denote the potential outcome that would be observed if exposed to the intervention. Under the assumption that the intervention does not affect the outcome before implementation, we have $Y_{i,t}^N = Y_{i,t}^I = Y_{i,t}$ for $t \leq T_0$ and all $i = 0, \ldots, N$. In many applications and ours included, the treatment may have an effect before implementation via announcement or anticipation, and T_0 should be redefined accordingly. We assume implicitly that the treatment does not affect the outcome for the control units (see Rosenbaum (2007) for a thorough discussion on this). Let $W_{i,t}$ be an indicator taking value one if the intervention happens at time t for unit i and zero otherwise. As treatment happens solely for the first unit in the post-treatment period, the treatment indicator $W_{i,t}$ satisfies

$$W_{i,t} = \begin{cases} 1 & \text{if } i = 0 \text{ and } t > T_0, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The observed outcome for unit i at time t is then

$$Y_{i,t} = Y_{i,t}^N + \tau_{i,t} W_{i,t}, \tag{2}$$

[1] We will use *treatment* and *intervention* interchangeably.

where we define $\tau_{i,t} = Y^I_{i,t} - Y^N_{i,t}$ as the effect of the intervention for unit i at time t. The causal effects for the treated unit in the post-treatment period are then $\boldsymbol{\tau}_0 = (\tau_{0,T_0+1}, \ldots, \tau_{0,T})'$. Because $Y^I_{0,t} = Y_{0,t}$ for $t > T_0$, we have $\tau_{0,t} = Y_{0,t} - Y^N_{0,t}$ and we need only estimate the counterfactual $Y^N_{0,t}$ for $t > T_0$. Our main goal is then to estimate flexibly the average treatment effect (ATE) as the average of $\tau_{0,t}$ over the post-treatment periods, i.e.

$$\tau = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \tau_{0,t} \tag{3}$$

In the most general form, we describe the (no intervention) outcome for the treated unit as given by $Y^N_{0,t} = \mathbb{E}\left[Y^N_{0,t} | \mathcal{F}_t\right] + \varepsilon_{0,t}$, where $\varepsilon_{0,t}$ are unobserved transitory shocks at the unit level with zero mean, and $\mathcal{F}_t$ denotes all information available (not to the econometrician) at time t. Denote by $X_{0,t}$ the $(p \times 1)$ vector of potential covariates relevant for $Y^N_{0,t}$. Our method is scale-invariant and can handle both categorical and continuous covariates. Note that $X_{0,t}$ may include covariates other than the control units as long as they are not affected by the intervention. For instance, we would not be able to include stock market indicators for Israel and Palestine. For simplicity, however, we follow Abadie et al. (2010) and focus on using the control units as covariates by letting $X_{0,t} = \left(Y^N_{1,t}, \ldots, Y^N_{N,t}\right)'$ denote the observed $(N \times 1)$ vector of outcomes for all the N control units at any time t. We assume the conditional expectation as a flexible function of the control units, i.e. $\mathbb{E}\left[Y^N_{0,t} | \mathcal{F}_t\right] = f^\star(X_{0,t})$. Thus, our objective is to estimate $f^\star$ as a function of the control units using only $t \leq T_0$ such that *if* the intervention did not take place, the model would still approximate well the treated unit in the post-treatment periods, $t > T_0$.

2.2 *Related literature*

This paper builds on a growing literature on treatment effects. Abadie et al. (2010) also consider the estimable object $\tau_{0,t} = Y_{0,t} - Y^N_{0,t}$ for $t > T_0$. Assume that there exists a set of perfect weights $\boldsymbol{\omega}^* = (\omega^*_1, \ldots, \omega^*_N)'$ such that $\sum_{i=1}^N \omega^*_i Y_{i,t} = Y_{0,t}\ \forall t \leq T_0$. Considering

$Y_{0,t} - \sum_{i=1}^{N} \omega_i Y_{i,t}$, Abadie et al. (2010) prove that its mean is approximately zero under standard conditions, which suggests using $\hat{\tau}_{0,t} = Y_{0,t} - \sum_{i=1}^{N} \hat{\omega}_i Y_{i,t}$ as an estimator for $\tau_{0,t}$ in periods $t > T_0$. The weights are then estimated by

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega} \in \mathbb{R}^N} \left\{ \sum_{t=1}^{T_0} \left(Y_{0,t} - \sum_{i=1}^{N} \omega_i Y_{i,t} \right)^2 \right\} \quad \text{st.} \ \sum_{i=1}^{N} \omega_i = 1, \ \omega_i \geq 0 \forall i. \tag{4}$$

This boils down to assuming linearity of $f^\star$ in $X_{0,t}$. The synthetic control method is mainly tailored for empirical settings with relatively more time periods than control units, i.e. $T \gg N$.

Doudchenko and Imbens (2016) propose a regularized extension to synthetic controls, namely the elastic net estimator. The optimization problem is similar to (4) but adds a regularization term to the objective function with inspiration from shrinkage estimation. Let $(\lambda, \alpha) \in \mathbb{R} \times \mathbb{R}$ be a given pair of hyper-parameters to be tuned and let $\mu \in \mathbb{R}$ be an intercept, capturing the possibility that the outcomes for the treated unit are systematically different from the other units. Then, Doudchenko and Imbens (2016) propose to estimate the weights by

$$(\mu, \hat{\boldsymbol{\omega}}) = \arg\min_{\mu, \boldsymbol{\omega}} \left\{ \sum_{t=1}^{T_0} \left(Y_{0,t} - \mu - \sum_{i=1}^{N} \omega_i Y_{i,t} \right)^2 + \lambda \left(\frac{1-\alpha}{2} \sum_{i=1}^{N} \omega_i^2 + \alpha \sum_{i=1}^{N} |\omega_i| \right) \right\}. \tag{5}$$

Note that (5) neither requires zero intercept, weights summing to one nor non-negative weights. The elastic net estimator enjoys the selection property known from lasso by the L_1-penalty term (Tibshirani, 1996; Zou and Hastie, 2005). Essentially, some weights are likely to be zeroed out, meaning that some control units are not predictive of the treated unit.

Both the synthetic control and the elastic net estimator may be viewed as cross-sectional regressions in which the outcome of the treated unit is regressed on the outcomes of the control units in the pre-treatment period. Assuming stability over time, the cross-sectional pattern is then carried over into the post-treatment period, based on which the counterfactual outcome for the treated unit is predicted using the control units. This form of regression in causal panel data models is known as vertical regressions, a term coined by Athey et al. (2018). The (almost) symmetric formulation is known as horizontal regressions, where the post-treatment outcomes are regressed on the pre-treatment outcomes using only the control units. This time-series approach estimates a relationship, which is then applied to the treatment unit assuming stability across units and requires $N \gg T$. It is not a symmetric problem because the order of T matters in contrast to the order of N.

However, both methods have a disadvantage in cases with $T \approx N$ as they do not fully exploit the panel structure by running either cross-sectional or time-series regressions. A recent approach to causal panel data models that takes both sources of variation into account is the matrix completion method by Athey et al. (2018), treating $Y_{0,t}^N$ for $t > T_0$ as missing. In Section 4, we compare all methods introduced.

2.3 *The Tree-based Synthetic Control Method*

Our method is conceptually similar to the idea of Abadie et al. (2010) to the extent that we also use vertical regressions to estimate the relationship between the treated unit and the control units in the pre-treatment period and assume that the estimated relationship continues into the post-treatment period. But contrary to using the weighted *control* outcomes, we take a more direct approach by using a weighted average of the outcomes for the *treated* unit in different pre-treatment subperiods. In particular, we use the control outcomes to stratify the pre-treatment periods into homogenous subgroups in which the outcomes for the treatment unit are similar. This immediately removes the risk of extrapolation. Note that subgroups need not be equidistant or consecutive. Then, we apply the estimated stratification scheme to divide the post-treatment period into these subgroups, and for each of these subgroups, we finally estimate the potential outcome as the average of the pre-treatment outcomes of the treated unit that fall into the same subgroup. The stratification rules are estimated in a nonparametric manner based on the original random forests method in Breiman (2001), allowing us to estimate $f^\star$ as a flexible relationship between the treated unit and the control units. Various theoretical studies (see for instance Biau et al. (2008), Ishwaran and Kogalur (2010), Biau (2012), and Scornet et al. (2015)) have been performed, analyzing the consistency of random forests. The theoretical justification of our method is provided by Scornet et al. (2015) who prove the consistency of random forests. The cornerstone of random forests is a single decision tree.

Decisions trees recursively segment the input space into simpler subspaces and then assign a constant output value to all samples within each terminal subspace. After the segmentation, each observation belongs uniquely to one particular category, and to predict the outcome variable at an unseen sample, the model uses the average outcome based on the observations falling into the same category. Figure 1 shows an example related to our application. In the example, we divide the weekly level of conflicts in Israel and Palestine at each period $t \leq T_0$ into bins based on the weekly level of conflicts in Bahrain, Jordan, and Qatar. Given observations on the weekly level of conflicts in Bahrain, Jordan, and Qatar at a new point in time, say $t' > T_0$, we decide which of the four categories t' belongs to. As an example, suppose we end in category 1. Our prediction of the weekly level of conflicts in Israel and Palestine is then the average of all observations that fall into category 1 in the pre-treatment period. Hence, the outcomes for Bahrain, Jordan, and Qatar enter only in the stratification, and thus, our approach also allows the inclusion of other covariates, e.g. stock market indicators or news data from the control countries.

Next, we explain the model in greater detail. Recall that our goal is to predict the potential outcome $Y_{0,t}^N$ for $t > T_0$ given observed outcomes for both the treated and the control units in the pre-treatment period. Hence, we estimate the fundamental relationship for $t \leq T_0$

$$Y_{0,t}^N = f^\star \left(X_{0,t}\right) + \varepsilon_{0,t}, \tag{6}$$

where $\{\varepsilon_{0,t}\}$ are zero mean and assumed to be stationary and weakly dependent. After learning $\hat{f}\left(\cdot\right)$ from the pre-treatment period, we estimate $Y_{0,t}^N = \hat{f}\left(X_{0,t}\right)$ for each $t > T_0$, giving us $\hat{\tau}_{0,t} = Y_{0,t} - \hat{Y}^N_{0,t}$. Our estimate of the ATE comes from the sample analog to (3), namely $\hat{\tau} = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\hat{\tau}_{0,t}$.

Formally, we use $\mathcal{X}$ to denote the input space for X_0 and $\mathcal{Y}$ for the output space for Y_0. Any node η represents a subspace $\mathcal{X}_\eta \subseteq \mathcal{X}$ starting from root node η_0 that represents $\mathcal{X}$ itself. Internal nodes η are associated with a split s_η taken from a set of binary questions, e.g. questions of the form "*Does* $X_0 \in \mathcal{X}_A$?", where $\mathcal{X}_A \subset \mathcal{X}$ or "*Were there more than five conflicts in Bahrain?"*. The split s_η divides the input space $\mathcal{X}_\eta$ into two disjoint subspaces $\mathcal{X}_\eta \cap \mathcal{X}_A$ and $\mathcal{X}_\eta \cap (\mathcal{X} \setminus \mathcal{X}_A)$ known as children nodes. The terminal nodes are associated with our best guess of the output value for the treated unit $\hat{Y}_{0,\eta}$. Here, we take splits as given and refer to the standard CART algorithm in Breiman et al. (1984) for details. Let now the global generalization error be given by

$$
\begin{aligned}
\mathcal{L}(f^\star) &= \mathbb{E}_{X_0,Y_0}\left[\ell\left(Y_0, f^\star(X_0)\right)\right] \\
&= \sum_{\eta \in \mathcal{R}} P\left(X_0 \in \mathcal{X}_\eta\right) \mathbb{E}_{X_0,Y_0|\eta}\left[\ell\left(Y_0, \hat{Y}_{0,\eta}\right)\right],
\end{aligned}
\tag{7}
$$

where ℓ is some loss function and $\mathcal{R}$ denotes the set of disjunct terminal nodes. The loss associated with the prediction error for a branch is often called impurity. The inner expectation in (7) is the local generalization error of model $f^\star$ at node η. Minimizing the global generalization error corresponds to minimizing the inner expectation pointwise for all terminal nodes. Hence, the optimal decision tree finds the best constants $\hat{Y}_{0,\eta}$ at each terminal node. Given the squared error loss, the inner expectation in (7) is minimized in η by

$$
\begin{aligned}
\hat{Y}^*_{0,\eta} &= \arg\min_{\hat{Y}_0} \mathbb{E}_{X_0,Y_0|\eta}\left[\left(Y_0 - \hat{Y}_0\right)^2\right] \\
&= \mathbb{E}_{X_0,Y_0|\eta}\left[Y_0\right],
\end{aligned}
\tag{8}
$$

and the feasible solution to (8) can be approximated by the sample analog, i.e.

$$\begin{aligned}\hat{Y}_{0,\eta} &= \arg\min_{\hat{Y}_0} \frac{1}{N_\eta} \sum_{X_0, Y_0 \in \mathcal{D}_\eta} \left(Y_0 - \hat{Y}_0\right)^2 \\ &= \frac{1}{N_\eta} \sum_{X_0, Y_0 \in \mathcal{D}_\eta} Y_0, \end{aligned} \tag{9}$$

where $\mathcal{D}_\eta$ is the subset of the samples falling into node η, that is all pairs (X_0, Y_0) such that $X_0 \in \mathcal{X}_\eta$, and where N_η denotes the number of observations in node η. This leads to the prediction rule as $\hat{f}_{\text{dt}}(X_0) = \sum_{\eta \in \mathcal{R}} \hat{Y}_{0,\eta} \mathbb{1}\{X_0 \in \mathcal{X}_n\}$. Put differently, we are interested in approximating the conditional mean of the output variable at a value of the regressors by taking the average of the output variable over observations that fall into the same category.

Albeit intuitive, decisions trees tend to perform inferiorly in terms of prediction accuracy due to overfitting to sample noise. That is, although decision trees usually have a low bias, the cost is high variance across different realizations of data. Breiman (2001) propose *random forests* as an ensemble extension to decisions trees using bootstrap aggregation to reduce the overfitting. The idea is to draw B bootstrap samples with replacement and grow a deep tree for each sample. However, in each sample, we only consider a subset of covariates, which corresponds to only considering a subset of control units in our model. More precisely, when growing a tree on bootstrap data $\mathcal{D}^b$ $\forall b \in \{1, \ldots, B\}$, only $m \leq N$ of the control units are chosen at random as candidates for each split. Growing all B trees leaves a sequence $\left\{\hat{f}_{\text{dt}}(X_{0,b}, \Theta_b)\right\}_{b=1}^{B}$, where Θ_b summarizes the bth tree in terms of split variables, split points, and values at the terminal nodes. The final step in the random forests algorithm is to average over the B bootstrap samples, i.e.

$$\hat{f}(X_0) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{\text{dt}}(X_{0,b}, \Theta_b), \quad X_{0,b} \in \mathcal{D}_b. \tag{10}$$

This results in a consistent estimator of $f^\star$ in the sense that $\mathbb{E}\left[\hat{f}(X_0) - f^\star(X_0)\right]^2 \to 0$ as $T \to \infty$, where expectation is taken over X_0 and the training data (Theorem 1, p. 7, Scornet et al., 2015). To get confidence intervals around the average of the estimated treatment effects, we recommend a standard nonparametric bootstrap or block bootstrapping.

To continue the example from Section 1, a possible data-generating process (DGP) that falls under the overarching model in (6) would be

$$Y_{IP,t}^{N} = \beta_1 Y_{SA,t} + \beta_2 Y_{IR,t} + \beta_3 Y_{SA,t} Y_{IR,t} + \varepsilon_{0,t}, \tag{11}$$

where $Y_{\cdot,t}$ denotes the conflict level in period $t < T_0$, IP abbreviates Israel-Palestine, SA Saudi Arabia, and IR Iraq. A model that does not take the interaction into account would suffer from omitted variable bias. On the other hand, if we consider a linear DGP as $Y_{IP,t}^{N} = \beta_1 Y_{SA,t} + \beta_2 Y_{IR,t} + \varepsilon_{0,t}$, the random forests model is asymptotically able to recover the linear model as it is essentially a sum of piecewise linear models (averages).

Choosing the best parametrization of the highly flexible tree-based model is essential to avoid overfitting to the pre-intervention period. To see this, imagine a single decision tree that is fully grown. Hence, every leaf contains only one observation. Using this particular tree in the pre-intervention period delivers a mean squared error of exactly zero because it can fit every single observation perfectly, which is not ideal. The same applies to random forests. Therefore, we split further the pre-intervention period into an estimation sample and a validation sample of relative sizes equal to 80% and 20% respectively, keeping the temporal ordering. We estimate the model on the estimation sample and select the model complexity on the validation sample by tuning hyperparameters. We tune the number of control units selected for each tree, namely m. By this data splitting approach, we control the bias and variance of the model. Similar ideas of sample splitting have been suggested by Chernozhukov et al. (2018a) and Chernozhukov et al. (2018b). We note, however, that we obtain essentially identical results using default settings, which is $m = \sqrt{N}$. We grow $B = 500$ trees and implement our tree-based method using the **sklearn** library in Python. Similarly, one could implement the method using the **randomForest** or **ranger** package in

R, and the **TreeBagger** class in MATLAB.

2.4 *Extensions*

First, recent work on synthetic controls focuses on the case of multiple treated units, given its relevance in empirical applications (see for instance Hainmueller (2012), Cavallo et al. (2013), or Robbins et al. (2017)). Incorporating multiple treated units into our framework entails to extending the univariate random forests model with a loss function is expressed by the multivariate nature of the treated units. For instance, De'ath (2002) defines multivariate regression trees analogously to a decision tree with the extension that the loss function is the multivariate sum of squared error losses. The idea of partitioning the space of the explanatory variables into disjoint regions and assigning a constant to each region remains intact. Another extension is provided by Segal and Xiao (2011), who propose multivariate random forests. Again, the core idea is the same and the extension entails to minimizing a covariance weighted loss of the multivariate sum of squared error losses, where the covariance matrix is based on the multivariate response function. The multivariate random forests have for instance been applied by Pierdzioch and Risse (2018) to forecasting multiple metal returns. To estimate the treatment effects on multiple units, we suggest applying the multivariate random forests directly instead of the random forests. This would lead to a vector of counterfactual outcomes for the treated units in each of the post-treatment periods.

Second, a key advantage of regression-based estimators and, in particular, classical synthetic controls is the transparency of the resulting counterfactual prediction due to the estimated weights. In the case of synthetic controls, the counterfactual is a convex combination of control units and a natural generalization of difference-in-differences. In contrast, many nonparametric methods optimized for prediction and, in particular, machine learning methods no not come with such transparency and are often viewed as non-interpretable black boxes. We briefly explain two approaches that would allow one to recover part of the transparency. Particularly for forests, the first approach is based on relative importance measures, where the basic idea is to accumulate over each tree the improvement conducted by each variable in the loss function for every split. For instance, one could compute the conditional variable importance measure proposed by Strobl et al. (2008) that reflects the true impact of each predictor variable. A second approach along the same lines would be to compute SHAP values for each variable as suggested by Lundberg and Lee (2017). Both approaches would allow researchers to assess which of the control units that drive the counterfactual prediction. Note that because our tree-based counterfactual prediction is not a weighted average of control units but an average of treated outcomes in the pre-treatment period, the two approaches would rather assess which of the control units that are important drivers for computing the similarities between subperiods and eventually group them.

Last, we comment on the ability of the model to recover treatment effects beyond the mean. Using random forests, the conditional mean $\mathbb{E}\left[Y_0|X_0 = x\right]$ is approximated by the averaged prediction of B decision trees, which is essentially a weighted mean over the observations of Y_0 with weights depending on (X_0, Y_0). Likewise, one could define an approximation to $\mathbb{E}\left[\mathbb{1}\left\{Y_0 \leq y\right\}|X_0 = x\right]$ by the weighted average over observations of $\mathbb{1}\left\{Y_0 \leq y\right\}$. This approximation is suggested by Meinshausen (2006), leading to quantile regression forests. Quantile regression forests is a consistent estimator of the conditional distributions and the quantile functions. To estimate treatment effects beyond the mean using tree-based controls, we recommend to replace random forests by the quantile random forests and estimate the treatment effects over a range of quantiles.

3 Estimating the Effects of moving the Embassy

3.1 *Background*

Monday afternoon December 6, 2017, the US President fulfilled a major campaign promise by announcing the move of the embassy from Tel Aviv to Jerusalem, which took place May 14, 2018. Many international media reported intensively on the move that broke with decades of US policy by recognizing Jerusalem as the capital of Israel, although former US presidents have also been commenting on the move. For instance, Bill Clinton supported recognizing Jerusalem as the capital and the principle of moving the embassy there. George W. Bush said before taking office that he intended to move the embassy, and Barack Obama spoke of Jerusalem as the capital of Israel that ought to remain undivided. However, the former presidents all consistently signed waivers to postpone the move.

The move should be viewed as the most recent event in the ongoing Israeli-Palestinian conflict, dating back to the mid-20th century in which the Jewish immigration and the sectarian conflict in Mandatory Palestine between Jews and Arabs took place. In 1948, the establishment of the State of Israel alongside the State of Palestine was proclaimed and US President of the time Harry S. Truman recognized the new nation. Since 1967, Israel has held all of the pre-war cities of West and East Jerusalem, and in addition, the Gaza Strip has been under Israel's control. Ever since, several wars have been fought between the Arab countries and Israel, and a permanent solution is still to be found. For a complete review and analysis of the Israeli-Palestinian conflict, see Frisch and Sandler (2004) and Eriksson (2018).

3.2 *Data and Sample*

We use daily country-level panel data in the period December 28, 2015, to November 3, 2018, on conflicts reported by the Armed Conflict Location & Event Data Project (Raleigh et al., 2010). The conflicts cover riots, protests, strategic development, remote violence, violence against civilians, various types of battles, and headquarter or base establishments. We consider the aggregate of all conflicts and leave the disaggregating for further research. The data consist of multiple daily observations, which we aggregate into weekly observations. We have no other data on a daily or weekly frequency. The treated countries considered are Israel and Palestine, which we aggregate into one treated unit to take into account the interdependency of the two countries (Arnon and Weinblatt, 2001). Aggregating them into one treated unit rather than having one of them, say Israel, as a potential control is necessary to meet the assumption of no interference between units. One may be interested in the effects on Israel and Palestine separately, leaving out completely the other country to avoid interference. An interesting hypothesis is whether the conflicts in Palestine accelerate earlier than the conflicts in Israel. However, this is hard to measure as the conflicts in both countries may be initiated by people from either where, making it difficult to disentangle the effect in Israel from the effect in Palestine. As we are interested in the overall effect in the area, we aggregate the countries for now and leave the other hypothesis for future research. We sometimes refer to Israel and Palestine as Israel-Palestine. The control countries we consider are all the remaining countries in the Middle East but Syria and Iran, which include Bahrain, Iraq, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Turkey, United Arab Emirates, and Yemen, giving us a total of 11 control countries. The data coverage for Syria starts from January 2017, and instead of restricting our sample to begin here, we choose to exclude Syria. We also exclude Iran because of its involvement in the Israeli-Palestinian conflict and its relation to the US, which make it too difficult to justify the assumption of no inference between units (see Buonomo (2018) for an analysis of the Iran-US relation). In fact, if we compare the trends in the weekly level of conflicts in Iran and Israel-Palestine before and after the move of the embassy, the co-movement is clear. We document the trends in the weekly number of conflicts for all countries in the Middle East except Syria in Appendix A. The pre-intervention period covers 101 weeks, starting December 28, 2015, and ending December 3, 2017, just before the announcement. The post-intervention period begins on December 4, 2017, and ends on November 3, 2018, leaving 48 weeks for estimating the average level of conflicts in Israel and Palestine in the counterfactual situation where the US embassy is not relocated. Summary statistics for the weekly number of conflicts across the Middle East countries are provided in Table 1. Further, we show the distribution of the weekly number of conflicts in Israel-Palestine in both the pre-treatment and post-treatment period in Figure 2. It follows from Figure 2 that the distribution is shifted to the right in the post-treatment period, which tentatively suggests that violent weeks tend to occur more often in the post-treatment period. Last, Figure 3 shows the level of conflicts over time in Israel-Palestine as well as the average of the remaining countries. Note that when $Y_{0,t}^N$ is stationary, a simple before-after comparison is sufficient to identify the average treatment effect, which in this case would be 23.5 weeks. This simple yet unbiased estimate is roughly in line with the results we show next.

3.3 *Results*

Our application is motivated by Figure 3, showing the weekly number of conflicts in Israel-Palestine over the entire sample period. The two vertical lines indicate the date when the move of the US embassy was announced and the date of the actual move, respectively, and also, we plot the average of the remaining countries. A couple of observations are worth noting. First, visual inspection suggests that the average weekly number of conflicts in Israel-Palestine has in fact increased subsequent to the announcement. In contrast, the average number of weekly conflicts over the remaining countries in the Middle East does not appear to follow the same upward shift after the announcement. We formalize this shortly. Second, the volatility of the weekly number of conflicts in Israel-Palestine seems much higher after the announcement, supporting the histogram in Figure 2. This has important economic implications as it indicates that conflicts tend to cluster and misfortunes never come singly. Considering the conflicts more closely, for instance analyzing the degree of violence in the clusters, is interesting but we postpone this for future research. Finally, note the large spike in the average number of conflicts across the remaining countries in the Middle East around July 2016. Specifically, the week with the highest average number of conflicts runs from July 18 to July 24, which is just after the military coup was attempted in Turkey on July 15 against state institutions, including the government and President Erdoğan. During the coup, more than 2,100 people were injured and over 300 were killed. This rare event shows up in the estimation for some methods that are exposed to outliers.

Figure 4 displays the weekly number of conflicts for Israel-Palestine and its estimated counterpart during the period December 28, 2015, to November 3, 2018. The observed level of conflicts in Israel-Palestine is closely followed by the estimated counterpart in the entire pre-intervention period until the move was announced on December 4, 2017. This suggests that the time periods before the announcement can be grouped together into homogeneous subgroups based on the level of conflicts in the neighboring countries, and for these subgroups of time periods, the level of conflicts in Israel and Palestine is relatively constant. In fact, the average of the observed weekly number of conflicts in the pre-intervention period is 25.32, whereas the estimated counterpart is 25.41, indicating an accurate fit on average. Note that the estimated counterpart to Israel-Palestine is always closer to the average level of weekly conflicts instead of capturing the spikes to the fullest extent. The is an attractive feature of the averaging that happens in our model.

Altogether, we take this as evidence that the tree-based synthetic control method can be used to predict a counterfactual Israel-Palestine, which provides a sensible approximation to the true level of conflicts that would have occurred in that region in absence of the move. Thus, we next use the tree-based synthetic control method to estimate the average treatment effect of moving the embassy.

We estimate the effect of the move of the US embassy for each of the 48 weeks after the announcement as the difference between the observed level of conflicts in Israel-Palestine and its counterfactual analog. The differences follow as the discrepancies between the two lines in the shaded area of Figure 4. Immediately after the move is announced, both the observed and counterfactual level of conflicts increase but to very different degrees, and in fact, the observed level of weekly conflicts in Israel and Palestine reaches its maximum level across the entire sample within the first week of the announcement. For the rest of the post-announcement period, the observed level of conflicts experiences a higher base level with distinctly conflict-ridden weeks, whereas the counterfactual Israel-Palestine maintains the lower base level from the pre-announcement period. Specifically, the average of the observed number of weekly conflicts in the post-intervention period is 48.88, whereas the estimated counterpart is 22.78, indicating a significant difference. This suggests that the move of the embassy has a numerically positive effect on the level of conflicts in Israel and Palestine, meaning that the level generally increases in the entire post-announcement period.

We assess the weekly estimates of the impact directly in Figure 5, where we plot the differences between the observed and estimated number of weekly conflicts in Israel and Palestine. Figure 5 unveils the same story as Figure 4. The gap of approximately zero on average in the pre-intervention period indicates that the tree-based synthetic control method is able to approximate well the true level of conflicts albeit very fluctuating. To be precise, the average difference between the observed and estimated weekly number of conflicts in the pre-intervention period is only -0.09. Using all 48 weeks after the announcement, our results show that the level of conflicts in Israel and Palestine is increased by an average of more than 26 incidents per week, which corresponds to an increase of approximately 103%. The estimated average effect is associated with a bootstrapped standard error of 2.67 using 10,000 block bootstrap samples with block length equal to 3. That is, the 95% bootstrap confidence interval of the weekly increase is between 20.88 and 31.36. This translates into a percentage point change between roughly 82-124%. We acknowledge that the confidence interval is rather wide, which is not surprising due to the volatility in the number of conflicts across weeks. The results are insensitive to the choice of block length.

Naturally, the assumption of no interference between the treated and control units can be violated in several ways in the context of analyzing the effect of moving the US embassy. The Israeli-Palestinian conflict is an issue in all of the region, and the ties between the countries are complex to understand. For instance, we choose to exclude Iran in the sample, because the Iranian government has played an active role in the conflict. The results with and without Iran are, however, not significantly different, because the tree-based synthetic control method averages over the number of conflicts in Israel-Palestine and uses only the neighboring countries to partition the time periods. This feature of the method makes it more robust to the potential violations compared to methods that base the estimates on the outcomes for the control units. Further, the average weekly number of conflicts across all control countries does not differ between the pre- and post-intervention period. In particular, the average over the control countries in the pre-intervention period is 32.80, whereas the same figure is 30.82 in the post-intervention period. The small difference is likely to be driven by the coup attempt in Turkey. The placebo tests we review shortly reveal that no other relevant country experienced the same effect of the move of the US embassy. Last, the conformal inference test in Section 3.4 provides evidence that our model is correctly specified and that the increase is statistically significant. Taken altogether, it is our judgment that the potential violations do not appear to be severe in this context.

3.4 *Inference*

We want to assess how much our results are driven by mere chance. If we are able to obtain estimated effects of the same magnitude for the control countries as for Israel-Palestine by relabeling treatment and control unit, we would not be able to interpret our analysis as providing any significant effects. To make inference about the effect of the embassy move, we follow the strategy outlined in Abadie et al. (2010), Bertrand et al. (2004), and Abadie and Gardeazabal (2003) and run placebo tests. Placebo tests re-do the original analysis but switch the roles between the treated unit and a randomly chosen control unit, the rationale being that using the control unit not exposed to treatment should lead to an estimated effect of approximately zero. By applying the tree-based synthetic control method individually to all the countries in the donor pool, we can therefore evaluate the significance of our analysis. We expect one of two outcomes. If the placebo tests deliver estimates of the average effect of similar magnitude as for Israel-Palestine, we cannot rightfully interpret our results as evidence for a significant effect. If, on the other hand, that none of the placebo tests for the countries in which the US embassy was not moved lead to similar estimated effects, then we take this as evidence that our tree-based analysis documents a significant effect of moving the US embassy in terms of an increased level of conflicts. One condition, however, is that the pre-intervention fit to the weekly number of conflicts is precise for the country in question when we run the placebo test.

To assess the significance of our estimates, we perform a series of placebo test for which we create a counterfactual state of the world. That is, we iteratively treat each control country in the remaining Middle East as if it had experienced a move of the US embassy at exactly the same time as the move in Israel, while we also reassign both Israel and Palestine to the control group. In each iteration, we apply tree-based controls to the respective country to estimate the impact of the fictive embassy move on the weekly number of countries. The series of placebo tests gives us a distribution of differences between the observed and estimated number of conflicts over the countries.

Figure 6 plots the differences in the observed and estimated number of conflicts for all the placebo analyses and the original analysis. The blue line shows the case for Israel-Palestine, reproducing Figure 5. The other lines show the same differences estimated by the tree-based synthetic control method but for each of the 11 control countries in the donor pool. Figure 6 indicates that the tree-based synthetic control method provides an accurate fit in the pre-intervention period for Israel and Palestine as well as for most of the control countries. In particular, the pre-intervention root mean squared prediction error (RMSPE) for Israel-Palestine is 5.77, where RMSPE is computed as the root average of the squared differences between the observed and estimated weekly number of conflicts. The pre-intervention median RMSPE for the control countries is 1.71. This should not be taken as evidence that the ability to fit the pre-intervention is higher for the control countries than for Israel-Palestine. In fact, mean RMSPE over the control countries is 9.51, indicating that a few control countries stand out in terms of high RMSPE while for most control countries, we achieve a very low RMSPE. This is supported by Figure 6 from which it is apparent that the pre-intervention fit is very imprecise for some countries. The country with the worst fit is Turkey with an RMSPE of 61.88. This result, however, is not surprising due to the attempted military coup in 2016 that led to an extreme spike in the number of conflicts. As this coup attempt was, of course, unanticipated, the conflict situation in the other countries was normal, and therefore, no statistical method would be able to capture this outlier. Similar problems arise for Iraq and Yemen, which are the countries with the overall highest variation in the weekly number of conflicts. This high variation makes it difficult for the tree-based synthetic control method, and likely any other method, to produce a valid fit in the pre-intervention period without imposing too much flexibility. As a result, the RMSPE for Turkey, Iraq, and Yemen are all more than double that of Israel-Palestine and any other control country.

To handle the countries for which the tree-based synthetic control method gives a poor fit, we follow an argument provided in Abadie et al. (2010) as they encounter the same issue for some of the states. If the tree-based synthetic control method had failed to deliver a reasonable fit to the observed weekly level of conflicts in the pre-intervention period for Israel-Palestine, we would treat the lack of fit as evidence that the estimated increase in the weekly number of conflicts in the post-intervention period was arbitrary and not caused by the move of the US embassy. Analogously, we cannot take into account the estimated effects in the post-intervention period for Turkey, Iraq, and Yemen when assessing the degree of chance in our results for Israel-Palestine. Consequently, we provide another version of Figure 6 in which we have excluded the placebo tests for Turkey, Iraq, and Yemen. This effectively corresponds to removing countries for which the RMSPE is more than double the one for Israel-Palestine. Figure 7 provides the restricted version of Figure 6 from which we have excluded Turkey, Iraq, and Yemen. The median RMSPE over the remaining countries in the Middle East drops to 0.35, and the corresponding mean drops to 1.37. Removing the countries for which the tree-based synthetic control method would be ill-advised tells a clear message. The largest estimated effect on the weekly number of conflicts in the post-intervention period is to be found for Israel-Palestine. More precisely, while the average estimated effect for Israel-Palestine is 26.12 in the post-intervention period, the corresponding figure over the placebo tests is 1.38. For the pre-intervention period, the estimated gaps are -0.09 and -0.02, respectively.

We emphasize that placebo tests as a mode of inference evaluates significance *relative* to a benchmark distribution for the given assignment mechanism in the data as opposed to a random assignment mechanism. This is important because our intervention is not randomly assigned, which is rarely the case in comparative case studies. This makes it impossible to impose the cross-unit exchangeability condition that underlies the sampling-based statistical tests (Abadie, 2019). In our case, however, the design-based inference mitigates this by conditioning on the data and exploiting the assignment mechanism.[2]

We consider another approach to assessing the significance of our results, namely computing ratios of post/pre-intervention measures both for Israel-Palestine and the control countries. As Abadie et al. (2010), we compute the ratios in terms of RMSPE. Arguably, the advantage of comparing ratios relative to post-intervention gaps is that we do not necessarily have to exclude ill-fitting placebo runs in an iterative way as demonstrated by figures 6 and 7. For instance, although the RMSPE for Turkey is the highest across all in the pre-intervention period, it is similarly high in the post-intervention period, and the ratio will be more robust to this. The only countries with a higher ratio of post/pre-intervention RMSPE than Israel-Palestine are Jordan and Oman. This observation, however, does not cause much concern when we take into account the gaps in both periods. For Jordan, the pre-intervention gap between the observed and estimated weekly number of conflicts is -0.02, whereas the same figure is 0.46 in the post-intervention period. Likewise, the figures for Oman are -0.00 and 0.06, respectively. Thus, the high ratios of post/pre-intervention RMSPE for the two countries are likely driven by a few very conflict-ridden weeks after the intervention. In addition to the ratios of post/pre-intervention RMSPE used in Abadie et al. (2010), we also compute the ratios of post/pre-intervention mean absolute error (MAE) between the observed and estimated weekly number of conflicts. Using either the ratio of post/pre-intervention RMSPE or MAE have different advantages. RMSPE penalizes large errors more than MAE,

[2]We thank Alberto Abadie for pointing this out.

but MAE is more interpretable. We provide both ratios for each country in Table 2, in which we also provide the respective pre- and post-intervention measures. Note from Table 2 than Oman is the only country with a higher ratio of post/pre-intervention MAE than Israel-Palestine. In absolute terms, again, the result for Oman is not too disturbing for our analysis.

3.4.1 *Exact and Robust Conformal Inference*

We consider one last approach to draw inference about our results. Recall that our proposed method as well as the other methods considered rely on cross-sectional regressions. Whenever the joint distribution of the data is not well-approximated by cross-sectional regressions, the model will provide a poor global fit in the sense that not all N controls will fit the model, which is exactly the case in our application as well as in Abadie et al. (2010). In this situation, Chernozhukov et al. (2017b) propose an exact and robust conformal inference method. The method requires only a good *local* instead of a good *global* fit as it relies solely on a suitable model for the treated unit and it focuses on the time-series dimension. Essentially, the procedure postulates a null trajectory $\boldsymbol{\tau}^o = \{\tau_t^o\}_{t=T_0+1}^T$ and test the sharp null hypothesis $\mathcal{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}^o$. For the test to be valid, the estimator of the counterfactual outcome for the treated unit needs to be consistent and stable, and be able of providing residuals that are exchangeable. To assess the plausibility of the key assumptions, Chernozhukov et al. (2017b) provide placebo specification tests. The conditions result in non-asymptotic validity of the test, meaning that the p-value is approximately unbiased in size (Theorem 1, p. 23, Chernozhukov et al., 2017b). The proposed inference method is valid for stationary and weekly dependent data.

We are interested in testing the hypothesis that the trajectory of the policy effects in the post-treatment is zero. Hence, our main hypothesis is

$$\mathcal{H}_0 : \boldsymbol{\tau} = \boldsymbol{\tau}^o, \quad \text{where } \boldsymbol{\tau}^o = \underbrace{(0, \dots, 0)'}_{|T-T_0|\times 1} \tag{12}$$

The test statistic S is based on the $((T-T_0)\times 1)$ vector of residuals of our model $\hat{\mathbf{u}} = (\hat{u}_{T_0+1},\ldots,\hat{u}_T)'$. The test statistic is then defined by

$$S(\hat{\mathbf{u}}) = S_q(\hat{\mathbf{u}}) = \left(\frac{1}{\sqrt{T-T_0}}\sum_{t=T_0+1}^{T}|\hat{u}_t|^q\right)^{1/q}, \tag{13}$$

where we set $q=1$. To compute p-values, the test relies on two different sets of permutations, the i.i.d permutations denoted $\Pi_{\text{i.i.d}}$ and the moving block permutations denoted $\Pi_{\rightarrow}$. The moving block permutations are necessary if the sequence of residuals exhibits serial dependence. The p-value is estimated as $\hat{p} = 1-\hat{F}(S(\hat{\mathbf{u}}))$, where

$$\hat{F}(x) = \frac{1}{|\Pi|}\sum_{\pi\in\Pi}\mathbb{1}\left\{S(\hat{\mathbf{u}}_\pi < x)\right\}. \tag{14}$$

To assess the validity of the assumptions underlying the test, the first step is to perform a placebo specification test. Based on the outlined procedure, the idea is to test the null hypothesis that

$$\mathcal{H}_0 : \tau_{T_0-\kappa+1} = \cdots = \tau_{T_0} = 0, \tag{15}$$

for a given $\kappa \geq 1$ based on pre-treatment data. The null hypothesis (15) is true if the underlying assumptions are correct. Thus, rejecting the null provides evidence against a correct specification. For proofs and additional details, we refer to Chernozhukov et al. (2017b).[3]

We begin the analysis by testing the underlying assumptions of our proposed method, i.e. consistency, stability, and exchangeability of the residuals. We apply both i.i.d. permutations and the moving block permutations. We use $\kappa = 10$ and randomly sample 10,000 elements from the set of all permutations with replacement for the i.i.d. permutations. The resulting p-values follow from Table 3. All p-values from both permutation schemes are above 60% and most of them are above 80%, and thus, we fail to reject the null hypothesis. This serves

[3]Note that Chernozhukov et al. (2017b) also provide a test for the average effect over time. However, this requires the total number of periods to be much larger than the post-treatment periods, which is not

as evidence for a correct model specification. We further see that the p-values differ slightly between the i.i.d. permutations and the moving block permutations, where the p-values tend to be lower using moving block permutations. This provides evidence for some serial dependence in the residuals.

Next, we turn to our main hypothesis in (12). We consider again both the i.i.d. permutations with 10,000 random samples as well as the moving block permutations. The p-value based on the i.i.d. permutations is 0.000, whereas the p-value based on the moving block permutations is 0.007. We reject the null hypothesis in both cases given both p-values are below 1%, providing evidence that the trajectory of the policy effects from the embassy move is different from zero. The formal test results thus appear to be in agreement with the other inference results provided in this section.

4 Comparing Methods

In Section 3, we provide evidence that the decision to move the US embassy from Tel Aviv to Jerusalem has resulted in a significant increase in the weekly number of conflicts in Israel and Palestine. We assess the robustness of our results in several ways, including performing formal inference tests, conducting a series of placebo runs, and evaluating the fit on different measures such as ratios of post/pre-intervention RMSPE and MAE. In this section, we compare the tree-based synthetic control method to three state-of-the-art methods in the econometric literature. First, we apply the synthetic control method, serving as a baseline model. Then, we apply the regularized counterpart, i.e. the elastic net estimator. Recall that in addition to the systematic selection of comparison groups, the synthetic control group improves upon difference-in-difference approaches by accounting for the effects of confounders changing over time (Abadie et al., 2015). The elastic net generalizes the synthetic control by allowing the weights to be negative and their sum to differ from one. Both methods can be viewed as vertical regressions as pointed out by Athey et al. (2018), where vertical regressions refer to models that regress the outcomes of the treated unit on the outcomes of the control unit in the pre-treatment period and use the estimated relationship in the post-treatment period. Alternatively, one could regress the post-treatment outcomes on the pre-treatment outcomes using only the controls, known as a horizontal regression. The matrix completion method combines elements from vertical and horizontal regressions, and it is the last method we include.

Figure 8 shows the observed and estimated number of weekly conflicts in Israel-Palestine for all four methods, and two features of the methods are noticeable. First, the fit in the pre-intervention period gives an idea of the ability to approximate the weekly level of conflicts in Israel-Palestine, which is highly fluctuating. The synthetic control method, the elastic net estimator, and the matrix completion method are comparable in terms of pre-intervention fit, the matrix completion method being marginally in the lead. The reason the elastic net estimator performs slightly better compared to the synthetic control method is likely because the elastic net is less restrictive when estimating weights. None of the comparison methods, however, are able to approximate the weekly level of conflicts in the pre-intervention period as well as the tree–based control method.

Second, the variation in the estimated counterfactuals in the post-intervention period hints at the degree of overfitting, and both the elastic net estimator and the tree-based synthetic control method appear to deliver reasonable variation in the estimates. They are able to fit the shape and pattern but not the level of the observed conflicts. The ability to fit shape not level is exactly what leads us to estimate a significant effect of the embassy move. In contrast, the estimates by the synthetic control method and the matrix completion method have little variation and are closely centered around the average weekly number of conflicts in the pre-intervention period. This is a sign of overfitting. However, given the data available and in particular the number of control units, this is not surprising. Recall the matrix completion method combines elements from vertical and horizontal regressions.

---

the case in our application.

For the horizontal part, it tries to fit the post-intervention outcomes to the pre-intervention outcomes using only 11 control countries. As the number of weeks is much greater than the number of control countries, it is not surprising that horizontal regressions do perform better.

Figure 9 conveys the same insights as Figure 8, but instead of showing the observed and estimated number of weekly conflicts separately, it displays the differences between the two. Considering the differences instead of actuals provides an easier approach to evaluating pre-intervention fit. Again, a good ability to approximate the pre-intervention level of conflicts corresponds to differences closely around zero. As apparent in Figure 9, the tree-based synthetic control method delivers the best pre-intervention fit, followed by the matrix completion method, the elastic net estimator, and the synthetic control method. It is, however, impossible to assess the overfitting indicated by little post-intervention variation from Figure 9.

From figures 8 and 9, we have argued that the tree-based synthetic control method performs at least as well as state-of-the-art methods. Supporting this, Table 4 provides the various measures that follow from the figures. In particular, we compute the RMSPE and MAE in the pre-intervention period for all the methods considered. Both measures capture the ability to approximate the observed weekly level of conflicts in Israel-Palestine. The tree-based synthetic control method outperforms all other methods on these metrics. We also report the standard deviation of the estimated number of weekly conflicts in the counterfactual Israel-Palestine absent of the embassy move. The elastic net estimator is the only comparison method that delivers higher variation than the tree-based synthetic control method. The matrix completion method delivers almost no variation in the estimates.

Evaluating the degree of overfitting by computing standard errors is somewhat insufficient. One final approach to assessing simultaneously the ability of the methods to approximate the weekly number of conflicts in Israel-Palestine and the degree of overfitting is to repeat the analysis but hold out a subsample of the pre-intervention period and compute the RMSPE and MAE on this subsample. The hold-out sample serves as a test sample, but in contrast to the post-intervention period, we observe $Y_{0,t}^N$ as the intervention has not yet occurred. This allows us to evaluate the predictive ability. Specifically, we hold out the last 10% of the observations in the pre-intervention period, resulting in an estimation sample and a validation sample. Then, we re-run all methods on the estimation sample. For the methods that require tuning of hyperparameters, namely the tree-based synthetic control method, the elastic net estimator, and the matrix completion method, we further split the estimation sample using an 80/20% split as in the original analysis. We use the 20% to select the hyperparameters rather than selecting hyperparameters on the full estimation sample. For the synthetic control method, we use the whole estimation sample to estimate the weights for each country as it does not require any hyperparameters. Having estimated all parameters, we apply all the methods on the validation sample for which we know the true outcome and compute RMSPE and MAE.

Table 5 shows the results of the hold-out sample approach. The elastic net estimator performs best in terms of both metrics, followed by the tree-based synthetic control method, the synthetic control method, and lastly the matrix completion method. Our suspicion that the matrix completion method overfits as seen in Figure 8 appears to be confirmed. We emphasize that this is not an objection to the method rather than a result of the structure of the data, namely $T \gg N$. The elastic net estimator performs very well on the validation sample, and in fact, better than evaluated on the entire pre-intervention period. Normally, we would take this as a sign of underfitting, but as we run more than 20 different specifications of the elastic net estimator in the pre-intervention period, it is more likely caused by the validation sample being too small. The tree-based control performs comparably in the validation sample as in using the entire pre-intervention period, which indicates that neither overfitting nor underfitting takes place. Being a nonparametric method, however, it requires more data and the fact that we only estimate the hyperparameters using roughly 70% of the pre-treatment data seems critical in this assessment of the fit. Ideally, we would use a larger validation sample to compare the methods on validation RSMPE and MAE.

5 Conclusion

The synthetic control method is an effective method in comparative case studies in which relatively more time periods than potential control units are available. The main advantage is the data-driven approach to control unit selection. Since the estimation of the synthetic controls is performed to maximize the pre-treatment fit to the treated unit, however, the fit may not carry over into the post-treatment period. One can argue that synthetic controls are not designed to balance bias for variance, which may lead to overfitting to the pre-treatment period despite the importance of high predictive performance in the post-treatment period. The elastic net estimator is an extension that regularizes the weights on the control units to improve the post-treatment fit. Both methods, however, impose a linear model that may not be guided theoretically. In addition, if interactions and higher-order terms of the control units are important to approximate the treated unit but difficult to anticipate, the estimators may suffer from bias. We recast the problem of estimating a counterfactual state as a prediction problem. Specifically, we provide a data-driven method that balances bias and variance to achieve post-treatment accuracy and is able to capture nonlinearities without the need for a researcher specifying them. Our method can be applied in domains without theoretical guidelines and is also able to recover linear models. We achieve predictive accuracy because we replace the linear component of the synthetic controls with a powerful model inspired by machine learning, namely the random forests model. The ability to capture nonlinearities in a data-driven way is a special feature of this model. This makes the tree-based synthetic control method powerful yet simple.

To demonstrate the applicability of the tree-based synthetic control method, we evaluate the move of the US embassy from Tel Aviv to Jerusalem. Specifically, we estimate the weekly number of conflicts in Israel and Palestine in the counterfactual state of the world absent of the embassy move. The estimates cover the period from the announcement of the move on December 6, 2017, until November 3, 2018. Comparing the estimates to the observed numbers, we find that the average number of weekly conflicts in Israel and Palestine has increased by more than 26 incidents since the move was announced. By placebo tests, we show that the estimated effect of the embassy move is very unlikely to be replicated if one were to relabel arbitrarily the treated unit in the data given that the pre-treatment fit is reasonable. To formally justify our results, we apply exact and robust conformal inference tests and find statistical significance at the 1% level. We further compare the tree-based controls to state-of-the-art methods and conclude that our method is data-driven and needs no linearity assumptions, while it is not dominated even by the best of the comparison methods. All comparison methods agree on the magnitude of the effect.

References

Abadie, A. (2019). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*.

Abadie, A. and Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10:465–503.

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

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2):495–510.

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

Arnon, A. and Weinblatt, J. (2001). Sovereignty and Economic Development: the Case of Israel and Palestine. *The Economic Journal*, 111(472):291–308.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2018). Matrix Completion Methods for Causal Panel Data Models. *NBER Working Paper no. 25132*.

Athey, S. and Imbens, G. (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Biau, G. (2012). Analysis of a Random Forests Model. *J. Mach. Learn. Res.*, 13:1063–1095.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*, volume 8 of *The Wadsworth and Brooks-Cole statistics-probability series*. Taylor & Francis.

Buonomo, T. (2018). Iran's Supreme Leader: An Analysis of His Hostility Toward the US and Israel. *Middle East Policy*, 25(1):33–45.

Card, D. (1990). The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial and Labor Relations Review*, 43(2):245–257.

Cavallo, E., Galiani, S., Noy, I., and Pantano, J. (2013). Catastrophic Natural Disasters and Economic Growth. *The Review of Economics and Statistics*, 95(5):1549–1561.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017a). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, 107(5):261–265.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018b). Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. *NBER Working Paper no. 24678*.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2017b). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. *arXiv e-prints*, page arXiv:1712.09089v7.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2017). Practical and robust t-test based inference for synthetic control and related methods. *arXiv e-prints*, page arXiv:1812.10820v4.

De'ath, G. (2002). Multivariate Regression trees: A New Technique for Modeling Species-Environment Relationships. *Ecology*, 83(4):1105–1117.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. *NBER Working Paper no. 22791*, (22791).

Eriksson, J. (2018). Master of None: Trump, Jerusalem and the Prospects of Israeli-Palestinian Peace. *Middle East Policy*, 25(2):51–63.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science*, 345(6203):1502–1505.

Frisch, H. and Sandler, S. (2004). Religion, State, and the International System in the Israeli–Palestinian Conflict. *International Political Science Review*, 25(1):77–96.

Guha, R. and Ng, S. (2019). *A Machine Learning Analysis of Seasonal and Cyclical Sales in Weekly Scanner Data*. University of Chicago Press.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423, International Convention Centre, Sydney, Australia. PMLR.

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

Ishwaran, H. and Kogalur, U. B. (2010). Consistency of Random Survival Forests. *Statistics & Probability Letters*, 80(13-14):1056–1064.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–95.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Meinshausen, N. (2006). Quantile Regression Forests. *J. Mach. Learn. Res.*, 7:983–999.

Montgomery, J. M. and Olivella, S. (2018). Tree-Based Models for Political Science Data. *American Journal of Political Science*, 62(3):729–744.

Pierdzioch, C. and Risse, M. (2018). Forecasting Precious Metal Returns with Multivariate Random Forests. *Empirical Economics*, pages 1–18.

Raleigh, C., Linke, A., Hegre, H., and Karlsen, J. (2010). Introducing ACLED: an Armed Conflict Location and Event Dataset: Special Data Feature. *Journal of Peace Research*, 47(5):651–660.

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. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, 102(477):191–200.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of Random Forests. *The Annals of Statistics*, 43(4):1716–1741.

Segal, M. and Xiao, Y. (2011). Multivariate Random Forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):80–87.

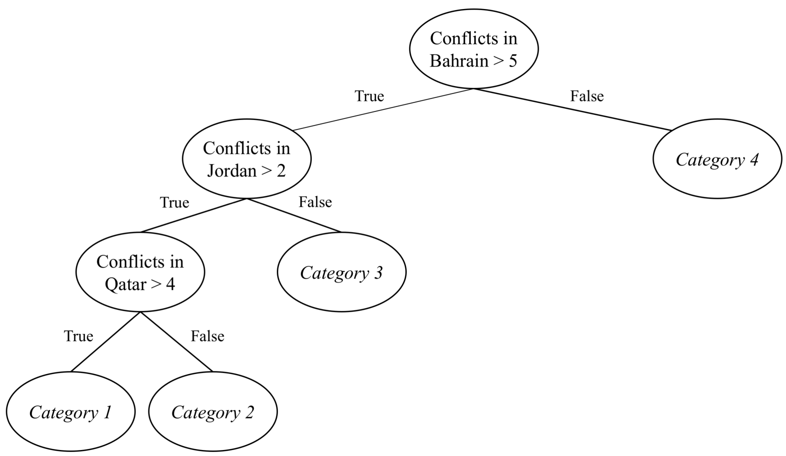
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1).

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, pages 1–15.

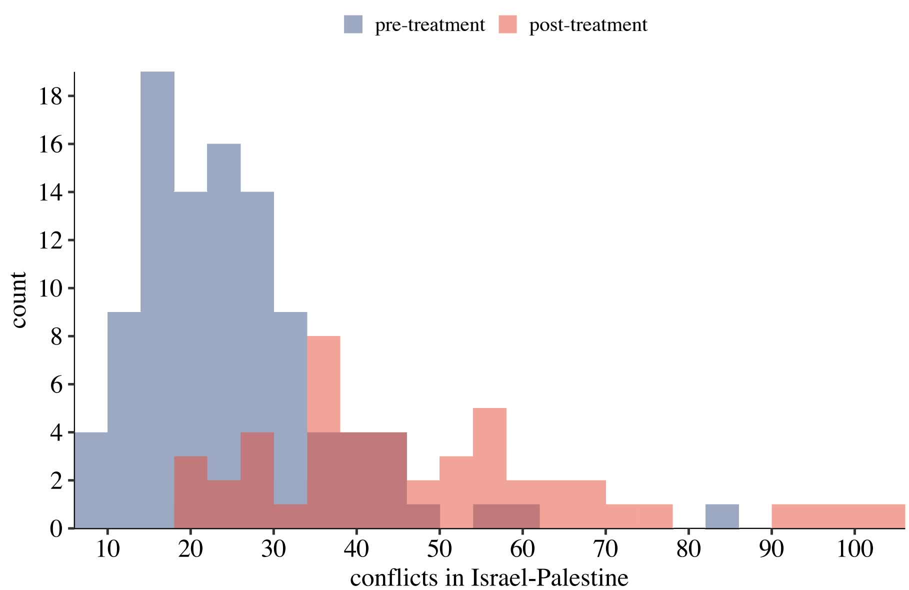
Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.

Figure 1: An example of a decision tree.



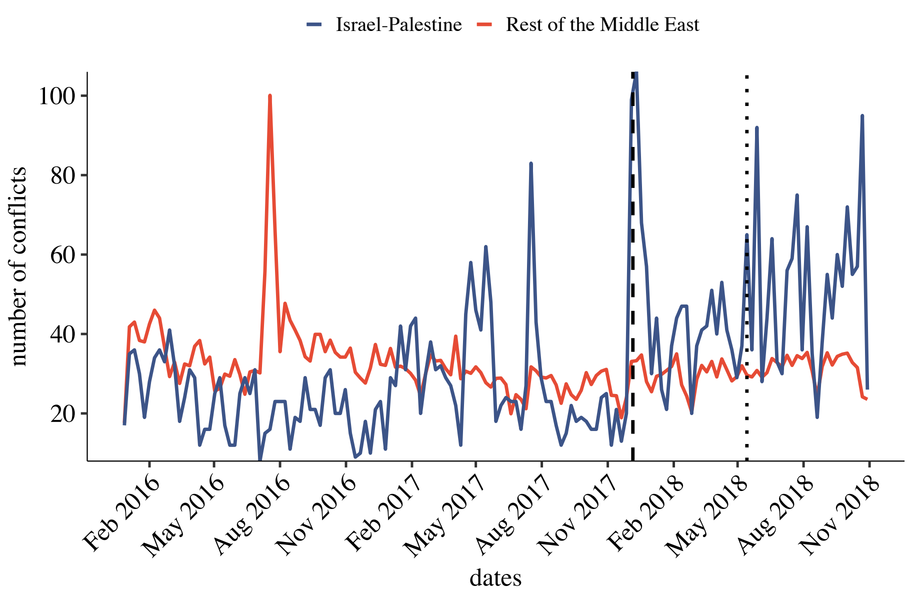
Notes: As input variables, we consider the level of weekly conflicts in Bahrain, Jordan, and Qatar. First, we stratify observations depending on whether or not the level of weekly conflicts in Bahrain is above five. This will place any observation in one of two halves. Next, we partition the subset into whether or not the weekly level of conflicts in Jordan is above two, etc. The recursive stratification leaves us with four distinct categories in which each point in time belongs to exactly one.

Figure 2: Distribution of weekly conflicts in Israel-Palestine.



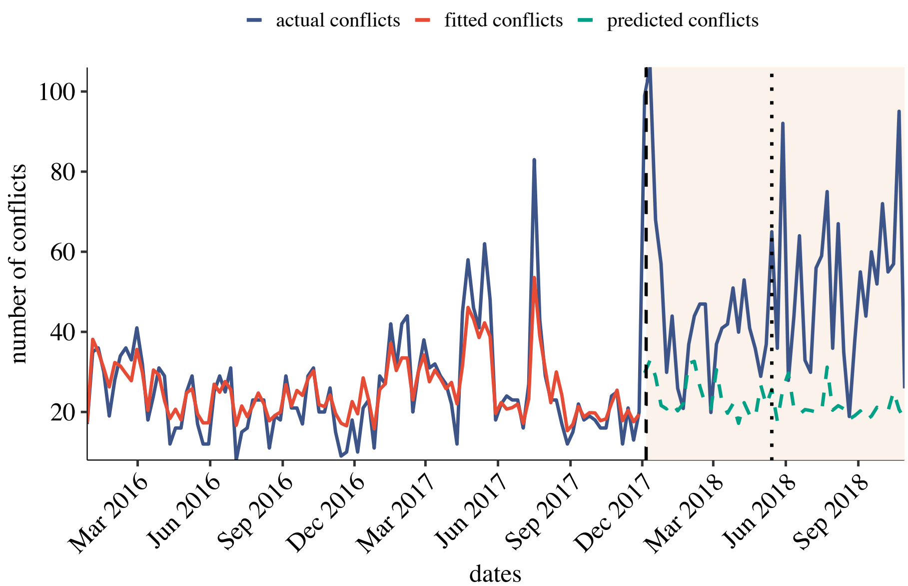
Notes: Distribution of weekly conflicts in Israel and Palestine pre-treatment (blue) and post-treatment (red). The conflicts cover riots/protests, strategic development, remote violence, violence against civilians, various types of battles, and headquarters or base established.

Figure 3: Weekly number of conflicts in Israel-Palestine and the Middle East.



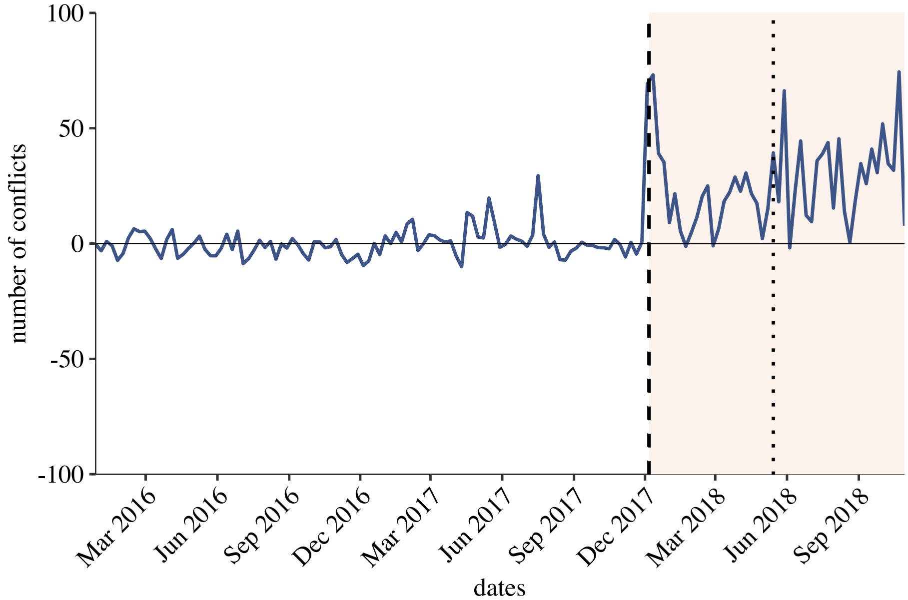
Notes: Weekly number of conflicts in Israel and Palestine (blue line) in addition to the average of the remaining countries in the Middle East (red line). The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively.

Figure 4: Weekly number of conflicts in Israel-Palestine and its estimated counterpart.



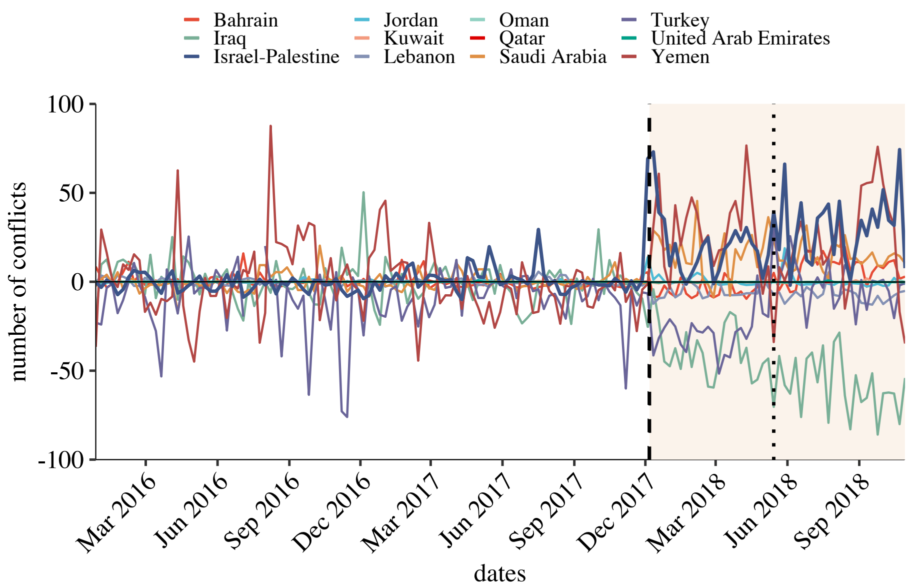
Notes: Weekly number of conflicts in Israel and Palestine (blue line) and its estimated counterpart in the pre-intervention period (red line) and post-treatment period (green dashed line). The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively.

Figure 5: Discrepancies between the observed and estimated conflicts in Israel-Palestine.



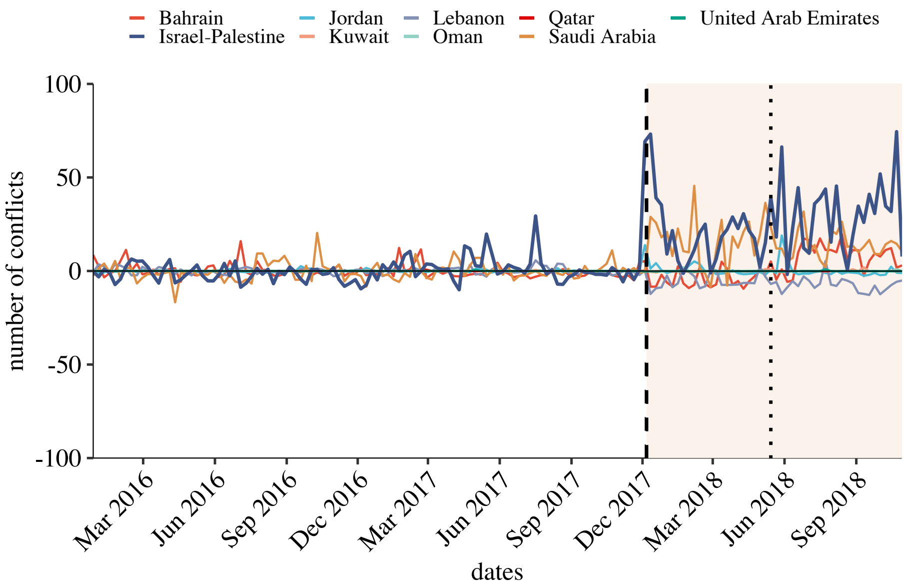
Notes: Weekly gaps between the number of observed and estimated conflicts in Israel and Palestine. The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively.

Figure 6: Discrepancies between the observed and estimated conflicts in the Middle East.



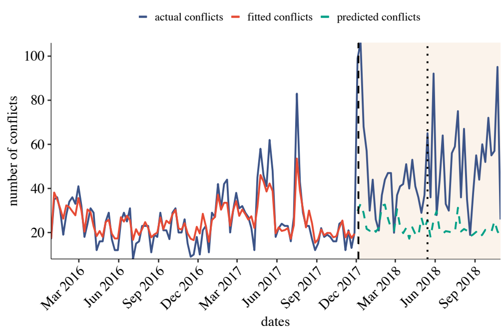
Notes: Weekly gaps between the number of observed and estimated conflicts for all countries considered in the placebo tests. The blue line represents the differences for Israel and Palestine, whereas the other lines represent the differences for the control units defined temporarily as treated units. The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively.

Figure 7: Discrepancies between the observed and estimated conflicts in the Middle East except Turkey, Iraq, and Yemen.

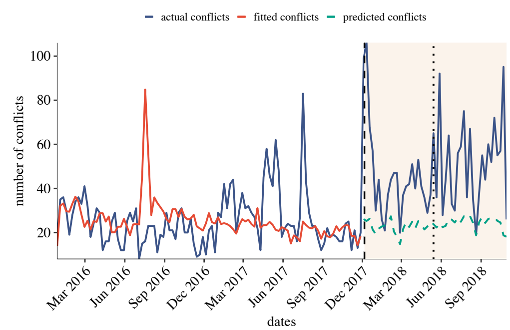


Notes: Weekly gaps between the number of observed and estimated conflicts for all countries considered in the placebo tests except Turkey, Iraq, and Yemen. The blue line represents the differences for Israel and Palestine, whereas the other lines represent the differences for the control units defined temporarily as treated units. The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively.

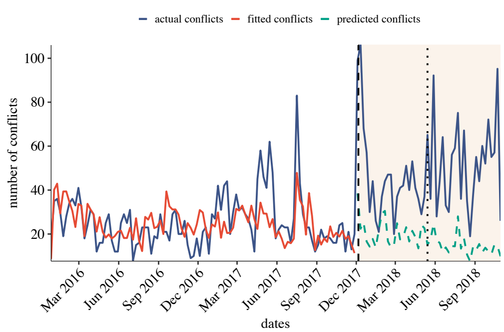
Figure 8: Comparison of the four methods based on the observed and estimated conflicts in Israel-Palestine.



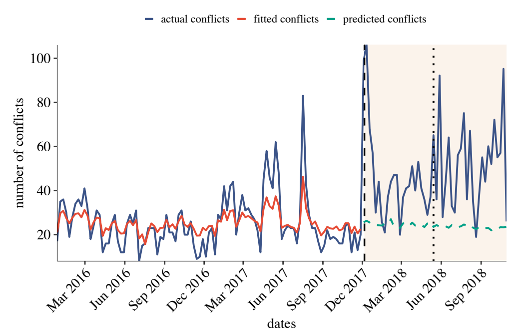
(a) Estimation by tree-based controls



(b) Estimation by synthetic controls



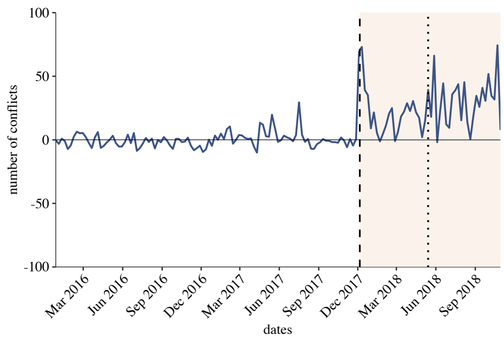
(c) Estimation by elastic net



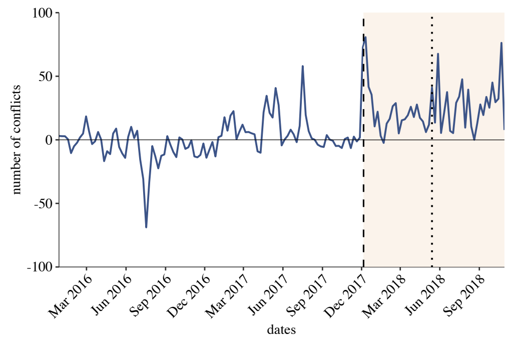
(d) Estimation by matrix completion

Notes: Comparison of the four methods showing the weekly number of conflicts in Israel and Palestine (blue line) and its estimated counterpart in the pre-intervention period (red line) and post-intervention period (green dashed line). The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively. (a) shows the result of the tree-based controls, (b) for the synthetic controls, (c) for the elastic net, and (d) for the matrix completion.

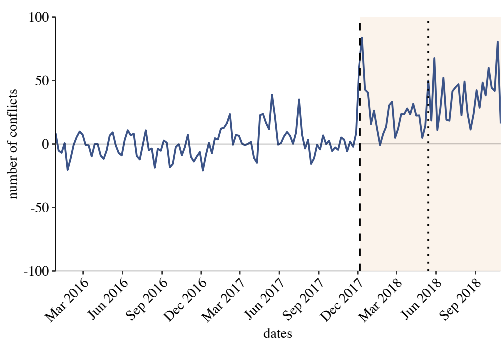
Figure 9: Comparison of the four methods based on discrepancies between the observed and estimated conflicts in Israel-Palestine.



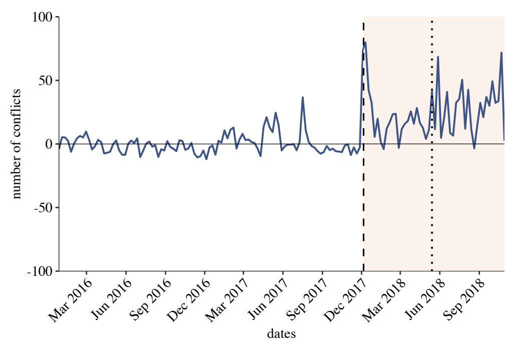
(a) Estimation by tree-based controls



(b) Estimation by synthetic controls



(c) Estimation by elastic net



(d) Estimation by matrix completion

Notes: Comparison of the four methods showing gaps between the observed and estimated weekly number of conflicts in Israel and Palestine (blue line). The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively. (a) shows the result of the tree-based controls, (b) for the synthetic controls, (c) for the elastic net, and (d) for the matrix completion.

Table 1: Summary statistics of the weekly conflicts in the Middle East, excl. Iran and Syria

Country	**Mean**	**Sd.**	**Min**	**Q1**	**Median**	**Q3**	**Max**
Israel-Palestine	32.9	18.7	8.0	20.0	29.0	41.0	106.0
Bahrain	6.8	6.9	0.0	1.0	5.0	11.0	31.0
Iraq	96.8	33.8	32.0	65.0	97.0	120.0	186.0
Jordan	1.4	2.6	0.0	0.0	1.0	2.0	21.0
Kuwait	0.1	0.4	0.0	0.0	0.0	0.0	2.0
Lebanon	6.2	4.8	0.0	3.0	5.0	9.0	25.0
Oman	0.0	0.2	0.0	0.0	0.0	0.0	2.0
Qatar	0.0	0.1	0.0	0.0	0.0	0.0	1.0
Saudi Arabia	27.8	15.8	0.0	17.0	27.0	39.0	75.0
Turkey	46.0	75.4	6.0	22.0	34.0	51.0	777.0
United Arab Emirates	0.0	0.1	0.0	0.0	0.0	0.0	1.0
Yemen	168.7	39.5	72.0	137.0	173.0	197.0	313.0
Average (excl. Israel-Palestine)	32.2	8.4	17.6	28.8	31.1	34.3	100.1

Notes: Summary statistics of the weekly conflicts in the Middle East, excl. Iran and Syria. Measures in order of appearance include mean, standard deviation, minimum, first quartile, median, third quartile, and maximum. The countries other than Israel-Palestine are grouped as *Average (excl. Israel-Palestine)*.

Table 2: Summary of performance measures across countries pre-treatment and post-treatment

	Ratio		Pre-intervention		Post-intervention	
	MAE	RMSPE	MAE	RMSPE	MAE	RMSPE
Israel & Palestine	6.59	5.61	3.99	5.77	26.28	32.38
Bahrain	3.03	2.40	7.40	3.58	2.44	8.61
Iraq	5.84	4.57	48.64	11.36	8.33	51.89
Jordan	5.12	6.78	2.09	0.57	0.41	3.89
Kuwait	3.60	3.47	0.22	0.13	0.06	0.43
Lebanon	4.92	4.32	6.60	1.71	1.34	7.37
Oman	8.40	7.38	0.07	0.04	0.00	0.32
Qatar	3.15	0.97	0.05	0.07	0.02	0.07
Saudi Arabia	4.40	3.76	15.46	4.80	3.52	18.06
Turkey	0.86	0.37	18.44	61.88	21.41	22.94
UAE	5.31	2.23	0.08	0.08	0.01	0.17
Yemen	1.91	1.68	28.48	20.35	14.95	34.20

Notes: Summary of measures used to assess the significance of the results obtained for Israel and Palestine. Measures include mean absolute error and root mean squared prediction error between the observed and estimated weekly number of conflicts for both the pre- and post-intervention period. We also include the ratios of post/pre-intervention measures. All measures are reported for Israel and Palestine, and for each of the placebo runs.

Table 3: Placebo specification test

Placebo Specification										
κ	1	2	3	4	5	6	7	8	9	10
i.i.d. Perm.	0.902	0.664	0.850	0.678	0.832	0.883	0.902	0.933	0.952	0.974
Moving Block Perm.	0.901	0.594	0.782	0.614	0.762	0.812	0.851	0.891	0.901	0.941

Notes: Placebo specification test p-values over varying κ from 1 to 10 based on both the i.i.d. and the moving block permutations. We fail to reject the null hypothesis at any significance level above 60\%. Failure to reject the null hypothesis provides evidence for correct specification. In the i.i.d. case, we randomly sample 10,000 elements from the set of all permutations with replacement.

Table 4: Summary of performance measures across models pre-treatment and post-treatment

	Pre-intervention		Post-intervention	
	MAE	RMSPE	Std	Ave. gap
tree-based controls	3.99	5.77	4.34	26.12
synthetic controls	9.62	14.73	2.86	25.14
elastic net	7.88	10.67	6.08	31.32
matrix completion	5.53	7.65	1.05	24.80

Notes: Summary of measures used to assess the performance of the results obtained for Israel and Palestine. Measures include mean absolute error and root mean squared prediction error between the observed and estimated weekly number of conflicts for both the pre- and post-intervention period. We also include the estimated standard deviation of the estimates and the average gap in the post-intervention period. We include the measures for the tree-based control method and the comparison methods.

Table 5: Summary of performance measures in validation sample

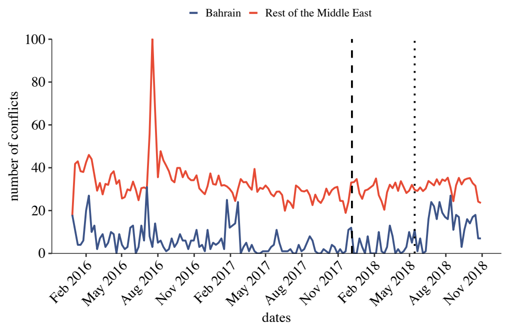
	RMSPE	MAE
tree-based controls	6.60	5.02
synthetic controls	7.72	6.96
elastic net	4.81	4.33
matrix completion	8.95	7.99

Notes: Summary of measures used to assess the performance of the results obtained for Israel and Palestine. Measures include mean absolute error and root mean squared prediction error between the observed and estimated weekly number of conflicts on a validation sample from the pre-intervention period. We include the measures for the tree-based control method and the comparison methods.

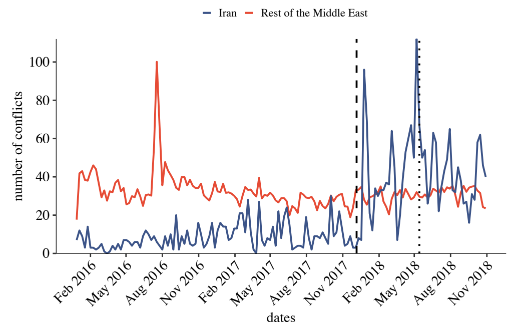
Appendices

A Common trends in the Middle East

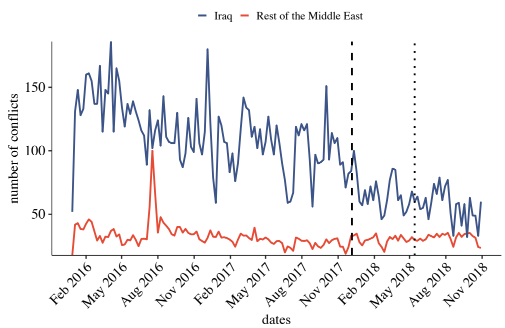
Figure 10: Weekly number of conflicts in the Middle East.



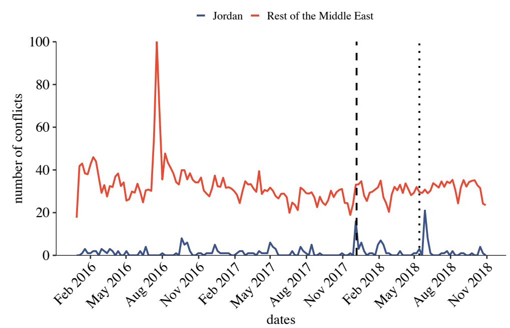
(a) Conflict trends in Bahrain



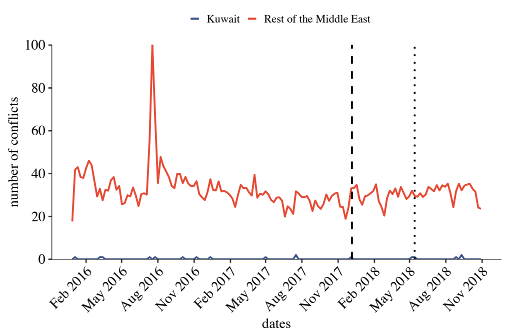
(b) Conflict trends in Iran



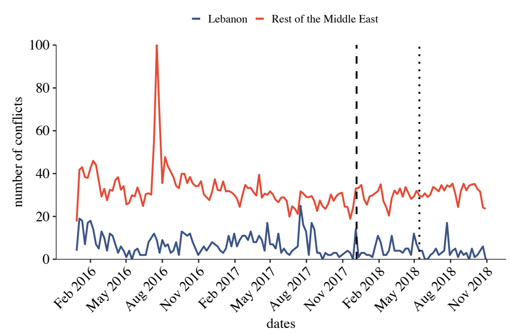
(c) Conflict trends in Iraq



(d) Conflict trends in Jordan



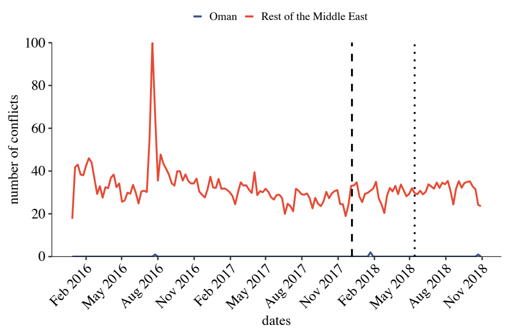
(e) Conflict trends in Kuwait



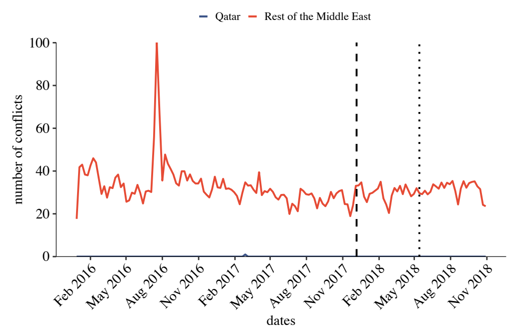
(f) Conflict trends in Lebanon

Notes: Weekly number of conflicts in each of the control countries in the Middle East together with Iran (blue line) in addition to the average of the control countries in the Middle East (red line). The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively.

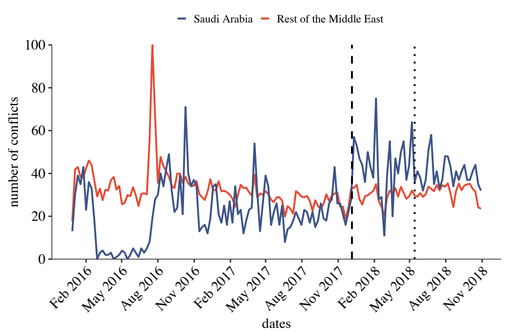
Figure 10: Weekly number of conflicts in the Middle East.



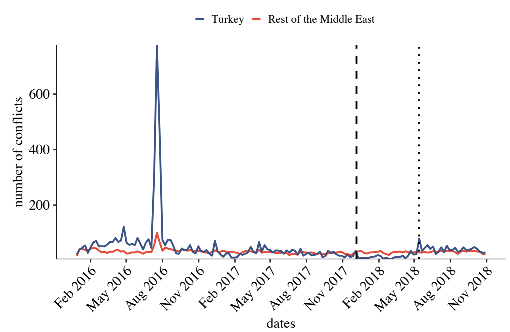
(a) Conflict trends in Oman



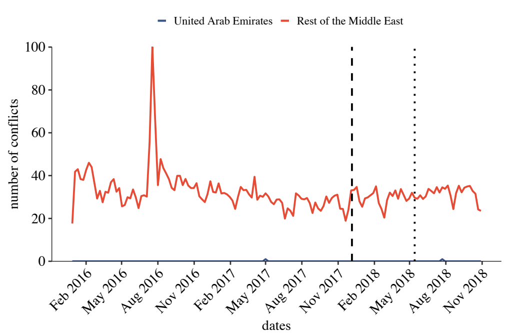
(b) Conflict trends in Qatar



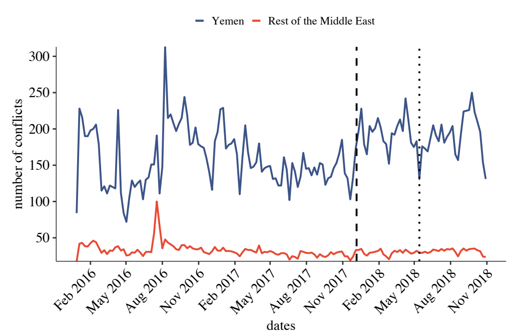
(c) Conflict trends in Saudi Arabia



(d) Conflict trends in Turkey



(e) Conflict trends in UAE



(f) Conflict trends in Yemen

Notes: Weekly number of conflicts in each of the control countries in the Middle East together with Iran (blue line) in addition to the average of the control countries in the Middle East (red line). The vertical dashed and dotted lines represent the date when the move of the US embassy was announced and the date of the actual move, respectively.

Research Papers 2020

2019-12: Vanessa Berenguer-Rico, Søren Johansen and Bent Nielsen: Uniform Consistency of Marked and Weighted Empirical Distributions of Residuals

2019-13: Daniel Borup and Erik Christian Montes Schütte: In search of a job: Forecasting employment growth using Google Trends

2019-14: Kim Christensen, Charlotte Christiansen and Anders M. Posselt: The Economic Value of VIX ETPs

2019-15: Vanessa Berenguer-Rico, Søren Johansen and Bent Nielsen: Models where the Least Trimmed Squares and Least Median of Squares estimators are maximum likelihood

2019-16: Kristoffer Pons Bertelsen: Comparing Tests for Identification of Bubbles

2019-17: Dakyung Seong, Jin Seo Cho and Timo Teräsvirta: Comprehensive Testing of Linearity against the Smooth Transition Autoregressive Model

2019-18: Changli He, Jian Kang, Timo Teräsvirta and Shuhua Zhang: Long monthly temperature series and the Vector Seasonal Shifting Mean and Covariance Autoregressive model

2019-19: Changli He, Jian Kang, Timo Teräsvirta and Shuhua Zhang: Comparing long monthly Chinese and selected European temperature series using the Vector Seasonal Shifting Mean and Covariance Autoregressive model

2019-20: Malene Kallestrup-Lamb, Søren Kjærgaard and Carsten P. T. Rosenskjold: Insight into Stagnating Life Expectancy: Analysing Cause of Death Patterns across Socio-economic Groups

2019-21: Mikkel Bennedsen, Eric Hillebrand and Siem Jan Koopman: Modeling, Forecasting, and Nowcasting U.S. CO2 Emissions Using Many Macroeconomic Predictors

2019-22: Anne G. Balter, Malene Kallestrup-Lamb and Jesper Rangvid: The move towards riskier pensions: The importance of mortality

2019-23: Duván Humberto Cataño, Carlos Vladimir Rodríguez-Caballero and Daniel Peña: Wavelet Estimation for Dynamic Factor Models with Time-Varying Loadings

2020-01: Mikkel Bennedsen: Designing a sequential testing procedure for verifying global CO2 emissions

2020-02: Juan Carlos Parra-Alvarez, Hamza Polattimur and Olaf Posch: Risk Matters: Breaking Certainty Equivalence

2020-03: Daniel Borup, Bent Jesper Christensen, Nicolaj N. Mühlbach and Mikkel S. Nielsen: Targeting predictors in random forest regression

2020-04: Nicolaj N. Mühlbach: Tree-based Synthetic Control Methods: Consequences of moving the US Embassy