Matched Wake Analysis: Finding Causal Relationships in Spatiotemporal Event Data*

Sebastian Schutte
International Conflict Research
ETH Zurich
8092 Zurich, Switzerland
schuttes@ethz.ch

Karsten Donnay
Sociology, Modeling and Simulation
ETH Zurich
8092 Zurich, Switzerland
kdonnay@ethz.ch

July 18, 2014

Abstract

This paper introduces a new method for finding causal relationships in spatiotemporal event data with potential applications in conflict research, criminology, and epidemiology. The method analyzes how different types of interventions affect subsequent levels of reactive events. A sliding spatiotemporal window and statistical matching are used for robust and clean causal inference. Thereby, two well-described empirical problems in establishing causal relationships in event data analysis are resolved: the modifiable areal unit problem and selection bias. The paper presents the method formally and demonstrates its effectiveness in Monte Carlo simulations and an empirical example by showing how instances of civilian assistance to US forces changed in response to indiscriminate insurgent violence in Iraq.

*We would like to thank Marco Steenbergen, Skyler Cranmer and Alrik Thiem for valuable feedback on an earlier version of this paper. Philipp Hunziker and Thomas Chadefaux also provided extremely helpful advise.

1 Introduction

The study of political violence has benefited in recent years from a rapid increase in the availability of conflict event data sets (Raleigh et al. 2010; Sundberg et al. 2010). In these data, single instances of violence are coded together with their geographic coordinates and the date they occurred on. Several recent publications have successfully shed light on some of the micro-dynamics of civil conflict by analyzing such data (for example Raleigh & Hegre 2009; Hegre et al. 2009; Buhaug 2010; O'Loughlin & Witmer 2011). However, while progress has been made in relating conflict intensity to geographic conditions, more complex endogenous mechanisms that drive conflict at the micro-level remain largely elusive to quantitative analysis, despite their theoretical prominence (e.g. Kalyvas 2006).

To fill this gap, we introduce a novel approach to causal inference in disaggregated event data that combines two techniques for ensuring robust and clean causal inference: sliding spatio-temporal windows (Kulldorff 1997; Braithwaite & Johnson 2012) and statistical matching (Rubin 1973; LaLonde 1986; Iacus et al. 2012). The presented approach clears the path for answering a whole class of high-profile research questions regarding the causal effects of specific types of events on future events. To demonstrate this approach and its capabilities, we show that the experience of indiscriminate insurgent violence in Iraq has led civilians to collaborate with the US military.

While presented in the context of conflict research, this method could be equally applied in other quantitative fields of research that rely on georeferenced data on specific events:

Criminologists might investigate the effects of law enforcement activities on subsequent levels of crime. Epidemiologists could analyze the spread of infectious disease as a function of specific types of interaction between individuals.

This paper proceeds as follows: After discussing the existing research and its shortcomings in the next section, we introduce our methodological contribution in detail and use a series of Monte Carlo simulations to test its capabilities and limitations. After that, we demonstrate the use of the method in an empirical example by analyzing the effects of indiscriminate insurgent violence on civilian collaboration with US troops in Iraq.

2 Abilities and limitations of existing approaches

The theoretical prominence of endogenous conflict dynamics (Kalyvas 2006) has motivated a number of empirical studies in recent years. In order to understand how past conflict events shape future levels of violence, a rapidly growing number of studies rely on newly available event data (see: Raleigh et al. 2010; Sundberg et al. 2010; SIGACT 2010; Leetaru & Schrodt 2013).

In principle, event data reflect changes in the trajectory of conflicts brought about by specific incidents. Along these lines, research into the causes and effects of violence against civilians in civil war (Kalyvas 2006; Lyall 2009) and escalation dynamics (Jaeager & Paserman 2008; Haushofer et al. 2010; Linke et al. 2012) has drawn on conflict event data. Several studies have used village-level counts of violent events to investigate whether indiscriminate incumbent violence has a deterrent or escalating effect on subsequent insurgent activity. Especially Lyall (2009) and Kocher et al. (2011) pioneered this type of analysis with innovative matching designs and villages as units of analysis.

However, in many situations such natural spatial units of analysis are missing. Some studies have circumvented this problem by relying on artificial units of analysis, such as grid-cell months, and aggregated event counts and covariates accordingly. While introducing these artificial units conveniently clears the way for econometric analysis, it also leads to two problems widely described in the methodological literature. First, if cells of arbitrary sizes are the units of analysis, the number of available observations directly scales with the chosen cell size: the smaller the cells, the more observations. Of course, regular null hypothesis tests crucially depend on the number of available observations. As N increases, the standard errors tend to decrease and even the smallest empirical signals becomes statistically "significant". A second problem extensively described in the geographic literature is the "modifiable areal unit problem" (MAUP), i.e. the fact that the selection of artificial cell sizes drives spatial inference (Openshaw 1984; Cressie 1996; Dark & Bram 2007).

Approaches to overcoming the MAUP have been proposed in the past and also been applied in conflict research (O'Loughlin & Witmer 2011). A commonly used method called "SaTScan" (Kulldorff 1997) relies on sliding spatial and temporal windows to reveal clusters of events on different levels of aggregation.[1] Applied to epidemiology, SaTScan was

[1] For another approach to identifying event clusters see Leslie & Kronenfeld (2011).

originally introduced as a tool for testing whether a certain region faces an elevated per capita risk of disease. The method provides a fast assessment of whether event clusters could have been brought about by chance under corresponding distributional assumptions. To establish a baseline level of clustered events, SaTScan applies a simulation technique: For each size of the spatiotemporal window under consideration, the software allocates events randomly in space and time. Repeating this process in multiple iterations generates a distribution of simulated events under baseline assumptions. Significant empirical deviations from this baseline can then be identified for different cell sizes. In other words, comparing the distribution of artificial events to the empirical record yields an estimate of how likely is it that observed clustering was brought about by chance.

In the epidemiological case of Kulldorff (1997), this baseline is well justified as it assumes a constant per capita rate of instances of non-infectious disease. In conflict settings, however, finding suitable baselines is usually much more difficult. Instances of insurgent violence, for example, are likely to result from a host of factors, including geographic exposure and reaction to previous violence. Randomly allocating events in space and time might not adequately capture plausible counterfactual scenarios: Instances of violence against civilians, for example, might be simulated to take place in uninhabited areas and a simulated baseline would not reflect the causal order of events found in the empirical record.

Relaxing the assumption of a uniform spatial distribution of events, Braithwaite & Johnson (2012) apply a permutation test within the framework of sliding spatiotemporal windows to the analysis of violent events in Iraq. In this setup, a random baseline is also simulated, but not by relocating conflict events in space and time. Instead, events remain in their original positions but event categories are randomly assigned. By holding constant the location and timing of events while changing event categories, a baseline scenario can be established in which event types are independent of one another. Comparing this simulated baseline to empirical distributions of event categories shows whether or not specific classes of events tend to occur together, i.e. in clusters that are unlikely to have been brought about by chance. However, this measure of systematic co-occurrence, as well as SaTScan's identification of event clusters, does not establish a clear causal relationship between the event types.[2] We therefore decided to introduce a new framework for inferential analysis in conflict event data.

In the following section we describe a new method for finding causal relationships in event data that combines the best of the two most promising techniques reviewed above: sliding spatio-temporal windows to overcome the MAUP and statistical matching to allow for clean causal inference.

3 Matched wake analysis

Any attempt to overcome the discussed methodological shortcomings in the analysis of causal relationships in conflict event data must start with a theoretical understanding of the data generating process. A first crucial insight is that events come into existence through a

[2]It should be mentioned, however, that SaTScan permits the simulation of non-uniform baselines which makes it a very versatile tool for the analysis of spatial event clusters.

variety of different mechanisms. In conflict research, there is the widely described effect of exogenous geographic conditions that drive overall levels of violence (McColl 1969; Hegre et al. 2009; Raleigh & Hegre 2009; O'Loughlin & Witmer 2011). For example, strategic locations might see higher levels of violence. Ethnic settlement patterns have been linked to conflict events in Iraq (Weidmann & Salehyan 2013) and in Israel (Bhavnani et al. 2014). For conceptual clarity one can refer to these factors as the a priori *exposure* of any location to violence. Furthermore, levels of violence generally vary over time. A negotiated ceasefire and seasonal cycles may drive the intensity of conflict across a war zone. These aspects can be referred to as the *momentum* of a conflict at any given time. Isolating the effects of exposure and momentum is a crucial prerequisite for cleanly analyzing the third mechanism driving levels of violence: *reaction* to specific events, i.e. the causal effect of specific interventions. Figure 1 illustrates the logic of this empirical strategy.

In this conceptual sketch, three types of conflict events are depicted. The rectangular symbol in the center of the left cylinder represents an instance of violence assigned to the "control" category. The triangle in the right cylinder represents a "treatment" event and the star-shaped symbols represent events in the dependent category, which are possibly affected by treatment. In general, context information can be obtained with regard to *exposure* for both control and treatment events: spatial information such as local elevation (Gesch et al. 1999), natural land-cover (Hansen et al. 2000), the proximity of strategic locations such as the nearest international border (Weidmann et al. 2010), and the predominant ethnic group in the region (Wucherpfennig et al. 2011) can be calculated based

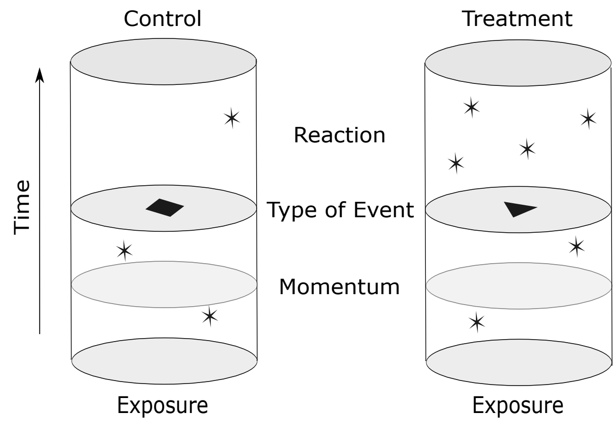

Figure 1: Illustration of the empirical strategy. Conflict events are divided into two classes of "treatment" and "control" events. For each event, previous levels of "dependent" events and their temporal trends and subsequent levels are established in an automated GIS analysis.

on geocoded data.

Similarly, *momentum* of violence for all conflict events can be established by counting the number of previous dependent events. As figure 1 indicates, the lower half of the cylinder is subdivided into two halves. A trend in the number of dependent events can be calculated. It is flat in both cases depicted here (one conflict event in each of the first two quarters of the cylinders). Of course, the quantity of interest in this setting is the number of subsequent events, i.e. the reaction to instances of treatment and control.

3.1 Sliding window design

In principle, associating observations with static spatial covariates and dynamic counts of previous and subsequent dependent events would be sufficient to generate a statistical sample for subsequent analysis. This setup, however, still does not account for the MAUP since the size of cylinders in space and time cannot be identified based on theoretical expectations: Why should events at a distance of 20 kilometers be counted while events at a distance of 30 kilometers be excluded? It is exactly this type of arbitrary coding that Openshaw & Taylor (1979) have shown to obscure quantitative inference.[3]

As pointed out in the previous section, solutions to this problem have been identified in terms of sliding spatiotemporal windows. In this setup, the entire procedure of counting previous and subsequent events for every intervention is repeated for multiple sizes of spatiotemporal cylinders. This helps us to overcome the problem of inference hinging on arbitrary cell sizes and to distinguish among small- and large-scale effects empirically. For example, the effect of a treatment event on the level of dependent events might be stronger in its direct spatial and temporal vicinity and not affect more distant locations. Moreover, averaging the effects for different window sizes allows us to calculate a bottom-line effect.

3.2 Statistical matching

In the previous step, interventions were associated with counts of previous and subsequent dependent events for different spatiotemporal windows. Moreover, spatially referenced data – such as distances to major cities and population numbers in the area – were used

[3]Of course, applied researchers are not always in the comfortable position to have exact data on the locations of the events they study. Some data are only available on the level of administrative units or pre-aggregated into artificial cells. This methodological discussion is no way intended to discredit the corresponding studies, but merely an attempt to encourage researchers to use the full geographic information that is available to them.

to provide context information for each event. However, without explicitly accounting for confounding factors, causal inference in this setup can still suffer from selection bias. In line with studies using natural spatial units of analysis (Lyall 2009; Kocher et al. 2011), we apply statistical matching in order to compare treated and untreated observations under otherwise comparable conditions.

The general idea behind matching is to approximate as closely as possible experimental conditions in observational data (Rubin 1973). Matching has become an important tool in the social scientific toolbox, although its effectiveness has been disputed (LaLonde 1986). In experimental settings, treatment is applied randomly and its effects are observed in comparison to an untreated control group. Exactly this type of randomization that is so critical for unbiased inference is frequently absent in observational data. To emulate randomization, several techniques have been proposed. In the most simple setting, a large quantity of observations for both treatment and control are available and *exact matching* can be applied. In exact matching, only those observations are retained in the treatment group for which a corresponding observation can be found in the control group with identical numerical values for all relevant confounding variables. Exact matching entails that these observations only differ with regard to treatment being applied or not. Clearly, under such ideal conditions, the treatment effect can be directly estimated through the difference in means between the groups for the dependent variable (Iacus et al. 2012:1). Unfortunately, such conditions are hard to find in practice. Usually, the confounding variables between treatment and control observations are comparable, but not completely identical. Several strategies exist to alleviate this problem. One approach is to capture the effect of the confounding factors on the probability of treatment assignment in a propensity score model (Rosenbaum & Rubin 1983). Propensity score matching essentially amounts to predicting the probability of treatment assignment with a binary dependent variable regression model. The predicted probabilities of treatment assignment for each observation are used as the "propensity score" and observations from treatment and control group with similar scores are used in the subsequent analysis.

There is a practical problem associated with this technique for sliding spatiotemporal windows. A propensity score model requires as much care in post-estimation analysis as any other binary dependent variable model. Moreover, since the goal of matching is to increase *balance*, i.e. to make the empirical distributions of the covariates more similar, the balance has to be assessed for each covariate before and after matching. In practice, researchers have to go back and forth between propensity score model specifications and assess the improvements in balance. Poorly performing propensity score models can very well decrease the overall balance and therefore completely defeat the purpose of matching.

Clearly, a more robust and automated technique is needed for MWA: Due to the sliding window design, matching has to be performed repeatedly for all spatial and temporal parameter combinations, and manual readjustments after post-estimation analysis are not an option. A very recent and computationally efficient automated matching technique alleviates this problem: *Coarsened Exact Matching* (CEM) (Iacus et al. 2012). In CEM, substantially identical but numerically slightly different values are collapsed into bins of variable sizes for each covariate. Matching is then performed for observations belonging to the same bins. Finally, a subsequent analysis can be performed for matched observations, but with the original numerical values. CEM generates well-balanced data sets by choosing bin sizes for different variables based on their empirical distributions. This method is much faster and more transparent than its alternatives and we therefore rely on CEM for automated matching.

3.3 Estimation of causal effects

Several methods exist that are commonly used to estimate the causal effect of the treatment after matching is performed. For example, a Difference-in-Differences design (DD) (Angrist & Pischke 2009:227-243) has been proposed and used in related empirical studies (Lyall 2009). To assess the within-subject before and after change, DD performs an OLS regression on the matched data set to estimate changes in the number of dependent events brought about by the treatment. The dependent variable in this model is the number of dependent events after interventions. The number of dependent events before the intervention is also necessarily included in the model. Note that counts were aggregated for each of the pre- and post-intervention period, solving the problem of serial correlation that DD designs are otherwise prone to (Bertrand et al. 2004:252). Moreover, the setup accounts for changing conflict dynamics unrelated to the interventions by matching on the trend in the dependent variable before interventions. The trend itself is calculated simply by subdividing the lower half of the spatiotemporal cylinder into two periods (see figure 1).

The resulting DD specification is then:

$$n_{post} = \beta_0 + \beta_1 n_{pre} + \beta_2 treatment + u \tag{1}$$

In this model, β_2 is the estimated average treatment effect of the treated, i.e. the quantity of interest in the analysis. In the result presentation below, estimates for β_2 are shown for each spatiotemporal window under investigation. We further provide detailed summary statistics for the matching procedure in terms of the multivariate *L1* imbalance measure and the percentage of common support (Iacus et al. 2012). *L1* is a multivariate distance metric expressing the dissimilarity between the joint distributions of the covariates in treatment and control groups. To calculate this statistic, the joint distributions are approximated in fine-grained histograms. Average normalized differences between these histograms are expressed in the *L1* statistic ranging from complete dissimilarity (1) to full congruence (0). A similarly intuitive measure is common support: It expresses the overlap between the distributions of matching variables for treatment and control groups in percent (Iacus et al. 2012). 100% common support refers to a situation where the exact same value ranges can be found for all matching variables in both groups. A formal description of *L1* and common support can be found Iacus et al. (2012).

In summary, a suitable setup for the causal analysis of conflict events has been sketched out in four steps. Intervention events are associated with geographic context information and counts of previous and subsequent events. After that, they are matched with regard to previous event counts, trends, and geographic variables. Finally, they are analyzed in a Difference-in-Differences regression design. Figure 2 provides a graphical representation of this procedure.

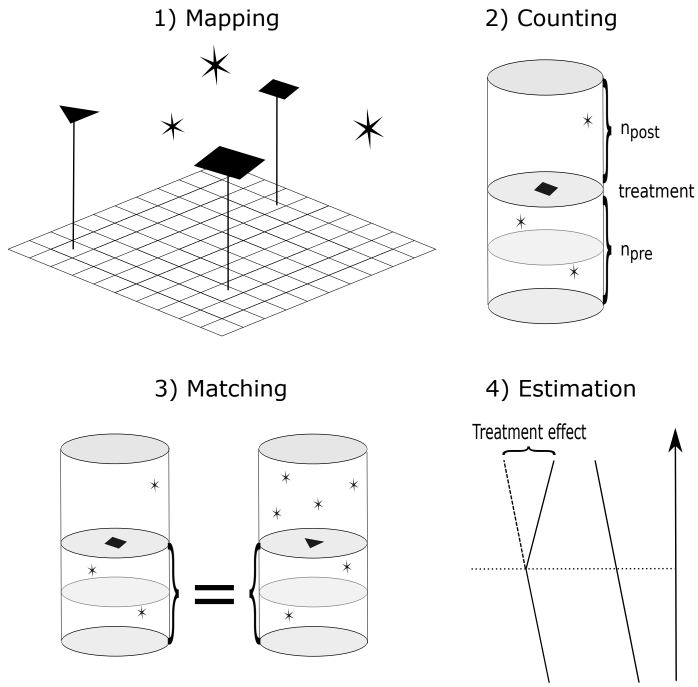

Figure 2: Graphical overview of the MWA procedure: In a first step, observations are associated with geographic information via nearest neighbor mapping. After that, previous and subsequent instances of "dependent" events are counted. In step three, observations are matched with regards to previous events, event trends, and geographic information. The method of choice in this procedure is coarsened exact matching. Finally in step four, the treatment effect on the dependent variable is established in a Difference-in-Differences regression design for the matched sample.

3.4 Limitations of the approach

While the underlying logic of matching designs is sound and widely used in empirical social science (see Abadie & Imbens 2006; Herron & Wand 2007; Diprete & Engelhardt 2004), spatiotemporal data introduce potential pitfalls. Most importantly, the spatiotemporal cylinders around interventions can overlap partially. If they do, the "Stable Unit Treatment Value Assumption" (SUTVA) inherent to matching is violated. It states that the treatment effect of any observation should be independent of the assignment of treatment to other units (Cox 1958). Violating this assumption can lead to biased estimates. Two MWA scenarios are imaginable in which the SUTVA assumption would be clearly violated. First, multiple treatment events could overlap. Assuming a positive treatment effect, the corresponding estimates are likely to be biased upward in this scenario. Second, treatment and control events could overlap in space and time and thereby "water down" the treatment effect. In this case, the estimate for the treatment effect would be biased downward. To address this problem, we match on the number of intervention events that precede each intervention. This remedy and its effectiveness will be discussed in more detail below.

While SUTVA violations may indeed pose a problem to clean causal inference in MWA, there are ways to mitigate this problem. First, spatiotemporal overlaps are easily identified in empirical data. As described above, counting previous and subsequent instances of violence is part of the data preprocessing, and multiple instances of overlapping treatment and control events can be counted as well. The simplest way to avoid drawing false inference is therefore to check the data for overlaps of treatment and control events and select subsets that are not affected by this problem. For example, a civil war might go through phases of intense violence (e.g. summer offensives) and calmer periods, and researchers could test the causal effects of different types of events in the calmer periods to avoid false inference from overlapping events. However, empirical insights into the conflict dynamics would then, of course, be exclusively limited to such calmer periods instead of the entire conflict.

Second, if substantial numbers of overlapping cylinders cannot be avoided, data can still be analyzed using MWA. In this situation, the following problem has to be accounted for: Interventions of different types prior to the intervention under investigation can affect subsequent levels of dependent events. As a result, the causal effect attributed to the intervention would be in fact the product of a specific mix of different interventions (a double treatment, for example). A simple remedy in this situation is to match on the numbers of previous treatment and control events. This ensures that the interventions retained in the post-matching sample have similar histories of treatment and control events.

Another effect of matching on previous interventions is that non-overlapping treatment and control events have a higher probability of being selected into the post-matching sample. This is due to the fact that overlapping cylinders tend to differ with regard to the previous number of treatment and control events because the earlier event will be counted as a previous event for the later one. This effect leads to a matched data set with fewer overlapping events. A side effect of this approach is that it decreases overall balance between the treatment and control groups with regard to *exposure*, since overlapping events yield similar values for the related spatial confounding factors.

A third strategy is to simply remove overlapping observations from the sample. The obvious problem with this approach is the potential bias arising from non-random deletion itself. In a benchmark analysis using simulated data, we show that this strategy still performs better than the baseline method for smaller overlaps, but for larger overlaps the problems associated with non-random deletion are very noticeable. The strategy also appears to lead to less robust estimates for overlapping cylinders than matching on the number of previous treatment and control events. We demonstrate quantitatively in the next section how these remedies perform.

4 Monte Carlo simulations

In this section, we demonstrate the performance of MWA based on simulated event data. We rely on artificial data to maximize the transparency of the setup and generate benchmarks under controlled conditions that include simulated causal effects, but also random noise that can be expected in any empirical application.

Two scenarios were used for the simulations. First, as a proof-of-principle, a treatment effect was established under ideal circumstances: Cleanly separated “treatment” and “control” events were analyzed under otherwise comparable conditions. Second, data with increasingly stronger overlaps were analyzed to illustrate the resulting biases. Remedies such as deletion of overlapping events and matching on previous intervention events were tested.

4.1 Data generating process

In order to emulate some of the empirical complexity of event data, we constructed artificial samples using three types of events. One type of event represents the "dependent" category and our quantity of interest was changes in the frequency of these events after interventions. The other two types are intervention events, which are labeled "treatment" and "control" in compliance with the matching terminology. The artificial causal effect was modeled in two steps. Events of the "dependent" category were placed prior to interventions and exhibited varying trends. Dependent events following interventions were placed in fixed temporal and spatial distances from the interventions.

The frequency of dependent events increased such that one more dependent event occurred after treatment than before. For events of the "control" category, the number of dependent events following interventions remained unchanged in comparison to the number of preceding events. An increase of one event is the smallest possible effect for discrete event counts and provides a difficult test situation: the larger the effect, the more easily it is recovered by the the method. Absolute counts and trends in dependent events were varied to increase the realism of the simulations. The data contained 200 "controls" and 100 "treatments". This imbalance was intentionally chosen to emulate the complications of empirical data. We account for this difference by using weighted regressions for the DD analysis in the simulations and the empirical section.

For each intervention event, we also assigned two stylized confounding variables which were simply numerical values drawn from the same random distributions. For the simulation, we ignored the potential effects of confounding factors on the probability of treatment being applied, since they would be mitigated by the matching if they were present. [4] Artificial intervention events were distributed over a geographical region of 2 by 2 degrees around the Equator, which corresponds to an area of roughly 220 km by 220 km. Figure 3 depicts the spatial setup for the simulations. Of course, intervention events were separated temporally. By varying the simulated time period, the probability of events overlapping in this simulated setup could be adjusted: the longer the simulated time span, the smaller the probability of overlaps. By varying the time span under investigation, we could assess the effects of increasing overlaps on the estimation of the treatment effect.

4.2 Simulation results

To overcome MAUP, MWA establishes event counts and estimates for the treatment effect for different spatial and temporal cylinder sizes. The corresponding insights can be communicated graphically as a contour plot:

The lighter the color the larger the estimated treatment effect (β_2 in formula 1). The corresponding standard errors are indicated by shading out some of the estimates: No shading corresponds to $p < 0.05$ for the treatment effect in the DD analysis. Dotted lines indicate p-values between 0.05 and 0.1 and full lines indicate $p > 0.1$. The cells indicating

[4] For more details on the generation of our test data, please refer to the supplementary information.

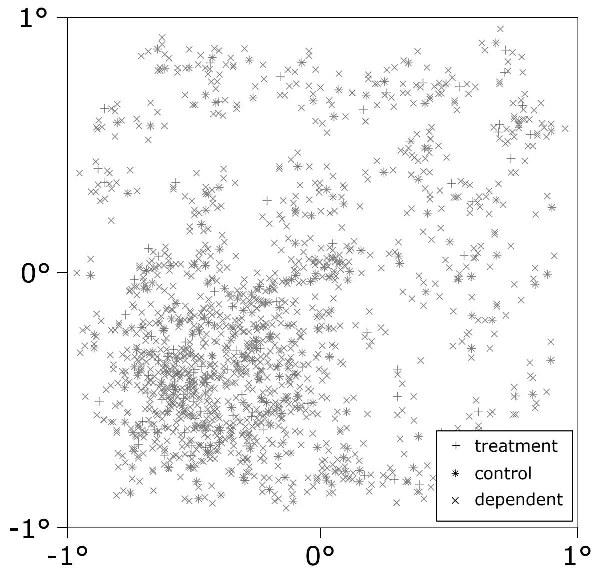

Figure 3: Map of the simulated data distributed over the region within the 1st degree latitude North and South and the 1st degree longitude East and West, an area that corresponds to roughly 220 km by 220 km. This generic spatial setup was used for all Monte Carlo simulations.

effect size and significance level are arranged in a table where each field corresponds to one specific combination of spatial and temporal sizes of the cylinders depicted in figure 1 (see figure 4).

To illustrate the ability of MWA to reveal the spatiotemporal distances at which reaction to intervention occurs, the dependent events after interventions (i.e. reactive events) were placed at distances of eight days and eight km. Figure 3 shows how the resulting clusters of events are distributed randomly in space. The probability of clusters overlapping was minimized as they were spread out over a temporal span of 20 years. In this case, clean causal inference is possible and the method clearly recovers exactly the simulated causal

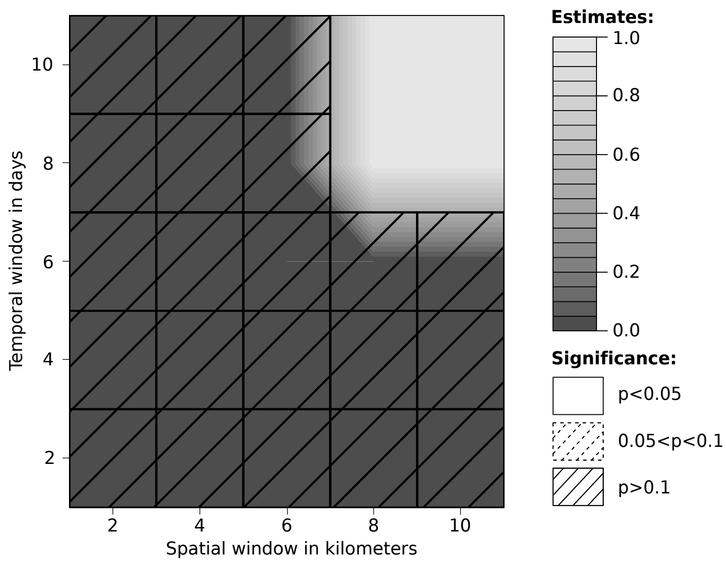

Figure 4: Estimates and significance levels for simulated data. Significance levels are indicated graphically. No shading corresponds to p<0.05, dotted lines to p<0.1, and full lines to p>0.1.

effect in the number of dependent events at eight days, eight km (figure 4). Note that larger spatial and temporal window sizes yield the same results (for example, 10 days and 10 km). This is because for the special case of non-overlapping cylinders larger windows still only contain the same number of dependent events as the smaller windows. For smaller spatial and temporal window sizes, the estimates are not significant.

4.3 Robustness of the method

We ran a series of tests to assess the effects of overlapping interventions on the causal inference and to demonstrate the effectiveness of the proposed remedies. To generate overlaps, we distributed simulated events in the same simulated space as shown in figure 3 and with the same reactive pattern as before, but within increasingly shorter time periods (from 1 year down to 10 days). For each time interval we generated 100 random test data sets and applied the method for each one.

Biased results as a function of overlapping interventions should make it more difficult to infer the true treatment effect. In our simulated example, this effect appears at spatial distances of eight kilometers and temporal distances of eight days from the interventions. Therefore, we used corresponding cylinder sizes of eight days and eight kilometers to capture the simulated causal effect. Figure 5 shows the average estimates and confidence intervals for the estimated causal effect as a function of growing overlaps of the interventions. The standard matching procedure is compared to a setup where matching is performed on previous interventions.

In the figure, the overlap of spatiotemporal cylinders is expressed as the percentage of observations for which at least two treatment events overlap. The "% overlaps" in figure in 5 indicates the percentage of observations for which at least two treatment events are in the same cylinder. [5]

The true treatment effect in all simulations is 1 and indicated with a dotted line. Estimates for this true effect vary for the different simulation runs: Mean values are shown as circles and 95% confidence intervals are shown as whiskers. The asterisk above many data points indicates that all simulation runs yielded p-values smaller than 0.05.

[5] Whether SUTVA violations are measured in double treatments, double controls, or treatment and control overlaps does not strongly affect the results as shown in the supplementary information.

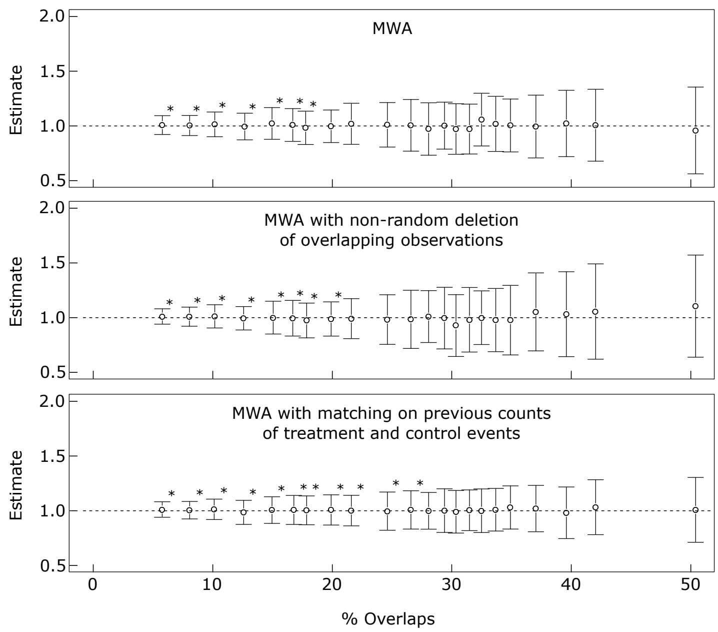

Figure 5: Average estimates with 95% confidence intervals as a function of the overlaps of the spatiotemporal cylinders. The graph shows estimates for MWA (top), MWA with non-random deletion of overlapping observations (middle), and MWA with matching on counts of previous treatment and control events (bottom). Asterisks indicate that all estimates for all simulated data sets were significant at the 0.05 level and the dotted line marks the true effect.

The figure clearly indicates that all three methods produce correct estimates on average also for larger overlaps, but substantial differences exist when it comes to the reliability of the different approaches. For the normal MWA procedure, overlaps affecting up to about 20% of the observations still yield consistently significant results. For slightly higher levels of overlaps, deletions of overlapping observations more reliably produces correct p-values for the treatment effect, as shown in the middle panel. However, for highly clustered data, non-random deletion performs worse than standard MWA. The best results for all ranges of overlaps can be achieved by matching on counts of previous interventions. This approach is demonstrated in the lowest panel: For overlaps of up to 28%, all 100 analyses of simulated data correctly reveal a positive and significant treatment effect. Moreover, confidence intervals are smallest for this procedure.

This analysis shows that the method robustly identifies the true causal effect for a given spatiotemporal lag for situations of moderate overlaps (up to 20%). In the cases of stronger overlaps, matching on the number of previous treatment and control events improves the accuracy of the estimated treatment effect, in line with our theoretical arguments in section 3.4 but only to a point: Beyond 25-30% overlaps, inference becomes less robust.

In the next section we turn to our analysis of an empirical example and investigate the effects of insurgent violence on civilian cooperation with the US military in Iraq. Based on the results of our Monte Carlo simulations we use MWA with additional matching on previous treatment and control events for our empirical analysis.

5 Empirical case: civilian collaboration in Iraq

This section demonstrates that MWA can provide substantive insights into the turmoil of civil conflict and the causal effects of specific types of events. The ongoing war in Iraq was identified as a suitable test case as it lends itself both conceptually and empirically to testing micro-level hypotheses.

After the 2003 US-led invasion, the country went through several phases of intense political violence. Following the initial occupation in 2003, a low-level insurgency developed and grew in subsequent years. This sequence of macro-events is typical of a wider class of cases: A government is replaced through outside intervention and subsequent occupation of the country. The new government faces a problem of legitimacy and is heavily reliant on outside support. Elements loyal to the former administration start a protracted campaign to topple the new incumbent.

An additional source of violence in Iraq were sectarian clashes between Sunni and Shia that intensified after the Al-Askari Mosque bombing in February 2006. In the following 24 months, sectarian violence escalated dramatically. During 2007, 20,000 additional US troops were deployed in the country to contain the escalating civil war and to strengthen the Iraqi security apparatus. During the same period, the Sons of Iraq movement began assisting incumbent forces in fighting foreign insurgents. During 2008 and 2009, violence against incumbent forces steadily declined, while sectarian tension continued to claim civilian lives.

In 2010, a large number of temporally and spatially referenced conflict events recorded by the US military were released to the general public through the online platform `wikileaks.org` (SIGACT 2010).[6] Several inquiries into the conflict dynamics in Afghanistan and Iraq have been published recently that focus on the spatial and temporal distribution of conflict

[6] We decided that these illegally distributed data could be used in a responsible manner for basic research, given that the empirical analysis would not in any way harm or endanger individuals, institutions, or involved political actors. To ensure this, our analysis only focuses on the events in the statistical aggregate. Moreover, the matching design entails that no marginal effects are estimated for confounding factors, which further strengthens the anonymity of the findings. Based on these precautions, the ethics committee of ETH Zurich reviewed a proposal for this study carefully and then allowed it to proceed.

events (O'Loughlin et al. 2010), conflict dynamics (Linke et al. 2012), the clustering of conflict events in space and time (Braithwaite & Johnson 2012), and violence-induced migration (Weidmann & Salehyan 2013). However, micro-level conflict dynamics and causal relationships between events remain heavily understudied.

Following a line of argument that predicts increased civilian collaboration with the strategic adversary in reaction to indiscriminate violence by either side, we assume indiscriminate insurgent violence to increase civilian collaboration with the US military in Iraq (see Kalyvas (2006:144), Kocher et al. (2011); Linke et al. (2012); Ellsberg (1970); Mason & Krane (1989)). More specifically, we assume that civilians are more likely to deny insurgents access to explosives in response to indiscriminate violence. But how can such an expectation be tested empirically?

First, it is important to understand how a substantial fraction of insurgent violence was applied in Iraq. To compensate for the lack of heavy weaponry, Improvised Explosive Devices (IEDs) have been used against both military and civilian targets. In many cases, IEDs are military-grade explosives obtained from unexploded ordnance. These explosives are combined with improvised trigger mechanisms. Unlike landmines, many IEDs are attacker activated and can therefore be used both selectively against adversary combatants or indiscriminately against civilians.

Due to these technical particularities, obtaining unexploded ordnance is a crucial prerequisite for generating a constant supply of new IEDs. Confronted with unexploded ordnance, civilians face a strategic choice: They can either remain passive and thereby allow explosives to be obtained by insurgents, or they can turn in explosive remnants of war. Arguably, civilians will be more inclined to do so if other civilians have been harmed with IEDs in their spatial and temporal vicinity. We therefore test the following hypothesis: *Indiscriminate insurgent violence using IEDs increases civilian handover of unexploded ordnance to US troops compared to selective insurgent violence using IEDs.*

Testing this hypothesis based on MWA requires three event categories to be specified. First, the dependent variable has to be selected. In this case, instances of civilians *turning in* unexploded remnants of war is the dependent variable. The treatment category is *IED Explosions* that have led to civilian casualties, while events that have not claimed civilian lives are used as the control category. Instead of relying on exact casualty counts which might be difficult to obtain under wartime conditions, we relied on so-called "friendly force information requirements" that are associated with many SIGACT observations. We used this information to focus the analysis on events that the reporting unit classified as severe.

5.1 SIGACT data and event categories

The version of SIGACT (Significant Activity) files used for this study cover the time period from 2004 to 2009 and amount to 391,832 records. However, the data provide different spatial resolutions for different parts of the country: Events coded in the Baghdad region are coded with a spatial resolution of approximately 1 km while events for the rest of the country are only accurate to about 10 km.

We decided to analyze the Baghdad subset of the data in MWA and focused on the last two recorded years (2008 and 2009). As mentioned above, the conflict went through numerous phases that can be roughly divided into an initial insurgency (2003-2006), sectarian civil war and the rise of pro-government militias (2006-2008), and a mixture of all of these conflicts with reduced intensity since 2008.

Especially the last phase of the war covered by the data (2008 and 2009) is suitable for testing the proposed hypothesis as events are not as densely clustered as during the most intense violence in 2006 and 2007. Moreover, collaboration with incumbent forces is more frequent than during the initial insurgency. In total, 2,484 events were used for testing the proposed mechanism in the 2008-2009 period for the Baghdad area. The substantive findings generalize well for the rest of the country, as shown in the supplementary information in a separate analysis.

Civilian collaboration with US forces can be measured directly in the data set. Three event categories reflect direct civilian assistance in terms of civilians passing on information or turning in evidence or weapons.[7] We used instances of "turn in" (667 events in the sample) as the dependent type. To distinguish among two types of events that affect subsequent levels of civilian collaboration, IED explosions that harmed (killed or injured) at least one civilian were coded as "treatment" (254 incidents), and those that did not were used as "controls" (177 incidents). Figure 6 shows the geographic locations of events in the treatment, control, and dependent categories.

[7] These categories are tagged as "turn in", "explosive remnants of war/turn in", and "erw/turn-in" in the SIGACT data.

Generally, casualty reports in military data collection might be affected by biases. For example, soldiers might underreport civilian casualties that they have caused themselves, or give too optimistic accounts of enemy casualties. When it comes to civilian casualties caused by insurgents, there are no obvious incentives for misreporting in an internal data collection. We nevertheless use these data conservatively by focusing on reportedly severe incidents. This information was obtained from another field in the SIGACT data, the "friendly force information requirements". We also used information on casualties conservatively and only checked whether or not civilian were harmed to code "treatment" and "control" events.

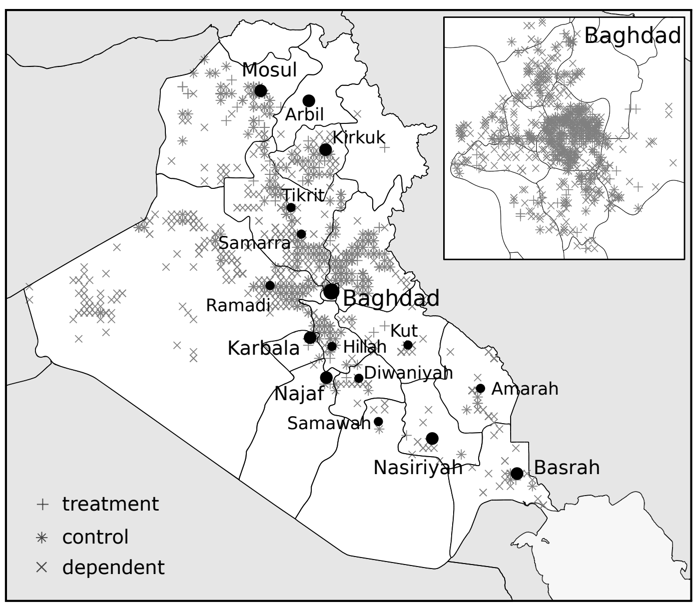

Figure 6: Map of Iraq and Baghdad showing the location of all events (treatment, control, and dependent) included in the analyses

Geographic matching variables were coded for all SIGACT events under investigation. We obtained geocoded data on approximate population figures for the year 2000 (CIESIN 2005), distances to Baghdad's "Green Zone", and stable nighttime light emissions for the year 2008 as a proxy for infrastructural development (NGDC 2012). The ethnic composition of the neighborhood under attack could not be established based on existing data sources. For central Baghdad, Weidmann & Salehyan (2013) have coded time variant data on ethnic groups, but their data only cover a fraction of the greater Baghdad area under investigation.[8] Summary statistics for the matching variables can be found in the supplementary information.

The spatial variables were coded through nearest neighbor mapping between SIGACT observations and the mentioned data sets. Beyond these variables, we also matched on the pretreatment trend in civilian assistance and previous "treatment" and "control" counts, which is in line with the previous discussion.

5.2 Empirical results

The results allow a nuanced insight into how violence changes patterns of collaboration at specific temporal and spatial distances from the intervention. Figure 7 gives an overview of the central findings. Almost all estimates for all cylinder sizes are positive. As visible in the center of the plot, significant increases in collaboration occurred in response to IED attacks with civilian casualties in comparison to attacks that did not harm civilians. For distances

[8] For the analysis of the whole country, we used data on ethnic settlement regions from Wucherpfennig et al. (2011)

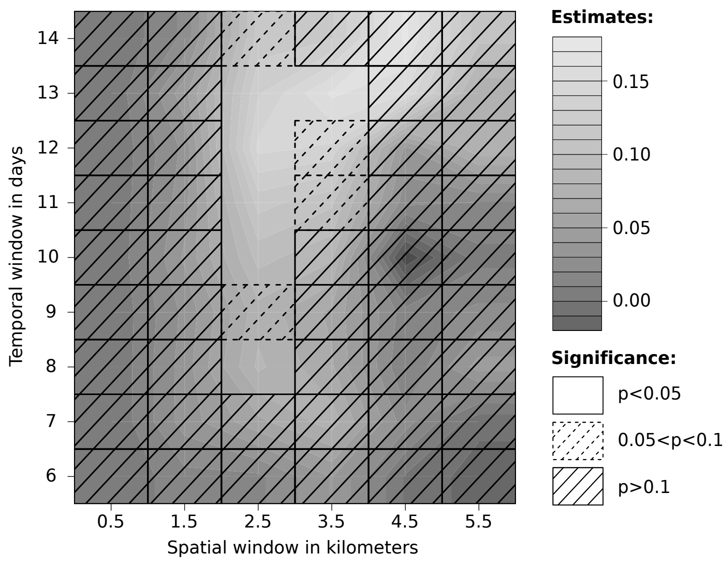

Figure 7: Empirical results of the MWA analysis of civilian collaboration in Baghdad for the 2008-2009 period. The underlying contour plot shows the estimated effect of insurgent violence against civilians on civilian collaboration with the incumbent. Non-shaded areas are significant at $p<0.05$, dotted lines indicate $p<0.1$, and full lines indicate $p>0.1$.

of up to 2.5 kilometers from the incident, a robustly significant effect can be found for a range of temporal offsets from 8 to 14 days. Again, p-values are communicated as shaded areas in the plot. Table 1 also communicates the effects, as well as the fraction of incidents that have seen previous interventions numerically. Based on the almost exclusively positive estimates, we conclude that indiscriminate insurgent violence led to increased civilian collaboration with US ground forces in the later phases of the war in Iraq. This effect is present in the close spatial vicinity of the attack, but with a delay of one to two weeks.

While the effect is significant and robust, it is only moderately strong: For small spatial distances (between 1.5 and 2.5 kilometers) and temporal distances between 1 to almost

Time (days)	Space (km)	Treatment effect	P-value	SO	MO
8	2.5	0.08	0.03	0.20	0.13
10	2.5	0.09	0.03	0.23	0.15
11	2.5	0.11	0.01	0.24	0.16
12	2.5	0.14	0.01	0.26	0.17
13	2.5	0.13	0.03	0.27	0.18
13	3.5	0.16	0.03	0.37	0.30

Table 1: Summary statistics for the interpretable areas of the contour plot in figure 7. The estimated treatment effect for these statistically significant areas averages to 0.12. The acronym SO ("same overlap") refers to situations where either the cylinders of two or more treatment events or two or more control events overlap. MO ("mixed overlap") refers to situations where treatment and control cylinders overlap.

2 weeks after the event, levels of civilian support of the treatment group are significantly higher than in the control group. The estimated treatment effect peaks at 0.16 (for 13 days and 3.5 km). Averaging over the interpreted effects, for every 100 IED attacks against civilians one would expect up to 12 more instances of civilian assistance to US ground forces. Of course, this insight only holds for the Baghdad area and the time period under investigation.

This moderate effect size is empirically plausible. Not every IED attack with civilian casualties would directly lead to an instance of collaboration. Civilians that are inclined to assist US forces would also have to know where unexploded ordnance can be found to actively assist US troops. Clearly, this condition is not met in all situations. It is more plausible that only some incidents happen under circumstances that allow civilians to actively support US troops. Moreover, the results indicate that reactions to insurgent attacks take place with a certain temporal delay that may result from the lack of opportunity to collaborate with US forces but may also reflect risk aversion. In order to conceal their assistance to incumbent forces, civilians might let a few days go by before approaching US troops.

Summary statistics for the matching procedure are presented in table 2. The upper section of the table refers to the empirical sample before matching is applied. The lower section refers to the matched sample. The summary statistics that express the similarity of the joint distributions of the matching variables show a substantive improvement after matching. *Common support* doubles from approximately 25% to approximately 50% and the *L1* distance metric changes in similar magnitude. In summary, the automated matching procedure based on *Coarsened Exact Matching* proves very efficient in this case and substantively improves the balance of the sample. In the area of the substantive effect, the data include more instances of IED attacks that harmed civilians (∼140) than those that did not lead to civilian casualties (118), but this slight difference in the number of corresponding observations is accounted for by the weighted regression.[9]

In summary, we find that there was a significant increase in civilian collaboration with US troops in Iraq during 2008-2009 as a result of insurgent IED attacks with civilian casualties: Up to 12 more instances of civilian assistance for every 100 indiscriminate IED attacks can be attributed to the presented mechanism. This effect is present in the close spatial vicinity of the attack, but with a delay of about one week.

[9] A robustness check reported in the supplementary information without weighted regression leads to almost identical results.

Time (days)	Space (km)	Controls_{pre}	Treatments_{pre}	L1_{pre}	$\%\text{Support}_{pre}$
8	2.5	171	244	0.50	26.20
10	2.5	171	242	0.52	25.40
11	2.5	171	242	0.52	27.40
12	2.5	170	242	0.53	27.50
13	2.5	169	242	0.53	25.70
13	3.5	169	242	0.57	22.90
Time (days)	Space (km)	Controls_{post}	Treatments_{post}	L1_{post}	$\%\text{Support}_{post}$
8	2.5	118	160	0.30	51.00
10	2.5	118	153	0.32	51.90
11	2.5	118	151	0.32	57.70
12	2.5	118	149	0.33	57.10
13	2.5	120	152	0.34	56.40
13	3.5	116	128	0.32	50.00

Table 2: Summary statistics for the matching procedure showing results for the interpretable areas of figure 7. The upper half of the table refers to the original sample and the lower half shows summary statistics for the matched sample.

6 Discussion and conclusion

In this paper, we have discussed the need for better methodology in the analysis of causal relations in conflict event data. Existing approaches based on inferential methodology only work reliably when data are available in natural spatial units of analysis. In many scenarios, such data are absent, and relying on artificial units bears the risk of generating false inference. Sliding window designs have been previously applied in these contexts. While adequately accounting for the MAUP, corresponding studies are rather weak on the inferential side: Usually, sliding window designs can only show that spatial and temporal clustering in empirical data significantly deviates from the clustering that can be expected under simulated baseline conditions.

Combining the best of both worlds, MWA applies a sliding window and an automated matching technique, offering an analysis of the causal connections between different types of events for different spatial and temporal distances from a given intervention. The sliding windows entail that pre-aggregated events cannot be easily analyzed, but the matching procedure is generic enough to work with fixed spatial cells, such as administrative units or settlement regions of ethnic groups. In numerical simulations, the method has revealed artificially constructed causal relationships. We have also shown that substantive inference can still be performed when small fractions of interventions overlap in space and time. Higher levels of overlaps (that indicate SUTVA violations) can still be analyzed – albeit less reliably – if numbers of previous treatment and control events are included in the matching procedure.

Applying these lessons to an empirical example yielded novel insights into the ongoing conflict in Iraq. Instead of being mere fence-sitters, civilians in Iraq actively supported incumbent forces in reaction to indiscriminate insurgent violence. This result is a strong reminder of the importance of civilian agency in asymmetric, population-centric conflicts and the negative repercussions that can result from indiscriminate violence.

All results reported in this study were produced using custom *R* code designed to automatically and efficiently perform all steps of MWA, including the sliding window analysis, automated matching using CEM, and the graphical presentation of the results. A corresponding "mwa" package for the *R* programming language will be released to the public and made available at http://cran.r-project.org/web/packages/mwa/index.html.

A number of applications of this method for future research also spring to mind. The effectiveness of different kinds of peacekeeping interventions on subsequent levels of conflict could be analyzed, for example. In criminological studies, different containment strategies could be tested against one another with regard to subsequent crime rates. A prerequisite for such analyses is detailed data on locations and timings of events and relevant geographic information for the matching procedure. If such information is available, the presented method can be used to generate relevant insights.

References

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

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics (pp. 227-243)*. Princeton, N.J.: Princeton University Press.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119 (1), 249–275.

Bhavnani, R., Donnay, K., Miodownik, D., Mor, M., & Helbing, D. (2014). Group segregation and urban violence. *American Journal of Political Science*, 58 (1), 226–245.

Braithwaite, A., & Johnson, S. D. (2012). Space-time modeling of insurgency and counterinsurgency in iraq. *Journal of Quantitative Criminology*, 28 (1), 31–48.

Buhaug, H. (2010). Dude, where's my conflict? LSG, relative strength, and the location of civil war. *Conflict Management and Peace Science*, 27 (2), 107–128.

CIESIN (2005). Gridded population of the world, version 3 (gpwv3). Columbia University; and Centro Internacional de Agricultura Tropical (CIAT). 2005. Gridded Population of the World Version 3 (GPWv3): Population Grids. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. Available at http://sedac.ciesin.columbia.edu/gpw.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45 (3-4), 562–565.

Cressie, N. (1996). Change of support and the modifiable areal unit problem. *Geographical systems*, 3 (2-3), 159–180.

Dark, S. J., & Bram, D. (2007). The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography*, 31 (5), 471–479.

Diprete, T. A., & Engelhardt, H. (2004). Estimating causal effects with matching methods in the presence and absence of bias cancellation. *Sociological Methods and Research*, 32 (4), 501–528.

Ellsberg, D. (1970). Revolutionary judo. Available online at www.rand.org/content/dam/rand/pubs/documents/2006/D19807.pdf.

Gesch, D. B., Verdin, K. L., & Greenlee, S. K. (1999). New land surface digital elevation model covers the earth. *Eos, Transaction, American Geophysical Union*, 80 (6), 69–70.

Hansen, M., DeFries, R., Townshend, J. R., & Sohlberg, R. (2000). Global land cover classification at 1km resolution using a decision tree classifier. *International Journal of Remote Sensing*, 21 (6-7), 1331–1365.

Haushofer, J., Biletzki, A., & Kanwisher, N. (2010). Both sides retaliate in the israeli-palestinian conflict. *Proceedings of the National Academy of Sciences*, 107 (42), 17927–17932.

Hegre, H., Østby, G., & Raleigh, C. (2009). Poverty and civil war events: a disaggregated study of liberia. *Journal of Conflict Resolution*, 53 (4), 598–623.

Herron, M. C., & Wand, J. (2007). Assessing partisan bias in voting technology: the case of the 2004 new hampshire recount. *Electoral Studies*, 26 (2), 247–261.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: coarsened exact matching. *Political Analysis*, 20 (1), 1–24.

Jaeager, D. A., & Paserman, M. D. (2008). The cycle of violence? An empirical analysis of fatalities in the palestinian-israeli conflict. *American Economic Review*, 98 (4), 1591–1604.

Kalyvas, S. (2006). *The logic of violence in civil wars*. New York: Cambridge University Press.

Kocher, M. A., Pepinsky, T. B., & Kalyvas, S. N. (2011). Aerial bombing and counterinsurgency in the vietnam war. *American Journal of Political Science*, 55 (2), 201–218.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26 (6), 1481–1496.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76 (4), 604–620.

Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global data on events, location and tone, 1979-2012. Paper presented at the International Studies Association meetings, San Francisco, April 2013, 1-49.

Leslie, T. F., & Kronenfeld, B. J. (2011). The colocation quotient: A new measure of spatial association between categorical subsets of points. *Geographical Analysis*, 43 (3), 306–326.

Linke, A., Wittmer, F., & O'Loughlin, J. (2012). Space-time granger analysis of the war in iraq: a study of coalition and insurgent action and reaction. *International Interactions*, 38 (4), 402–425.

Lyall, J. (2009). Does indiscriminate violence incite insurgent attacks? Evidence from chechnya. *Journal of Conflict Resolution*, 53 (3), 331–362.

Mason, D. T., & Krane, D. A. (1989). The political economy of death squads: toward a theory of the impact of state-sanctioned terror. *International Studies Quarterly*, 33 (2), 175–198.

McColl, R. W. (1969). The insurgent state: territorial bases of revolution. *Annals of the Association of American Geographers*, 59 (4), 613–631.

NGDC (2012). Nighttime lights time series (version 4). Available online at: http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html.

O'Loughlin, J., & Witmer, F. (2011). The localized geographies of violence in the north caucasus of russia, 1999-2007. *Annals. Association of American Geographers*, 101 (1), 178–201.

O'Loughlin, J., Witmer, F., Linke, A., & Thorwardson, N. (2010). Peering into the fog of war: The geography of the wikileaks afghanistan war logs 2004-2010. *Eurasian Geography and Economics*, 51 (4), 472–95.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Norwich (England): Geobooks.

Openshaw, S., & Taylor, P. J. (1979). A million or so correlated coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley (Ed.), *Statistical Applications in the Spatial Sciences* (pp. 127–144). London: Pion.

Raleigh, C., & Hegre, H. (2009). Population size, concentration, and civil war. a geographically disaggregated analysis. *Political Geography*, 28 (4), 224–238.

Raleigh, C., Linke, A., & Hegre, H. (2010). Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research*, 47 (5), 651–660.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1), 41–55.

Rubin, D. (1973). Matching to remove bias in observational studies. *Biometrics*, 29 (1), 159–183.

SIGACT (2010). Wikileaks: The iraq war logs. Available online at http://wikileaks.org/.

Sundberg, R., Lindgren, M., & Padskocimaite, A. (2010). UCDP GED codebook version 1.5-2011. Available online at http://www.ucdp.uu.se/ged/.

Weidmann, N. B., Kuse, D., & Gleditsch, K. S. (2010). The geography of the international system:the CShapes dataset. *International Interactions*, 36 (1), 86–106.

Weidmann, N. B., & Salehyan, I. (2013). Violence and ethnic segregation: A computational model applied to baghdad. *International Studies Quarterly*, 57 (1), 52–64.

Wucherpfennig, J., Weidmann, N. B., Girardin, L., Cederman, L.-E., & Wimmer, A. (2011). Politically relevant ethnic groups across space and time: introducing the GeoEPR dataset. *Conflict Management and Peace Science*, 28 (5), 423–437.