Applying Spatial Causal Inference on Induced Seismicity

Yuchen Xiao [1], Corwin Zigler[3], Peter H. Hennings [4], Alexandros Savvaidis [4], Michael J. Pyrcz[1,2]

[1]Hildebrand Department of Petroleum and Geosystems Engineering, University of Texas at Austin, Austin, Texas, USA
[2]Jackson School of Geosciences, University of Texas at Austin, Austin, Texas, USA
[3]Department of Statistics and Data Sciences, University of Texas at Austin, Austin, Texas, USA
[4]Bureau of Economic Geology, University of Texas at Austin, Austin, Texas, USA

Key Points:

- The separation of causal and statistical conditions force consideration on important assumptions like the strong ignorability
- Sensitivity analysis of grid configuration on statistical results is necessary for raster-based spatial problems
- There is a stable and significant causal relationship between saltwater disposal and induced seismicity in the Fort-Worth Basin

Corresponding author: Yuchen Xiao, xiao.jack@utexas.edu

Corresponding author: Corwin Zigler, cory.zigler@austin.utexas.edu

Abstract

Saltwater disposal has been identified as the dominant causal factor that contribute to induced seismicity. Physical models rely on mechanistic understanding to infer causality where they evaluate various conditions for fault slips albeit with a high degree of uncertainty due to sparse data and subsurface heterogeneity. Given these uncertainties, statistical analysis is designed to measure statistical associations in the observed data with parametric regression models and interpret the significance of specific coefficient as evidence of causation. However, it is often difficult to interrogate the coefficients between different statistical models as the coefficients hold different implications. We propose a causal inference framework with the potential outcomes perspective to explicitly define what we meant by causal effect and declare necessary assumptions to ensure consistency between models for model comparison. The proposed workflow is applied to the Fort-Worth Basin of North Central Texas with the area of interest is discretized into non-overlapping grid blocks. Two statistical methods are employed to test the significance of the causal effect between the presence or absence of saltwater disposals and the number of the earthquakes and to estimate the magnitude of the average causal effect. In addition, our analysis is repeated for different grid configurations to directly assess the sensitivity of statistical results. We have identified a stable and statistically significant causal relationship between the presence of saltwater disposals and the number of earthquakes and have estimated there are, on average, 13 more earthquakes occurring in grids with saltwater disposals.

Plain Language Summary

Causal inference, a sub-field of statistics, has gained popularity across other quantitative fields of medicine, epidemiology, and social sciences to provide evidence of causality but has not been previously explored in geoscience. We apply a causal framework with the potential outcomes perspective, the outcomes we would observe under a counterfactual scenario, to analyze the effect of saltwater disposal on earthquakes. We found there is a statistically significant causal relationship between saltwater disposal and the number of earthquakes and estimated, on average, there are 13 more earthquakes occurring in grid with saltwater disposal. We performed sensitivity analysis on the effect of grid configuration on statistical results that is unique in raster-based spatial analysis.

1 Introduction

1.1 Background

Saltwater disposal (SWDs) has been linked to the recent increase of earthquakes in various regions of the United States (Ellsworth, 2013; Frohlich et al., 2016a; Grigoratos et al., 2020b; Hennings et al., 2019; Justinic et al., 2013; Keranen et al., 2013; Langenbruch & Zoback, 2017; McClure et al., 2017; Walsh & Zoback, 2015; Weingarten et al., 2015). In Texas, the development of shale hosted hydrocarbon resources in the Permian Basin, Eagle Ford Basin and Barnett Basin has resulted in a rapid expansion in both the number of SWDs and the cumulative injection volume, along with an abrupt increase in the number of earthquakes in respective basins (Hennings et al., 2019; Hornbach et al., 2015; Ogwari et al., 2018; L. Quinones et al., 2019; Scales et al., 2017; Zhai & Shirzaei, 2018). Of particular importance is the Fort-Worth Basin which hosts Barnett Shale in the North Texas that include most of the Dallas-Fort Worth (DFW) metropolitan area. Although the rate of earthquake activity in the DFW region has decreased since its peak in 2015, the potential linkages to oil and gas activity continuous to be a concern and put the social license of developing oil and gas resources in Texas at stake.

In response to this concern, the TexNet Seismological Observatory and the Center for Integrated Seismicity Research (CISR) at The University of Texas at Austin were

established to monitor potentially induced seismicity and to better understand the earthquake activities across the State of Texas (Hennings et al., 2019; Savvaidis et al., 2019). One of the overarching goals of TexNet-CISR is to improve causative understanding of the relationship between SWDs and onset earthquakes and the quantification of any identified causal relationship.

Advanced physics-based modeling has indicated the significant increases in pore pressure from large-scale SWD activities, which reduces frictional resistance of critically stressed faults, can induce fault slips (Fan et al., 2019; Zhai & Shirzaei, 2018; Keranen et al., 2014; Lund Snee & Zoback, 2016). However, the physical models do not provide direct evidence of whether an instance of earthquake is coincidental or whether there exist a clear causal relationship between larger number of earthquake and large number of SWDs (Hornbach et al., 2016; Fan et al., 2019; Langenbruch & Zoback, 2016; McClure et al., 2017).

To complement deterministic physical models, statistical analyses can provide additional evidence of the causal relationship between SWD activity and earthquakes which has practical and policy related to SWD regulation. In particular, causal inference, a subfield of statistics, has gained popularity across other quantitative fields of medicine, epidemiology, and social sciences to provide evidence of causality but has not been previously explored in geoscience (C. M. Zigler & Dominici, 2014; C. M. Zigler et al., 2018; Dominici & Zigler, 2017; Papadogeorgou et al., 2019; C. M. Zigler & Papadogeorgou, 2021; Reich et al., 2020; Imbens & Rubin, 2015; Hahn et al., 2020). An integral component of causal inference is the notion of potential outcomes where we conceive of different outcomes for a unit (e.g., a particular location in the study region) under different treatment options, noting that only one outcome can be ultimately observed for that unit (Imbens & Rubin, 2015). Using the notion of potential outcomes, causal inference methodology allows practitioner to explicitly define the causal effect at the unit-level as the difference between the potential outcomes (Imbens & Rubin, 2015). More specifically, the causal effect of interest in this work, formalized with potential outcomes, is the difference between what earthquake activity would potentially be at a location if SWDs were present and what earthquake activity would potentially be at the same location absent SWDs. The "fundamental problem of causal inference", where only one outcome can be observed at a given location, motivates the use of average comparison across multiple locations within the area of interest where SWDs are and are not present to compute the average causal effect (Holland, 1986). Importantly, the assumption of strong ignorability that there are unmeasured confounding features is essential in causal inference (Imbens & Rubin, 2015). The strong ignorability assumption, in this context, clarifies that the confounding factors would be unmeasured variables that jointly dictate SWD activity and earthquakes.

Although an existing body of work has used parametric regression models to establish spatiotemporal correlations between SWD fluid injection and earthquakes and has interpreted the statistical significance of specific coefficients as evidence of causation, the caveat here is "correlation does not imply causation" (Hornbach et al., 2015; Fasola et al., 2019; McClure et al., 2017; Grigoratos et al., 2020a; Aldrich, 1995; Langenbruch & Zoback, 2016). More specifically, we argue the causal validity from statistical analysis is not completely determined by statistical model specification, but rather related to explicit or implicit assumptions about the study design (C. M. Zigler & Dominici, 2014; Dominici & Zigler, 2017). For example, McClure et al. (2017) first describe their model specifications with modeling assumptions (i.e., assume the number of earthquakes is generated from a Poisson distribution) and then discuss the causal assumptions, strong ignorability assumption, to ensure the associations can be formally interpreted as demonstrating causality within their longitudinal study design (McClure et al., 2017).

We expand the workflow in McClure et al. (2017) and propose a new spatial causal inference workflow that integrates the notion of potential outcomes and relevant assump-

tions for the assessment of causality for induced seismicity. We apply two statistical methods for two specific aspects of the average causal effect of interest. First, we offer a randomization-based test of the null hypothesis of no causal effect of SWD placement on earthquakes, tailoring the null distribution of the test to the specifics of the study design. Second, we estimate the average causal effect and its uncertainty of SWDs on earthquakes with the average difference between the potential outcomes across multiple grids within the area of interest. Our focus on the effects of presence or absence of SWDs versus, for example, other work's focus on the effects of distributed SWD volume, is meant to simplify the problem and focus on key features of the causal framework. With new developments in spatial causal inference, exposure models that link the influence of particular SWD to earthquakes through distributed volume will be incorporated in future analysis (C. Zigler et al., 2020).

In addition to explicit causal considerations, we also offer an assessment of sensitivity of the statistical results to decisions about how to process the spatial data into a raster layer for analysis, specifically, the size and offset of spatial grids. Studies typically select a single grid configuration, chosen based on underlying knowledge or convenience, and then condition all inference on the chosen configuration (McClure et al., 2017; Grigoratos et al., 2020a). Lack of a universally accepted grid configuration, even for the same study area, invites an assessment of how sensitive a given study's results are to a chosen configuration to determine the possibility of analysis artifacts that are attributable to different grid configurations. Rather than condition inference on a single configuration, we conduct statistical analyses and summarise the statistical results under a variety of configurations to gauge sensitivity to the grid size and placement of the raster layer. Results point towards the potential for sensitivity to grid configuration that warrants careful consideration in raster-based spatial analysis.

In Section 2, we expound our design decisions and describe the specifics of two designated statistical methods. More specifically, we detail the implications of two causal conditions and highlight how formulating the problem with potential outcomces forces deliberate considerations on the placement of SWDs to approximate an randomized experiment (Section 2.2 and Section 2.3). We argue more emphasis should be placed on the proper construction of the null distribution for hypothesis testing and demonstrate how our approach arrives at proper null distribution for our study and for previous studies (Section 2.4.2 and Appendix Appendix E). We further differentiate between the marginal interpretation and the conditional interpretation in raster-based spatial problems (Section 2.4.3). Lastly, we perform sensitivity analysis to directly assess the impacts of grid configuration on the statistical results (Section 2.6). We discuss our results in Section 3 and motivate future researches in Section 4.

2 Methods

2.1 Data Assembly and Parameterization

Our study area is the Dallas Fort-Worth (DFW) Basin in North-Central Texas. Numerous studies have documented the evolution of earthquake sequences, collected extensive compilations of mapped faults, and conducted numerical simulations of hydrological modeling and fault activation in the area of interest (Hennings et al., 2019; Frohlich et al., 2016a, 2020; Fan et al., 2019; Hornbach et al., 2016; Scales et al., 2017; L. A. Quinones et al., 2018; Lund Snee & Zoback, 2016; Gao et al., 2019). We refer to above references for complete background information for the study area.

We use the North Texas Earthquake Study (NTXES) catalog (2008-2018), collected at the South Methodist University (SMU), in this study (L. Quinones et al., 2019; DeShon et al., 2019). We have not screened the earthquake catalog because there are significantly less earthquakes that have magnitudes above 2.0 compared to those in Oklahoma. There

are only about 103 earthquakes after declustering assuming the magnitude of completeness is 2. We are aware that the SMU had few earthquake monitoring stations back in 2008 and the temporary stations have mostly captured the aftershocks, not the main shocks. We aim the demonstrate the merits of our statistical framework and avoid being hampered by data-related issues. We use the operator reported SWDs injection volume data in DFW area from 2000s to 2017. It is available to download from Texas Railroad Commission website. The study area is within 32.07 degree to 33.68 degree latitude and -98.38 degree to -96.74 degree longitude. A specific coordinate reference system is used to convert from latitude and longitude coordinates to Cartesian coordinates. The perimeter of the study area is selected to best encompass all available SWDs and earthquakes while constraining the total area.

2.2 Causal Quantities of Interest

We discretize the study area into non-overlapping grid blocks, each block representing an observational unit of analysis. Each grid block is indexed i, taking on values $1, ..., N$. The presence or absence of SWDs in grid block i is denoted as W_i, taking on value 0 if grid i does not have SWDs and 1 if the grid i does have SWDs. Hence $\mathbf{W}$ indicates the presence or absence of SWDs across all grids within the area of interest and it is a representation of the spatial placement of SWDs in the study area. Let $Y_i(0)$ denote the potential outcome, that is, the number of earthquakes that would occur at grid i if there were no SWDs. Define $Y_i(1)$ analogously to be the potential outcome for grid i if there were SWDs present in that grid block. The individual-level causal effect of the presence of SWDs on the number of earthquakes in block i is defined as $Y_i(1) - Y_i(0)$. The average causal effect over the study area is defined as $\bar{Y}(1) - \bar{Y}(0)$, which is the average over the sample of the individual-level effects. The above potential outcomes notation implicitly assumes that there is "no interference" between grids, where the presence or the absence of SWDs in one grid block does not impact the number of earthquakes in other grid blocks and vice versa. This assumption is also employed in McClure et al. (2017) and Grigoratos et al. (2020a) where injection volume in one grid is assumed to not impact the modeled outcome of the number of earthquakes in other grids. The validity of this assumption may warrant more careful consideration in studies of induced seismicity, a point to which we return in Section 4.

With the above definition of causal effect, we can explicitly state a sharp null hypothesis of no causal effect of the presence of SWDs on earthquakes in any grid block as $Y_i(1) = Y_i(0)$ for all i, corresponding to the hypothesis that the presence of SWDs does not causally affect the number of earthquakes in any grid of the study area. We develop an appropriate null distribution and test for this null hypothesis using a randomization distribution that considers all plausible values of W. In addition to a statistical test of the sharp null hypothesis of no causal effect, we also estimate the magnitude of the average causal effect across the study area, $\bar{Y}(1) - \bar{Y}(0)$. Ideally, if SWDs were randomly allocated in the area of interest, then testing the null hypothesis and estimating the average causal effect would be trivial (Imbens & Rubin, 2015; McClure et al., 2017). In reality, using observations across a study area where some locations have SWDs requires careful considerations of why SWDs are placed in their observed locations (Imbens & Rubin, 2015), so one can judge the extent to which this placement could be reasonably assumed to be random with respect to earthquakes. This judgment will be dictated in large part by a) assumption about the mechanism determining the placement of the SWDs, which we elaborate as the strong ignorability assumption in Section 2.3; and b) assumption about the spatial distribution of SWD placement, which will dictate the construction of an appropriate null distribution for a hypothesis test of no causal effect in Section 2.4.

2.3 Strong Ignorability and the "Assignment" of SWDs

The main assumption dictating the extent to which the study can reasonably approximate the design of a randomized experiment is that of strongly ignorable treatment assignment, or strong ignorability (Imbens & Rubin, 2015). This assumption states that, whatever the mechanism dictating the presence or absence of SWDs across the study area, it can be regarded as "random" in the sense that it is unrelated to the potential outcomes of the number of earthquakes for any grid. Formally, this assumption specifies conditional independence between W_i and $Y_i(0), Y_i(1)$, conditional on other grid features. In other words, there are no unobserved confounding factors, such as human attribution or geologic factors, that dictate both the placement of SWDs and the occurrence of earthquakes(McClure et al., 2017). One potential threat to the validity of this assumption is confounding due to the location of geologic faults. It is reasonable to suggest that locating SWDs closer to or farther from geologic faults might make it more or less likely their fluid injection triggers fault slips (McClure et al., 2017; Keranen et al., 2013; Hincks et al., 2018; Gao et al., 2019). Intentional placement of SWDs in relation to fault locations would violate the ignorability assumption and indicate poor approximation of a controlled experiment that randomly place SWDs. We expect this threat in our analysis to be minimal, since operators typically did not have complete information on fault locations, which are typically mapped after the occurrence of earthquakes (Hennings et al., 2019; Horne et al., 2020), which themselves may be induced at long time lags following the initiation of SWD (McClure et al., 2017; Fasola et al., 2019; Schoenball & Ellsworth, 2017).

2.4 Randomization-Based Hypothesis Test

Beyond the assumption of ignorability, the notion of approximating a randomized experiment also points towards consideration of alternative values of $\mathbf{W}$ that might have arisen from a similar design to serve as the basis of a null distribution for the test of the sharp null hypothesis. In our proposed workflow, the plausible $\mathbf{W}$ correspond to plausible arrangement of SWDs in the study area, which should correspond to the unique spatial characteristics evident in the observed placement of SWDs. The key idea of a randomization-based test of the null hypothesis is to compare the observed relationship between the presence of SWDs and earthquakes against what would be observed under the observed distribution of earthquakes but under various probabilistically-generated alternative values of $\mathbf{W}$ corresponding to alternative random assignments in a randomized experiment. To construct such a null distribution of plausible alternative values of $\mathbf{W}$, we need to model the mechanism that simulate the randomization of $\mathbf{W}$ which matches with the unique spatial characteristics in the observed placement of SWDs. In particular, the randomization of $\mathbf{W}$ grants every grid (i.e., even those grids without observed SWDs or observed earthquakes) to be eligible for having SWDs and constructs just one null distribution corresponding to plausible SWD placement across the entire study area.

To reflect the spatial structure inherent to the observed placement of SWDs, we select the Log-Gaussian Cox Process, (LGCP), to reproduce different SWD point patterns that assemble the observed SWD point pattern, Figure 1, where the number of SWDs is fixed for every reproduction. The comparison between the set of first- and second-order spatial summary functions of the observed SWD point pattern and those of the fitted LGCP model are displayed in the right columns of Figure 2 and Figure 3, respectively. We observe all empirical summary functions (i.e., black lines) are within the confidence intervals of the fitted LGCP model (i.e., shaded grey regions) and are near the expectations of the fitted LGCP model (i.e., red dotted lines). For illustrations, eight simulated SWD point patterns are shown in Figure 4, where some are indistinguishable from the observed SWD point pattern (i.e., Figure 1) in terms of inter-distances and spatial correlations. In short, while one might envision a controlled experiment where each grid is randomly assigned to either have or have not an SWD or where a fixed number of SWDs are placed in a manner that reflects complete spatial randomness, we advocate instead

for the approximation of an experiment where a fixed number of SWDs are placed across the entire study region in a manner that reflects basic spatial features of the observed SWD distribution.

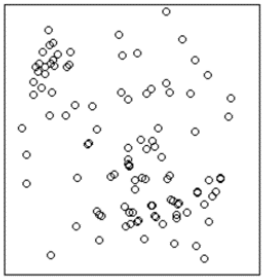

Figure 1. The observed SWD point pattern is shown where the black bounding box is the perimeter of the study region.

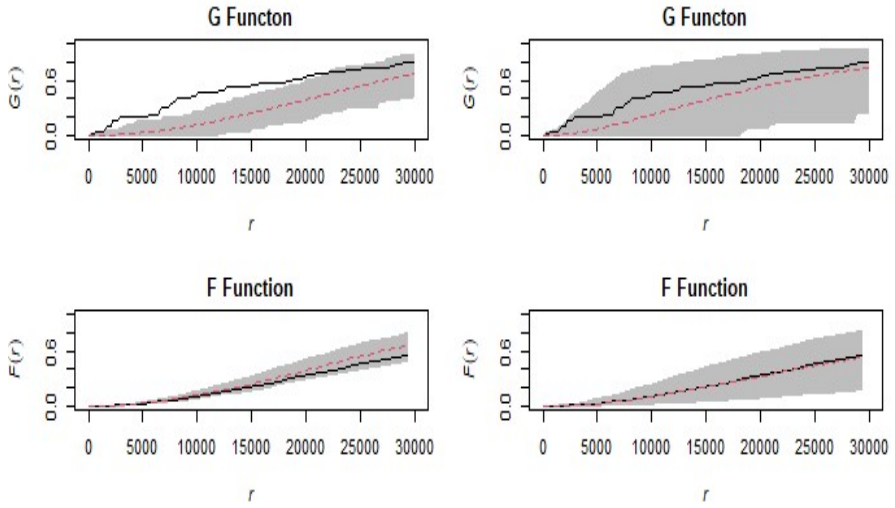

Figure 2. The first-order empirical summary functions (i.e., the cumulative nearest-neighbor distances, the G function, and the cumulative empty-space distances, the F function), shown in black line, are compared to that of the Complete Spatial Random (CSR) (left column) and the LGCP (right column), respectively. The grey intervals are the confidence intervals constructed from 3000 Monte Carlo simulations and the red dotted lines are the expectations with respect to each point process models (Baddeley et al., 2015; Illian et al., 2008). It is obvious the empirical summary functions run outside the confidence intervals of CSR, indicating the observed SWD point pattern is not of CSR origin. In comparison, the empirical summary functions nearly match the expectation of the summary functions of the fitted LGCP model, indicating the LGCP model fits well with the observed SWD point pattern in term of inter-distances.

To test the sharp null hypothesis of no causal effect, we repeat the generation of $\mathbf{W}$ with the LGCP model 1000 times and calculate 1000 average causal effects under different realized $\mathbf{W}$ to constitute a reliable null distribution for hypothesis testing. To test the null hypothesis, we calculate the test statistics described below for each randomly-

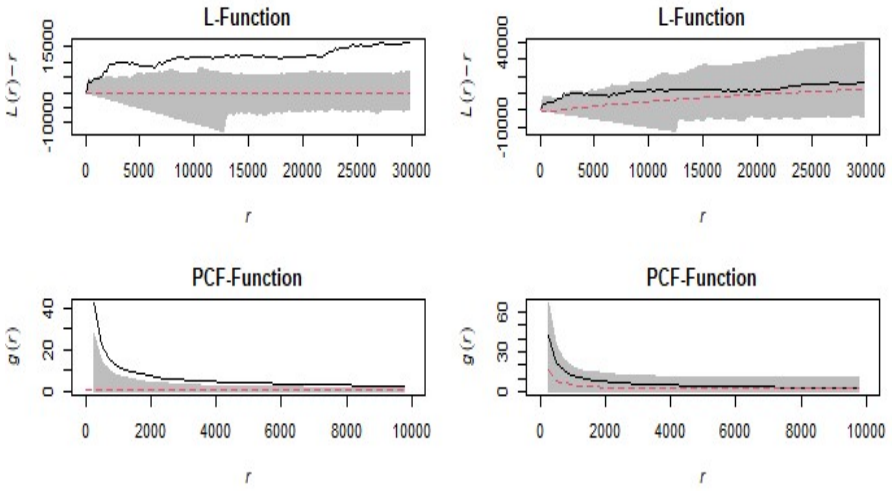

Figure 3. The second-order empirical summary functions (i.e., the L-function and the Pair Correlation Function (PCF)) are compared to that of the CSR (left column) and the LGCP (right column), respectively. Again, the left column shows the empirical summary functions run outside the confidence intervals of CSR, indicating the observed SWD point pattern is not of CSR origin. In comparison, the empirical summary functions nearly match the expectations of the summary functions of the LGCP, indicating the LGCP fits well with the observed SWD point pattern in term of correlations.

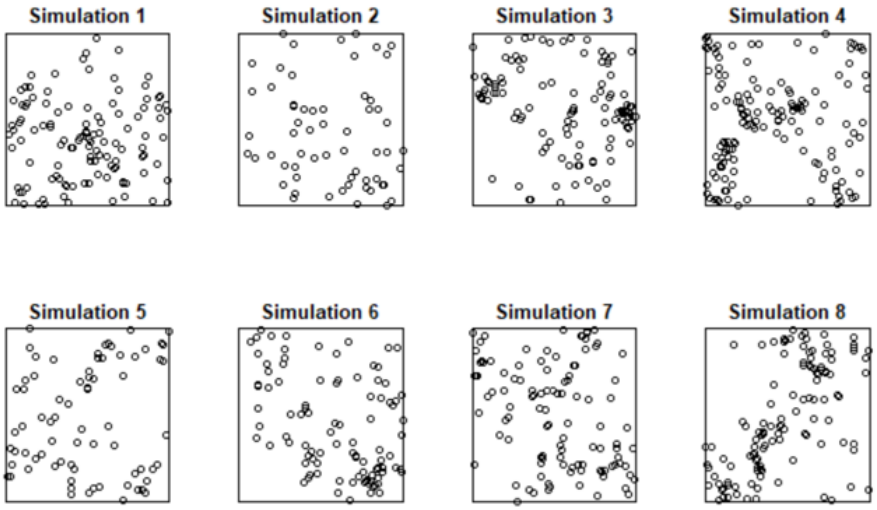

Figure 4. Eight simulated SWD point patterns are compared to the observed SWD point pattern in Figure 1. Some are very similar to the observed SWD point pattern in terms of inter-distance and spatial correlation.

generated value of $\mathbf{W}$, calculating a p-value to describe the observed value of the test statistic relative to the distribution of simulated values under the assumed sharp null hypothesis. A large p-value indicates an observed relationship between SWD placement and earthquakes that is consistent with no causal effect and ignorable SWD placement;

a low p-value indicates evidence to reject the sharp null and is interpreted as evidence of a causal effect.

2.4.1 Rank Transformation

A unique feature in raster-based spatial analysis is there might be excessive zeros. For example, consider a 30 by 30 discretization scheme where there are 900 grid blocks but only less than 80 grid blocks have non-zero counts of earthquakes. Consequently, a standard t-statistic may not be appropriate because of the excessive zero counts; in any particular discretization, the vast majority of grids do not have earthquakes. To address this, we implement a ranked T statistic to directly minimize the impacts of the zero-counts (Imbens & Rubin, 2015).

2.4.2 The Proper Null Distribution for Hypothesis Testing

Following above, we highlight it is the explicit definition of the causal effect and the transparent characterization of the sharp null hypothesis prompt us to utilize the LGCP model to approximate a randomized experiment and to construct proper null distribution that best reflect the intended null hypothesis. Furthermore, we naturally calculate one p-value for the entire area of interest following the study design and the causal assumptions instead of for every grid. We now distinguish our approach with the previous studies where we focus on the construction of the null distribution and the implications of the marginal interpretation and the conditional interpretation.

Approximating a randomized experiment where the SWDs are assigned across the study area clarifies two points that distinguish with previous work. First, the null distribution is constructed with respect to the entire study region instead of specific to each grid (McClure et al., 2017; Grigoratos et al., 2020a). Second, the inclusion of all grid cells within the area of interest focuses on marginal interpretation instead of conditional interpretation (McClure et al., 2017; Grigoratos et al., 2020a). McClure et al. (2017) propose to use the resampling method to construct null distribution for hypothesis testing specific to each grid. The resampling method requires the identically and independently distributed (*iid*) assumption to guarantee the resampling is done in a way that reflects the intended null hypothesis (Carsey & Harden, 2013; Hall & Wilson, 1991; McClure et al., 2017). However, the sampled distributed volumes for each grid are not *iid* because they are spatially correlated which produce narrower null distribution and lead to overly-optimistic p-value for each grid. In Appendix Appendix E, we further illustrate that our approach even constructs proper null distribution with distributed volumes which is directly applicable to previous studies (McClure et al., 2017; Grigoratos et al., 2020a).

2.4.3 The Inclusion of Zero-Counts

Both McClure et al. (2017) and Grigoratos et al. (2020a) have only analyzed grids that had hosted at least one earthquake and have made an implicit statistical condition that grids with zero observed earthquakes are implausible to have nonzero predicted earthquakes. In fact, knowing locations with low predicted earthquakes that actually have zero earthquakes would provide very useful information on the causal relationship we want to investigate, as would knowing locations with zero observed earthquakes but have predicted to have many (Panzeri et al., 2008). Our general causal formulation of the problem with the potential outcomes perspective focus considerations on the randomization of $\mathbf{W}$. For every randomization, we allow the SWDs to be allocate to any locations within the area of interest as long as they match the spatial characteristics of the observed SWD point pattern. Our definition of the causal effect and the hypothesis testing procedure would be ill-conceived if the SWDs are only allowed to be allocate in the grids with observed earthquakes. We underline that the selection of the study area should be made before any statistical analysis and not dictated by the statistical analysis to avoid selec-

tion biases where the latter does not provide population-level summary (i.e., conditional interpretation vs. marginal interpretations).

2.5 Estimating the Average Causal Effect

Testing the causal effect between the presence or absence of SWDs and the number of earthquakes is the first step towards understanding causality, quantifying such causal effect is the second step. Given the explicit definition of the causal effect with the potential outcomes perspective, we are interested to know what would the average number of earthquakes be if all grids were to have SWDs, $\bar{Y}(1)$ in notation format. Similarly, we ask what would the average number of earthquakes be if all grids were to have no SWDs, $\bar{Y}(0)$ in notation format. More importantly, what is the difference between the average potential outcomes? We define such difference between the average potential outcomes as the average causal effect and denote it as τ_{fs} where:

$$\tau_{fs} = \bar{Y}(1) - \bar{Y}(0) = \sum_{i=1}^{N} (Y_i(1) - Y_i(0))/N \tag{1}$$

We interpret τ_{fs} as the average increase in the number of earthquakes for any grid in the area of interest with SWDs compared to without SWDs and it serves as an approximation to the average causal effect (Imbens & Rubin, 2015). Provided with the causal assumptions, we employ the LGCP model in a similar fashion to compute the average causal effects for every randomization of $\mathbf{W}$. We further derive the expectation and variance of the average causal effect. Stepping away from hypothesis testing marks an important milestone to deepen our understanding of the causal relationship between SWDs and the number of earthquakes. It is often trivial to prove, in terms of statistically significant p-values, that the onset earthquakes are linked to SWDs. Perhaps, it is more consequential to quantify the effects of SWDs on the number of earthquakes. The specifications of the method are provided in Appendix D.

2.6 Assessing the Sensitivity of Grid Configuration

In previous studies, McClure et al. (2017) divided the State of California and Oklahoma into uniform grid blocks of 0.2 latitude and 0.2 longitude (roughly 22,5 km by 18 km) and performed one grid offset which found no significant difference in results. Grigoratos et al. (2020a) calculated p-values in 20 km grid blocks and took the median value from the sixteen 20 km grid blocks as the p-value for a 5 km grid block. The above approaches either somewhat disregard the impacts of grid sizes and grid offsets or failed to properly capture the variations resulting in erratic behaviors of p-values (McClure et al., 2017; Grigoratos et al., 2020a). Consider a study area in a 4 by 4 discretization scheme (left of Figure 5), where the pivot, defined as the bottom left corner, is allowed to move within in a grid block of the same size as the grid blocks in the study area and the center of that grid block coincides with the original pivot. The pivot is randomly shifted 100 times (right of Figure 5) to generate 100 slightly different raster layers. For every unique raster layer, we repeat the statistical analyses. In addition, we calculate the average across the 100 raster layers and repeat the statistical analyses. This process is repeated for a range of grid sizes where the statistical analyses are repeated 101 times for every grid size. Although grid configurations have been partially informed by domain expertise in previous studies, they are still arbitrary and provoke instability in statistical results. Sensitivity analysis of grid configuration is therefore critical because different areas of interest with different data availability might require different grid sizes. For example, the State of Oklahoma has hosted thousands of $M \geq 3$ earthquakes where DFW has hosted significantly less. The average number of earthquakes in 5 km grid blocks could vary drastically depending on the residing States. To our knowledge, there has not been comprehensive sensitivity analysis on grid configuration and we aim to bridge this gap in the

literature. We repeat our analyses described in Section 2.4 and Section 2.5 for 3,131 times (i.e., 101 grid offsets for every grid size with total 31 different grid sizes), respectively, and summarize the results across all grid offsets for every grid size to gauge sensitivity of the analysis about grid configuration.

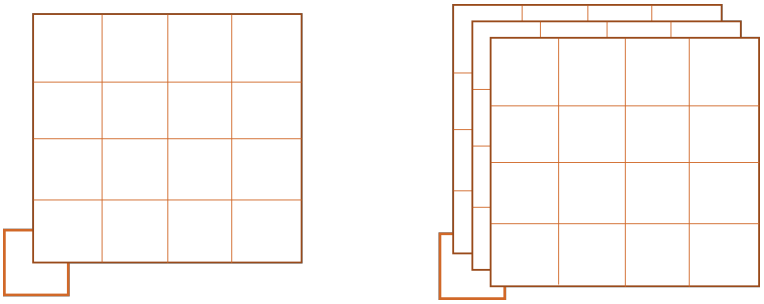

Figure 5. Illustration of Grid Offsets

3 Results and Discussion

3.1 Results from the Randomization-Based Hypothesis Test

Figure 6 displays the p-values from the sharp null hypothesis test where the y-axis shows the p-values transformed into log-scale to help better examining clusters near zero and the x-axis shows all the grid sizes implemented in the study. Every p-value represents a unique combination of grid size and grid offset. Each green point is a statistically significant p-value calculated from performing the hypothesis test on one grid offset for a particular grid size and the blue point is the p-value calculated from performing the hypothesis test on the average data values of 100 raster layers from grid offsets. Because only the statistically significant p-values are marked with dots, the overall distribution of the p-values for every grid size is rendered with a violin plot. Our results should be differentiated from Grigoratos et al. (2020a), where they summarized median p-values from the 20 km grids for the 5 km grids and did not perform independent hypothesis test on the 5 km grids (Grigoratos et al., 2020a).

Overall, Figure 6 indicates there is a stable and statistically significant causal relationship between the presence of SWDs and the number of earthquakes over the entire study area across a range of grid sizes. There is a trend from larger grid sizes to smaller grid sizes where there are less extremely small p-values that are below 0.0010. This is potentially caused by the reduction in the ratio of non-zero counts and zero-counts as the grid size diminishes, when there are many grids and the overwhelming majority are without earthquakes.

To recognize the impacts of grid offset, we focus on grid size 10.2 km by 12.3 km. This grid size has a wider and a more uniform distribution of p-values as displayed in the violin plot, which indicates grid offsets can have a large impact on the p-values for this specific grid size. The wide distribution of p-values suggests small shifts in the segmentation of the study area might separate critical clusters differently and result in a divergence of statistical results.

Following above, there is another observed trend where larger grid sizes have flatter distribution of p-values and smaller grid sizes have more concentrated distribution of p-values. Because the pivot is allowed to move within a grid block of the same size as the grid blocks in the study area, larger grid blocks are more likely to experience more distinct placements of the raster layer due to different initial grid offsets and have more

divergent results. In comparison, smaller grid blocks are more constrained to obtain divergent results.

Through the compelling visualization, we conclude there is a stable and statistically significant causal relationship between the presence of SWDs and the number of earthquakes for the entire study area. In addition, we argue comprehensive sensitivity analysis of grid configuration is necessary in raster-based spatial analysis to demonstrate the stability of statistical results and to arrive at objective conclusions.

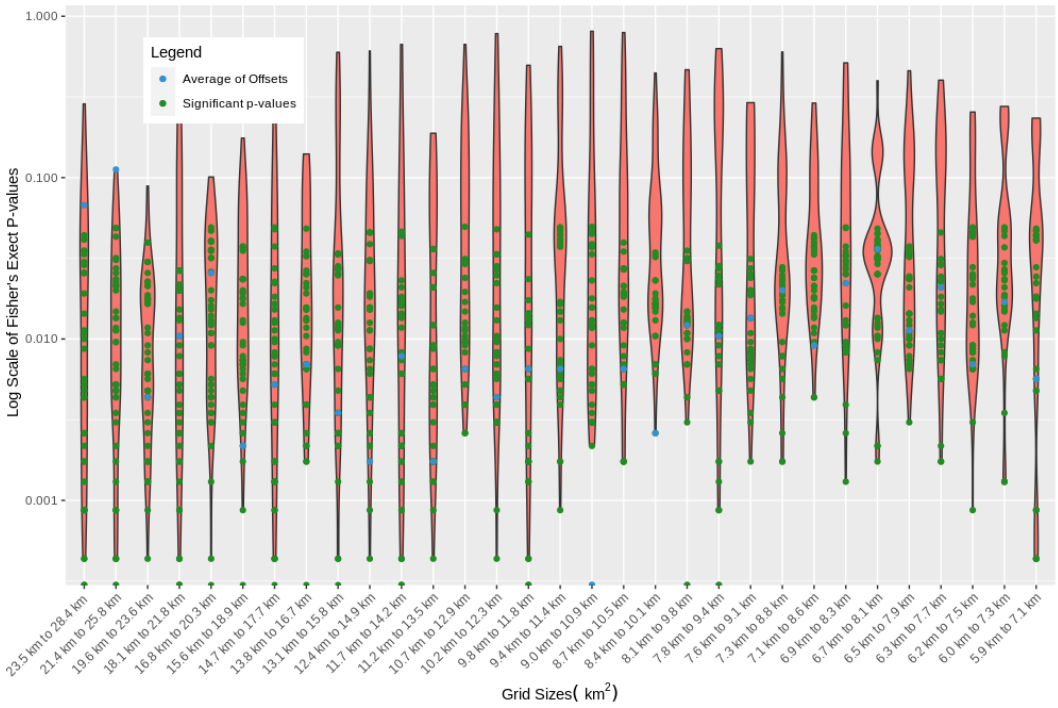

Figure 6. The p-values from hypothesis test are displayed. Only the statistically significant p-values are marked with green dots and the p-value computed from the average of 100 raster layers is marked with blue dot.

3.2 Results from Estimating the Average Causal Effect

Figure 7 shows the average causal effects in boxplot for different grid configurations. The solid bar is the median of the average causal effects summarised over all grid offsets for every grid size. There are two noteworthy observations. First, there is a decreasing trend in the average causal effect as the grid size diminishes. This is an unique artifact for raster-based spatial problems and it is expected since both the SWDs and earthquakes are measured at points, so increasingly finer grids will eventually separate each earthquake and each SWD into its own grid cell, resulting in a null effect estimate. Second, there is a general trend that larger grid sizes exhibit greater variations in the average causal effect from grid offsets. Larger grid sizes experience more distinctive discretization schemes from grid offsets where they have more ways to divide critical clusters and undoubtedly result in more diverse statistical results.

Table 1 shows the average height of the confidence intervals of the average causal effects and the ratio of the 90% confidence intervals that overlap zero. Because we generate a pair of the average causal effect and the corresponding confidence interval for ev-

ery grid offset, it is hard to visualize all the confidence intervals for every grid size. Alternatively, we calculate the height of every confidence interval (i.e., subtracting the lower bound from the upper bound) and summarize the average height over all grid offsets for every grid size. The average confidence interval height is providing some sense of the typical uncertainty around a point estimate for a given grid size.

Table 1 illustrates yet another unique artifact to raster-based spatial problems where the uncertainty around the point estimates goes down with decreasing grid size, primarily because of the increasing number of observations. The rightmost column in Table 1 shows the ratio of the 90% confidence intervals that overlaps 0 for every grid size. If a confidence interval overlaps 0, it serves as evidence that it is not significantly different from 0. For the largest grid size, 50% of the confidence intervals include zero. This percentage is increasing for smaller grid sizes where the confidence intervals overlap 0 for all grid offsets for grid sizes smaller than 10.7 km by 12.9 km. Furthermore, all 95% confidence intervals overlap 0 for all grid offsets for every grid size thus they are not shown.

We conclude using the average difference between the potential outcomes across grids as an approximation to the average causal effect is very sensitive to grid configurations. Unless there are specific grid sizes of interest, it is difficult to make any interpretation. We select 19.6 km by 23.6 km and 18.1 km by 21.8 km as the closet grid sizes to the grid size used in McClure et al. (2017) (i.e., 18 km by 22.5 km) where we find the expectation of the average causal effects for the two grid sizes to be 13. In other words, we expect there are, on average, 13 more earthquakes occurring in any grid with SWDs versus without SWDs within the area of interest provided the selected grid size is sound from domain expertise (McClure et al., 2017). Importantly, we note the choice of grid size(s) of interest should be made based on physical understanding of the problem, and then sensitivity to grid offsets within those relevant grid sizes should be gauged. We provide the wide range of grid sizes for illustration only - some grid sizes could presumably be ruled out as irrelevant for our analysis, and we underline the choice of grid size should not be based on the convenience of statistical methods.

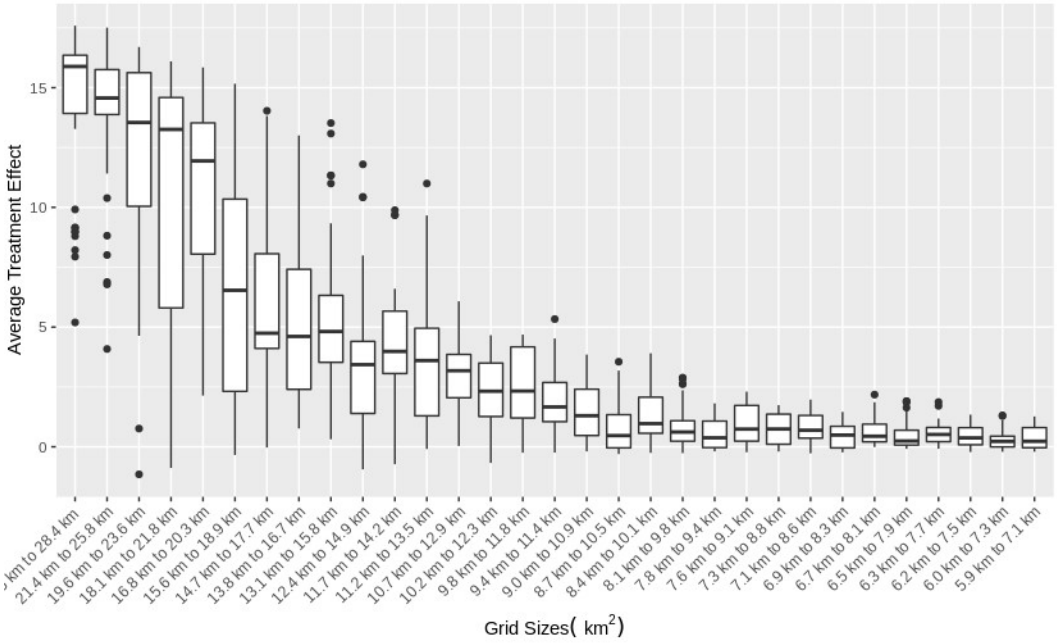

Figure 7. The boxplot displays the average causal effects across all grid offsets for every grid size. The average over the grid offsets is indicated by the solid horizontal bar.

Table 1. Summary of the Average Magnitude of the Average Causal Effect and the Ratio of the Confidence Intervals (CI) that Overlaps Zero

Gridsize	Average90CI	Average95CI	Ratio of 90%CIs that Overlaps Zero
23.5 km to 28.4 km	30.28	34.54	0.50
21.4 km to 25.8 km	28.92	33.21	0.50
19.6 km to 23.6 km	25.89	29.52	0.56
18.1 km to 21.8 km	23.36	26.30	0.72
16.8 km to 20.3 km	22.92	26.08	0.66
15.6 km to 18.9 km	17.13	18.68	0.92
14.7 km to 17.7 km	16.04	17.76	0.92
13.8 km to 16.7 km	13.81	15.37	0.94
13.1 km to 15.8 km	14.15	15.71	0.82
12.4 km to 14.9 km	9.78	11.10	0.98
11.7 km to 14.2 km	11.32	12.80	0.88
11.2 km to 13.5 km	9.61	11.04	0.94
10.7 km to 12.9 km	8.67	10.13	1.00
10.2 km to 12.3 km	7.13	8.36	1.00
9.8 km to 11.8 km	7.53	8.70	1.00
9.4 km to 11.4 km	6.05	7.36	1.00
9.0 km to 10.9 km	5.04	6.38	1.00
8.7 km to 10.5 km	3.28	4.64	1.00
8.4 km to 10.1 km	4.21	5.61	1.00
8.1 km to 9.8 km	3.24	4.63	1.00
7.8 km to 9.4 km	2.51	3.90	1.00
7.6 km to 9.1 km	3.26	4.75	1.00
7.3 km to 8.8 km	2.77	4.25	1.00
7.1 km to 8.6 km	3.12	4.57	1.00
6.9 km to 8.3 km	2.13	3.63	1.00
6.7 km to 8.1 km	2.46	3.98	1.00
6.5 km to 7.9 km	1.93	3.48	1.00
6.3 km to 7.7 km	2.20	3.74	1.00
6.2 km to 7.5 km	1.79	3.37	1.00
6.0 km to 7.3 km	1.40	2.98	1.00
5.9 km to 7.1 km	1.58	3.19	1.00

4 Conclusion and Future Work

Improvements in the understanding of the causal relationship between SWD and induced seismicity, more importantly, the quantification of such relationship, require advancement in statistical analysis that bypass certain limitations in deterministic approaches. Traditional parametric regression models presume the specified regression models accurately reflect the true relationship between the variables of interest which, when coupled with the (often implicit) assumption of strong ignorability, can provide evidence of causation. In contrast, we propose a general causal formulation of the spatial problem where we explicitly define the casual estimand with the potential outcomes perspective and implement appropriate statistical methods for subsequent testing and estimation. The causal conditions are deliberately separated from the statistical conditions so that the causal estimand is purposefully chosen rather than inherited from the specified parametric models with the expectation that the chosen estimand is more directly relevant to address the scientific question. Note that this perspective does not preclude the usefulness of re-

gression modeling strategies, it only serves to separate key determinations of causal validity from the specification of such models.

Using the potential outcomes perspective, we explicitly define what is meant by a causal effect, and then use the framing relative to the approximate design of a randomized experiment as a benchmark to guide the analysis and interpretation of threats to validity. In particular, this led to different choices about grid configuration to include in the analysis and the construction of an appropriate null distribution. We perform inferences on two specific aspects of the average causal effect. First, we perform a sharp null hypothesis to test the statistical significance of the causal effect between the presence of SWDs and the number of earthquakes. We find a stable and statistically significant causal relationship between the presence or absence of SWDs and the number of earthquakes for the entire study area across a range of grid sizes. This result is consistent with the results from other studies which found strong evidence of wastewater-induced seismicity in the DFW region of North-Central Texas. Second, we estimate the average causal effect and observe there are, on average, 13 more earthquakes for any grid with SWDs versus without SWDs for grid sizes 19.6 km by 23.6 km and 18.1 km by 21.8 km. We emphasize grid configuration has a material consequence on the statistical results. Grid configuration should be studied empirically because different areas of interest have different data availability and acquire different grid configurations. Domain knowledge should guide the choice of grid configuration but it can not replace empirical experimentation.

We highlight the statistical analyses that adopt causal inference framework with causal inference terminology are not superior by default (C. M. Zigler & Dominici, 2014). For example, the work from McClure et al. (2017) demonstrates causality under the strong ignorability assumption in a longitudinal design. The causal inference methodology only provides a formal structure to frame the question, whether our analysis demonstrates causality depend on how much we believe the LGCP model accurately reproduce the observed placement of SWDs conditional on all potential confounding variables - namely, the exclusion of geologic faults. We view the abovementioned two approaches are more different in styles and less different in substances. Observational studies face confounding problems that are difficult to fully account for: oversimplification of the complexity of the problem (e.g. adding the strong ignorability assumption and no interference assumption) can potentially invalidate the causal portion of the analysis (Carone et al., 2020). Nevertheless, we argue a more general causal formulation of the problem with the potential outcomes perspective could continuously reflect the complexity of the problem and improve the clarity and transparency regarding the most important tenets for discerning whether empirical statistical analyses provide evidence of causality between SWD and seismicity (Imbens & Rubin, 2015; Carone et al., 2020).

Causal inference methodology has become popular and led to important contributions in a variety of other disciplines including education, psychology, economics, epidemiology, medicine, sociology (Friedrich & Friede, 2020; Glass et al., 2013; Imbens & Rubin, 2015). It has been mostly unexplored in the areas of geoscience and engineering where the objective is to infer causality between spatial variables. Spatial causal inference is a fast-growing field and has contributed to air pollution epidemiology that support important regulatory policies. A major obstacle of applying spatial causal inference in raster-based spatial problems is the assumption of no interference. For example, it is reasonable to suggest that the presence or absence of SWDs in some neighboring grids could all contribute to the occurrence of earthquakes for a grid. There is an inherent many-to-one and one-to-many relationship where many SWDs reside in different grids all contribute to the occurrence of earthquakes for a grid and one SWD might affect the occurrence of earthquakes in many different grids. Similarly, it is logical to postulate that distributed volume in some neighboring grids affect the occurrence of earthquakes for a reference grid. Difficulties arise when extending the causal inference framework to ac-

knowledge the interference that would arise when earthquakes at a given location might depend on the distributed volume at the location from multiple SWDs. These are subject to future works with a recently developed bipartite interference network in spatial causal inference where the target is to investigate the causal effect while relaxing the independence assumption between grids (C. M. Zigler & Papadogeorgou, 2021; C. Zigler et al., 2020; Giffin et al., 2020; Marrett et al., 2018).

Appendix A A Brief Introduction to Causal Inference

As Rubin (1974) points out, the problem of causal inference is a missing data problem: given any treatment assigned to an individual unit, the potential outcome associated with any alternate treatment is missing (Rubin, 1974; Imbens & Rubin, 2015). The assignment mechanism, therefore, plays a key role and answers questions such as: how is it determined which units get which treatments or, equivalently, which potential outcomes are realized and which are not.

We will now allude necessary notations in a general case. Let us index units in a population of size N by i, taking on values $1, ..., N$, and denote W_i as the treatment indicator for unit i, taking on values 0 (control treatment) and 1 (active treatment). Let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes of unit i for the control and active treatments, respectively. Recall only one of the potential outcomes will ultimately be realized and therefore possibly observed. Let Y_i^{obs} denotes this realized and possibly observed outcome:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases} \tag{A1}$$

Analogously, let Y_i^{mis} denotes the missing potential outcome:

$$Y_i^{mis} = Y_i(1 - W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 0, \\ Y_i(0) & \text{if } W_i = 1. \end{cases} \tag{A2}$$

By the same token, $\mathbf{Y}^{obs}$ and $\mathbf{Y}^{mis}$ are the corresponding N-vectors. Usually we want to characterize the potential outcomes in terms of the observed and missing outcomes therefore we invert these notations:

$$Y_i(0) = \begin{cases} Y_i^{mis} & \text{if } W_i = 1, \\ Y_i^{obs} & \text{if } W_i = 0. \end{cases} \text{ and } Y_i(1) = \begin{cases} Y_i^{mis} & \text{if } W_i = 0, \\ Y_i^{obs} & \text{if } W_i = 1. \end{cases} \tag{A3}$$

We define the assignment mechanism to be the function that assigns probabilities to all 2^N possible values for the N-vector of assignments $\mathbf{W}$, given the N-vectors of potential outcomes $\mathbf{Y(0)}$ and $\mathbf{Y(1)}$ (Imbens & Rubin, 2015). The assignment mechanism is then a row-exchangeable function $Pr(\mathbf{W}|\mathbf{Y(0)}, \mathbf{Y(1)})$, taking on values in $[0, 1]$, satisfying

$$\sum_{\mathbf{W} \in [0,1]^N} Pr(\mathbf{W}|\mathbf{Y(0)}, \mathbf{Y(1)}) = 1 \tag{A4}$$

for all $\mathbf{Y(0)}$ and $\mathbf{Y(1)}$.

A1 The Stable Unit Treatment Value Assumption (SUTVA)

Besides potential outcomes and assignment mechanism, additional assumptions are needed for causal validity under RCM. Here, we only touch one component of the Stable Unit Treatment Value Assumption (SUTVA), the No Interference assumption, which will be sufficient for our purposes. The No Interference assumption states the treatments

applied to one unit do not affect the outcome for another unit(Imbens & Rubin, 2015). Put simply, the potential outcomes and assigned treatments for any unit do not vary with the treatments assigned to other units.

Appendix B Log-Gaussian Cox Point Process

The LGCP is chosen because it scored lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) values compared to other alternatives, shown in Table B1 (Baddeley et al., 2015; Illian et al., 2008).

Table B1. Comparison of AIC and BIC Values Between Some Point Process Models

	LGCP	Thomas	MatClust	Cauchy
AIC	45500.29	45604.65	45601.63	45630.82
BIC	45601.97	45607.34	45604.32	45633.50

Appendix C Fisher's Exact Test with Sharp Null Hypothesis

Given data from a completely randomized experiments with SUTVA, Fisher's Exact Test (FET) is assessing the sharp null hypothesis of no effect of the treatment versus no treatment, that is, the null hypothesis under which, for each unit in the experiment, both values of the potential outcomes are identical (Mehta & Patel, 1983; Imbens & Rubin, 2015). Consider any test statistic T: a function of the stochastic assignment vector, $\mathbf{W}$; the observed outcomes, $\mathbf{Y}^{obs}$. The sharp null hypothesis allows us to determine the distribution of T, generated by the complete randomization of units across treatments. The test statistic is stochastic solely through the stochastic nature of the assignment. We refer to the distribution of the statistic determined by the randomization as the randomization distribution of the test statistic T. Using this distribution, we can compare the actually observed value of the test statistic, T^{obs}, against the distribution of T under the null hypothesis. An observed value that is "very unlikely", given the null hypothesis will be taken as evidence against the null hypothesis using p-value (Imbens & Rubin, 2015). Hence, the FET approach entails the two steps: (i) the choice of a sharp null hypothesis, and (ii) the choice of test statistic.

C1 Ranked Statistics

Ranked Statistic is an important class of test statistics involves transforming the data to ranks before calculating the test statistics. Such a transformation is attractive when the data have a distribution with a substantial number of outliers (Imbens & Rubin, 2015). We consider the large-portion of zero-counts as outliers and use normalized rank to reduce the impacts of zero-counts in the test statistics. The definition for the normalized rank with ties is:

$$R_i = R_i(Y_1^{obs}, ..., Y_N^{obs}) = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{obs} < Y_i^{obs}} + \frac{1}{2}\Big(1 + \sum_{j=1}^{N} \mathbf{1}_{Y_j^{obs} = Y_i^{obs}} - \frac{N+1}{2}\Big) \tag{C1}$$

Given the N ranks R_i, $i = 1, ..., N$, an obvious test statistic is the absolute value of the difference in average ranks for the treated and control units:

$$T^{rank} = |\bar{R}_t - \bar{R}_c| = \Big|\frac{\sum_{i:W_i=1} R_i}{N_t} - \frac{\sum_{i:W_i=0} R_i}{N_c}\Big| \tag{C2}$$

where we denote $\bar{R}_t$ to be the ranked statistic in the active treatment group (Imbens & Rubin, 2015).

Figure C1 shows the improvements in the statistical results after the rank transformation. The p-value under the regular T-statistic distribution (right) is drastically reduced by an order of magnitude after taking normalized rank transformation (left), minimizing the impacts of zero-counts.

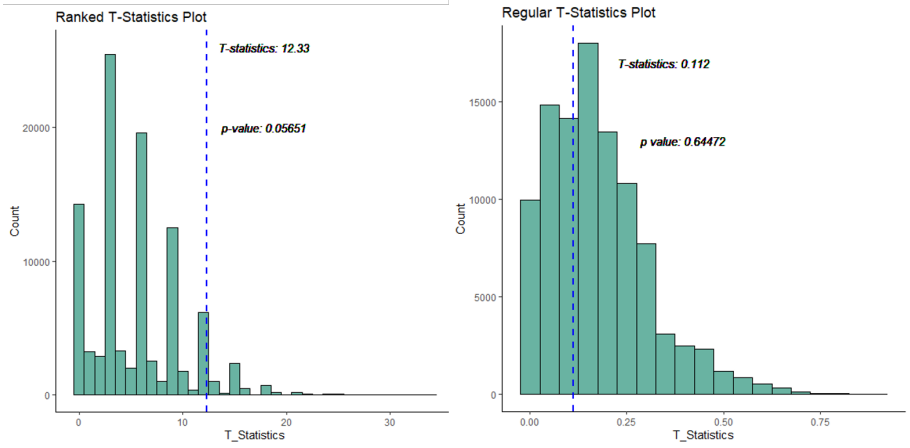

Figure C1. Comparison Between Ranked T-Statistics and Regular T-Statistics

Appendix D Neyman's Repeated Sampling

The Fisher's Exact Test provides limited information besides establishing the causal link between earthquakes and the presence of SWDs which is trivial to most researchers and policy makers (Grigoratos et al., 2020a; McClure et al., 2017; Frohlich et al., 2016b; Fan et al., 2019; Hennings et al., 2019). To broaden the scope of interpretation, Neyman's Repeated Sampling is employed to find the average causal effects of the presence of SWDs on the number of earthquake in any grid over the entire study area. Neyman's two basic questions are: (1) what would the average outcome be if all units were exposed to the treatment, $\bar{Y}(1)$? (2) How did that compare to the average outcome if all units were exposed to the control treatment, $\bar{Y}(0)$? Most importantly, what is the difference between these averages ? Neyman's approach was to develop an estimator of the average causal effect and derive its expectation and variance under repeated sampling.

D1 Unbiased Estimation of the Average Causal Effect

The population average causal effect τ_{fs} has the form:

$$\tau_{fs} = \bar{Y}(1) - \bar{Y}(0) = \sum_{i=1}^{N} (Y_i(1) - Y_i(0))/N \tag{D1}$$

where $\bar{Y}(0)$ and $\bar{Y}(1)$ are the averages of the potential control and treated outcomes, respectively:

$$\bar{Y}(0) = \frac{1}{N} \sum_{i=1}^{N} Y_i(0) \tag{D2}$$

$$\bar{Y}(1) = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) \tag{D3}$$

The LGCP enables generating thousands of random assignment vectors while preserving spatial correlation, where $N_t = \sum_{i=1}^{N} W_i$ are the number of grid blocks have SWDs and the remaining $N_c = \sum_{i=1}^{N}(1 - W_i)$ are the number of grid blocks absent SWDs. Because of the randomization, a natural estimator for the average causal effect is the difference in the average outcomes between those assigned to treatment and those assigned to control:

$$\hat{\tau}^{dif} = \bar{Y}_t^{obs} - \bar{Y}_c^{obs} \tag{D4}$$

where:

$$\bar{Y}_c^{obs} = \frac{1}{N_c}\sum_{i:W_i=0} Y_i^{obs} \tag{D5}$$

$$\bar{Y}_t^{obs} = \frac{1}{N_t}\sum_{i:W_i=1} Y_i^{obs} \tag{D6}$$

To prove our estimator $\hat{\tau}^{dif}$ is unbiased for τ_{fs}, we use the fact that $Y_i^{obs} = Y_i(1)$ if $W_i = 1$, and $Y_i^{obs} = Y_i(0)$ if $W_i = 0$, to rewrite the estimator $\hat{\tau}^{dif}$ as:

$$\hat{\tau}^{dif} = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{W_i \cdot Y_i(1)}{N_t/N} - \frac{(1-W_i)\cdot Y_i(0)}{N_c/N}\Big) \tag{D7}$$

The potential outcomes are treated as fixed, the only component in this statistic that is random are the treatment assignments obtained from our simulated realizations that preserve spatial correlation (Imbens & Rubin, 2015). Thus, $Pr(W_i = 1|\mathbf{Y}(0), \mathbf{Y}(1)) = \mathbb{E}_W[W_i|\mathbf{Y}(0), \mathbf{Y}(1)] = N_t/N$ and $\hat{\tau}^{dif}$ is unbiased for the average causal effect τ_{fs}:

$$\begin{aligned}\mathbb{E}_W\Big[\hat{\tau}^{dif}\Big|\mathbf{Y}(0), \mathbf{Y}(1)\Big] &= \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{\mathbb{E}_W[W_i]\cdot Y_i(1)}{N_t/N} - \frac{\mathbb{E}_W[1-W_i]\cdot Y_i(0)}{N_c/N}\Big)\\ &= \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big) = \tau_{fs}\end{aligned} \tag{D8}$$

D2 The Sampling Variance of the Neyman Estimator

We develop an estimator for the sampling variance and appeal to a central limit argument for the large sample normality of $\hat{\tau}$ over its randomization distribution and use its estimated sampling variance to create a large-sample confidence interval for the average causal effect τ_{fs}.

$$\mathbb{V}_W(\hat{\tau}^{dif}) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{ct}^2}{N} \tag{D9}$$

where

$$S_c^2 = \frac{1}{N_c - 1}\sum_{i:W_i=0}\Big(Y_i^{obs} - \bar{Y}_c^{obs}\Big)^2 \tag{D10}$$

and

$$S_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=0} \left(Y_i^{obs} - \bar{Y}_t^{obs} \right)^2 \tag{D11}$$

The third term, S_{ct}^2, is the population variance of the unit-level treatment effects and is generally impossible to estimate empirically. Recall potential outcomes $Y_i(1)$ and $Y_i(0)$ cannot be both observed. Assuming the treatment effects are constant and additive ($Y_i(1) - Y_i(0) - \tau_{fs}$ for all units), then the third term is equal to zero and we have the reduced version of an unbiased estimator for the sampling variance:

$$\hat{\mathbb{V}}^{neyman} = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} \tag{D12}$$

This estimators is widely used for two reasons. First, by implicitly setting the third term equal to zero, the expected value of the $\hat{\mathbb{V}}^{neyman}$ is at least s large as the true sampling variance equal of $\bar{Y}_t^{obs} - \bar{Y}_c^{obs}$, irrespective of the heterogeneity in the treatment effect, because the third term is non-negative. Hence, the confidence interval generated using this estimator in large sample would be greater or equal to the nominal coverage and is statistically conservative. Second, using $\hat{\mathbb{V}}^{neyman}$ as an estimator for the sampling variance of $\bar{Y}_t^{obs} - \bar{Y}_c^{obs}$ is that it is always unbiased for the sampling variance of $\hat{\tau}^{dif}$ as an estimator of the infinite super-population average causal effect. From above derived estimator for sampling variance, we construct 90% and 95% confidence intervals below, respectively (Imbens & Rubin, 2015).

$$\mathbf{CI}^{0.90}(\tau_{fs}) = \left(\hat{\tau}^{dif} - 1.645 \cdot \sqrt{\hat{\mathbb{V}}}, \hat{\tau}^{dif} + 1.645 \cdot \sqrt{\hat{\mathbb{V}}} \right) \tag{D13}$$

$$\mathbf{CI}^{0.95}(\tau_{fs}) = \left(\hat{\tau}^{dif} - 1.96 \cdot \sqrt{\hat{\mathbb{V}}}, \hat{\tau}^{dif} + 1.96 \cdot \sqrt{\hat{\mathbb{V}}} \right) \tag{D14}$$

Appendix E Extending to Studies of Distributed Volume

We highlight that the spatial point process model can be applied to previous studies to construct proper null distribution specific to each grid. More specifically, we generate a permutation of the distributed volume over the entire area of interest for every reproduction of the SWD point pattern. In particular, the permutation of the distributed volume is governed by a deterministic equation where the cumulative injection volume for all SWDs are sampled from the empirical histogram of the cumulative injection volume, shown in Figure E1. Given enough repetitions, there is an empirical distribution of distributed volume specific to each grid. We argue the resulting empirical distribution, specific to each grid, is a more appropriate null distribution than the one obtained from resampling for the following reasons:

1. Every permutation of distributed volume is simulated using the same deterministic physical model under realistic SWD allocation scheme where the cumulative injection volume for every SWD is drawn from the empirical distribution. Every permutation of distributed volume is realistic, more specifically, the distributed volume specific to each grid is realistic.
2. The null distribution of distributed volume for a particular grid block is constituted from different permutations of the distributed volume of the same grid. Because every permutation of the distributed volume is independent, every permutation of distributed volume of that grid is independent. This conforms with the *iid* sampling assumption.
3. Given enough permutation of distributed volume, those *iid* samples of distributed volume construct non-spatial and unbiased null distribution for every grid block.

4. The null distributions for all grids are converging under the Law of Large Numbers (LLN) with enough permutations (Casella & Berger, 2001). This is crucial if p-value comparisons are needed between grid blocks, since p-values are not comparable across different null distributions. Because the resampling method does not guarantee *iid* samples, it is less straightforward to apply LLN for convergence.

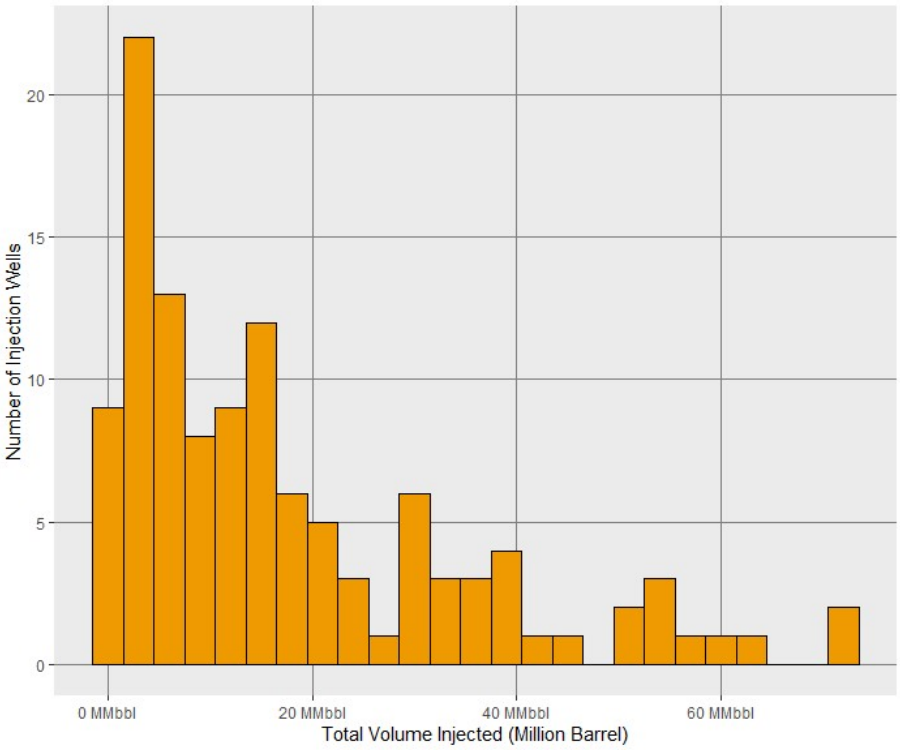

Figure E1. The empirical histogram of cumulative injection volume

Figure E2 shows the entire workflow from a random placement of SWDs to a permutation of distributed volume. We use inverse-distance method to calculate distributed volume for the DFW region. Grigoratos et al. (2020a) implemented a more appropriate physics-based diffusion model that includes time to calculate distributed volume across 12 years period. We are precluding time in our current analysis, thus the inverse-distance method is sufficient to produce modest distributed volume across some time period

Acknowledgments

Thank you to Louis Quinones and Heather DeShon for providing their databases of the Dallas Forth-Worth region. Thank you to Iason Grigoratos for helping to decluster the earthquake catalog. The financial support of the TexNet Research and the Center for Integrated Seismicity Research is gratefully acknowledged. The earthquake data reported in this paper is available from DeShon et al. (2019) and the saltwater disposal data is available from Texas Data Repository (`https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/CEQEDF`)

References

Aldrich, J. (1995, 11). Correlations genuine and spurious in pearson and yule. *Statist. Sci.*, *10*(4), 364–376. Retrieved from `https://doi.org/10.1214/ss/`

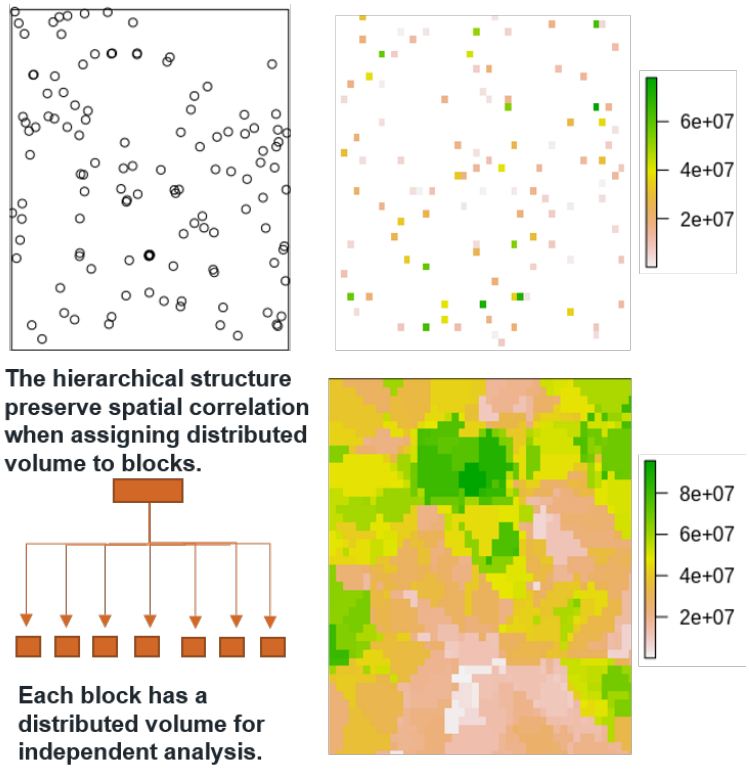

Figure E2. A SWD point pattern (top left panel) is simulated from the fitted LGCP and each SWD is assigned with a cumulative injection volume (top right panel) from the empirical histogram (Figure E2). Inverse-distance method is applied to calculate the distributed volume for the entire study area (bottom right panel). The hierarchical structure (bottom left) preserves the spatial correlation of every permutation of distributed volume at the high level where every permutation is independent. Under LLN, the empirical distribution of distributed volume specific to each grid converges to a proper null distribution.

1177009870 doi: 10.1214/ss/1177009870

Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial point patterns methodology and applications with r*. CRC Press.

Carone, M., Dominici, F., & Sheppard, L. (2020). In pursuit of evidence in air pollution epidemiology: the role of causally driven data science. *Epidemiology*, *31*(1), 1–6.

Carsey, T. M., & Harden, J. J. (2013). *Monte carlo simulation and resampling methods for social science*. Sage Publications.

Casella, G., & Berger, R. (2001). *Statistical inference*. Duxbury Resource Center. Textbook Binding.

DeShon, H. R., Hayward, C. T., Ogwari, P. O., Quinones, L., Sufri, O., Stump, B., & Beatrice Magnani, M. (2019). Summary of the North Texas Earthquake Study Seismic Networks, 2013–2018. *Seismological Research Letters*, *90*(1), 387–394. doi: 10.1785/0220180269

Dominici, F., & Zigler, C. (2017). Best practices for gauging evidence of causality in air pollution epidemiology. *American journal of epidemiology*, *186*(12), 1303–1309.

Ellsworth, W. L. (2013). Injection-induced earthquakes. *Science*, *341*(6142). Retrieved from https://science.sciencemag.org/content/341/6142/1225942

doi: 10.1126/science.1225942

Fan, Z., Eichhubl, P., & Newell, P. (2019). Basement Fault Reactivation by Fluid Injection Into Sedimentary Reservoirs: Poroelastic Effects. *Journal of Geophysical Research: Solid Earth*, *124*(7), 7354–7369. doi: 10.1029/2018JB017062

Fasola, S. L., Brudzinski, M. R., Skoumal, R. J., Langenkamp, T., Currie, B. S., & Smart, K. J. (2019). Hydraulic fracture injection strategy influences the probability of earthquakes in the eagle ford shale play of south texas. *Geophysical Research Letters*, *46*(22), 12958-12967. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085167` doi: https://doi.org/10.1029/2019GL085167

Friedrich, S., & Friede, T. (2020). Causal inference methods for small non-randomized studies: Methods and an application in covid-19. *Contemporary clinical trials*, *99*, 106213.

Frohlich, C., De Shon, H., Stump, B., Hayward, C., Hornbach, M., & Walter, J. I. (2016a). A historical review of induced Earthquakes in Texas. *Seismological Research Letters*, *87*(4), 1022–1038. doi: 10.1785/0220160016

Frohlich, C., De Shon, H., Stump, B., Hayward, C., Hornbach, M., & Walter, J. I. (2016b). A Historical Review of Induced Earthquakes in Texas. *Seismological Research Letters*, *87*(4), 1022–1038. doi: 10.1785/0220160016

Frohlich, C., Hayward, C., Rosenblit, J., Aiken, C., Hennings, P., Savvaidis, A., ... DeShon, H. R. (2020). Onset and Cause of Increased Seismic Activity Near Pecos, West Texas, United States, From Observations at the Lajitas TXAR Seismic Array. *Journal of Geophysical Research: Solid Earth*, *125*(1), 1–14. Retrieved from `https://doi.org/10.1029/2019JB017737` doi: 10.1029/2019JB017737

Gao, R., Pelletier, I., Horne, E., Nicot, J.-P., & Hennings, P. (2019). Basin-scale hydrogeological modeling of the fort worth basin ellenburger group for pore pressure characterization. In *Agu fall meeting abstracts* (Vol. 2019, pp. H53C–01).

Giffin, A., Reich, B., Yang, S., & Rappold, A. (2020). *Generalized propensity score approach to causal inference with spatial interference.*

Glass, T. A., Goodman, S. N., Hernán, M. A., & Samet, J. M. (2013). Causal inference in public health. *Annual Review of Public Health*, *34*(1), 61-75. Retrieved from `https://doi.org/10.1146/annurev-publhealth-031811-124606` (PMID: 23297653) doi: 10.1146/annurev-publhealth-031811-124606

Grigoratos, I., Rathje, E., Bazzurro, P., & Savvaidis, A. (2020a). Earthquakes Induced by Wastewater Injection, Part II: Statistical Evaluation of Causal Factors and Seismicity Rate Forecasting. *Bulletin of the Seismological Society of America*, *110*(5), 2483–2497. doi: 10.1785/0120200079

Grigoratos, I., Rathje, E., Bazzurro, P., & Savvaidis, A. (2020b). Earthquakes Induced by Wastewater Injection, Part I: Model Development and Hindcasting. *Bulletin of the Seismological Society of America*, *110*(5), 2466–2482. doi: 10.1785/0120200078

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, *15*(3), 965 – 1056. Retrieved from `https://doi.org/10.1214/19-BA1195` doi: 10.1214/19-BA1195

Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, *47*(2), 757–762. Retrieved from `http://www.jstor.org/stable/2532163`

Hennings, P. H., Snee, J. E. L., Osmond, J. L., Deshon, H. R., Dommisse, R., Horne, E., ... Zoback, M. D. (2019). Injection-Induced Seismicity and Fault-Slip Potential in the Fort Worth Basin, Texas. *Bulletin of the Seismological Society of America*, *109*(5), 1615–1634. doi: 10.1785/0120190017

Hincks, T., Aspinall, W., Cooke, R., & Gernon, T. (2018). Oklahoma's induced seis-

micity strongly linked to wastewater injection depth. *Science*, *1255*(March), 1251–1255.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. Retrieved from http://www.jstor.org/stable/2289064

Hornbach, M. J., Deshon, H. R., Ellsworth, W. L., Stump, B. W., Hayward, C., Frohlich, C., ... Luetgert, J. H. (2015). Causal Factors for Seismicity near Azle, Texas. *Nature Communications*, *6*, 1–11. Retrieved from http://dx.doi.org/10.1038/ncomms7728 doi: 10.1038/ncomms7728

Hornbach, M. J., Jones, M., Scales, M., DeShon, H. R., Magnani, M. B., Frohlich, C., ... Layton, M. (2016). Ellenburger wastewater injection and seismicity in North Texas. *Physics of the Earth and Planetary Interiors*, *261*, 54–68. Retrieved from http://dx.doi.org/10.1016/j.pepi.2016.06.012 doi: 10.1016/j.pepi.2016.06.012

Horne, E., Hennings, P., Osmond, J., & Deshon, H. (2020). Structural characterization of potentially seismogenic faults in the Fort Worth Basin. *j*, *8*(2). doi: 10.1190/INT-2019-0188.1

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). (Vol. Vol. 70). John Wiley & Sons.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction.* Cambridge University Press. doi: 10.1017/CBO9781139025751

Justinic, A. H., Stump, B., Hayward, C., & Frohlich, C. (2013). Analysis of the Cleburne, Texas, Earthquake Sequence from June 2009 to June 2010. *Bulletin of the Seismological Society of America*, *103*(6), 3083–3093. doi: 10.1785/0120120336

Keranen, K. M., Savage, H. M., Abers, G. A., & Cochran, E. S. (2013). Potentially induced earthquakes in oklahoma, usa: Links between wastewater injection and the 2011 mw 5.7 earthquake sequence. *Geology*, *41*(6), 699–702.

Keranen, K. M., Weingarten, M., Abers, G. A., Bekins, B. A., & Ge, S. (2014). Sharp increase in central oklahoma seismicity since 2008 induced by massive wastewater injection. *Science*, *345*(6195), 448–451. Retrieved from https://science.sciencemag.org/content/345/6195/448 doi: 10.1126/science.1255802

Langenbruch, C., & Zoback, M. D. (2016). How will induced seismicity in oklahoma respond to decreased saltwater injection rates? *Science Advances*, *2*(11). doi: 10.1126/sciadv.1601542

Langenbruch, C., & Zoback, M. D. (2017). Response to Comment on "How will induced seismicity in Oklahoma respond to decreased saltwater injection rates?". *Science Advances*, *3*(8), 1–10. doi: 10.1126/sciadv.aao2277

Lund Snee, J.-E., & Zoback, M. D. (2016). State of stress in texas: Implications for induced seismicity. *Geophysical Research Letters*, *43*(19), 10,208-10,214. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL070974 doi: https://doi.org/10.1002/2016GL070974

Marrett, R., Gale, J. F., Gómez, L. A., & Laubach, S. E. (2018). Correlation Analysis of Fracture Arrangement in Space. *Journal of Structural Geology*, *108*, 16–33. doi: 10.1016/j.jsg.2017.06.012

McClure, M., Gibson, R., Chiu, K.-K., & Ranganath, R. (2017). Identifying potentially induced seismicity and assessing statistical significance in oklahoma and california. *Journal of Geophysical Research: Solid Earth*, *122*(3), 2153–2172.

Mehta, C. R., & Patel, N. R. (1983). A network algorithm for performing fisher's exact test in r × c contingency tables. *Journal of the American Statistical Association*, *78*(382), 427-434. Retrieved from https://doi.org/10.1080/01621459.1983.10477989 doi: 10.1080/01621459.1983.10477989

Ogwari, P. O., DeShon, H. R., & Hornbach, M. J. (2018). The dallas-fort worth

airport earthquake sequence: Seismicity beyond injection period. *Journal of Geophysical Research: Solid Earth*, *123*(1), 553-563. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JB015003 doi: https://doi.org/10.1002/2017JB015003

Panzeri, S., Magri, C., & Carraro, L. (2008). Sampling bias. *Scholarpedia*, *3*(9), 4258. (revision #148550) doi: 10.4249/scholarpedia.4258

Papadogeorgou, G., Mealli, F., & Zigler, C. M. (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, *75*(3), 778–787.

Quinones, L., Deshon, H. R., Jeong, S., Ogwari, P., Sufri, O., Holt, M. M., & Kwong, K. B. (2019). Tracking Induced Seismicity in the Fort Worth Basin: A Summary of the 2008–2018 North Texas Earthquake Study Catalog. *Bulletin of the Seismological Society of America*, *109*(4), 1203–1216. doi: 10.1785/0120190057

Quinones, L. A., DeShon, H. R., Magnani, M. B., & Frohlich, C. (2018, 04). Stress orientations in the fort worth basin, texas, determined from earthquake focal mechanisms. *Bulletin of the Seismological Society of America*, *108*(3A), 1124-1132. Retrieved from https://doi.org/10.1785/0120170337 doi: 10.1785/0120170337

Reich, B., Yang, S., Guan, Y., Giffin, A., Miller, M. J., & Rappold, A. (2020). A review of spatial causal inference methods for environmental and epidemiological applications. *arXiv: Methodology*.

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

Savvaidis, A., Young, B., Huang, G.-c. D., & Lomax, A. (2019). Texnet: A statewide seismological network in texas. *Seismological Research Letters*, *90*(4), 1702–1715.

Scales, M. M., DeShon, H. R., Magnani, M. B., Walter, J. I., Quinones, L., Pratt, T. L., & Hornbach, M. J. (2017). A Decade of Induced Slip on the Causative Fault of the 2015 Mw 4.0 Venus Earthquake, Northeast Johnson County, Texas. *Journal of Geophysical Research: Solid Earth*, *122*(10), 7879–7894. doi: 10.1002/2017JB014460

Schoenball, M., & Ellsworth, W. L. (2017). A systematic assessment of the spatiotemporal evolution of fault activation through induced seismicity in oklahoma and southern kansas. *Journal of Geophysical Research: Solid Earth*, *122*(12), 10,189-10,206. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JB014850 doi: https://doi.org/10.1002/2017JB014850

Walsh, F. R., & Zoback, M. D. (2015). Oklahoma's recent earthquakes and saltwater disposal. *Science Advances*, *1*(5), 1–10. doi: 10.1126/sciadv.1500195

Weingarten, M., Ge, S., Godt, J. W., Bekins, B. A., & Rubinstein, J. L. (2015). High-rate injection is associated with the increase in U.S. mid-continent seismicity. *Science*, *348*(6241), 1336–1340. doi: 10.1126/science.aab1345

Zhai, G., & Shirzaei, M. (2018). Fluid Injection and Time-Dependent Seismic Hazard in the Barnett Shale, Texas. *Geophysical Research Letters*, *45*(10), 4743–4753. doi: 10.1029/2018GL077696

Zigler, C., Forastiere, L., & Mealli, F. (2020). *Bipartite interference and air pollution transport: Estimating health effects of power plant interventions.*

Zigler, C. M., Choirat, C., & Dominici, F. (2018). Impact of national ambient air quality standards nonattainment designations on particulate pollution and health. *Epidemiology (Cambridge, Mass.)*, *29*(2), 165.

Zigler, C. M., & Dominici, F. (2014). Point: clarifying policy evidence with potential-outcomes thinking—beyond exposure-response estimation in air pollution epidemiology. *American journal of epidemiology*, *180*(12), 1133–1140.

Zigler, C. M., & Papadogeorgou, G. (2021). Bipartite Causal Inference with Interference. *Statistical Science*, *36*(1), 109 – 123. Retrieved from `https://doi.org/10.1214/19-STS749` doi: 10.1214/19-STS749