FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Social Data Science

Josefine Bjørnholm and Anna Beck Thelin

Clustering of Mass Shootings in the United States of America

Supervisor: Andreas Bjerre-Nielsen

ECTS points: 7.5

Date of submission: 25/01/2020

Keystrokes: 55,820

Contributions
Common sections: 1, 7 and 8
Anna Beck Thelin: First half of literature review, 3, 3.1.1, 3.2, 3.4, 4, 5.1, 5.1.2, 6, 6.1
Josefine Bjørnholm: Second half of literature review, 3.1, 3.1.2, 3.3, 3.5, 5, 5.1.1, 5.2, 6.2, 6.3

Contents

1 Introduction **3**

2 Literature Review **5**

3 Methodology **7**

3.1 Kernel Density Clustering 7

3.1.1 Clustering across time 8

3.1.2 Spatial clustering 8

3.2 Nearest Neighbour 9

3.3 G-function 9

3.4 K-function 10

3.5 Autoregression (AR) Model 11

4 Data **11**

5 Descriptive Statistics **12**

5.1 Gun violence incidents and mass shootings 12

5.1.1 Incidents Across States 12

5.1.2 Incidents Through Time 14

5.2 Mass Shooting Characteristics 15

6 Clustering Analysis of Mass Shootings **17**

6.1 Spatial Clustering 17

6.2 Time Clustering 22

7 Discussion **26**

8 Conclusion **29**

References **31**

1 Introduction

In the United States an average of 344 mass shootings have occurred every year from 2014-2018 equivalent to almost a mass shooting a day [1]. These includes mass shootings gaining international attention such as the Pulse nightclub incident in Orlando (2016), the Santa Fe high school shooting (2018) and more recent the 2019 El Paso shooting in Texas where 20 people were killed by a right-wing extremist. This has resulted in a heated debate about gun laws with the United States gun lobby on one side and the citizens movements on the other. Despite the increased awareness on the issue there has been no signs of a decrease in mass shootings incidents in recent years.

When comparing the United States to other high income countries it is revealed that an overwhelming 84 percent of gun related deaths are located here (Grinshteyn and Hemenway, 2019). In the literature this striking results is primarily explained by the access to weapons (Reeping et al., 2019, Grinshteyn and Hemenway, 2019). This has resulted in citizens movements such as Everytown[2] and March for Our Lives[3] succeeding in implementing more restrictive gun laws in a number of states. In opposition to this movement stand the gun lobby[4] in the United States as well as pro gun politicians who fight to protect the Second Amendment of the United States Constitution. They claim that *"the people best positioned to stop mass shootings are those who also have guns"* which has in fact resulted in more permissive gun laws in other states [5]. Not surprisingly they argue that the problem is not the availability of guns but rather the state of the perpetrator. More specifically they point towards mental illness as well as the influence of violent computer games as possible explanations[6]. There are currently no studies that support these claim. However there is evidence that point towards factors such as alcohol, drug use, childhood abuse and male gender as valid explanations for mass shootings (Metzl and MacLeish, 2015).

Besides the exogenous factors mentioned above that contribute to the frequency of mass

[1] https://www.gunviolencearchive.org/past-tolls

[2] https://.everytown.org

[3] https://marchforourlives.com/

[4] A main actor in the gun lobby is the National Riffle Association (NRA), institute for legislative action http://home.nra.org

[5] https://newyorker.com/news/dispatch/conversations-about-mass-shootings-at-an-nra-expo-in-texas

[6] https://edition.cnn.com/2019/08/05/politics/violent-video-game-shooting-fact-check

shootings, the possibility that the mass shooter have been inspired, more or less consciously, by other violent events, including mass shootings, could also serve as an explanation. Studies have suggested that mass shootings attracting extensive media coverage, have given inspiration to new incidents replicating both time and place. This could lead to spatial and time clustering of mass shootings in the United States, which is the objective of this paper.

When investigating the spatial clustering we find that mass shootings appear significantly closer to each other than would have been expected had they been randomly distributed. Based on our finding, spatial clustering appears in California, North East Central[7] and Mid-Atlantic[8]. These three areas will serve as study areas through most of the analysis. More specifically we find that mass shootings in these areas cluster at small distances, suggesting that one mass shooting will happen within approximately 10 minute walking distance from another. As all three study areas includes some of the biggest cities in the United States with structural challenges such as poverty, high crime rates and racial polarization, this clustering pattern is not surprising. However it is noticeable that mass shooting reappear within such small distances indicating that mass shooters target smaller venues with a certain amount of people as well as easy access with weapons.

Further we find that mass shootings also show signs of time clustering. Here we both find evidence of seasonality in mass shootings i.e. that most incidents occur during the summer periods as well as a time dynamic effects of mass shootings. This supports, to some extend, the hypothesis that some mass shooters act in tribute to or inspired by previous incidents. This could also be explained by gang violence, where one incident may cause a revenge action resulting in another mass shooting in the near future. We find a significant time dynamic effect, suggesting that a mass shooting in one week will increase the probability of another mass shooting in the preceding 3-5 weeks.

From this analysis we are not able to determine the cause of mass shootings or why the clustering patterns across time and space emerge. However our results do confirm, in line with existing literature, that such patterns exist.

[7] Illinois, Indiana, Michigan, Ohio, Wisconsin

[8] Delaware, Maryland, New Jersey, New York, Pennsylvania, Virginia, Washington D.C, West Virginia

Taking this into account, the results of this paper suggest policy interventions aiming to increase security at specific venues where mass shootings has already taken place. Further the seasonality and time dynamic could suggest that an increase in awareness of the effects of massive media coverage could reduce the number of mass shootings. Finally reduction of gang violence will naturally also decrease the number of mass shooting incidents.

2 Literature Review

In this section we provide a brief overview of the existing literature investing mass shootings and the causes hereof. A number of studies have sought to explain the high number of mass shootings in the United States. Explanations such as mental illness, domestic violence and hated-fueled ideologies have been proposed by one branch of literature while others have focused more time and geographical dependence.

The media attention to mass shootings by individuals with mental illness has increase in recent years. McGinty et al. (2014) find that media coverage of mass shootings has a negative effect on public attitude towards people with mental illness. This has given rise to the notion that mental illness cause gun violence.
This notion is challenged by Metzl and MacLeish (2015) who argue that the hostile attitude towards people with mental illnesses reflect cultural stereotypes and anxieties about matters such a race/ethnicity, social class and politics. They further argue that the cause of mass shootings can instead be explained by an interplay of numerous factors such a alcohol, drug-use, childhood abuse and male gender.

A common belief about mass shootings is that they occur in public places and appears to be an act of random violence. However, studies suggest that mass shootings are in fact often related to domestic violence. A report by Gun Safety Support Fund (2019) finds that most mass shootings (61 percent) happen in private homes, but that public mass shootings are more deadly. While public mass shootings only account for 29 percent of all mass shootings, they resulted in more than half of the mass shooting deaths. They also find that in at least 54 percent of the mass shootings incidents, the perpetrator shot a current or former intimate partner or family member.

The United States is not the only country battling domestic violence and mental illness. However when comparing United States gun violence data with other high-income countries striking results emerge. The gun homicide rate is 25 times higher in the United States then in other high-income countries, and of all the gun related deaths in these countries, 83.7 percent of them happened in the United States (Grinshteyn and Hemenway, 2019). Thus, a new and compelling explanation for the high number of mass shootings and the general high level of gun violence is the mere access to weapons. Reeping et al. (2019) seek to determine if restrictiveness/permissiveness of state gun laws or gun ownership are associated with mass shootings in the United States. They find a 10 unit[9] increase in state gun law permissiveness and a 10 percent increase in state gun ownership is associated with a 12 and 35 percent higher rate of mass shootings respectively. Restricting the analysis to domestic and non-domestic mass shooting lead to similar results.

When considering the locations of mass shootings across the United States, signs of a non-random spatial patterns appear. Furthermore mass shootings also seem to emerge at non-random points in time. This has given rise to a different branch of studies investigating the spatial and time dependence of mass shootings.
One of these is the study by Barboza (2018) who investigates spatial concentration of gun violence around schools in Boston as an explanation for school shootings. He finds that up to six times as many shootings occur within 400 m of school than would be expected if shootings where spatially random. Next, Blum and Jaworski (2017) investigate spatial patterns of mass shootings in the US from 2013-2014. Carrying out point pattern analysis, they find that locations of mass shootings are not random, and are more likely to occur in areas of South and Upper Midwest United States and in Southern California. This finding suggests that structural factors that may contribute to mass shootings are more prevalent in some areas of the United States than in others.

Turning the attention towards time dependence of mass shootings Towers et al. (2015) find that mass shootings are incentivized by similar event in the immediate past. Specifically

[9] An annual rating between 0 (completely restrictive) and 100 (completely permissive for the gun laws of each state

they find that a mass shooting temporarily increases the probability of another mass shooting in the 13 preceding days, and that each incident leads to, on average, 0.3 new incidents. They argue that this trigger-effect is caused by the media coverage which inspires at-risk individuals to carry out similar acts. In line with this results, mass shooting events may also happen as means of tribute to past violent events. Murray (2017) even argues that mass shootings such as the Columbine High School incident (April 20th. 1999), the Virginia Tech incident (April 16th. 2007), the Aurora Movie incident (July 20th. 2012) and the incident at Pulse nightclub in Orlando (June 12th. 2016), that all attracted massive media attention, not only gave inspiration to replicate them, but even out-doing them.
Finally another evident time pattern of mass shooting and violence in general, is that of seasonality which is also explained by tribute killings. For instance, this could explain why both Virginia Tech and Columbine happened at schools during April as the former is inspired by the latter. Lastly Rotton and Cohn (2004) explain the seasonality of violence by the effect of climate on mental health, arguing that higher temperature can increase aggression.

3 Methodology

In order to determine whether there exists clustering patterns between mass shootings, both from a geographical perspective and across time, we use a range of different approaches that enable us to display and test the patterns in the data. We both consider first- and second order properties of the patterns, that we discover in the mass shooting data.

3.1 Kernel Density Clustering

Kernel density estimation is used in a non- and semi-parametric framework. It provides us with a probability estimate, and can be used when working with data that does not fit a normal distribution. In this case we use it to investigate clustering across time resulting in an one dimensional kernel density estimation.
In a point pattern analysis kernel density clustering is a first order property and is concerned with the variation of the observations' density within a certain study area (Gimond, 2019). Throughout our analysis our study area will change, as we zoom in on locations based on the discoveries we do.

3.1.1 Clustering across time

When investigating the clustering patterns across time. We use kernel density estimation to determine whether there are clustering patterns across time, i.e if there exists points with an increased probability of mass shootings. This can be expressed as:

$$\hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)$$

Here h is a scalar input, $K(\cdot)$ is a kernel function, and x_0 is a scalar input we loop through to get a continuous density plot. For Kernel density estimation we thus need to specify both a kernel function, $K(u)$, and a bandwidth, h, also known as the smoothing parameter. We use a Gaussian Kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$, where the weights in the model follow a standard normal distribution implying that we put more weight on observations close to x_0 (Cameron and Trivedi, 2005).

In kernel density estimation the bandwidth is the specification, that has the largest impact on the results. The bandwidth balances a trade-off between bias and the size of the variance in the model. A small bandwidth results in little smoothing and thereby more frequent variation in our density plot. This lead to a small bias and large variance. The opposite it true for a large bandwidth. Theoretically the optimal bandwidth minimizes the mean integrated squared errors. However, this requires the second derivative of the true density function of the data, which is per definition unknown.

In practice we apply other methods such as Silverman's plug-inn estimate:

$$\hat{h}^*_{Silverman} = 1.3643\delta N^{-0.2} \min(s, iqr/1.349)$$

Here δ is a constant related to the choice of kernel [10], s is the standard deviation and iqr is the interquartile range. When applying kernel density estimation to a low dimensional problem, it is common practice to compare the plug-in estimate with a manually chosen h to see if a better fit of h exists (Maire, 2017).

3.1.2 Spatial clustering

In a point pattern analysis framework, the kernel density computes a localized density for subsets of the study areas. These subsets can move and overlap, and are defined by the

[10] $\delta_{Gaussian} = 0.0885$

bandwidth of the kernel function. Within each subset, a grid is generated and the density of each of these grid cells is calculated on the basis of the kernel centered on that cell (Gimond, 2019).

In our analysis we use a quartic kernel: $K(u) = \frac{15}{16}(1-u^2)^2$ to get a smooth density map. This kernel assigns more weight to points that are inversely proportional to the kernel subsets' center. Here, we also use Silvermann's plug in estimation to determine the optimal level of smoothing, as a proxy for the bandwidth that minimizes integrated mean squared errors.

3.2 Nearest Neighbour

The nearest neighbour analysis is a second order property of the point pattern analysis. This type of distance based analysis measures the distance between some initial point and its nearest neighbour. Results are often summarized by the mean[11]. This can give us an initial idea of the clustering (Gimond, 2019).

Depending on the data input (points or coordinates) we need to specify how we want the distance to be calculated. In this analysis we have coordinates as inputs and therefore use an approach that takes the curvature of earth into consideration, as the distance will otherwise be underestimated. This is obtained by the Haversine formula expressed by:

$$d = 2r\sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_2-\phi 1}{2}\right)+\cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

where r is the radius of the sphere[12], (ϕ_1, λ_1) is the latitude and longitude of the first point and (ϕ_2, λ_2) is the latitude and longitude of the second point. The coordinates of both points should be in radian (Pedregosa et al., 2011).

3.3 G-function

From the average nearest neighbour analysis we cannot determine whether spatial clustering appear with any significance. However, by using a nearest neighbour distance function we can get closer to this type of result. Nearest neighbour distance functions enable us to compare the distribution function from the point pattern process in our data with

[11] Known as the Average Nearest Neighbor Analysis (ANN).

[12] In our case r is the radius of earth (6,371 km)

the point pattern process of complete spatial randomness (CSR) - in this case a Poisson process. The G-function measures the cumulative distribution of points as a function of the distance between each point and its neighbor; i.e. for a given d, $G(d)$ measures the proportion of nearest neighbor distances that are less than d:

$$G(d) = \sum_{i=1}^{n} \frac{\phi_i^d}{n} \quad , \quad \phi_i^d = \begin{cases} 1 & \text{if } d_{min}(s_i) < d \\ 0 & \text{otherwise} \end{cases}$$

Where s_i is the points of events, $i = 1, .., n$. For a CSR process following a Poisson distribution the G-function has an expectation of: $G(d)_{CSR} = 1 - exp(\lambda \pi d^2)$. This means that we can conclude significant clustering whenever the G-function is above this expected value (Serge and Kang, 2019).

3.4 K-function

To further investigate the point pattern we conduct a K-function analysis. The K-function is an interevent function accounting for higher order neighbours within a given radius r. Like the G-function the K-function is a cumulative distribution function, that can be described as follows:

$$K(r) = \frac{\sum_{i=1}^{n} \sum_{j}^{n} \psi_{ij}(r)}{n\hat{\lambda}} \quad , \quad \psi_{ij} = \begin{cases} 1 & \text{if } r_{ij} < d \\ 0 & \text{otherwise} \end{cases}$$

Here $\sum_j^n \psi_{ij}(r)$ is the number of events within a circle of radius r centered on the event s_i and $\hat{\lambda}$ is the overall point density. For a CSR process the K-function would be expected to be $K(r)_{CSR} = \pi r^2$ implying that for $K(r) < \pi r^2$ the underlying point process is a regular point process and for any $K(r) > \pi r^2$ the underlying process is clustered.

The K-function and the density estimations have a common problem, namely the bias along the edges of the study area. As we cannot obtain information or data points beyond the edge of the study area estimates near the edge will be biased. This can be accounted for in various ways; i) by simulating the outer points as a mirror of the points within the area at the edges, ii) by reducing the area and using the resulting outer points as part of the estimate but not part of the study areas or, iii) by using Ripleys edge correction formula. The implementation of these methods have proven somewhat comprehensive in

the Python framework. Further, the non-quadratic shapes of the different states we investigate proposes an additional problem as most of the packages will make a quadratic fit of the shapes (ArcMap, 2019). Hence, we do not implement edge-bias correction in our analysis.

3.5 Autoregression (AR) Model

As we expect mass shootings to have dynamic time effects we estimate an autoregressive (AR) model, to determine whether a mass shooting can increase the probability of another mass shooting in the following period. The AR model with p lags is defined by:

$$\text{MS}_t = \beta_0 + \beta_1 \text{MS}_{t-1} + ... + \beta_p \text{MS}_{t-p} + \epsilon_t, \qquad t = 1, .., T \quad \text{and} \quad \epsilon_t \sim IID(0, \sigma^2)$$

Here MS indicate a mass shooting at time t. In our analysis the time t indicates the duration of a week, as we assume a one-week implementation period, before another mass shooting occur. Thus MS_t is the number of mass shootings during week t as a function mass shootings in the previous $t - p$ weeks.

4 Data

The dataset used in this analysis is found on Kaggle.com, and is constructed by the user James Ko, [13] who has scraped the data from Gun Violence Archive (GVA)[14]. GVA was formed in 2013 and has since gathered data on gun violence throughout all of the United States, from 6,500 law enforcement, media, government and commercial sources daily.

The dataset consists of 239,677 incidents from January 2013 to March 2018. Each incident is described by a number of variables including date, location (state, city/county and coordinates), subject- and victim information and a number of incident characteristics. We restrict the data to only include incidents from January 2014 and on wards, as the 2013 data is defective. Furthermore we only consider the mainland of the United States and thereby exclude Alaska and Hawaii in order to maintain a relatively homogeneous sample[15]. The final dataset consists of 226,820 incidents of which 1,352 are mass shootings.

[13] `https://www.kaggle.com/jameslko/gun-violence-data`

[14] `https://www.gunviolencearchive.org`

[15] There are no mass shootings in either of the two stats during the period 2014-2018.

We are aware that using a dataset not directly provided by GVA may induce some pitfalls. We cannot be sure that the scraping method used by James Ko provides a correct dataset and some incident may be false or missing. However, having to obtain the dataset directly from GVA would be a potential dead-end, time consuming and comprehensive. Time and effort which we could otherwise use on the data analysis. Thus, we assume that the data is, on average, representative of the true gun violence data provided by GVA.

5 Descriptive Statistics

This sections contains descriptive statistics of mass shootings in the United States. First, we provide a map to get an initial idea of the distribution of gun violence incidents as well as mass shooting incidents across the United States. Secondly we consider the development of mass shootings and gun violence in general during the study period to detect potential time patterns. Finally we zoom in on mass shootings and provide some of the most frequent characteristics.

5.1 Gun violence incidents and mass shootings

Mass shooting are characterised as being incident with four or more victims killed or injured. Most mass shooting have between 4-6 victims, but some incidents involve up to a 103 [16]. In our analysis the definition of mass shootings includes gang-related mass shooting incidents. Some studies choose to exclude this type of incidents from their definition of mass shootings, as they are argued to have fundamentally different motivations. We choose to include them as we cannot reject the possibility that gang-related mass shootings exhibit the same type of spatial and time dependent properties as non gang-related mass shootings.

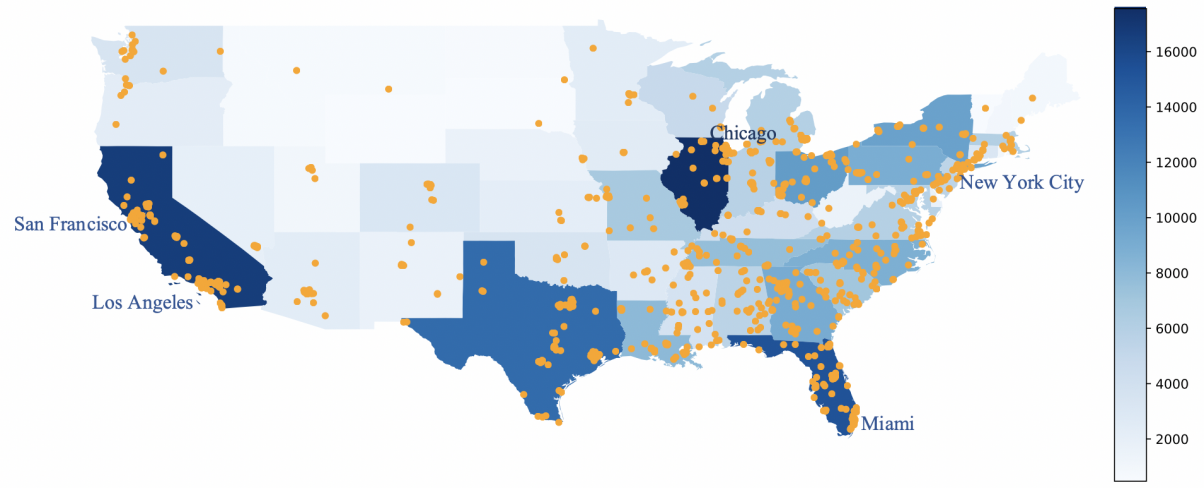
5.1.1 Incidents Across States

To get an initial idea of the spatial clustering pattern we have constructed a map of the United States that depict the number of mass shootings (orange dots) along with the amount

[16] June 12, 2016. Mass shooting incident at a bar in Orlando. `https://www.gunviolencearchive.org/incident/577157`

of gun violence. This is illustrated in figure 1. Mass shootings have occurred in almost all states (44/50) during the period. However, most mass shooting incidents are concentrated in California as well as states east of Texas. Especially, mass shootings appear to happen more frequently near bigger cities such as San Francisco, Los Angeles, Chicago, New York City and to some extend Miami. This is not surprising as these are cities with high population density, crime rate, poverty rate and racial heterogenityṠtructural factor that all contribute to an environment where mass shootings are more likely to occur (Blum and Jaworski, 2017). Further these cities host various gangs which also increase both spatial and time clustering.

Figure 1: Shootings



Note: Data consist of 226,820 gun violence incidents and 1,352 mass shootings from January 1st 2014 to March 31st 2018.

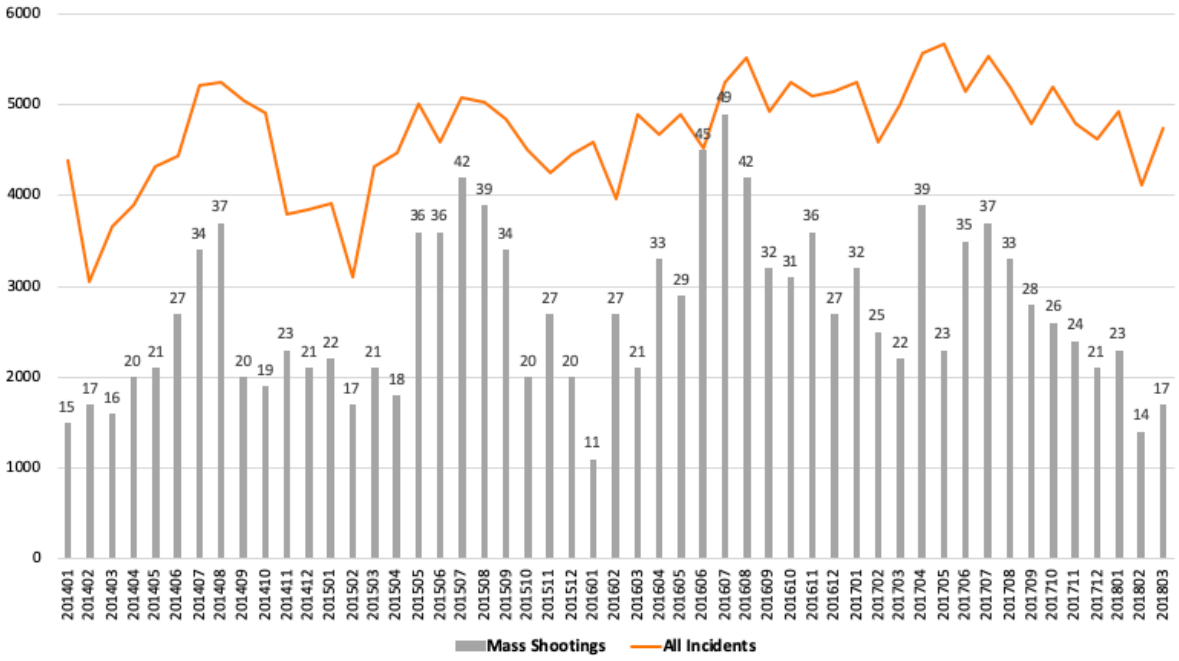
The states that have the highest number of incidents overall are Illinois, California, Florida and Texas. These four states are also those with the highest number of mass shootings during the period with 133, 155, 102 and 84 shootings respectively (table 3). At the opposite end of the scale are Montana, Wyoming, Idaho and Maine that all have few gun violence incidents as well as few mass shooting incidents (1-4) during the period. Hence, this could suggest that there is a correlation between the general level of gun violence and mass shootings across states. This correlation may simple be a matter of overlapping definitions between mass shootings and other types of gun violence, but could also be explained by violence leading to more extreme violence. It this is beyond the scope of this paper to in-

vestigate this relationship in further detail.

5.1.2 Incidents Through Time

As literature suggest time dependency in mass shootings and general gun violence, the total number of incidents and mass shootings from January 2014 to March 2018 are depicted in figure 2. For both graphs there seems to be seasonal variation, as the number of incidents within a month is highest during summer season and lowest during winter.

Figure 2: Mass shootings and total number of incidents by month



Note: Data consist of 226,820 gun violence incidents and 1,352 mass shootings from January 1st. 2014 to March 31st. 2018. Total number of incidents per month (orange line plot) are depicted on the y-axis. Total number of mass shootings per month (grey bar plot) are depicted above the bars.

As mentioned in section 2 seasonal variation in mass shootings and gun violence is a known phenomenon, and April marks what some call the beginning of the "killing-season" (Rosenwald, 2016). There are several possible explanations for this finding. Firstly, researchers argue that some subjects wish to pay tribute to other violent incident which leads to time clustering (Towers et al., 2015, Murray, 2017). Another explanation is the effect of climate on mental health. Among others Rotton and Cohn (2004) finds that moderately higher temperatures lead to an increase in aggression and thereby violence.

5.2 Mass Shooting Characteristics

In table 1 the 10 most common characteristics related to mass shootings are shown. Most incidents takes place in bars or club constituting 17 percent of the total number of mass shootings. This result is in line with the findings by Metzl and MacLeish (2015) who investigate what factors that have an effect on mass-shootings. They find that alcohol intake, among other factors, is highly correlated with mass shooting. As house parties are also closely related to alcohol consumption, as total of 21 percent of the incidents are supposedly alcohol-related. Next, drive-by's take up 17 percent of the mass shootings, followed by mass shooting with gang involvement which take up 12 percent. As it is reasonable to assume that drive-by's could also be related to gang involvement, a considerable amount of mass shootings seem to be gang-related.

Following this, domestic violence takes up 7 percent of the total number of mass shootings respectively. This is a considerably smaller number compared to the results of Gun Safety Support Fund (2019), who finds that in 54 percent of mass shootings, the perpetrator shot an intimate partner or family member. This under-representation of domestic violence incidence can most likely be explained by the definition of domestic violence, in the sense that it might no include incidence happening outside home, but still involving intimate partners or family members. Alternatively incidents may not be registered correctly as domestic violence in the database.

Table 1: Most frequent incident characteristics of mass shootings

	# Mass Shootings	**pct. of Total Mass Shootings**
Bar/club incident	228	17 %
Drive-by	213	16 %
Gang involvement	157	12 %
Domestic violence	93	7 %
Institution/Group/Business	87	6 %
Mass murder	80	6 %
Possession of gun by felon or prohibited person	79	6 %
Officer involved incident	70	5 %
Suicide	58	4 %
House party	57	4 %

Note: The incidents are described by 103 different characteristics, of which the 10 common are depicted in the table. All percentages a calculated relative to the total of 1,352 mass shooting incidents during the period January 1st 2014 till Match 31st 2018

Finally we found in section 2 that easier access to weapons is associated with mass shooting incidents (Reeping et al., 2019). We also find some evidence in this regard in table 1 as 6 percent of the mass shooting incidents involves felons or prohibited persons in possession of a gun.

In table 2 we consider the gender and status of those involved in mass shootings. First we notice that most subjects/suspects are men (89 percent). Metzl and MacLeish (2015) likewise found that there is a high correlation between being male and gun violence which seems to be apparent when considering mass shootings as well.

Table 2: Gender and status of suspects and victims of mass shootings

	Number of people
Subjects/Suspects	**1,210**
Female suspects	32 (3 pct)
Male suspects	1,073 (89 pct)
Injured suspects	25
Killed suspects	92
Unharmed/arrested suspects	550
Victims	**6,788**
Female victims	1,479 (22 pct)
Male victims	3,834 (56 pct)
Injured victim	5328
Killed victim	1426
Unharmed victim	3

Note: As not all the participants in the dataset have been assigned a gender, the male/female percentage distributions do not add up. The number of suspects is also smaller than number of incidents as not all suspects are caught.

Additionally 56 percent of the victims are men while 21 percent are female, the remaining 23 percent have not been assigned any gender in the data. Lastly we note that victims of mass shootings are more often injured than killed. The statistics of people involved in mass shootings are difficult to obtain, due to a number of reasons including confidentiality, on going investigation and so forth. The data at hand thus do not provide us with a very clear picture of the neither the victims nor the perpetrators.

6 Clustering Analysis of Mass Shootings

In the clustering analysis of mass shootings in the United States, we first consider a spatial- and then time clustering. In the spatial analysis we conduct an average nearest neighbour analysis in order to determine the mean distance between a mass shootings and its nearest neighbour. Secondly we estimate the spatial kernel density of mass shootings, in order to determine if and where mass shootings seem to cluster. Based on these findings, we choose our study areas of interest for which we compute the G- and K-functions that enable us to investigate the significance of the clustering within the study areas. In the time clustering analysis, we construct mass shooting timelines for each of the study areas and estimate the kernel densities across the time period. Finally we estimate AR models to detect any time dynamics in mass shootings.

6.1 Spatial Clustering

For each of the 1,352 mass shootings we carry out an nearest neighbor analysis measuring the distance from one mass shooting to another. These results are grouped by state and shown in table 3. The mean thus measures the average nearest neighbor distance within a state but because the distance is calculated for the entire area of United States the nearest neighbor is not restricted to be within state. This way we do not obtain edge effects between states. Further the range indicates the lowest and highest nearest neighbor distances within a state.

There is no clear patterns in terms of number of mass shootings and mean distance to nearest neighbour. For instance California and Illinois both have a high number of mass shootings, but the mean distance of California is twice the size of Illinois. It is however worth noticing that in most cases the mean distances are relatively low, compared to the upper bound of the range. This suggests that the nearest neighbour values are somewhat centered at the lower end of the interval, and could serve as evidence for spatial clustering between mass shootings.

Table 3: Nearest Neighbour Analysis

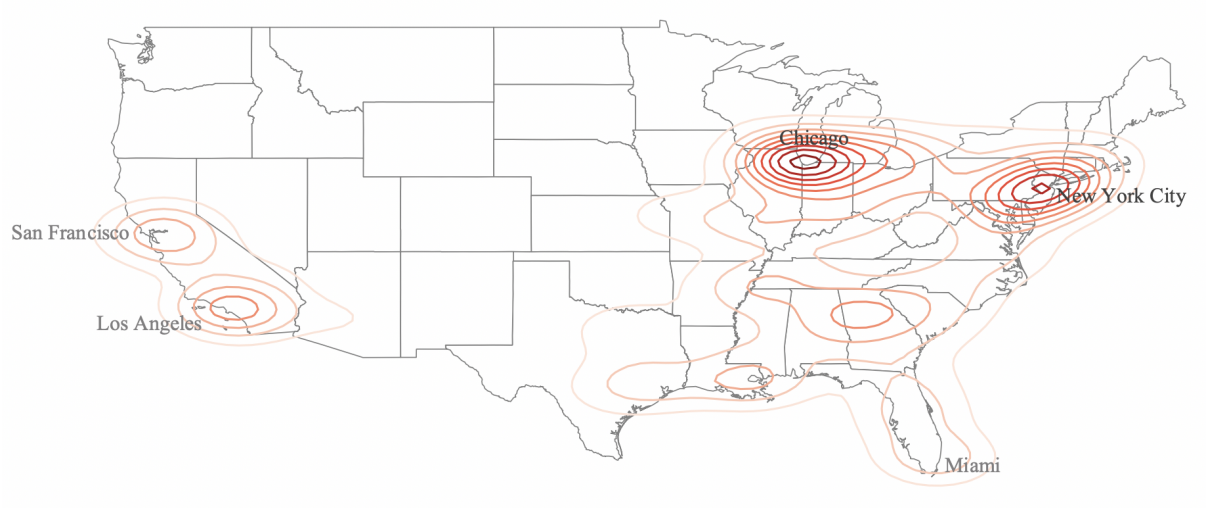
State	# Mass Shootings	Distance		State	# Mass Shootings	Distance	
		Mean	Range			Mean	Range
California	155	13.46	0.47-299.32	Minnesota	15	28.20	0.3-334.03
Illinois	133	6.96	0.22-112.31	Washington	15	60.67	1.27-360.15
Florida	102	16.12	0.28-129.1	Massachusetts	13	18.29	2.03-47.25
Texas	84	33.24	1.84-258.57	Connecticut	12	11.09	3.12-31.25
New York	64	7.39	0.15-117.63	Colorado	12	35.76	2.13-208.85
Georgia	63	21.53	0.48-107.45	Kansas	11	21.62	4.19-54.07
Louisiana	54	11.57	0.75-116.26	Arkansas	11	43.75	7.49-99.35
Ohio	54	17.01	0.89-97.54	Wisconsin	9	41.41	0.53-235.9
Pennsylvania	52	11.75	0.76-130.51	New Mexico	9	74.45	5.96-166.23
Tennessee	52	14.26	1.16-153.51	Nevada	9	5.66	2.49-10.92
Michigan	45	15.74	0.92-92.87	Oklahoma	8	27.77	1.23-129.38
Missouri	42	20.94	0.89-170.71	Oregon	6	62.04	11.57-213.05
New Jersey	38	7.81	0.76-59.78	Utah	4	33.38	19.77-72.46
Virginia	35	22.75	0.98-83.02	Iowa	4	33.60	3.62-121.75
Indiana	34	9.08	1-63.24	Delaware	4	22.49	0.81-86.61
Maryland	33	5.02	0.73-40.93	Nebraska	3	11.86	10.66-12.46
Alabama	33	34.72	2.98-82.54	Rhode Island	2	32.76	28.33-37.19
South Carolina	31	26.15	0.83-56.11	South Dakota	2	306.25	306.12-306.39
North Carolina	31	36.06	1.48-116.67	Maine	2	179.48	131.3-227.67
Mississippi	26	47.80	6.28-116.93	Montana	2	529.37	433.06-625.67
Kentucky	20	25.95	0.88-94.91	Vermont	1	188.95	188.95-188.95
Arizona	16	38.68	5.71-183.69	West Virginia	1	102.46	102.46-102.46

Note: All estimates are in kilometers. In order to account for the curvature of earth when measuring the distance between two points, the Haversine formula is used. The analysis include 1,352 mass shootings across 46 states from January 2014 to March 2018. Vermont and West Virginia only have one mass shooting each during the period. Thus the nearest neighbour to these incidents are in other states, as we do not restrict the study area to be state specific.

In figure 3 the spatial kernel densities are depicted. The kernel densities mark the areas with highest density of mass shootings; the closer the circles and the darker the color the higher the density. In line with the conclusion of section 5, figure 3 suggests a tendency towards clustering in most parts of East America and California. There seem to be strong clustering centered on Chicago (Illinois), New York City (New York) as well as some clustering in San Francisco (California) and Los Angeles (California). As mentioned in section 5.1.1 these clusters are possibly driven by underlying demographic factors such as racial composition, population density and so forth. To investigate this further it would require more spatially decomposition, considering much smaller areas such as counties, cities or even neighbourhoods. The objective of this paper is not to determine the underlying

drivers of mass shootings but rather to clarify whether spatially clustering exists. Thus we will not go into further details on this aspect.

Figure 3: Kernel density clustering of mass shootings



Note: The kernel density estimation is based on the 1,352 mass shootings. The kernel is the quartic kernel and the bandwidth is calculated based on Silverman's plug-in.

Based on the clustering patterns apparent in figure 3, we choose to move forward with the analysis on the following three study areas:

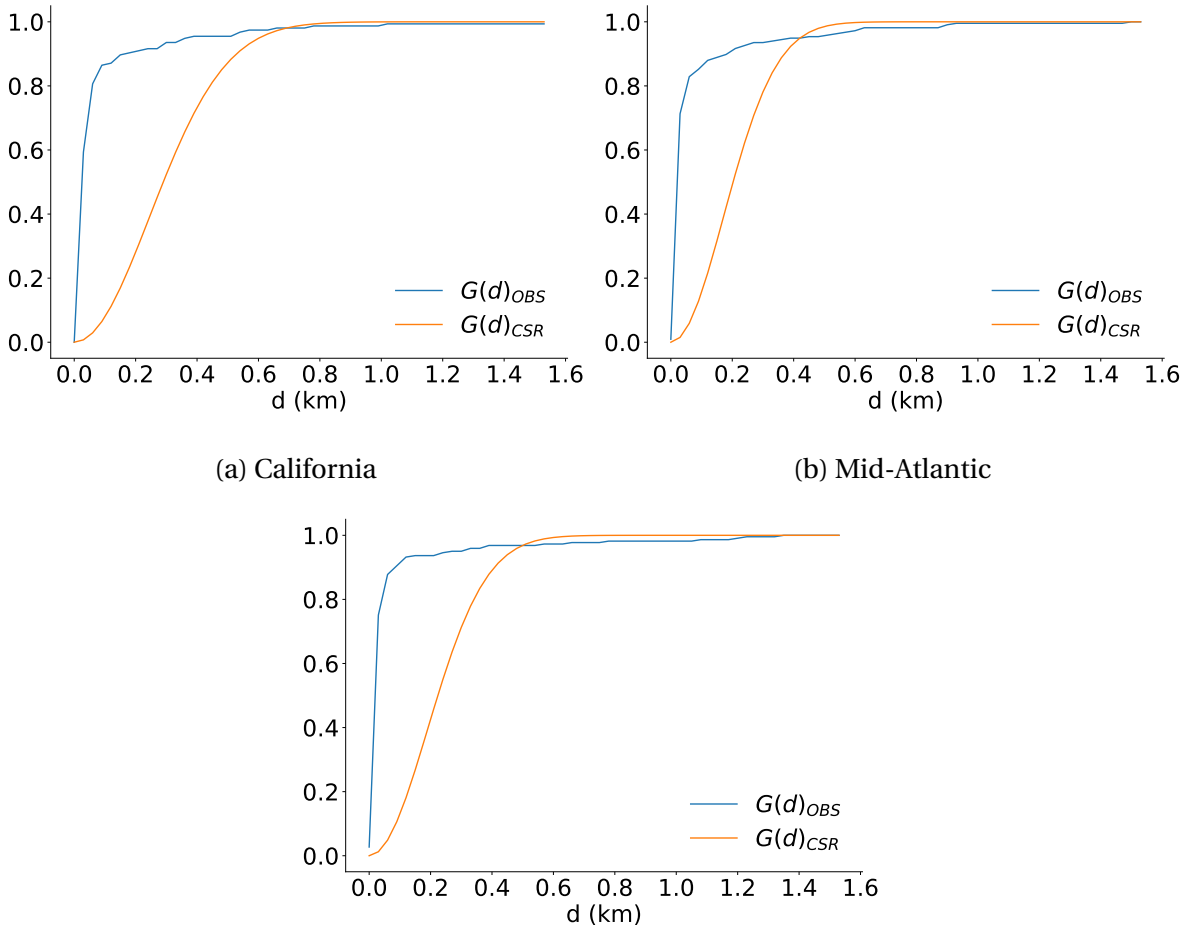
- **California**
- **North East Central**: Illinois, Indiana, Michigan, Ohio, Wisconsin
- **Mid-Atlantic**: Delaware, Maryland, New Jersey, New York, Pennsylvania, Virginia, Washington, D.C, West Virginia

As mentioned in section 3.4 density analyses are subject to edge biases. As we don't account for this in our estimation, this might bias our results along the coast. This inconsistency is visible in figure 3 as we get estimates of density probability beyond the coastline indicating a probability different from zero at points where it is highly unlikely that a mass shooting would occur. We do however, feel comfortable using the results for the purpose of selecting areas for further analysis.

In order to determine whether the spatial clustering that the ANN analysis and the kernel

density estimate suggested above is indeed significant, we compute G- and K-functions. In figure 4 the G functions of the three study areas are shown. The orange lines, $G(d)_{CSR}$, indicates complete spatial randomness (CSR) and the blue, $G(d)_{OBS}$, the cumulative distribution of the observed underlying point pattern. Whenever the $G(d)_{CSR}$ is below $G(d)_{OBS}$ the distance between the points in the underlying points pattern is smaller than if the points had been distributed randomly, implying that the clustering is significant at the distance shown on the x-axis. In all three cases we see significant clustering of mass shootings within relatively short distances (0.4-0.6 km).

Figure 4: G-functions



(a) California

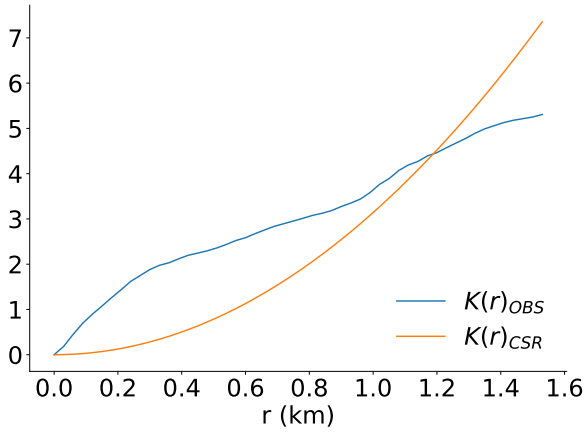
(b) Mid-Atlantic

(c) North East Central

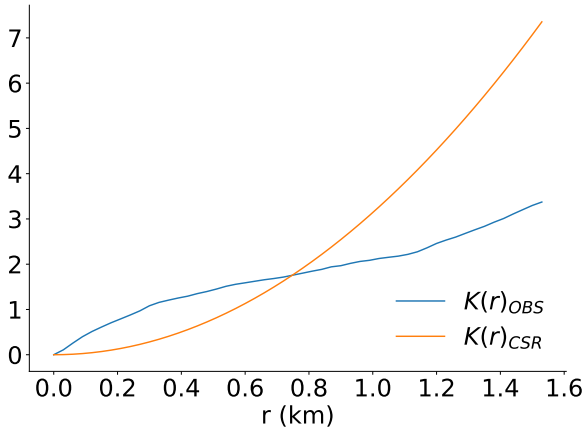
Figure 5 shows the results from the computed K-functions. For all mass shootings the observed $K(r)_{OBS}$ is above the expected value $K(r)_{CSR}$ for radius distances between 0 and 1.1 km at most. I.e. after adjusting for intensity, a typical shooting point pattern has more clustered neighbours than what would be observed if mass shooting incidents were lo-

cated randomly. Thus according to both the G- and K-function mass shootings cluster significantly at small distances.

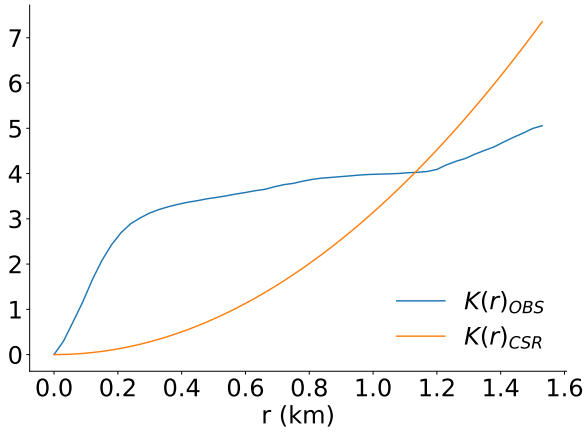
Figure 5: K-functions



(a) California



(b) Mid-Atlantic



(c) North East Central

Based on our findings in these different point pattern analysis approaches, we would conclude a significant degree of spatial clustering between mass shooting, particularly apparent at small distances. In line with what we saw in the descriptive analysis this suggests that mass shooting incidents are occurring at or around the same location multiple times. Barboza (2018) argues that a radius of 400 meters corresponds to a five minute walking distance. Based on this definition our results suggest that mass shootings occur within approximately 10 minutes walking distance from each other. Further it is worth noticing that the study areas have clustering at somewhat similar distances, implying that spatial clustering of mass shootings may not be area specific.

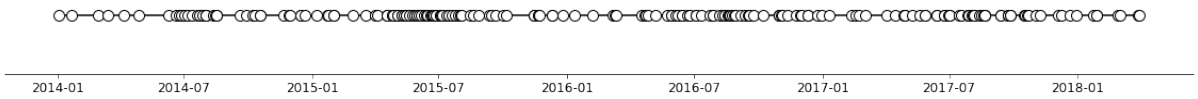
6.2 Time Clustering

Evidence suggest that one mass shooting can trigger the occurrence of subsequent mass shootings (Towers et al., 2015). Assuming this is true, we should be able to detect clustering of mass shootings at specific points in time. First, we plot the mass shootings incidents on timelines, in order to get a first impression of the distribution of the mass shootings across time. These are depicted in figure 6.

Figure 6: Mass Shooting Timelines

(a) California

Number of mass shootings in California: 155



(b) Mid-Atlantic

Number of mass shootings in Mid-Atlantic: 216

(c) North East Central

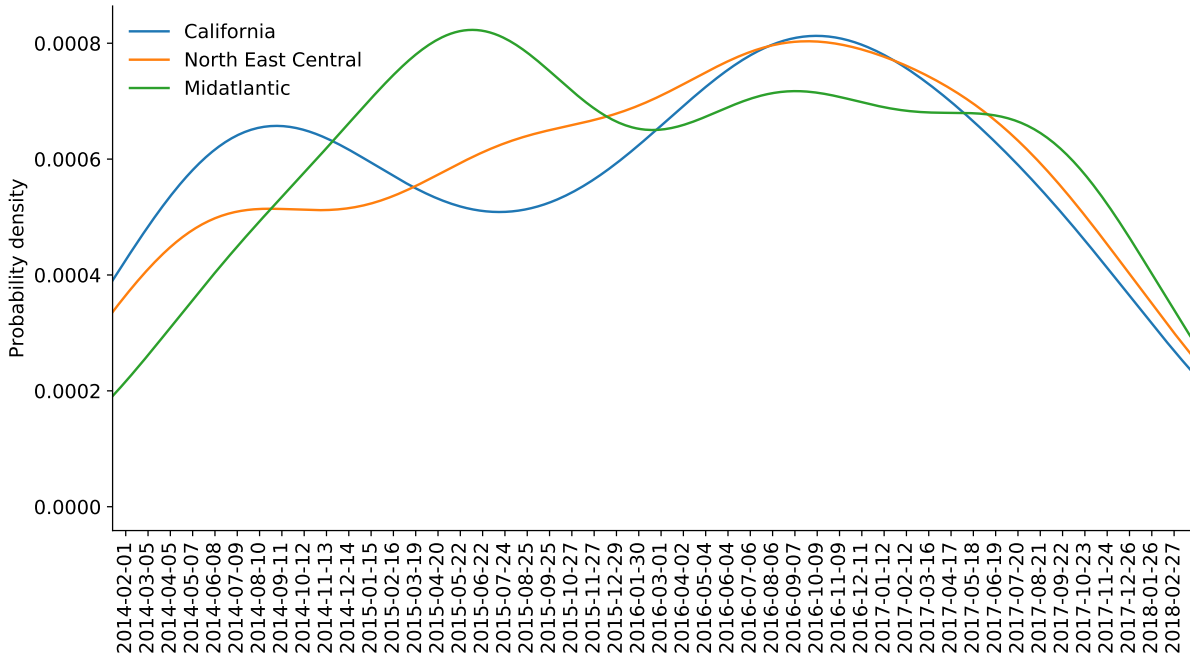
Number of mass shootings in North East Central: 221

In all three study areas mass shootings appear frequently throughout the entire period. However, it is possible to see that some periods stand out with a much higher frequency of events. In California, the first 'cluster' appear in the spring/early summer of 2014. This is around the same time that 7 people where killed in what is known as the "Isla Vista Massacre" (23rd of May 2014). The second big cluster has its off-set at around the same time that 103 people where shot dead in a gay-bar shooting in Orlando (12th of June 2016). If this event in Orlando (Florida) did indeed trigger subsequent events in California this trigger effect is apparent at quite long distances. Another explanation for the time clustering

pattern is in relation to gang related mass shootings, as one incidents may give rise to action of revenge causing more mass shooting in the future. Thus the hypotheses of mass shooting as a trigger event, might be relevant both in regard to gang related and non gang related mass shootings. Furthermore, the timelines also show sign of seasonality of mass shootings as discussed in the descriptive analysis. Mass shooting events thus seem to occur more frequently around the summer months.

The high frequency of mass shootings is most likely blurring the picture. It might thus be useful to conduct a density estimation across the time period. This is done in figure 7.

Figure 7: Kernel densities of Mass Shootings



Data consists of 155, 216 and 221 in California, Mid-Atlantic and North East Central respectively. We use a quartic kernel with bandwidth: California = 0.39, North East Central = 0.36, Mid-Atlantic = 0.36

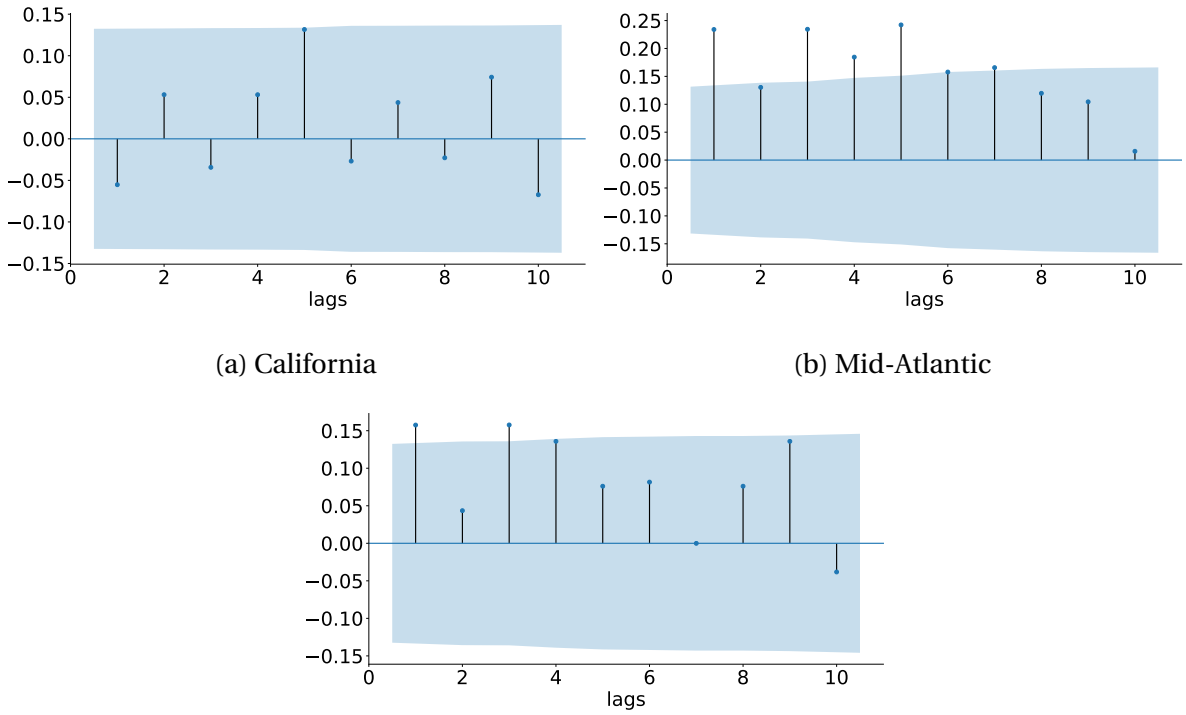
Figure 7 suggests some clustering across time. For California the density plot has two peeks suggesting increased probability of mass shootings at around August 2014 and November 2016. For Mid-Atlantic there is only one peek at July 2015 and finally North East Central has its peek at November 2016. For both Mid-Atlantic and North East Central the peeks are less dens compared to California. However, these two study areas cover several states and a larger study area, which may dilute the results hence the kernel density plot did not

reveal much additional information (see section 7).

The kernel density estimate of all areas, but most apparent in California, seem to suggest that mass shootings occur in waves with crests and troughs. However, whether these waves are triggered by a specific event cannot be concluded based on this analysis, as we do not control for any underlying factors that may contributed to this pattern.

In order to investigate whether a mass shooting can increase the probability of future mass shootings we set up an autoregression (AR) model.
When setting up an AR model we first need to determine the relevant number of lags. This is done by plotting the autocorrelation between mass shootings across weeks. The autocorrelation plots for each of the three study areas are depicted in figure 8.

Figure 8: Autocorrelation



(a) California

(b) Mid-Atlantic

(c) North East Central

Note: The blue area makes the 95% confidence interval. Each lag indicates the duration of a week. A positive sign indicates a positive correlation between lags.

Based on figure 8a we find no evidence of autocorrelation in California given that all lags

are insignificant. This implies that mass shootings have no dynamic time effects within California. Based on this we do not conduct an AR model for California. In Mid-Atlantic and North East central we do however, find significant lags for five and three weeks respectively. It worth noticing that the Mid-Atlantic (figure 8b) shows higher levels of autocorrelation for all lags.

In table 4 we have estimated the AR model for the two study areas; Mid-Atlantic and North East Central. An extra lag for each of the models are included in order to make sure that all dynamic effects are accounted for. We see that all significant lags has a positive sign. This implies that a mass shooting in one week increases the probability of a mass shooting occurring in the preceding weeks significantly. For instance in Mid-Atlantic a mass shooting in period $t-5$ significantly increase the probability of a mass shooting in period t by 17 percent. In North East Central a mass shooting in period $t-3$ significantly increase the probability of a mass shooting in period t by 14 percent. Hence, the dynamic effects of mass shootings are both more persistent and stronger in Mid-Atlantic compared to North East Central.

Table 4: AR Models of Mass Shootings

	Mid-Atlantic	**North East Central**
Lag	Coefficient	Coefficient
Mass shooting$_{-1}$	0.016**	0.14**
	(0.07)	(0.07)
Mass shooting$_{-2}$	-0.00	-0.00
	(0.07)	(0.07)
Mass shooting$_{-3}$	0.17***	0.14**
	(0.07)	(0.07)
Mass shooting$_{-4}$	0.06	0.09
	(0.07)	(0.07)
Mass shooting$_{-5}$	0.17***	-
	(0.07)	
Mass shooting$_{-6}$	0.03	-
	(0.07)	

Note: Each lag indicates the duration of a week. Standard errors are reported in parentheses. Significance is indicated by: *0.1, **0.5, ***0.01.

Given the fact that we work within specific study areas defined by geographical boundaries we might obtain biased estimates as we exclude potential incidents that could affect our estimates. More specifically one could imagine that an incident that occurred just outside the boarder of North East Central could potentially increase the probability of an incident inside the boarder. We do not capture this effect within this analysis framework. To mitigate this we could have expanded the study area to include the entire United States and controlled for state specific effects. This could also pose a potential explanation as to why California is excluded. Additional short comings of the this simple AR models are discussed in the following section.

7 Discussion

When working with large datasets data quality is always a matter of concern. For the dataset used in this paper there are multiple pit falls. We cannot be sure that all mass shooting incidents are registered or that the ones that are, are registered correctly. Mass shootings are defined as incidents with more than four victims. Thus, if by mistake, less than four victims are registered in an incident that is a mass shooting it will not be characterised as such. This makes the mass shooting variables particularly sensitive to registration errors. We cannot know how such data errors might impact our results as the number of mass shootings might as well be underestimated as overestimated. An underestimation of the number of mass shootings would mean that our results are a conservative estimate while an overestimation would mean that some of the clustering we detect is in fact explained by clustering of 'general violence' rather than mass shootings.

In our analysis we define three study areas based on the kernel density clustering for which we go into further investigation. The study areas were chosen based on where there were signs of clustering. These clustering patterns are not surprising as large cities such as San Francisco, Los Angeles, Chicago and New York City are all situated within these areas. According to Blum and Jaworski (2017) spatial clustering of mass shootings can be explained by structural factors such as population density, racial composition, poverty and crime-rate that are more apparent in large cities of the United States.

In contrast to the study area of California, North East Central and Mid-Atlantic consist of multiple states. This might delude our results, having the strongest impact on the time clustering density plot (figure 7). The spill-over effects, that are predominant in the time clustering analysis, are likely to be driven by media coverage and/or word of mouth. We suspect this to have the largest impact within, but not restricted to, relatively small geographical areas such as state or even county. Within the study areas North East Central and Mid-Atlantic several 'trigger mass shootings' might thus occur consecutively, that does not have any relation to each other, but impact the event of a another mass shooting within the smaller geographical distance. If this is the case, it could explain why there does not seem to be any evidence of time clustering in North East Central and Mid-Atlantic when considering the density plot as the high number of mass shootings are blurring the effect. On the other hand, in figure 6 we did see clustering in California which could potentially have been triggered by a mass shooting in Florida. These long distance effects could be explained by extensive national media coverage however only relatively few mass shooting attract this type of attention.

When estimating the AR-model we ignore this geographical aspect of time clustering, as we merely consider the impact of mass shootings occurring one week on the probability of mass shootings occurring in the subsequent weeks within the respective area. In this case we do find significant effect in Mid-Atlantic and North East but not in California. This is potentially driven by the higher frequency of mass shootings in the two former areas (216, 221 and 155 respectively). Further given that we do not account for the geographical aspect, the significant effect between mass shootings might be driven by mass shootings occurring very far from each other. Thus we still do not capture the time cluster effect of nearby mass shootings. Finally, we do not include any relevant control variables in the AR models implying that the estimates are subject to omitted variable bias. Relevant control variable could be demographics factors as well as restrictive gun laws in each state. Based on this, the predictive power of the model may not be very strong. However, we still believe that the signs and, potentially, significance of the estimates are somewhat valid as it is line with other studies in the field (Towers et al., 2015).

In figure 6 we suggest that there seem to be some potential trigger events, as we detect pe-

riods with high event-frequency. However, we are not able to actually determine which, if any, mass shootings are trigger event, and which features cause these events to increase the frequency of other mass shootings in the future. Such study would require other methods such as difference-in-difference or regression discontinuity estimation of mass shootings events.
A feature argued to cause a mass shooting to become a trigger event is the amount of media coverage. Thus the amount of media attention (number of related article, television-time etc.) to each mass shooting incident, could be an interesting proxy for how likely mass shootings are to trigger new incidents. Another trigger mechanism is related to gang mass shootings as one incident may give rise to action of revenge resulting in another mass shooting.

In the spatial analysis, figure 4 and 5, we find that mass shootings within the three study areas cluster at radius between 0-1.1 kilometers. Thus, though we cannot say anything about the specific characteristics of the locations in which mass shootings cluster, our evidence suggests that smaller venues are likely target locations. A potential mass shooter may target places such as schools, malls or theater as these are often densely populated and thus have a high number of potential victims, and relatively easy to access with weapons. However the geographical location of which this type of venue is targeted will probably be closely related to the structural demographic factors of the area.

The results from our analysis suggest both spatial and time clustering patterns of mass shootings. Our results from the spatial analysis suggests that policy interventions targeting specific areas, where mass shootings have previously occurred, are likely to be effective. Policies aiming at preventing mass shootings at specific locations have been implemented, but with various means of prevention. While some states has implemented more restrictive gun laws, preventing people from carrying weapons in public places such as schools, others have made it easier to do so. Which policy to implement is a question of whether one believe the weapons or the people carrying the weapons to be the problem. Based on our results from the time clustering analysis a way of mitigating the potential spill-over effect between mass shootings is to increase awareness of the negative effect of extensive media coverage. The effect being that high levels of media coverage inspire at-risk individuals to

carry out similar mass shootings, or even out-doing them (Murray, 2017).

8 Conclusion

In this paper we have investigated clustering of mass shooting across time and space. Carrying out a spatial point pattern analysis we find that mass shootings cluster within small radius of 0-1.1 kilometers. This indicates that mass shootings take place around small venues most likely in areas with structural demographic challenges. Through kernel density and autoregression estimation, we examine the time patterns in the data. Here we find vague evidence in favor of time clustering, meaning that mass shootings happen within small periods of time. More specifically we find that a mass shooting occurring in one week will increase the probability of another mass shooting in the subsequent 3-5 weeks.

Based on these findings potential policy implications could be to increase preventive action at locations where mass shootings have happened before. Furthermore as extensive media coverage can act as means of inspiration for at-risk individuals to carry out similar events, more awareness of this issue could mitigate some of these contagious effects of mass shootings.

References

ArcMap (2019). *Multi-Distance Spatial Cluster Analysis (Ripley's K Function)*. URL: https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/multi-distance-spatial-cluster-analysis.htm.

Barboza, Gia (2018). "A secondary spatial analysis of gun violence near Boston schools: a public health approach". In: *Journal of urban health* 95.3, pp. 344–360.

Blum, Dinur and Christian Gonzalez Jaworski (2017). "Spatial patterns of mass shootings in the United States, 2013–2014". In: *Environmental criminology: spatial analysis and regional issues*. Emerald Publishing Limited, pp. 57–68.

Cameron, A Colin and Pravin K Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Gimond, Manuel (2019). *Intro to GIS and Spatial Analysis*. URL: https://mgimond.github.io/Spatial/point-pattern-analysis.html.

Grinshteyn, Erin and David Hemenway (2019). "Violent death rates in the US compared to those of the other high-income countries, 2015". In: *Preventive medicine* 123, pp. 20–26.

Gun Safety Support Fund, Everytown for (2019). *Ten Years of Mass Shootings in the United States*. URL: https://everytownresearch.org/massshootingsreports/mass-shootings-in-america-2009-2019/#foot_note_5.

Maire, Daniel le (2017). "A short introduction to nonparametric econometrics". In: *University of Copenhagen*.

McGinty, Emma E et al. (2014). "News media framing of serious mental illness and gun violence in the United States, 1997-2012". In: *American Journal of Public Health* 104.3, pp. 406–413.

Metzl, Jonathan M and Kenneth T MacLeish (2015). "Mental illness, mass shootings, and the politics of American firearms". In: *American journal of public health* 105.2, pp. 240–249.

Murray, Jennifer L (2017). "Mass media reporting and enabling of mass shootings". In: *Cultural Studies Critical Methodologies* 17.2, pp. 114–124.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Reeping, Paul M et al. (2019). “State gun laws, gun ownership, and mass shootings in the US: cross sectional time series”. In: *bmj* 364, p. l542.

Rosenwald, Michael S. (2016). “The strange seasonality of violence: Why April is ‘the beginning of the killing season’”. In: *Independent*. URL: https://www.independent.co.uk/news/world/americas/the-strange-seasonality-of-violence-why-april-is-the-beginning-of-the-killing-season-a6969271.html.

Rotton, James and Ellen G Cohn (2004). “Outdoor temperature, climate control, and criminal assault: The spatial and temporal ecology of violence”. In: *Environment and Behavior* 36.2, pp. 276–306.

Serge, Rey and Wei Kang (2019). *Distance Based Statistical Method for Planar Point Patterns*. URL: https://pysal.org/notebooks/explore/pointpats/distance_statistics.html.

Towers, Sherry et al. (2015). “Contagion in mass killings and school shootings”. In: *PLoS one* 10.7, e0117259.