Spatial modeling of K-12 school shootings from 1990-2019 as a Matern clustered point process

J Steven Raquel

Introduction

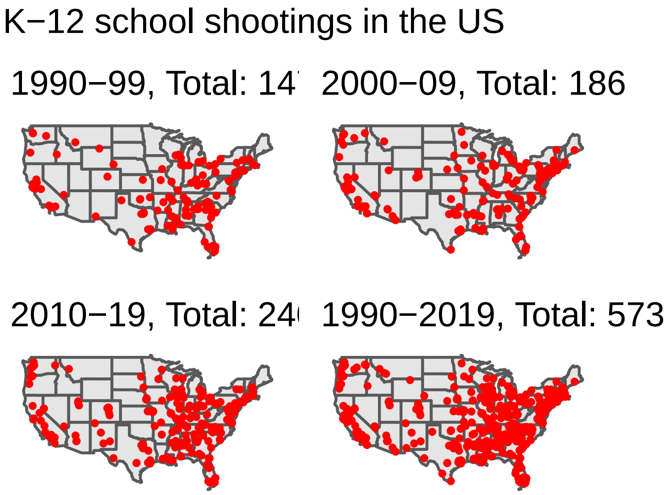
The prevalence of gun violence in schools in the United States has been referred to both as an epidemic and a public health crisis, and one that has steadily increased over the past several decades. Apart from the trauma that such an event can bring to a community, there is also resonant fear that such incidents inspire copycat events on a local and a larger scale. This spatial analysis attempts to model the incidence of these shootings as a Poisson point process, in order to ascertain whether the locations and events occur with complete spatial randomness, and thereafter create a model with which these events can be predicted. Ultimately, a Cox Matern cluster process model was decided upon, which lead us to conclude that in fact, school shooting events do give rise to future school shootings around them. This spatial analysis seeks to model on a spatial level these incidences, so that we can better understand where these tragedies are likely to happen such that we can avert them and make the country a safer place for young people.

Methods

Data

The school shooting data was sourced directly from the "K-12 School Shooting Database" made available by the Center for Homeland Defense and Security (CHDS), and was specifically subset to between the years of 1990-2019. The information that comprises the dataset was determined by a specific process which entailed asking what exactly comprises a school shooting. Although the original database contains shootings ruled accidental (from misuse of a firearm) as well as incidences of gang-related gun violence on school grounds, among other incidents, we did not opt to consider this data as relevant to this study in particular. Targeted events related to domestic situations, or the escalation of disputes (e.g. fistfights in which one person pulls out a firearm) were also ruled school shootings for the purposes of this study.

Exploratory Data Analysis



Source: The Department of Homeland Security.

As we can see from the plot above, the total number of shootings per decade has not only steadily increased over the past 30 years, but the events also occur in and around the same places, which gives credence to our hypothesis that the events exhibit a clustered pattern. We notice in fact, that there are areas that seem relatively untouched by school shootings in the western United States, whereas shootings all across the South, Midwest and the East Coast recur a great deal. While the West Coast is somewhat blighted by school shootings, particularly in the San Francisco and Los Angeles metropolitan areas, along with major cities in the Pacific Northwest), it is not nearly at the rate experienced by the other side of the US.

One thing that it is important to note, as with many spatial analyses that relate to events caused by humans, that the rate of these events do have a high correlation with population density, i.e. there are more observations of school shooting events in areas where many people live. While the implications of this unfortunately will not be well explored in this literature, it is important to take note of as a confounding factor when asking questions about the frequency of these events.

Point Pattern Analysis

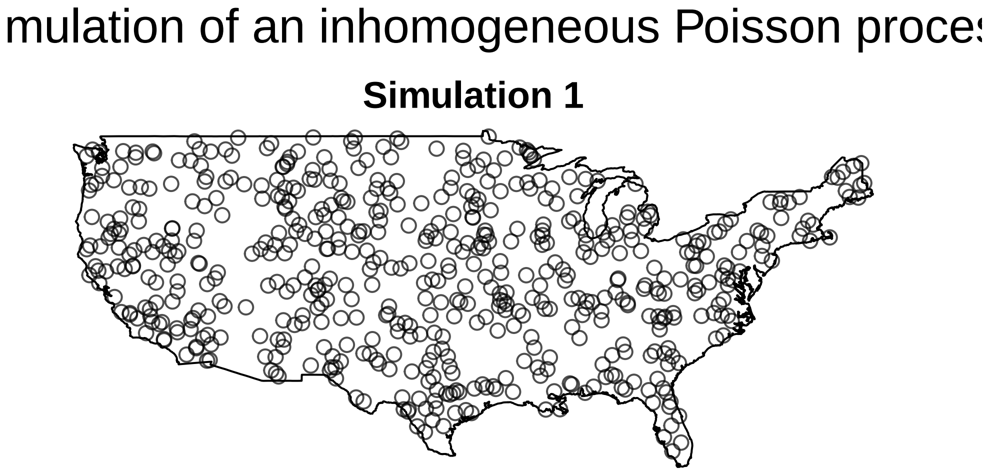
With respect to the coordinate-level data, as with any spatial point pattern analysis, we are concerned with the following three questions, 1) whether the points are located at random, 2) whether they are clustered, and 3) whether they are placed regularly. The hypothesis of *complete spatial randomness*, or a homogeneous Poisson process, asserts the following:

- The number of events in any region S with area $|S|$ follows a Poisson distribution with mean $\lambda|S|$, where λ is the intensity, i.e. λ does not change over S
- Given n events in S, the points s_i are independently located according to the uniform distribution on S, i.e. there is no interaction amongst events.

The intensity function $\lambda(s)$, also known as the first-order property of the spatial point process, is defined as

$$\lambda(s) = \lim_{|\Delta s| \to 0} \frac{E[N(\Delta s)]}{|\Delta s|}$$

Firstly we want to ascertain whether the incidences of school shootings are indeed a Poisson process, and if so, determine whether or not the process is homogeneous or inhomogeneous.



This plot demonstrates but an example of what a homogeneous Poisson process, which has complete spatial randomness, would look like. We note that the distribution of events is scattered all throughout the space such that there are no clusters anywhere, and moreover we do not see any specific pattern in their distribution. We want to test formally that the data does not follow such a distribution. This is in pursuit of our ultimate goal of creating a model that can accurately demonstrate the distribution of these events.

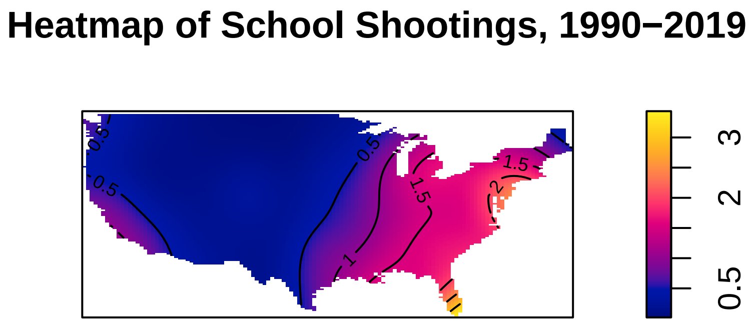
Kernel Density

One method by which we can visually ascertain as to whether the point pattern X is homogeneous is by looking at the plot of the Gaussian kernel smoothed intensity function, which appears as a heatmap. Density based mesasures look at the first order property of the process, which illustrate how observations vary from place to place due to the underlying property, whereas the second order property illustrates how observations vary from place to place due to interaction effects between observations themselves (Gimond).

The smoothing bandwidth σ, the standard deviation of the isotropic Gaussian kernel, is chosen to minimise the mean square error (MSE) as defined by Diggle (1985), which is calculated as follows

$$MSE(\sigma) = MSE(\sigma)/\lambda^2 - g(0)$$

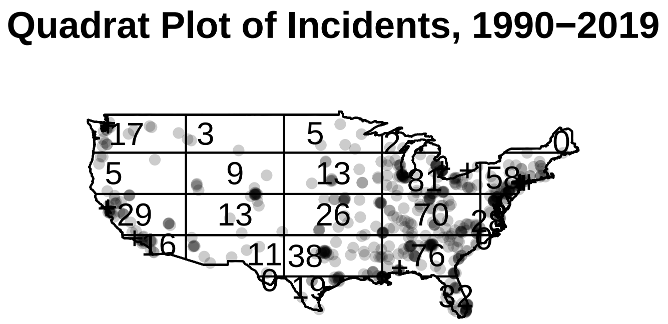
where λ is the mean intensity and g is the pair correlation function.



Based on this heatmap of the density function, we note that clustering is most strong around the south, namely Florida, as well as in the northeastern United States around New York. The same is also true for the Pacific Northwest.

Quadrats

While we could estimate the intensity function across the entire area, what we are interested in this particular analysis is to how the intensity varies across different regions contained therein. We do this by splitting up the area into what are referred to as *quadrats*, small subsets of the event space, and counting the number of events contained within each quadrat. We can test against the hypothesis of complete spatial randomness by generating a test statistic based on the number of expected vs observed events in each quadrat. The quadrat plot is given by the following plot:



It is important to note however that within a given quadrat, the number of events contained therein does not give us information about how clustered the events are inside that quadrat itself. Furthermoe,the count of events contained within each quadrat is also dependent on the definition of these dimensions, so they can be subject to misleading conclusions as a result. Since the continental United States is not shaped like a simple polygon, dividing it into a reasonable number of equitable and reasonably defined quadrats is not an easy task and this is a shortcoming we have to recognize.

One methodology for testing for clustering is a Monte Carlo quadrat test in which we take a number of simulations of patterns under the null hypothesis e.g. the homogeneous Poisson point pattern we observed, and compute a χ^2 test statistic based on the expected and observed counts in each quadrat, and their residuals. The alternative hypothesis varies depending on the test, but we want to determine specifically whether 1) the pattern is homogeneous or not and 2) whether the pattern is clustered.

Null hypothesis	Alternative hypothesis	Test Statistic	p-Value	Conclusion
X is a homogeneous Poisson process	X is not a homogeneous Poisson process	582.1008	4e-04	reject H_0
X is a homogeneous Poisson process	X is a clustered point pattern	582.1008	2e-04	reject H_0

Based on this table, we have the results of two separate quadrat tests for our point pattern X, with the null hypothesis of complete spatial randomness against the corresponding alternative hypotheses of 1) X not being a homogeneous Poisson process and 2) X being a clustered point pattern.

In both cases, our test returns a corresponding p-value of approximately zero, so we have evidence to reject H_0 in both cases and conclude that we have respectively in X, an inhomogeneous Poisson process, and a clustered point pattern.

Ripley's K-function and the G-function

Ripley's K-function of a point process is defined so that $\lambda K(r)$ equals the expected number of additional random points within a distance r of a typical random point of the point process X, and is determined by the second order moment properties of X. Deviations between the empirical and theoretical K function give us evidence of spatial clustering or regularity.

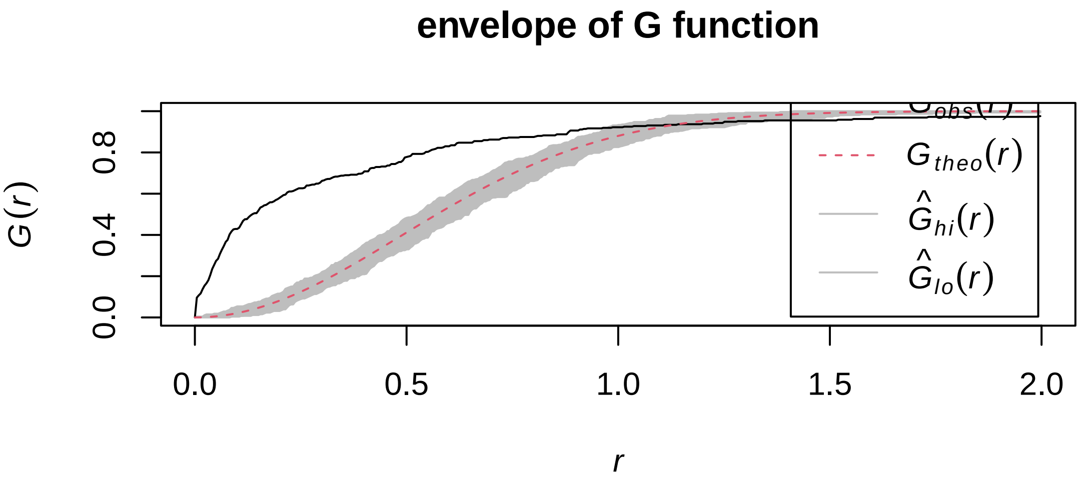
$$K(r) = \lambda r^2, \lambda_2(r) = \lambda^2$$

There are various transformations of the K-function, for example the L-function proposed by Julian Besag, which stabilizes the variance, or the G-function, which is the cumulative distribution function of the distance from a typical random point of X to the nearest other point of X. The G-function allows us to visually test for clustering on top of testing for deviation from complete spatial randomness (CSR). Moreover, the K and L-functions are rather computationally expensive relative to the G-function for this many events, so it will be our function of choice to demonstrate here.

The G-function is given by

$$G(r) = 1 - \exp(-\lambda \pi r^2) = 1 - \exp(-\pi K(r))$$

We can generate an "envelope" of simulated G-functions based on a random Poisson point pattern, and compare our empirical G-function from the data to ascertain whether the point pattern X has complete spatial randomness, and to diagnose whether the point pattern is clustered.



Contained in this figure we do see a marked difference between the respective G-functions of the empirical Poisson point process, and that of the theoretical point process which exhibits complete spatial randomness. Namely the empirical G-function exceeds the theoretical G-function up to approximately $r = 1$, after which point it crosses the envelope of the point process which exhibits CSR and emerges underneath it. A G-function predominantly underneath the envelope would be considered a regular point pattern, but we feel this plot moreso represents evidence towards it being clustered, which is corroborated by the results of the quadrat test earlier.

Modeling

The inhomogeneous Poisson point process

Given that we have confirmed via the quadrat tests and G-function that the Poisson process is inhomogeneous as well as clustered, our goal at this point is to develop a model with which we can estimate the intensity function $\lambda(s)$.

There are two different methods that we can model this clustered process: the first. the inhomogeneous Poisson process, assumes that the process varies spatially as a function of certain covariates, and assumes that the events themselves are independent. The second method, the Cox process, which is itself a generalization of the inhomogeneous Poisson process, treats the intensity function $\lambda(s)$ itself as a stochastic process that we can model in the same manner as the first method. The latter also assumes that the events are not independent of each other. A more benign example of such a process might be the growth of a forest, since trees leave seeds around them which then can grow into even more trees.

For an inhomogeneous Poisson process, given the number of events $N(B)$ in a subset B of the spatial domain S, the likelihood of an inhomogeneous point process is given by

$$P(N(B) = n)\Pi_{i=1}^{n} P(x_i = s_i) = \frac{1}{n!} \exp(-\int_B \lambda(s)ds)\Pi_{i=1}^{n} \lambda(s_i)$$

and the log-likelihood is proportional to

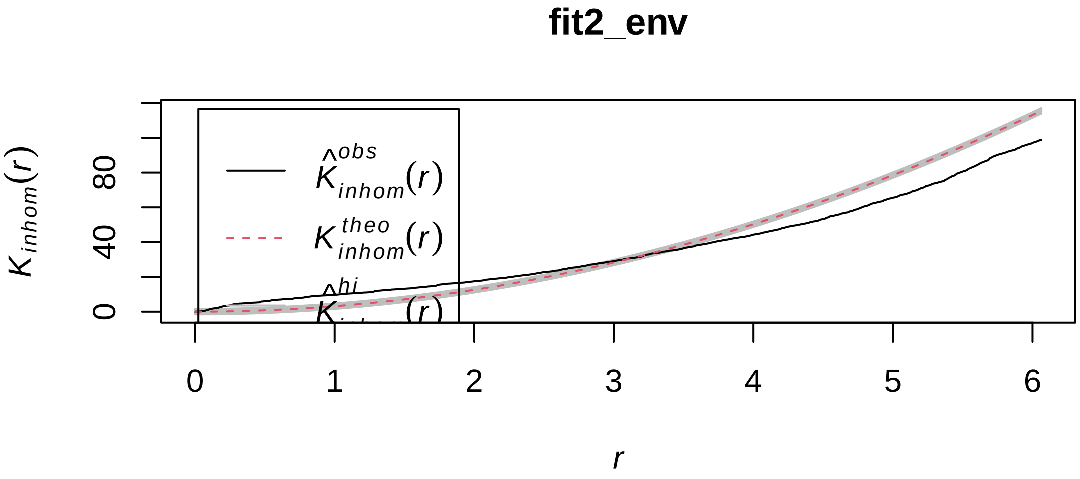
$$\log(\lambda(s)) = \sum_{j=1}^{p} \beta_j x_j(s)$$

We can then model the intensity function as

$$\log(\lambda(s)) = \sum_{j=1}^{p} \beta_j x_j(s)$$

where $x_j(s), j = 1, ...p$ are p covariates, such that the log-likelihood is a function of the parameter coefficients β_j.

Here we fit a clustered inhomogeneous Poisson point process model, using the Matern cluster algorithm. We do not use any other covariates other than the coordinates in this model. We want to analyze the K-function based on this fitted model so that we can diagnose it and proceed.



This figure depicts an observed K-function as well as an envelope of theoretical K-functions based on the intensity function $\lambda(s)$ fit in the cluster model that was just fit. As can be seen here, the level of clustering is actually greater in the observed point process up until the distance $r = 3$. We're not entirely satisfied by the results of this model due to how it underestimates the clustering relative to that expressed in the true data, hence we want to move on to the Matern cluster point process– although the results of this model do have implications for the subsequent model.

Cox processes and the Matern cluster process model

The Cox process model treats the intensity function $\lambda(s)$ as a stochastic process, adding a significant layer of complexity (and flexibility) relative to the somewhat inflexible inhomogeneous Poisson process model. More specifically we will be discussing the Matern cluster process model.

The Matern cluster point process is formed by taking a pattern of "parent" points, generated according to some Poisson process with intensity parameter κ, and then generating around it a random number of "offspring" which is itself a Poisson random variable with mean μ. The locations of the offspring are independent and identically distributed via a Uniform distribution in a radius around the parent defined by the parameter R, also known as the scale.

Our goal is to minimize the discrepancy between the estimated model and the data, given some constraints. This discrepancy criterion $D(\theta)$ is given by

$$D(\theta) = \int_0^{r_0} w(t)[(\hat{K}(t))^c - (K(t;\theta))^c]^2 dt$$

kappa	R	c	w(t)
0.0172795	1.322571	0.25	1

where we have some parameters r_0, c, and the weight function $w(r)$. Minimizing this function is known as the method of minimum contrast.

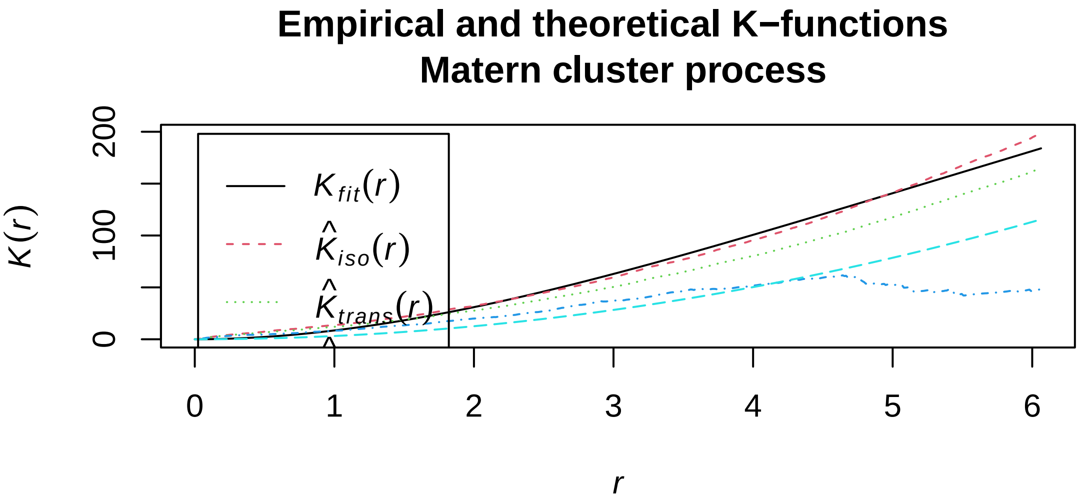
It works by first computing the K function, and then deriving the theoretical expected K value under the point process model. The model is then fit by tuning the optimal parameter values which minimizes the difference between the theoretical and empirical K-functions.

The theoretical K-function of this process is given by

$$K(r) = \pi r^2 + \frac{h(\frac{r}{2R})}{\kappa}$$

and the theoretical intensity of the process is $\lambda = \kappa\mu$.

Recall that earlier we fit an inhomogeneous Poisson point process model using the Matern cluster algorithm to define the clusters. The model that was fit returned estimates for not only the intensity function $\lambda(s)$, but also κ and μ, the intensity parameter of the parent points' pattern, and the mean parameter for the Poisson random variable of the number of offspring. For values of c and $w(t)$, we want to opt for $c = 0.25$ and a weight function of $w(t) = 1$, as these are well suited to well-clustered data.



This figure gives the theoretical and empirical K-functions of the Matern cluster process model. The cyan line is the homogeneous Poisson process, which we have long since established is ill-fitted to the data. The black line is our empirical K-function, and the red line is the theoretical K-function under the model.

Based off of this plot of the K-function, it appears in fact that the Matern cluster process model is fairly consistent at estimating the true underlying process given that the discrepancy between the theoretical and empirical K-functions is very low even up to very large distances. The following table gives the parameter values chosen for the Matern cluster model we have opted for.

Discussion

It's been debated at length as to whether incidences of school shootings give rise to future events, as long as they've been happening in the modern day. With respect to how this model applies in context, we have established that it's indeed possible that this is the case, as demonstrated by how well the proposed Matern process model fits to the empirical process. Ascertaining the reason as to why this happens would have to be left for some future work, perhaps incorporating some of the unused categorical variables that were included in the data, such as whether the event was preplanned. The data can also be looked at on an areal level and compared next to the accessibility of both mental healthcare and guns (as shown in the Appendix). This was considered as an initial objective, but was later sidelined in favor of focusing on analysis of the clustering process.

One also wonders how one might model these tragedies from a spatiotemporal perspective, as it is well documented how events have been on the rise in the past several years (also shown in the Appendix). The ultimate goal of course, would be to be able to effectively isolate the pattern and frequency of these events to pre-empt their occurrences and minimize further harm to others, especially young people.

Conclusion

This Matern cluster process model should be seen first and foremost as a building block upon which future work by individuals in public health, sociology, criminology, etc. can build. As with any instances of tragedy, many unresolved questions remain. Our hope is that some level of positive inspiration can happen from studying these events such that we can become better equipped at averting and dealing with these tragedies.

References

Adrian Baddeley, Ege Rubak, Rolf Turner (2015). Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press, 2015. URL https://www.routledge.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/

Bivand, Roger S. and Wong, David W. S. (2018) Comparing implementations of global and local indicators of spatial association TEST, 27(3), 716-748. URL https://doi.org/10.1007/s11749-018-0599-x

Hellebuyck, M., Halpern, M., Nguyen, T. and Fritze, D., 2018. The State of Mental Health in America. p.9.

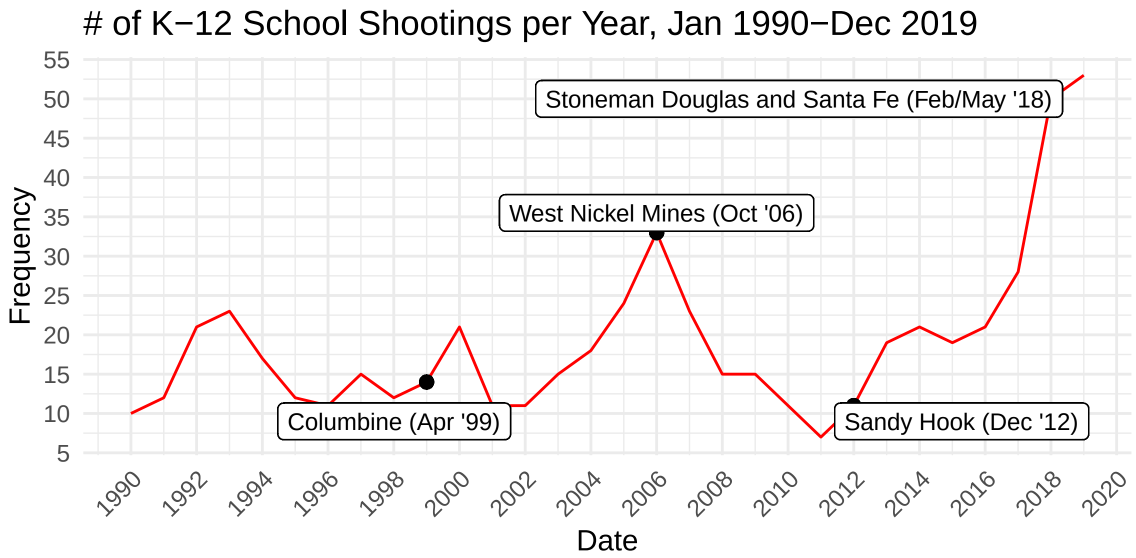
Paez A (2021). An Introduction to Spatial Data Analysis and Statistics: A Course in R. McMaster Invisible Press. ISBN: 978-1-7778515-0-7

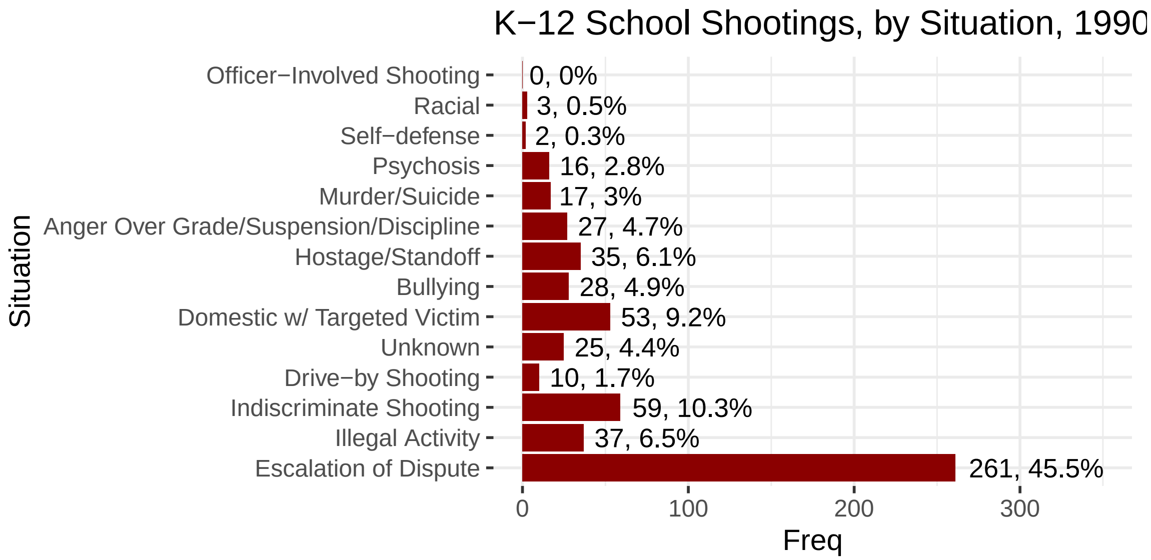
Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

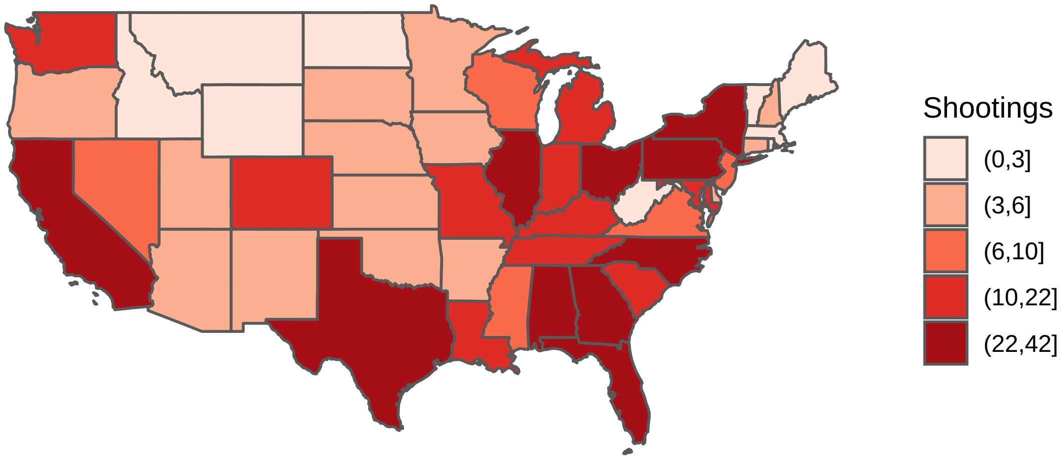
Riedman, D., Jernegan, E. and O'Neill, D., 2020. K-12 School Shooting Database. [online] Center for Homeland Defense and Security. Available at: https://www.chds.us/ssdb/ [Accessed 15 March 2022].

Siegel, M., 2022. State-by-State Firearm Law Data | State Firearm Laws. [online] Statefirearmlaws.org. Available at: http://www.statefirearmlaws.org/ [Accessed 15 March 2022].

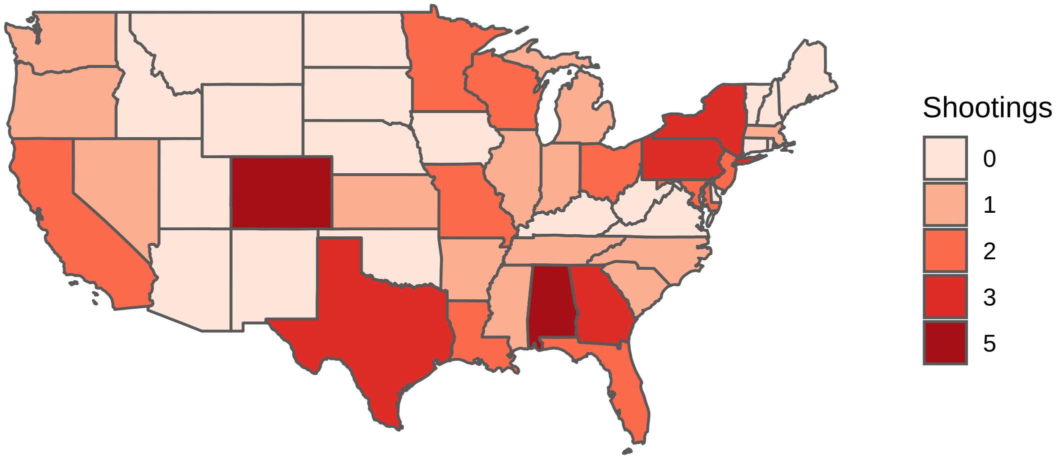
Appendix


# of K−12 School Shootings per Year, Jan 1990−Dec 2019
Frequency
55
50
45
40
35
30
25
20
15
10
5
Stoneman Douglas and Santa Fe (Feb/May '18)
West Nickel Mines (Oct '06)
Columbine (Apr '99)
Sandy Hook (Dec '12)
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Date

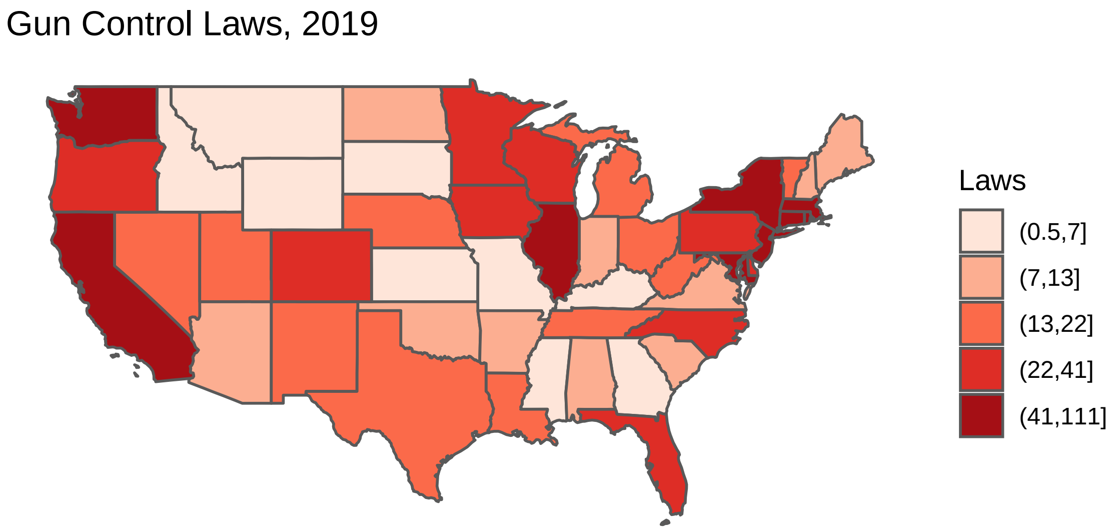

K−12 School Shootings, by Situation, 1990
Situation
Officer−Involved Shooting
Racial
Self−defense
Psychosis
Murder/Suicide
Anger Over Grade/Suspension/Discipline
Hostage/Standoff
Bullying
Domestic w/ Targeted Victim
Unknown
Drive−by Shooting
Indiscriminate Shooting
Illegal Activity
Escalation of Dispute
0, 0%
3, 0.5%
2, 0.3%
16, 2.8%
17, 3%
27, 4.7%
35, 6.1%
28, 4.9%
53, 9.2%
25, 4.4%
10, 1.7%
59, 10.3%
37, 6.5%
261, 45.5%
0
100
200
300
Freq



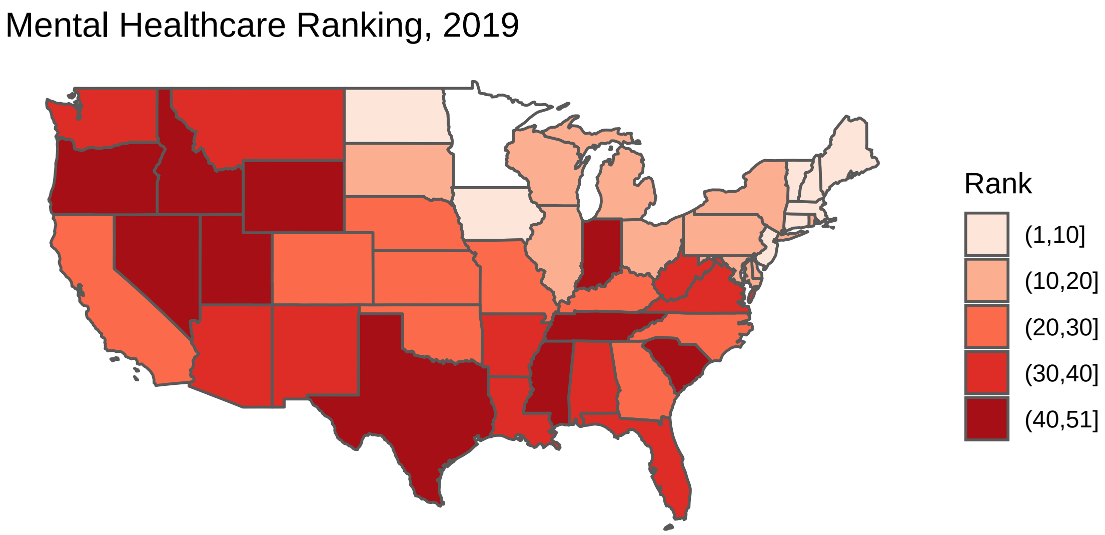
K−12 School Shootings, 1990−2019



K−12 School Shootings, 2019

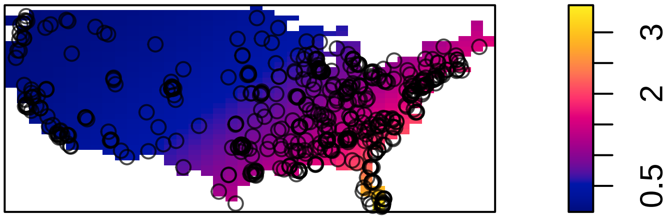


Gun Control Laws, 2019

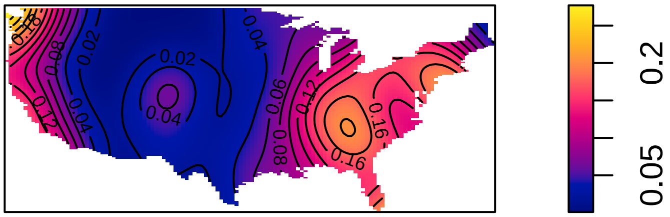


Mental Healthcare Ranking, 2019

fitted trend of the simple (x,y) model



Heatmap of Preplanned School Shootings, 1990−2019


0.18
0.08
0.02
0.02
0.04
0.04
0.12
0.04
0.06
0.08
0.12
0.16
0.16
0.2
0.05