Classification and cluster analysis on COVID-19 death trends in the United States

Feng, B. T.[1] , Kao, C. L.[2] , Hung, H. N.[2], Chen, Y. L.[2] , Wang, W.[2]

We combine two statistical methods (i.e. polynomial regression and hierarchical clustering algorithm based on Pearson's correlation coefficients) to classify and cluster the 50 states of the United States based on their COVID-19 death trends under a time frame that starts on each state's date of first confirmed case (referred to as 'days since first case' in our paper). Upon applying the correlation matrix to measure the similarity between states and taking into account the sign of the second-order coefficient in the polynomial regression model, three clusters are formed and named as "Initial Spike", "Two-Humped", and "Recent Spike". Inspired by observations suggesting a correlation between urban characteristics and our clusters, we propose an urban concentration (diversity) index to describe its relationship with COVID-19-related severity measures via regression analysis. Up till the 270th days since first case for each state, lower urban concentration corresponds to a higher percentage of confirmed cases (Pearson's correlation -0.57) but a lower mortality rate on the logarithmic scale (Pearson's correlation 0.52).

Since the beginning of the COVID-19 pandemic, many papers have studied factors affecting the spread and severity of the coronavirus. From human behavior, demographic and environmental factors, to governmental policymaking, there are copious amounts of variables that may intertwine and influence COVID-19 progression [1,2]. In the United States, which has recorded the most confirmed cases and deaths among all countries, the COVID-19 outbreak has rapidly moved from urban to rural areas [3,4]. Few attempts have been made to find out the factors influencing the spread of the pandemic across the United States. For example, Karim and Chen [5] classified U.S. counties into three categories (i.e., rural, micropolitan, and metropolitan) and found the categories to correlate significantly with the rates of COVID-19 deaths. From the work of Carozzi et al. [6], while population density affected the timing of the outbreak in each county, they found no evidence that population density is positively associated with COVID-19 cases and deaths.

Here we adopt an independent approach to analyzing the progression of the COVID-19 epidemic in the United States. First, we apply statistical methods to classify and cluster the 50 states based on their COVID-19 death trends. Then,

[1] Department of Computational Biology, The University of Texas at Austin, USA.
[2] Institute of Statistics, National Yang Ming Chiao Tung University, Taiwan.

inspired by the geographic locations of the clusters, we propose a new index to measure the level of urban concentration (diversity) in a state and find that it has strong correlations with the state's COVID-19-related severity variables. This novel urban index, which utilizes detailed county-level census data, offers a new and useful way of measuring a state's urban concentration level rather than just the state's population density. Our findings may generate insights into potential factors that influence the spread of the COVID-19 epidemic in the United States or other large countries with obvious urban–rural disparity.

Results

In this study, the target population includes the 50 states of the United States with each state treated as a unit of observation. All COVID-19 related data were obtained from the United States Central Disease Center database, while state geographic and demographic data came from the United States Census Bureau database. To measure the magnitude of the coronavirus epidemic in each state, we study the following variables: (1) the cumulative case percentage, which is the total state cases divided by the total state population; (2) the cumulative death percentage, which is the total state deaths divided by the total state population; (3) the monthly new death percentage, which is the monthly new state deaths divided by the total state population; and (4) the mortality rate, which is the total state deaths divided by the total state cases. The cumulative case percentage reflects the overall transmission of the coronavirus. However, it is subjected to different states' testing accessibilities and reporting policies, which may be changing over time. The cumulative death percentage yields insights into the overall impact of the coronavirus on a state but doesn't effectively reflect whether conditions of the coronavirus in a state are ameliorating or not, due to accumulated effects. The monthly new death percentage, which is essentially the derivative of the cumulative death percentage, can measure how quickly coronavirus situations are improving or worsening in a state. Monthly record is less subjected to short term fluctuations or reporting errors, as periodic oscillations have been observed in daily reported data [7]. Two time frames are under consideration. The usual time frame by calendar dates is a convenient choice for certain situations. However, different states have different COVID-19 starting dates. For example, the first state with a confirmed case is Washington on 2020-01-22, while the last state with a confirmed case is West Virginia on 2020-03-17. By such, the usual time frame is not suitable for comparing the pandemic evolution for each state. Therefore, in this article, we adopt a time scale called "*days since first case*", which is adjusted to each state's date of first confirmed COVID-19 case. This time frame has also been chosen when comparisons across different countries are made [8,9].

Classification by the trend of cumulative death percentages

By fitting a polynomial regression model to describe the curve of each state's trend, the sign of the second-order coefficient was used to classify the states into two groups (see details in Methods). In Figure 1a, the Negative group consists of states that exhibit a rise in COVID-19 deaths in the early stages but shows signs of containment and control around the 100th *days since first case* mark (such as New York and Illinois); while the Positive group consists of states that exhibit slow, gradual growth in COVID-19 deaths until the 150~200th *days since first case* mark, where a speedy rise is seen (such as Texas). In Figure 1b, with the y-axis denoting the daily new death percentage, we see that there exists a crossing point between the two groups around the 125th *days since first case* mark. To visualize whether the severity of the coronavirus in a state is alleviating or worsening, the monthly new death percentage measure is graphed in Figure 1c. The monthly new death percentage is computed on a 30-day interval from the 0th to the 270th *days since first case*; therefore, a total of 10 data points is collected from each state. The 270th *days since first case* is selected as the end point of the timescale to adjust to West Virginia, the state with the latest COVID-19 starting date; West Virginia was on its' 270th *days since first case* on the last day of our data collection (December 12th). Figure 1c seems consistent with our findings so far regarding the crossing point between the 100th~150th *days since first case*. However, upon inspecting the Positive group, we noticed that there were states that had remained relatively stable throughout this entire pandemic. In other words, these states behaved differently from those which were rapidly worsening in severity. This observation motivated us to adopt monthly new death percentages as the measure of epidemic severity in the next cluster analysis.

Clustering by Monthly New Death Percentages

The agglomerative hierarchical clustering algorithm produced four clusters as shown in Figure 2 and we used it to enhance the trend-based classification. Combining the results from two analyses, three clusters were formed and named as "Initial Spike", "Two-Humped", and "Recent Spike" based on the pattern they exhibited. The Positive group from trend-based classification is now split into cluster 1 and cluster 2; cluster 1 is labeled as "Two Humped" as its' trend exhibited a two humped pattern and cluster 2 is labeled as "Recent Spike" due to its later rise in COVID-19 deaths. Clusters 3 and 4, which both belong to the Negative group, exhibit similar trends with only slight differences in the timing of their early peaks. Therefore, we decided to merge them into the same cluster with the label "Initial Spike". Figures 3a and 3b, which are re-displays of Figures 1c and 1b but labeled according to the three new clusters, reveals more information about the turnover phenomenon in the COVID-19 pandemic observed earlier. From the map on Figure 3c, the geographic locations of the states in the three clusters reflect that the clusters may be associated with rural/urban characteristics. Based on Figure 4, we observed that the states in the "Recent Spike" cluster are least urbanized in both area and population. Figure 5 can be used to compare the magnitude of COVID-19-related severity variables on the 270th *days since first cases* in the three clusters. The "Recent Spike" cluster, which is composed of states with later rises in COVID-19 deaths, has the highest cumulative case percentage as shown in Figure 5a, but the lowest mortality rate as shown in Figure 5b. Lastly, the comparison based on death percentage is shown in Figure 5c. Although the differences are less significant, the "Recent Spike" cluster still shows the lowest death percentage. While the phenomenon of decreasing mortality rates can be explained by multiple factors, such as changing patient demographics in age and chronic conditions or improved treatments [10, 11], we will focus primary on the urban effect.

Relationships with Urban Concentration

The U.S. Census Bureau Database provides urban/rural area and population information on the county-level, so we could calculate the urban percentage, based on area or population, for each county. To quantify the urban effect, we proposed an index S_i^* for State *i* with a smaller value indicating higher level of urban concentration, and vice versa (For details, see Methods). Several linear regression analyses were performed to study the relationships between COVID-19-related measures and S_i^*. Table 1 summarizes the results which contain the R-squared and p-value of the fitted model and Spearman's correlation coefficient (denoted as r_s. Note that taking the square root of the R-squared leads to the absolute value of Pearson's correlation denoted as r_p. From Figure 6a, we see that higher urban area concentration from our index corresponds to higher mortality rate on the logarithmic scale ($r_p = -0.57$ and $r_s = -0.62$). We also observed that the "Recent Spike" cluster can be very well separated from the other two clusters at around $S_i^* \approx -19.5$ since most states in this cluster have an index value higher than -19.5. The states in "Initial Spike" and "Two-Humped" clusters tend to be more urbanized but the former had higher mortality rate. In Figure 6b, the urban index was calculated based on population and its relationship with the log mortality rate became slightly weaker ($r_p = -0.44$ and $r_s = -0.5$). Figure 6c indicates that higher urban diversity (larger S_i^*) corresponds to higher percentages of confirmed cases ($r_p = 0.52$ and $r_s = 0.53$). In terms of case percentages, the "Recent Spike" cluster is clearly higher than the other two clusters. Treating S_i^* as a response variable and the cluster type as a categorical regressor, the resulting R^2 were 40.78% and 34.84% calculated based on area and population respectively. The boxplots in Figure 7 indicate that the "Recent Spike" cluster contains less urbanized states but the value of S_i^* alone can't distinguish the difference between "Initial Spike" and "Two-Humped" clusters.

Discussion

Through the proposed entropy-based urban index, we have found urban area concentration to be moderately associated with cases percentage and strongly associated with the log mortality rate, with urban population concentration having slightly weaker yet similar associations. Furthermore, this index well explains whether a state had an early or a later rise in COVID-19 deaths. We conjecture that states with high urban concentrations have urban areas/populations that are less spread out, leading to earlier and more severe spikes. However, it also means that governmental policies make an impact faster and more effectively for these highly concentrated states, which could explain the index's negative association with case percentage. An obvious distinction between the Initial Spike and Two-humped clusters is the mortality rate (Figures 5b, 6a and 6b). We suspect the two-humped cluster has lower mortality since a large proportion of cases occurred in a later time.

We are excited about the insights generated from our research and hope that it could lead to even more detailed and insightful analysis into the relationship between urban concentration and the spread of a pandemic. We also acknowledge that, besides the urban factor, there are other relevant variables which affect the progressions and severity of the coronavirus. These variables can enhance a model's goodness of fit and shed light on policy making to control the epidemic of infectious diseases.

Methods

Classification based on the Trend of Cumulative Death Percentages

We fit the curve of cumulative death percentage by a parabolic function for each state. Specifically let Y_{ij} be the cumulative number of death in State *i* measured at day *j*, where $j = 0$ is the day of the first case in State *i*. Based on data $(Y_{i1,\dots,}Y_{im_i})$, where m_i is the number of days since first case to the end of study for State *i*, we fit the following second-order polynomial regression model such that

$$Y_{ij} = \beta_{0i} + \beta_{1i}j + \beta_{2i}j^2 + \varepsilon_i.$$

Denote $\hat{\beta}_{2i}$ as the estimate of β_{2i} for $i = 1, \dots, 50$. Based on the sign of $\hat{\beta}_{2i}$, the states are classified into two groups. States with $\hat{\beta}_{2i} > 0$ have concave up trends; while states with $\hat{\beta}_{2i} < 0$ have concave down trends.

Cluster Analysis using Correlation Matrix based on Monthly New Death Percentage

Let r_{ik} be the Pearson's correlation between State *i* and State *k* based on the monthly new death percentage. The matrix $R = (r_{ik})_{50\times 50}$ is a similarity measure so that $D = 1 - R$ can be viewed as a distance measure in the agglomerative hierarchical clustering algorithm.

Proposed Urban Diversity Index based on Entropy

The Shannon index [12] is often used to measure biodiversity. Its original definition is given by $H = -\sum_{i=1}^{K} P_i \ln(P_i)$, where K is the total number of species, P_i is the proportion of individuals belonging to the ith species and $\sum_{i=1}^{K} P_i = 1$. The negative Shannon index is the expectation of the natural logarithm of a species' proportion. Larger value of H corresponds to larger diversity (lower concentration). We adopt the concept of entropy to measure the level of urban concentration in each US state using the data from the United States Census Bureau Database. Here we consider the index, S_i for state *i*, which is the negative conditional expectation of the natural logarithm of a county's urban proportion. For the detail, let X_{ij} be the urban area (population) of County *j*, K_i be the number of counties and U_i be the total area (population) in State *i*. Let $P_{ij} = X_{ij}/U_i$ be the proportion of urban area (population) in County *j* of State *i*, then our S_i is defined as

$$S_i = -\sum_{j=1}^{K_i} \frac{P_{ij}}{C_i} \ln(P_{ij})$$

where $C_i = \sum_{j=1}^{K_i} P_{ij}$ is the proportion of urban area (population) in State *i*. Since the population and area of different states may vary many times, we need to modify S_i to account for this effect. Consider two states *i* and *k* such that State *i* is just *m* replicates of State *k* and their pandemic situations should be the same. It follows that $C_i = C_k$, and $S_i = S_k + \ln(m)$. Thus, we propose to remove this confounding effect due to the state size. The modified measures are $S_i^* = S_i - \ln(U_i)$. Finally, we will use S_i^* in our analysis.

Data and code availability

The datasets analysed in the manuscript and the code implementing the methods are available at https://github.com/yan9914/COVID-19.

References

1. Roy, S. & Ghosh, P. "Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking." *PLoS ONE* **15**,10; 10.1371/journal.pone.0241165 (2020)
2. Bherwani, H. et al. Exploring dependence of COVID-19 on environmental factors and spread prediction in India. *npj Clim Atmos Sci* **3**, 38; 10.1038/s41612-020-00142-x (2020).
3. COVID-19 Stats: COVID-19 Incidence, by Urban-Rural Classification — United States, January 22–October 31, 2020. *MMWR Morb Mortal Wkly Rep* **69**:1753;. 10.15585/mmwr.mm6946a6 (2020).

4. Leatherby, L. The Worst Virus Outbreaks in the U.S. Are Now in Rural Areas. *The New York Times.* October 22; www.nytimes.com (2020).
5. Karim, S.A. & Chen, H.F. Deaths From COVID□19 in Rural, Micropolitan, and Metropolitan Areas: A County□Level Comparison. *The Journal of Rural Health*, **37**, 124-132 (2020).
6. Carozzi, F., Provenzano, S., & Roth, S. Urban Density and COVID-19. *IZA Discussion Papers 13440, Institute of Labor Economics* (2020).
7. Bukhari, Q., Jameel, Y., Massaro, J.M., D'Agostino, R.B., & Khan, S. Periodic Oscillations in Daily Reported Infections and Deaths for Coronavirus Disease 2019. *JAMA Network Open* **3**, e2017521 (2020).
8. Hale, T. et al. Global Assessment of the Relationship between Government Response Measures and COVID-19 Deaths. medRxiv; 10.1101/2020.07.04.20145334 (2020).
9. Valcarcel, B. et al. The effect of early-stage public health policies in the transmission of COVID-19 for South American countries. *Rev Panam Salud Publica* **44**,e148; 10.26633/RPSP.2020.148 (2020).
10. Leora, I.H. et al. Trends in COVID-19 Risk-Adjusted Mortality Rates. *J Hosp Med. Published;* 10.12788/jhm.3552 (2020).
11. Ledford, H. Why Do Covid Death Rates Appear To Be Falling? Nature **587**, 190-192; https://www.nature.com/articles/d41586-020-03132-4 (2020).
12. Jaynes, E.T. Gibbs vs Boltzmann Entropies. *American Journal of Physics* **33**, 391-398 (1965).

Acknowledgements

This work was supported by,

Author contributions

Wang designed and led the study as well as the manuscript preparation. Feng collected data and contributed to data analysis and figures, with the support from Chen. Hung and Kao contributed to data interpretation and manuscript preparation. All authors contributed to writing the manuscript and revising the final version.

Competing interests

The authors declare no competing interests.

Figure Legends

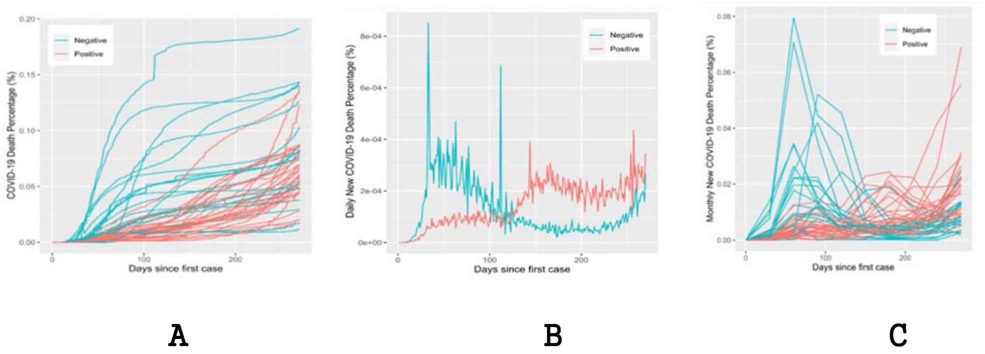


Figure 1. COVID-19 death trends marked by 2^{nd} order polynomial sign. (**a**) Cumulative COVID-19 death percentage in 50 States. (**b**) Daily new COVID-19 death percentage in two sign groups. (c) Monthly new death percentage in 50 States.

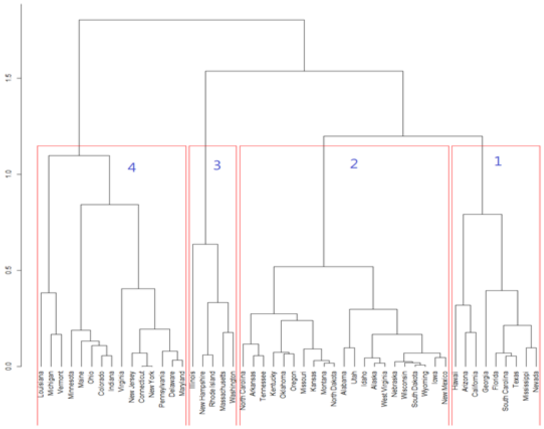


Figure 2. Agglomerative Hierarchical Clustering using Correlation as Similarity Measure based on Monthly New Death Percentages

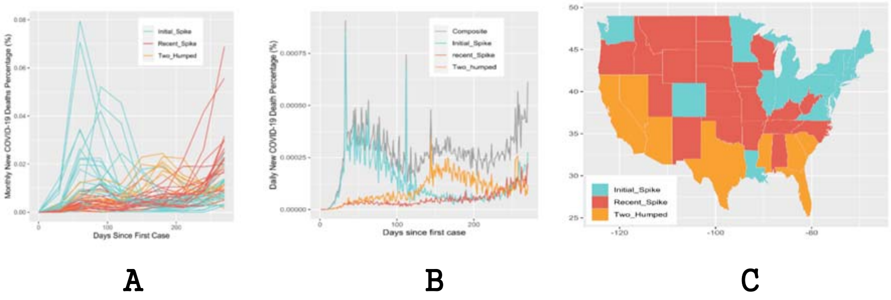


Figure 3. COVID-19 death trends marked by correlation-based clusters. (**a**) Monthly new death percentage in 50 States. (**b**) Daily new COVID-19 death percentage in correlation-based clusters. (**c**) Correlation-based clusters on US map.

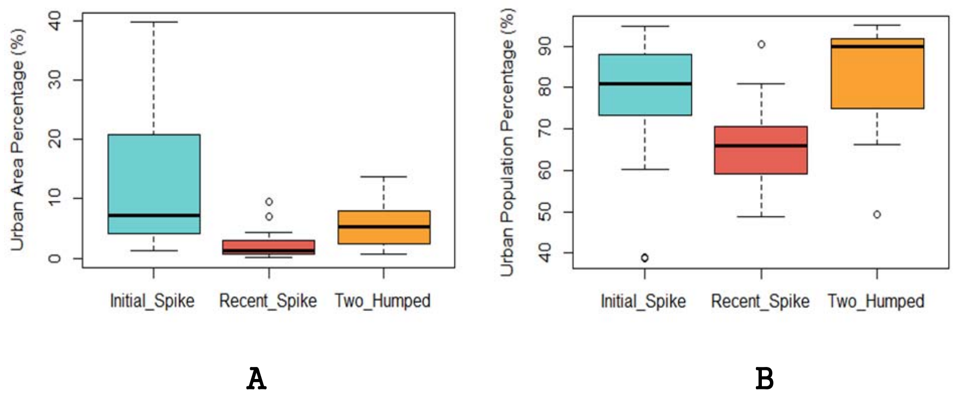


Figure 4. Box plots of urban area and population indices in correlation-based clusters. (**a**) Box plot of urban area percentage on 270th Days since first case. (**b**) Box plot of urban population percentage on 270th Days since first case.

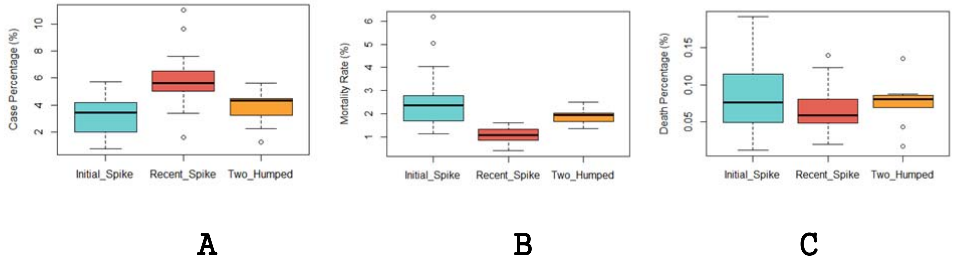


Figure 5. Plots of COVID-19 severity measures in correlation-based clusters. (**a**) Box plot of cumulative case percentage on the 270th Days since first case. (**b**) Box plot of mortality rate on the 270th days since first case. (**c**) Box plot of cumulative death percentage on the 270th days since first case.

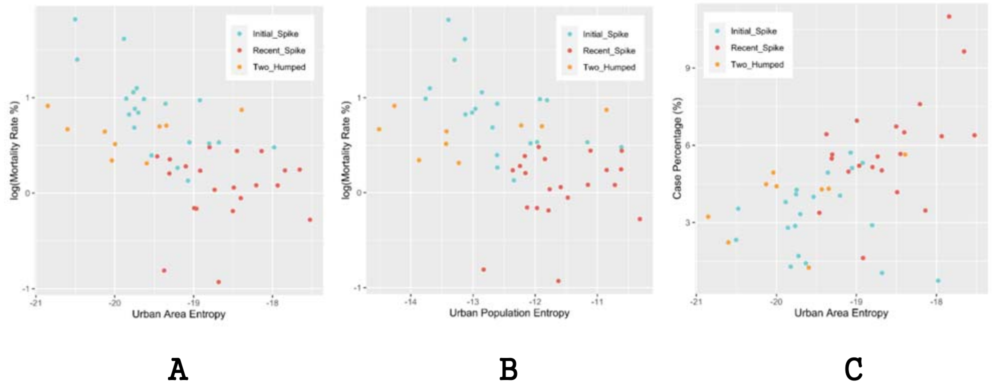


Figure 6. Scatterplots of urban entropy and COVID-19 severity measures on 270th days since first case. (**a**) Urban area entropy and log mortality rate. (**b**) Urban population entropy and log mortality rate. (**c**) Urban area entropy and case percentage.

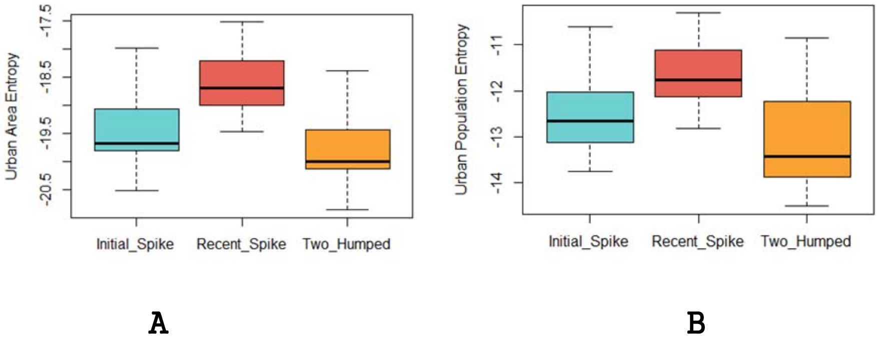
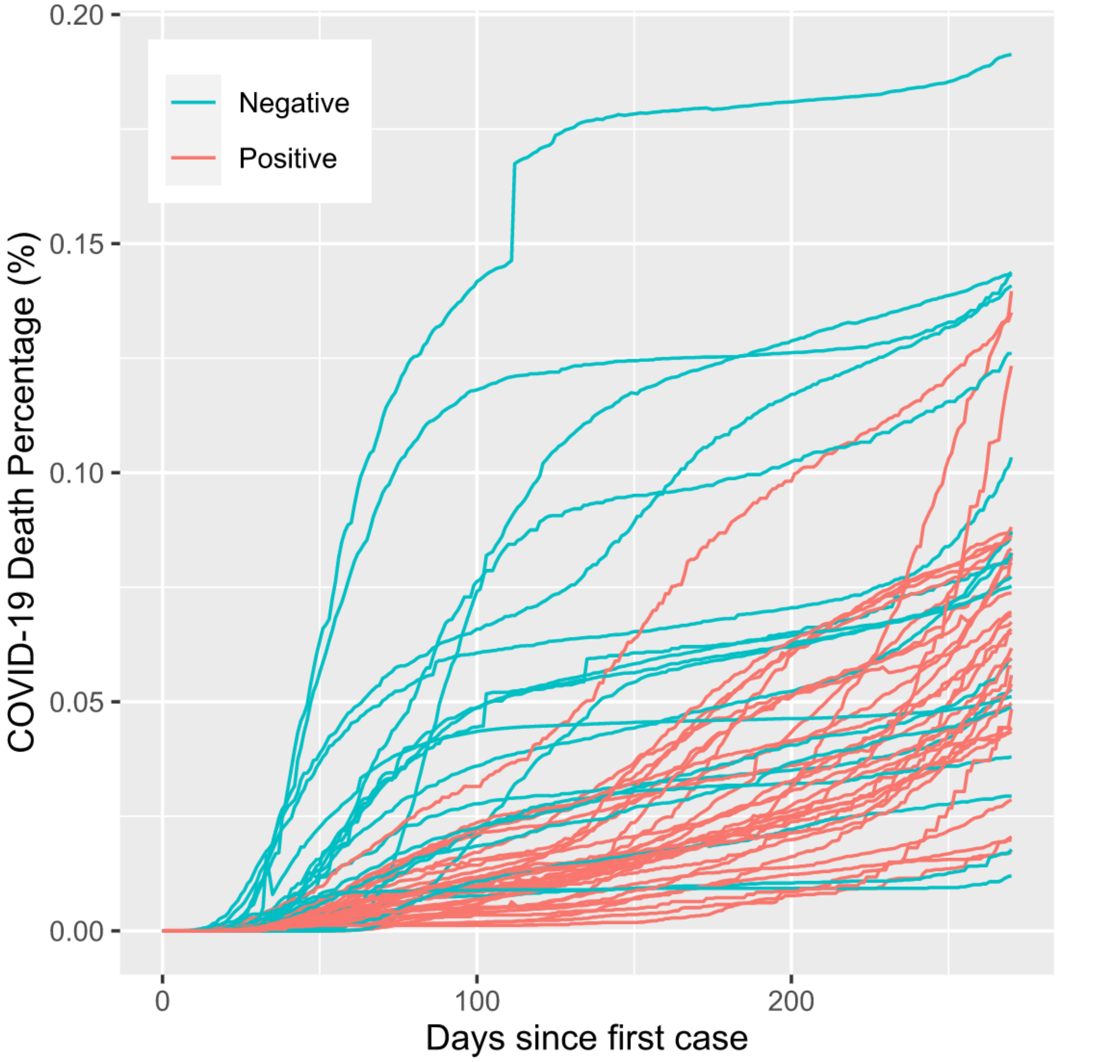


Figure 7. Box plots of urban area and population entropy in correlation-based clusters. (**a**) Distributions of urban area entropy. (**b**) Distributions of urban population entropy.

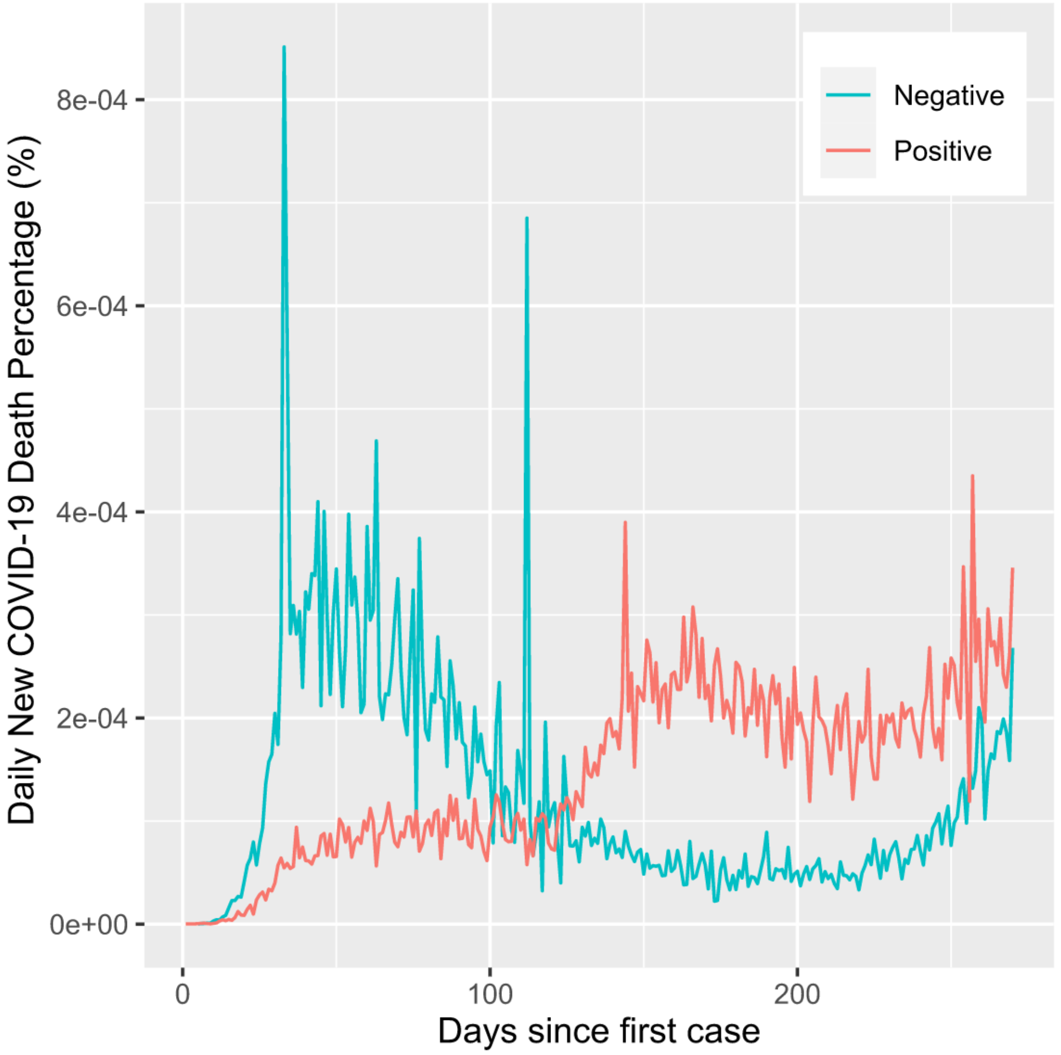
Table 1: Summary of Simple Regression Analysis

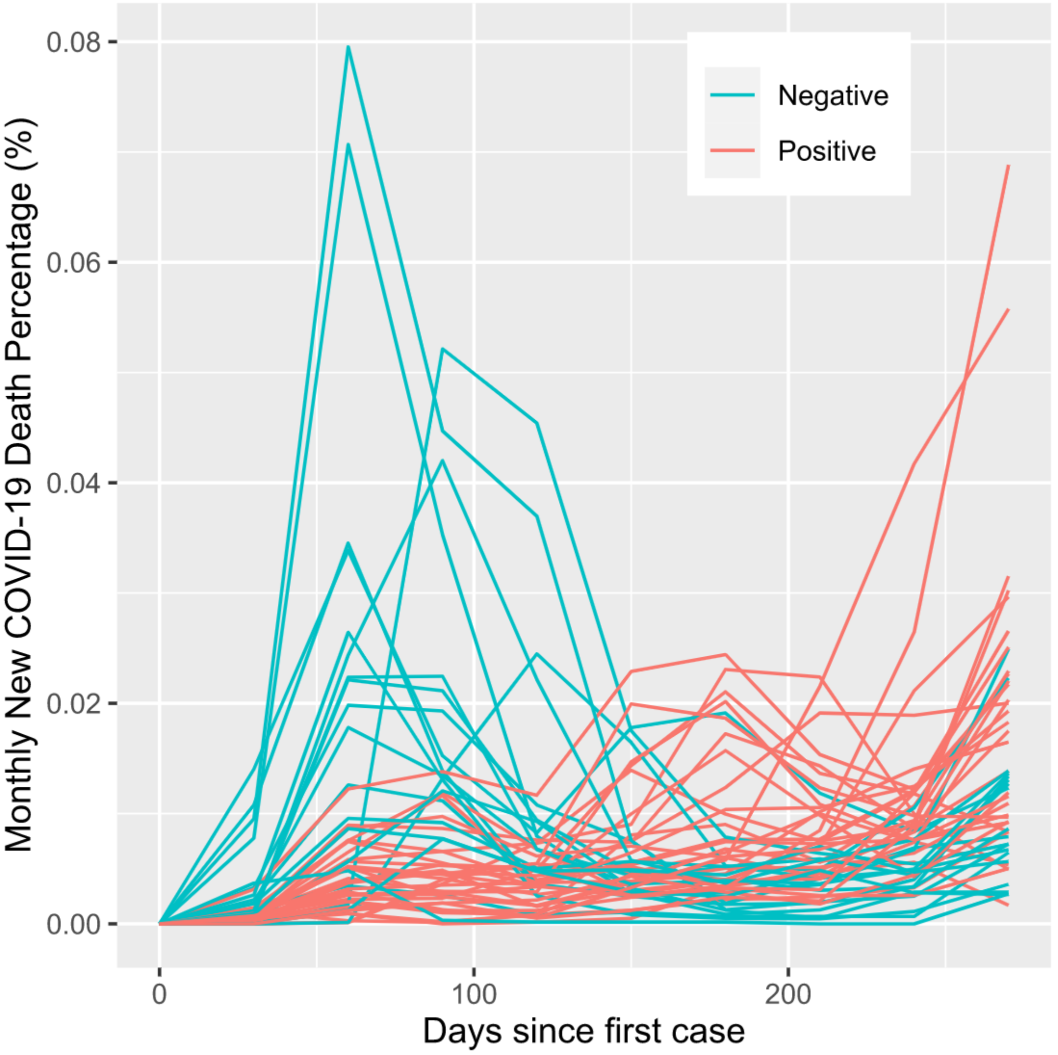
	Natural Log Mortality Rate	Natural Log Death Percentage	Cases Percentage	Correlation Clusters
Entropy based on Urban Area	LM $R^2 = 0.3183$ p = 1.98e-05 $r_s = -0.62$	LM $R^2 = 0.0275$ p = 0.25 $r_s = -0.16$	LM $R^2 = 0.2753$ p = 9.18e-05 $r_s = 0.53$	LM* $R^2 = 0.4078$ p = 4.493e-06
Entropy Based on Urban Population	LM $R^2 = 0.1936$ p = 0.00139 $r_s = -0.50$	LM $R^2 = 0.0099$ p = 0.492 $r_s = -0.083$	LM $R^2 = 0.1931$ p = 0.00083 $r_s = 0.49$	LM* $R^2 = 0.3484$ p = 4.247e-05

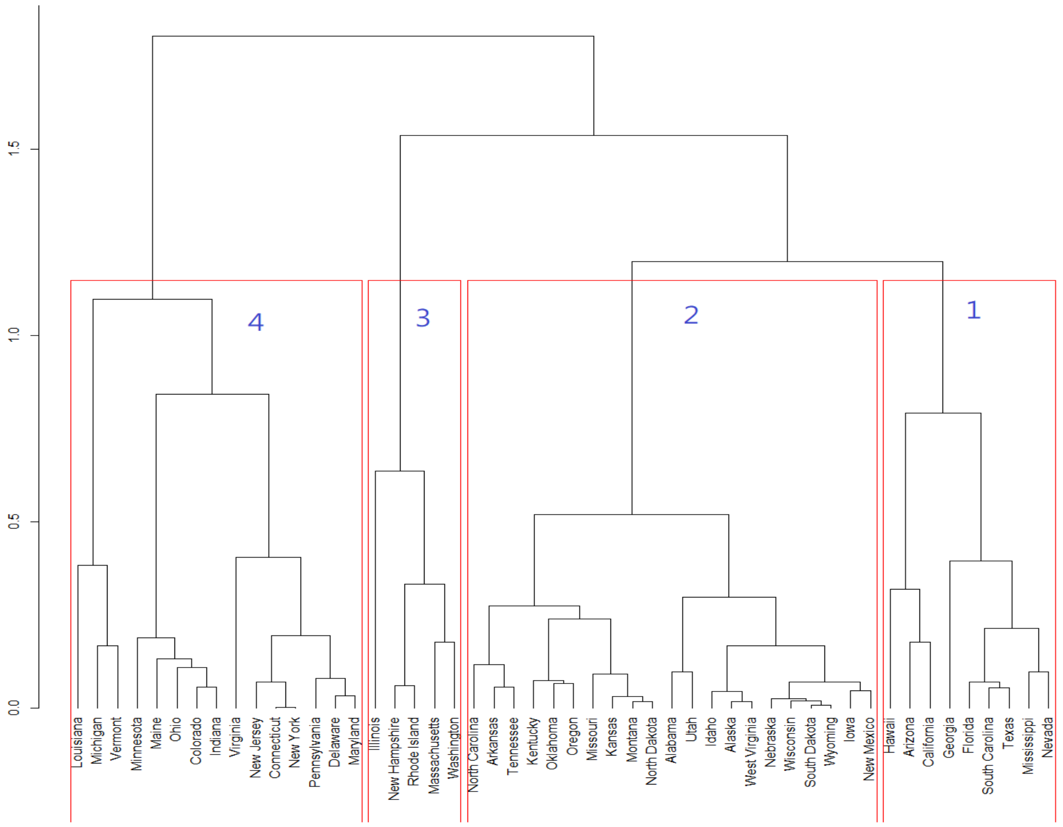
LM: Entropy as explanatory variable; LM*: entropy as response

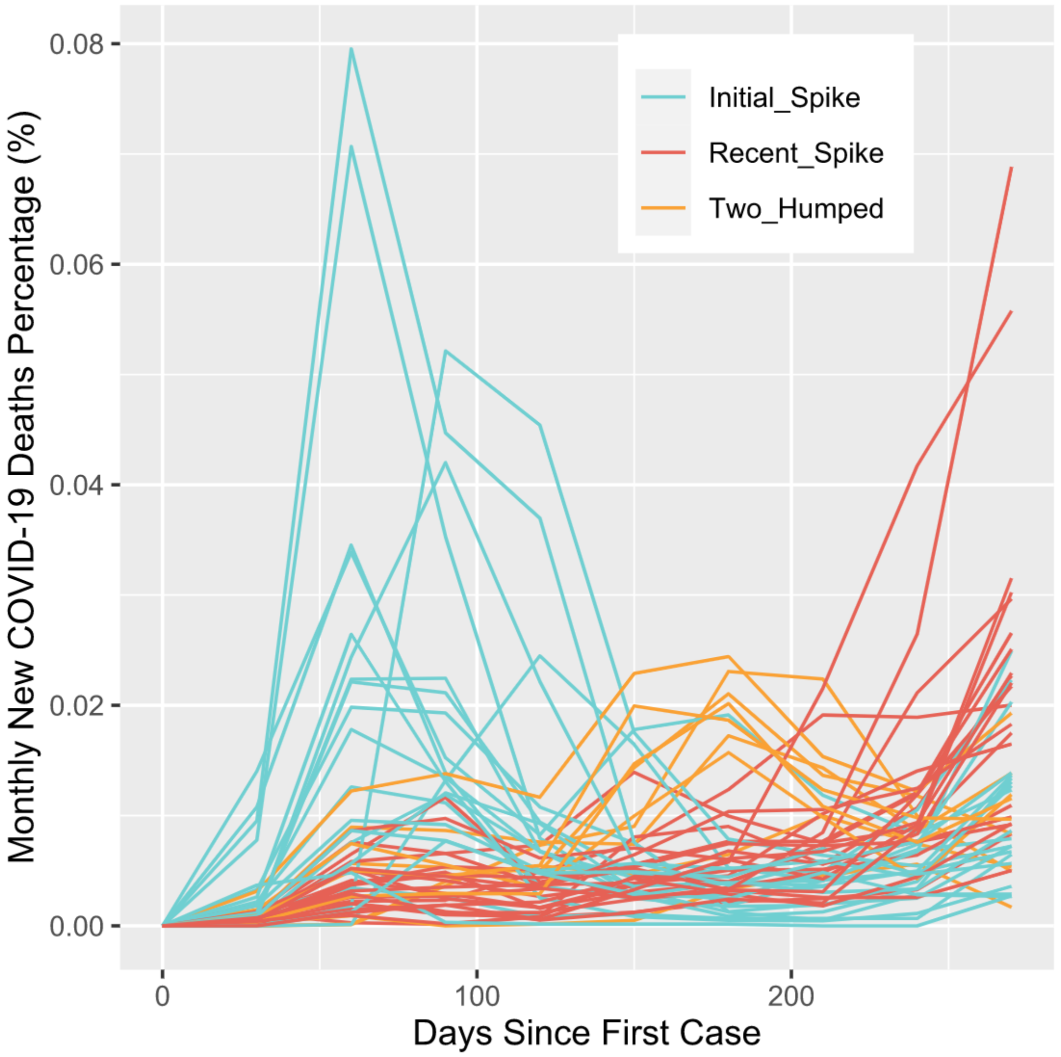


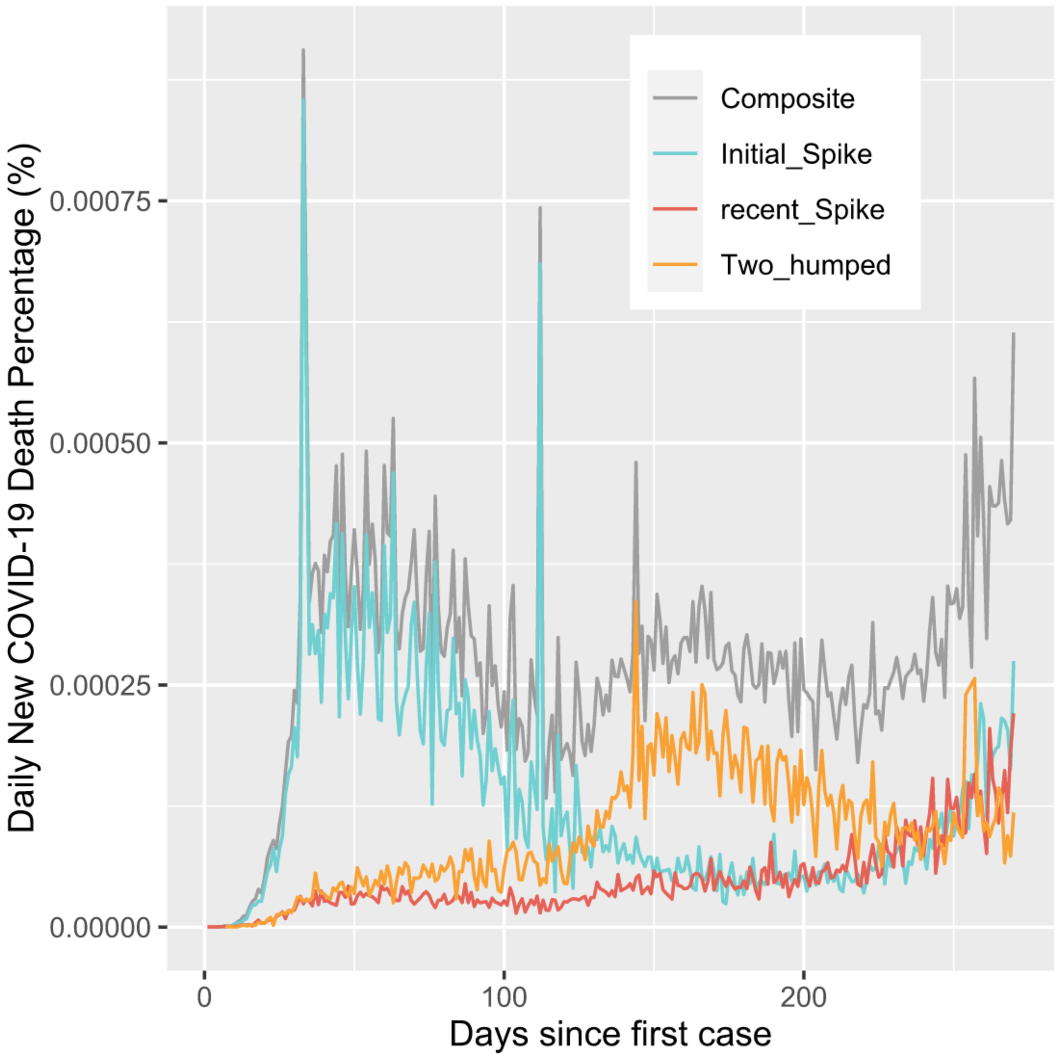
COVID-19 Death Percentage (%)
0.20
0.15
0.10
0.05
0.00
0
100
200
Days since first case
Negative
Positive

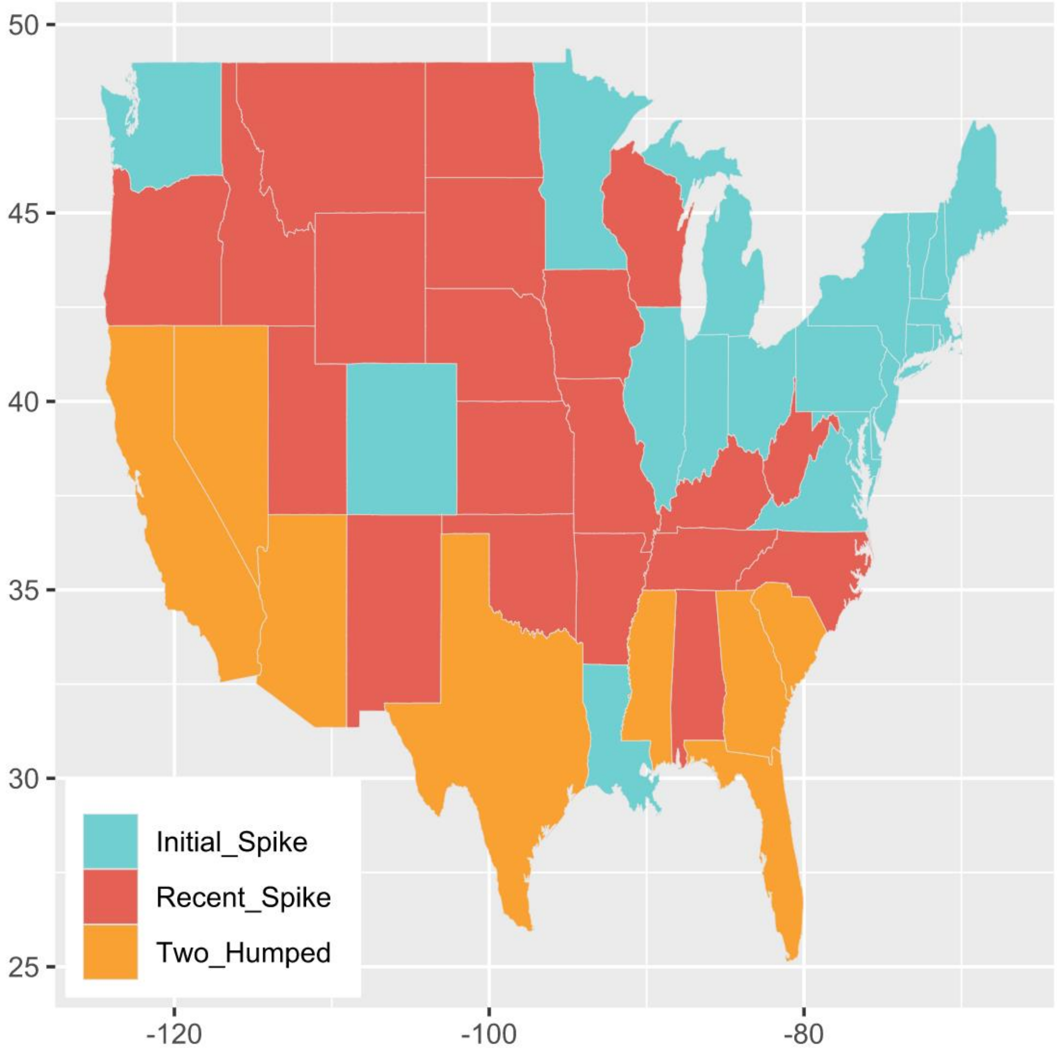

Negative
Positive
Daily New COVID-19 Death Percentage (%)
8e-04
6e-04
4e-04
2e-04
0e+00
0
100
200
Days since first case

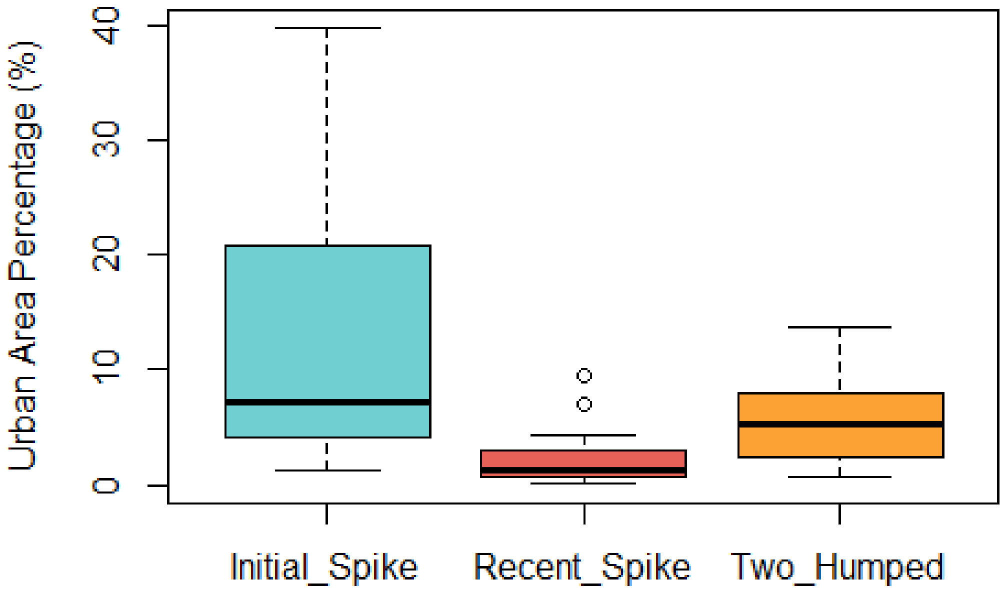

Monthly New COVID-19 Death Percentage (%)
0.08
0.06
0.04
0.02
0.00
0
100
200
Days since first case
Negative
Positive

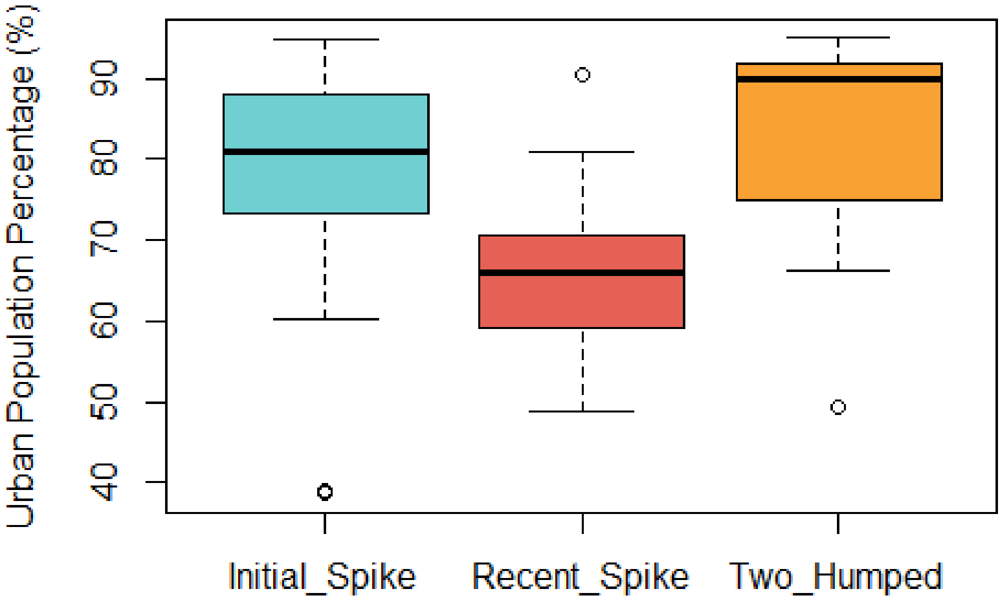

1.5
1.0
0.5
0.0
4
3
2
1
Louisiana
Michigan
Vermont
Minnesota
Maine
Ohio
Colorado
Indiana
Virginia
New Jersey
Connecticut
New York
Pennsylvania
Delaware
Maryland
Illinois
New Hampshire
Rhode Island
Massachusetts
Washington
North Carolina
Arkansas
Tennessee
Kentucky
Oklahoma
Oregon
Missouri
Kansas
Montana
North Dakota
Alabama
Utah
Idaho
Alaska
West Virginia
Nebraska
Wisconsin
South Dakota
Wyoming
Iowa
New Mexico
Hawaii
Arizona
California
Georgia
Florida
South Carolina
Texas
Mississippi
Nevada

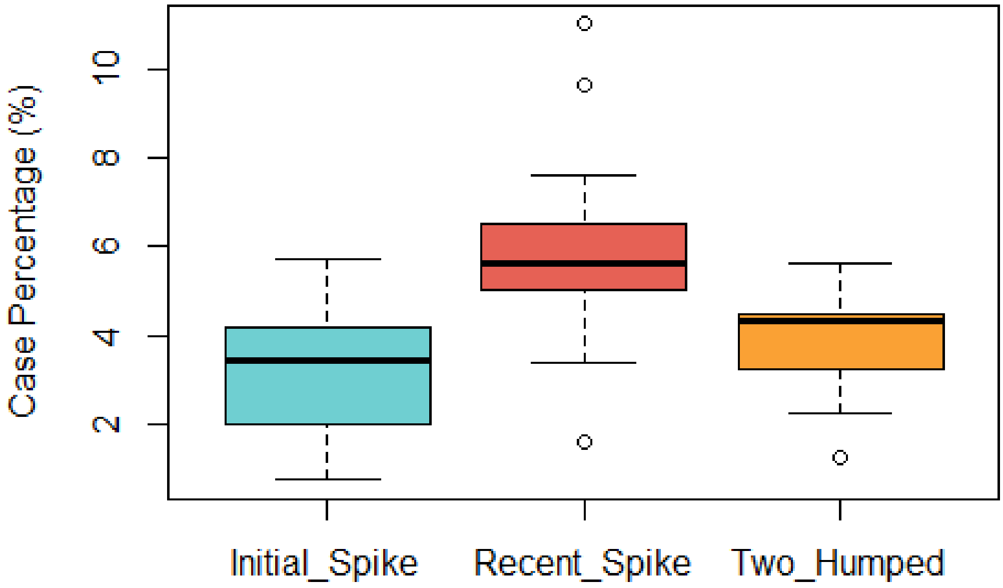

0.08
0.06
0.04
0.02
0.00
Monthly New COVID-19 Deaths Percentage (%)
0
100
200
Days Since First Case
Initial_Spike
Recent_Spike
Two_Humped

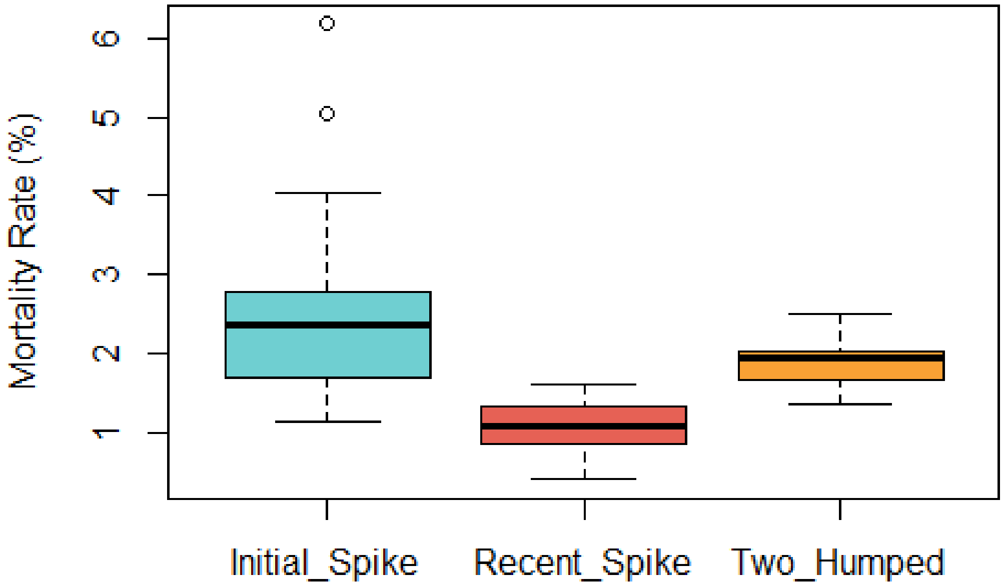

Composite
Initial_Spike
recent_Spike
Two_humped
Daily New COVID-19 Death Percentage (%)
0.00075
0.00050
0.00025
0.00000
0
100
200
Days since first case

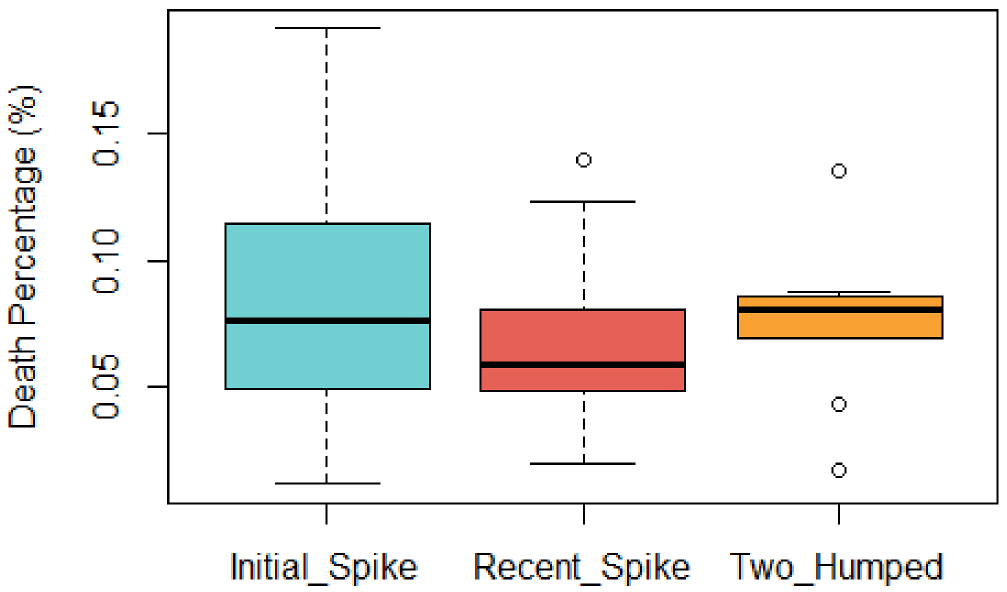

50
45
40
35
30
25
-120
-100
-80
Initial_Spike
Recent_Spike
Two_Humped

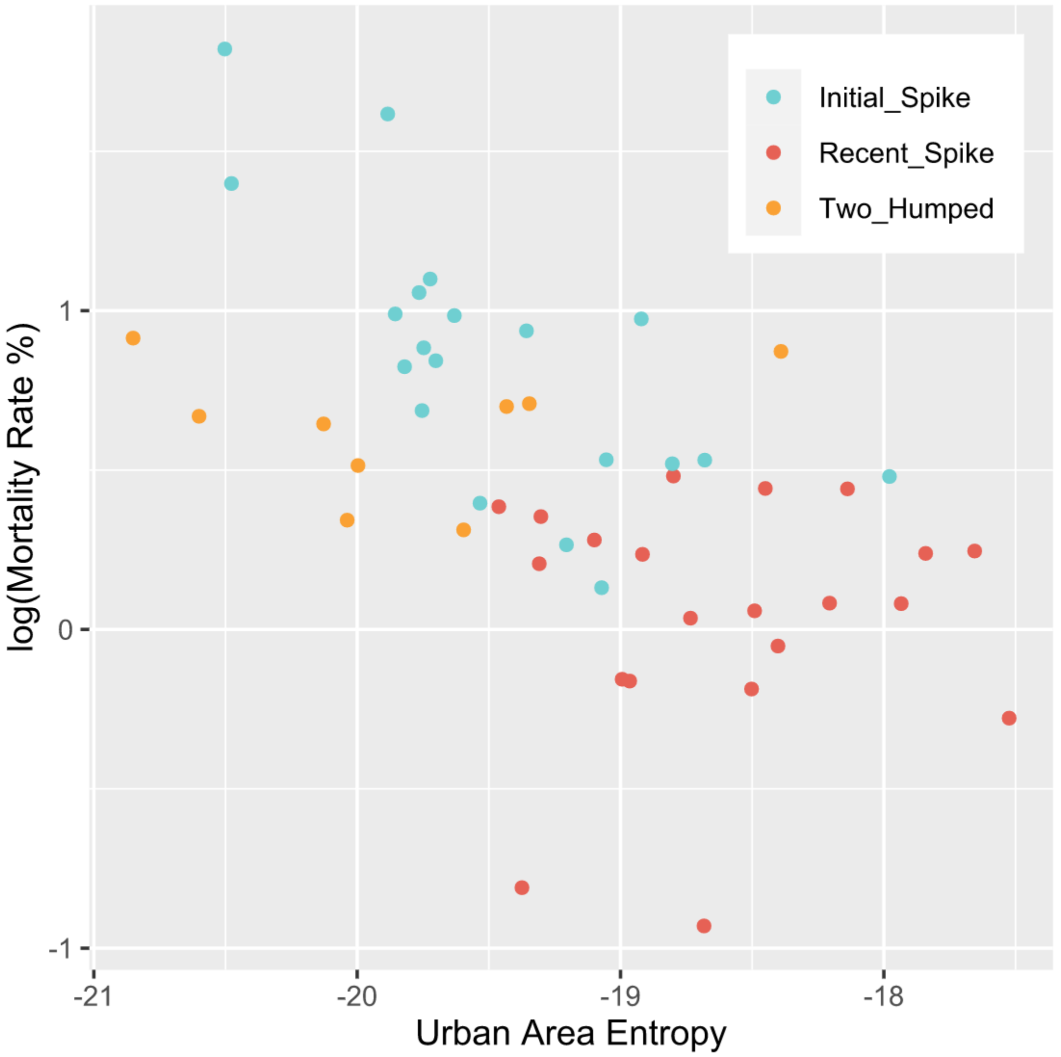

Urban Area Percentage (%)
0
10
20
30
40
Initial_Spike
Recent_Spike
Two_Humped

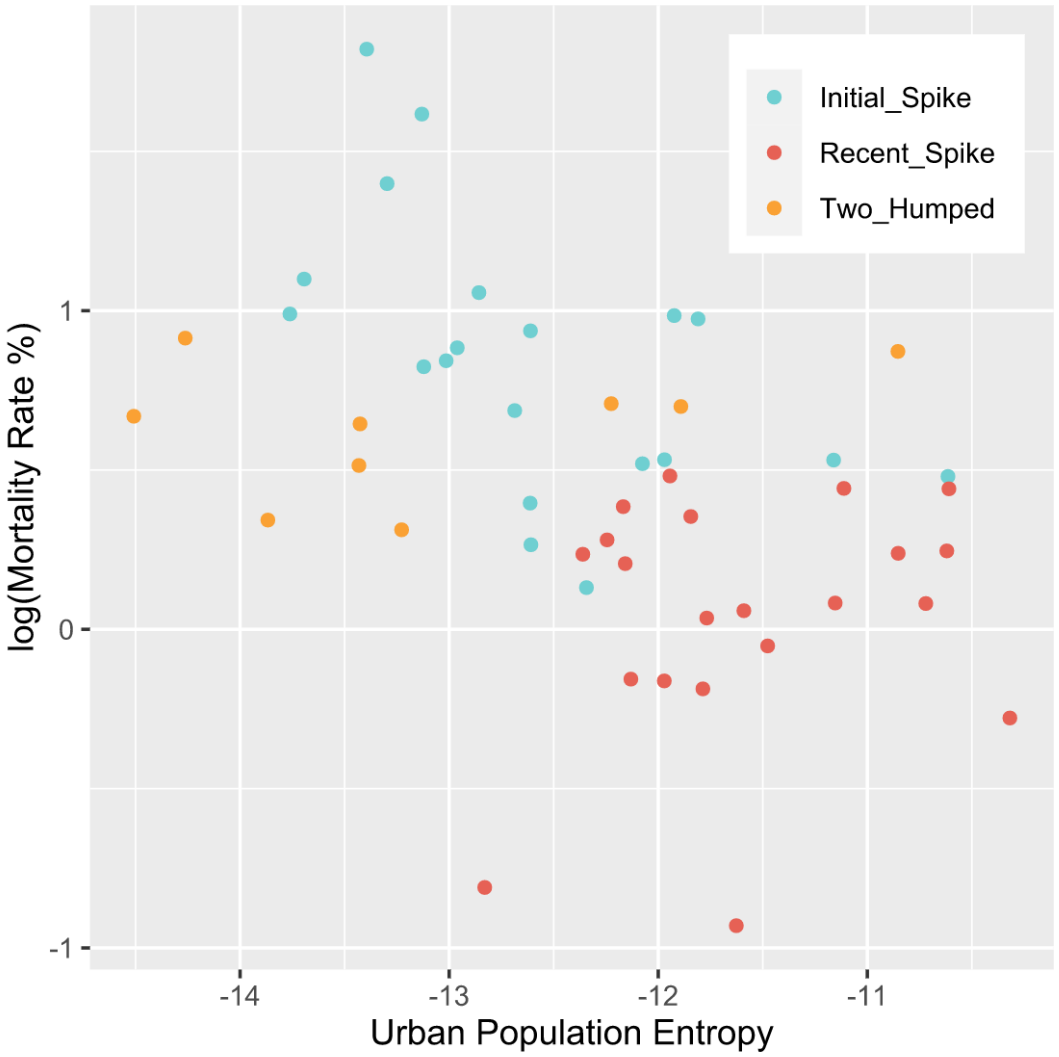

Urban Population Percentage (%)
40
50
60
70
80
90
Initial_Spike
Recent_Spike
Two_Humped


Case Percentage (%)
10
8
6
4
2
Initial_Spike
Recent_Spike
Two_Humped

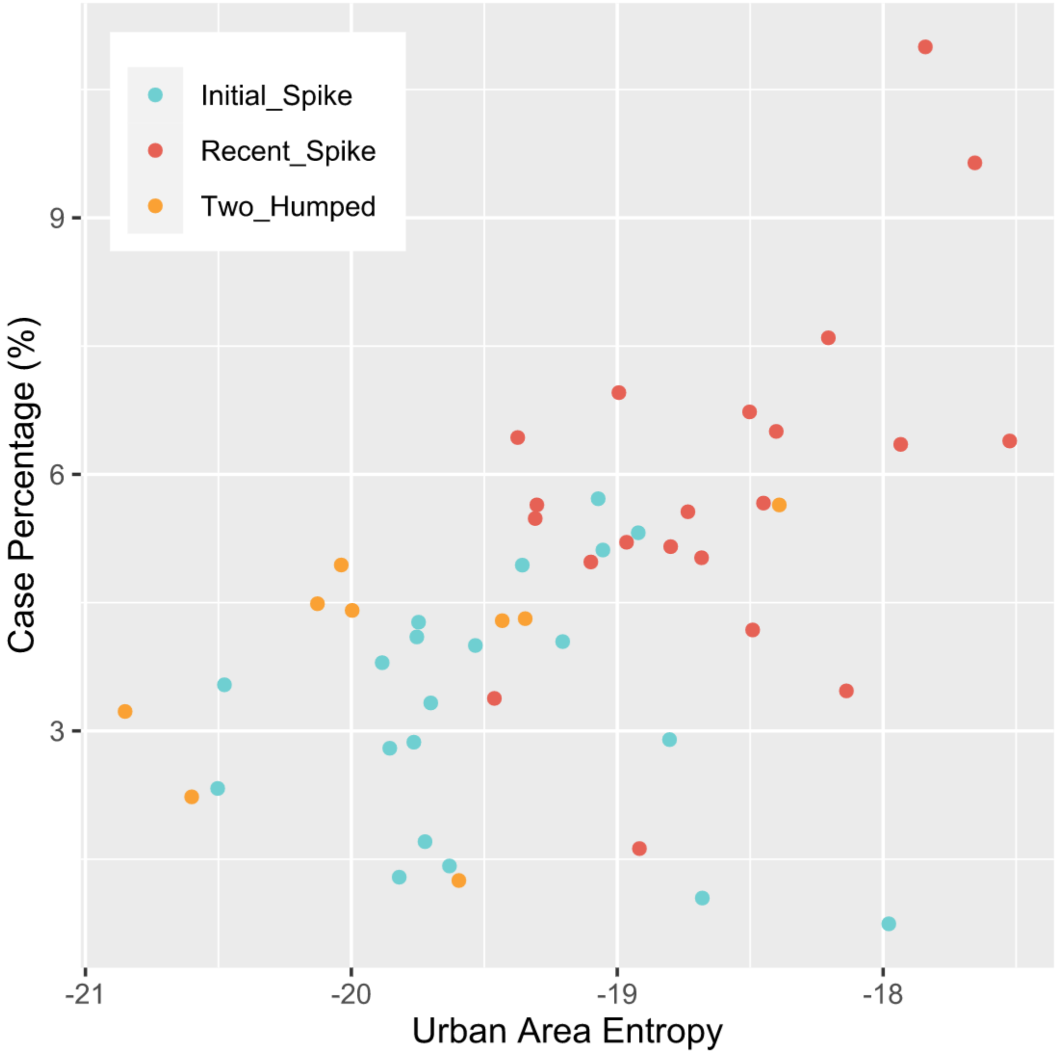

Mortality Rate (%)
6
5
4
3
2
1
Initial_Spike
Recent_Spike
Two_Humped

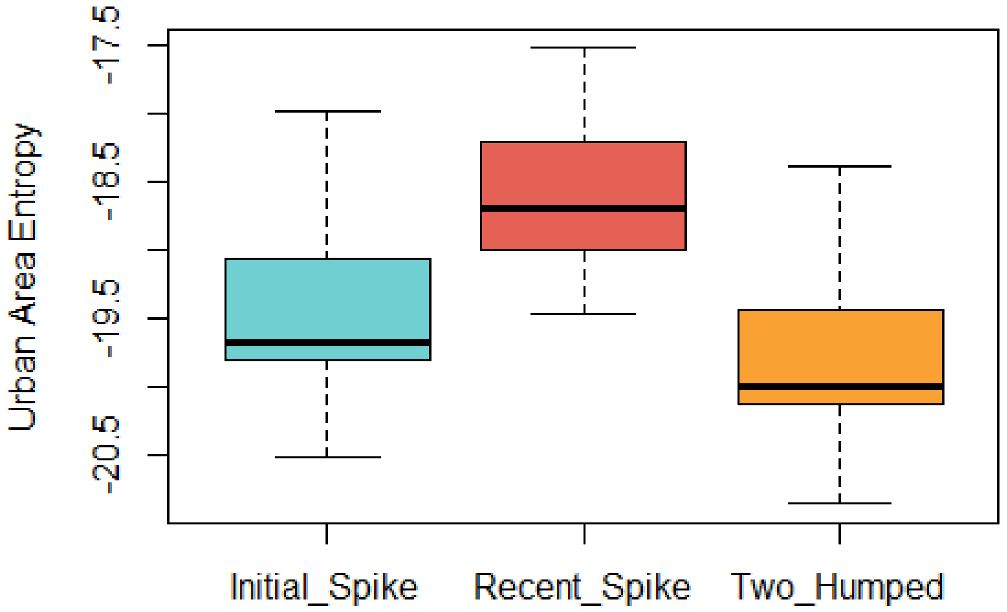

Death Percentage (%)
0.05
0.10
0.15
Initial_Spike
Recent_Spike
Two_Humped

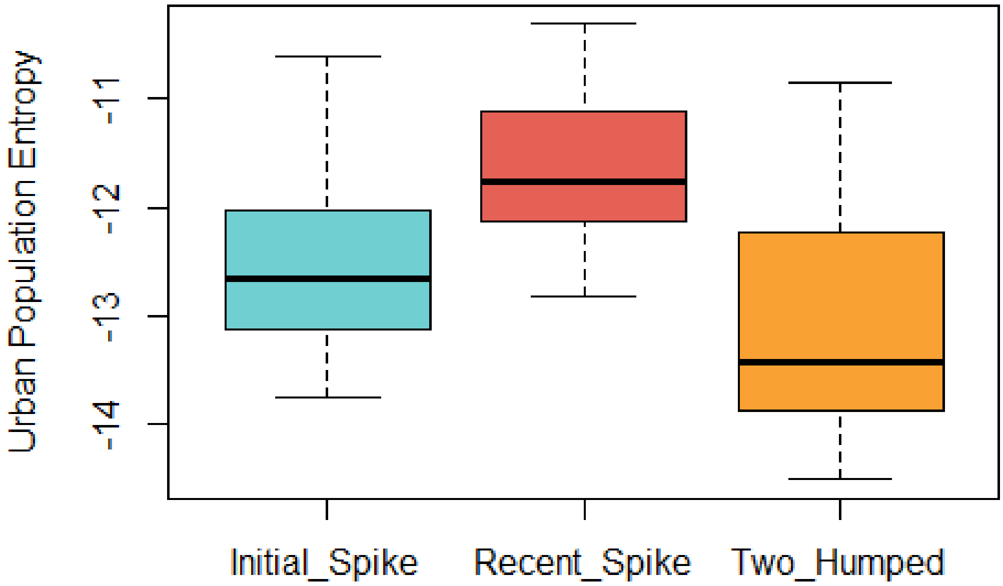

Initial_Spike
Recent_Spike
Two_Humped
log(Mortality Rate %)
1
0
-1
-21
-20
-19
-18
Urban Area Entropy



Initial_Spike
Recent_Spike
Two_Humped
1
0
-1
-14
-13
-12
-11
log(Mortality Rate %)
Urban Population Entropy


Initial_Spike
Recent_Spike
Two_Humped
Case Percentage (%)
9
6
3
-21
-20
-19
-18
Urban Area Entropy


Urban Area Entropy
-17.5
-18.5
-19.5
-20.5
Initial_Spike
Recent_Spike
Two_Humped


Urban Population Entropy
-11
-12
-13
-14
Initial_Spike
Recent_Spike
Two_Humped