The Spatiotemporal Interaction Effect of COVID-19 Transmission in the United States

Lingbo Liu
Wuhan University https://orcid.org/0000-0002-9876-8506
Tao Hu
Harvard University
Shuming Bao
China Data Institute
Hao Wu (wh79@whu.edu.cn)
Wuhan University https://orcid.org/0000-0001-7107-8081
Zhenghong Peng
Wuhan University
Ru Wang
Wuhan University

Research Article

Keywords: COVID-19, Moran's I index, K-means clustering, Spatiotemporal interaction effects, Spatial Lag Model, SIR

Posted Date: January 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-143786/v1

The Spatiotemporal Interaction Effect of COVID-19 Transmission in the United States

Author List

Lingbo Liu, School of Urban Design, Wuhan University, Wuhan, Hubei, China; lingbo.liu@whu.edu.cn;

Tao Hu, Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA; taohu@fas.harvard.edu

Shuming Bao, China Data Institute, Ann Arbor, MI, US; sbao@umich.edu

***Hao Wu**, School of Urban Design, Wuhan University, Wuhan, Hubei, China; wh79@whu.edu.cn

Zhenghong Peng, School of Urban Design, Wuhan University, Wuhan, Hubei, China; pengzhenghong@whu.edu.cn;

Ru Wang, School of Urban Design, Wuhan University, Wuhan, Hubei, China; wang_ru@whu.edu.cn;

*Corresponding author

Abstract

Background: Human mobility among geographic units is a possible cause of the widespread transmission of COVID-19 across regions. Due to the pressure of epidemic control and economic recovery, the states of the United States have adopted different policies for mobility limitations. Assessing the impact of these policies on the spatiotemporal interaction of COVID-19 transmission among counties in each state is critical to formulating the epidemic policies.

Methods: The study utilized Moran's I index and K-means clustering to investigate the time-varying spatial autocorrelation effect of 49 states (except the District of Colombia) with the daily new cases at the county level from Jan 22, 2020, to August 20, 2020. Based on the dynamic spatial lag model (SLM) and the SIR model with unreported infection rate (SIRu), the integrated SLM-SIRu model was constructed to estimate the inter-county spatiotemporal interaction coefficient of daily new cases in each state, which was further explored by Pearson correlation and stepwise OLS regression with socioeconomic factors.

Results: The K-means clustering divided the time-varying spatial autocorrelation curves of 49 states into four types: continuous increasing, fluctuating increasing, weak positive, and weak negative. The Pearson correlation analysis showed that the spatiotemporal interaction coefficients in each state estimated by SLM-SIRu were significantly positively correlated with median age, population density, and the proportion of international immigrants and the highly educated population, but negatively correlated with the birth rate. The voting rate for Donald Trump in the 2016 U.S. presidential election showed a weak negative correlation. Further stepwise OLS regression retained only three positive correlated variables: poverty rate, population density, and the highly educated population proportion.

Interpretation: This result suggests that various state policies in the U.S. have imposed different impacts on COVID-19 transmission among counties. All states should provide more protection and support for the low-income population, high-density populated states need to strengthen regional mobility restrictions, and the highly educated population should reduce unnecessary regional movement and strengthen self-protection.

Keyword:

COVID-19; Moran's I index; K-means clustering; Spatiotemporal interaction effects; Spatial Lag Model; SIR;

1 Introduction

COVID-19 is still rampaging around the world[1, 2], showing obvious spatial differences in global geographic distribution[3]. Countries with low incomes, incomplete health care capabilities and demographics with a large proportion of the elderly population have been facing the challenges of more serious disease output and health care burdens[4, 5]. However, the United States, as the country with the most developed economy and the highest level of medical care, has the largest number of infections and shows geographic differences in COVID-19 transmission, which has become an important global health research issue for pandemic control.

The spatial heterogeneity in the spread of infectious diseases comes from the social, economic, and environmental differences of the geospatial unit itself[6]. Compared with the potential climate correlations implied by some studies [7, 8], more studies indicate that population density[9], health measures, and mobility restrictions[10]have a greater impact on the spread of COVID-19. Wherein, mobility and connectivity[11], other than population density [12], mainly influence the pandemic transmission more in term of the spatial differences, which is also supported by related research based on US county daily commute data[13]and mobility data of Boston [14], consisted with the research on Italy's industrial spatial structure and epidemic distribution[15].

Inter-regional population movement is the main reason for the extensive spread of COVID-19 across regions [15]. Spatial distancing is considered to be the most effective way [16, 17], such a method has also been verified effective in both China[18, 19] and Europe[20, 21]. Due to the trade-off of epidemic control and economic recovery, the states in the United States have adopted different spatial regulatory policies at different stages to control regional

mobility. Assessing the spatiotemporal interaction of the inter-county spread of COVID-19 in the states is critical to the optimizing of the epidemic policy.

However，there are two main obstacles in the measurement of spatiotemporal interaction faces, appropriate model and data depression. Spatial interaction can be captured by a variety of methods, such as Geographic Weighted Regression (GWR) [22], Geographically Weighted Principal Component Analysis (GWPCA) [23], and Spatial Panel Models (SPD) which includes Spatial Lag Model (SLM), Spatial Error Model (SEM), Spatial Dubin Model (SDM) [24]. The exploration of the spatial interaction effect of COVID-19 transmission in previous studies faces two challenges. On the one hand, current studies mostly use static SPM with cross-sectional data [25], without considering the long-time effect on spatial interaction. On the other hand, related studies, often using infection and socioeconomic data as dependent and explanatory variables [26], ignoring the fact that the increase of infection is mainly driven by the values of infection variables in the Susceptible-Infective-Removal model (SIR) during previous time state.

Data suppression is another critical problem while applied by the spatial correlation model and SIR model [27]. The COVID-19 infection data officially released may be bias due to potential unreported infectives [28-30]. Studies have shown that asymptomatic infections and mildly infected people may be fully reported, resulting in an underestimate of infection data [31-33]. More importantly, the county-level data in the United States is not released, which may limit the reliability of the traditional SIR model. A SIR model integrated with unreported infection rate (SIRu) was recently proposed, providing an effective supplement to miscalculated data [34].

This paper thus proposed an integrated model of SLM and SIRu (SLM-SIRu) to calculate the inter-county spatial interaction effect of daily new cases in each state based on the county-level US COVID-19 data. As the uneven spatial correlation may derive from the inequity in the spatial units in terms of socioeconomic features [35, 36], the connection has been further explored between spatial effect and socioeconomic elements in each state, such as population density, income, political elements. The spatiotemporal correlation has also been tested by Morans' I index, which has been further used to identify spatial clusters before the appliance of the SLM-SIRu model.

This study provides an integrated model to capture the spatiotemporal interaction effect, which may help assess the impact of cross-regional mobility on epidemic transmission and assist the government in optimize COVID-19 control policy.

2 Methodology

The workflow can be divided into three parts: (1) spatial autocorrelation analysis of daily new infection; (2) spatial effect exploration based on SIRu model and Spatial Lag Model; (3) correlation analysis of spatial effect and socio-economic factors (Figure 1).

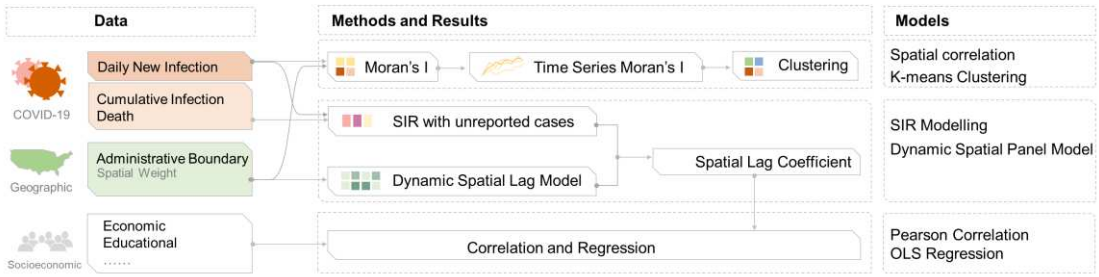


Figure 1 Research Workflow.

The research workflow explains the data, methods, and models used in the article. Moran's I and K-means clustering were introduced to capture the spatiotemporal feature of COVID-19daily new case changes in the US, then the spatial lag effect underlying the SIR model was further estimated by SLM-SIRu model and further used for correlation test with socio-economic variables.

2.1 Global Moran's *I* index and K-means Clustering

Except for extremely strict spatial restriction policies, any potential inter-county human movement between neighboring counties in a certain state may increase the contact chance, and then affect the amount of daily new infections in each geographic unit. Such a pattern of spatial interaction in the state could be defined as spatial autocorrelation, which could be calculated by the global Moran's *I* index based on spatial weights. The value of Moran's *I* is ranging from -1 to 1, -1 represents a negative spatial correlation, 0 is random, and 1 is a positive correlation.

For a certain attribute *x* of geographic units, the general formula of global Moran's *I* is

$$I = \frac{n}{W} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (1)$$

wherein, x_i is the attribute value of the i_{th} county in a state, and $\bar{x}$ is the mean value of all x_i in the state, x_j represents the value of the j_{th} county near to the i_{th} county. n is the total number of counties, and W is the sum of spatial weights w_{ij}.

The spatial weight W could be calculated by a pre-defined neighboring pattern such as Queen, Rook, K-near, or distances. Here the K-near algorithm was adopted with a max neighboring number of 4. As there is only one unit in the District of Columbia, so the calculation could not be applied.

If the total infections or daily new infections of a state are zero, making the calculation of Equation 1 unapplicable, a value of zero will be designated to Moran's I value of the corresponding state. Such a situation is also applied when the p-value of Moran's I calculation result is not significant (P>0.05).

2.2 K-means Clustering Algorithm

At different stages of COVID-19 transmission, the population movement among adjacent spatial units may change with the state's control policies or the epidemic situation, resulting in Moran's *I* index varying over time. The characteristics of time series Moran's *I* index may reflect the spatial homogeneity and heterogeneity among the states in the United States, which could be explored by clustering algorithms.

The K-means clustering algorithm is applied, as it can generate the Guarantees convergence and Relatively simple to implement, comparing to DBSCAN, GMM. The optimal group number is determined by the minimum value of AIC.

2.3 SIR with unreported infections

In the classic SIR model, the daily new infections (I_n) can be expressed as the product of the infected population(I), the susceptible population(S), the total population(N), and the transmission rate (β):

$$I_n = \beta SI\ /N \quad (2)$$

However, the official data cannot be directly used, as there may exist unreported infections. Moreover, the actual population of recovered patients has not been released in the county-level data, the parameter of I in Equation 2 could not be calculated directly. A SIR model integrated with unreported infections (SIRu) is proposed, adding two more parameters of φ and τ, wherein, φ is the average unreported/ reported rate of infections (UIR), and τ is the recovery/death rate (RDR). Equation 2 could be revised as:

$$\varphi I_n = \beta(N-\varphi I_c)(\varphi I_c - \tau R_d)/N \quad (3)$$

Wherein, I_n ,I_c and R_d is the official released COVID-19 data of daily new infections, cumulative infections, and death respectively, therefore, φI_n, φI_c and τR_d are the corresponding factual data.

A furthermore simplification of Equation 3 can be rewritten as:

$$I_n = \beta I_c - \frac{\beta\tau}{\varphi} R_d - \beta\varphi \frac{I_c^2}{N} + \beta\tau \frac{I_c R_d}{N} \quad (4)$$

Such an equation could be seen as a linear regression of variables of I_n, I_c, R_d, I_c^2/N, I_cR_d/N.

2.4 Dynamic Spatial lag model

Though dynamic spatial panel models consist of Spatial Lag Model (SLM), Spatial Error Model (SEM), and Spatial Dubin Model (SDM), the SLM is adapted to capture the spatial effect of the SIRu model.

The classic SLM model could be described as:

$$y = \lambda W y' + ax + \varepsilon \quad (5)$$

Wherein, y is the dependent variable of a certain geographic unit, y' represents the values of the adjacent geographic units, and x denotes the corresponding explanatory variables. W is the spatial weight, λ is the spatial lag coefficient.

The SLM-SIRu model can be combined by substituting Equation 4 into Equation 5:

$$I_n = \lambda W I_n + \beta I_c - \frac{\beta\tau}{\varphi} R_d - \beta\varphi \frac{I_c^2}{N} + \beta\tau \frac{I_c R_d}{N} \quad (6)$$

The COVID-19 data and the corresponding spatial weights at the county level were applied in Equation 6 to calculate the spatial coefficient λ, which would be taken as the dependent variable in the further correlation exploration with socioeconomic data.

2.5 Data

The data used in the article contain Covid-19 data, Geographic data, and Socioeconomic data.

(1) Covid-19 data of all counties in the United States from January 22, 2020, to August 20, 2020, including daily new infections, cumulative infections, deaths, and total population. The data was acquired from the GitHub repository of Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19).

(2) Geographic data of all counties and the states in the United States. The data is available in the Harvard Dataverse (https://dataverse.harvard.edu/dataverse/cdl_dataverse).

(3) Socioeconomic data of all states in the United States in 2019, including factors such as birth rate, death rate, international immigration rate, poverty rate, median age, average education level (high school graduation rate, undergraduate graduation rate, advanced education rate), population density. Such data is available on the US Census website (https://www.census.gov). Trump's vote rate in the 2016 U.S. presidential election in the states was also included in the correlation analysis as a potential political variable of residents.

3 Result

3.1 Time series Moran's *I*

The time series Moran's *I* index of each state was calculated by daily new cases and spatial weights at the county level from Jan 22, 2020, to August 20, 2020, which indicated that most states showed significant changes in terms of Moran's *I* index (Figure2). The spatial correlation within each state showed substantial differences over time, for example, New York was keeping a restricted state after the first wave, while Georgia and Illinois were still increasing, Florida has just passed the peak. Such a time series Moran's I may reflect the real effect imposed by spatial restriction policies.

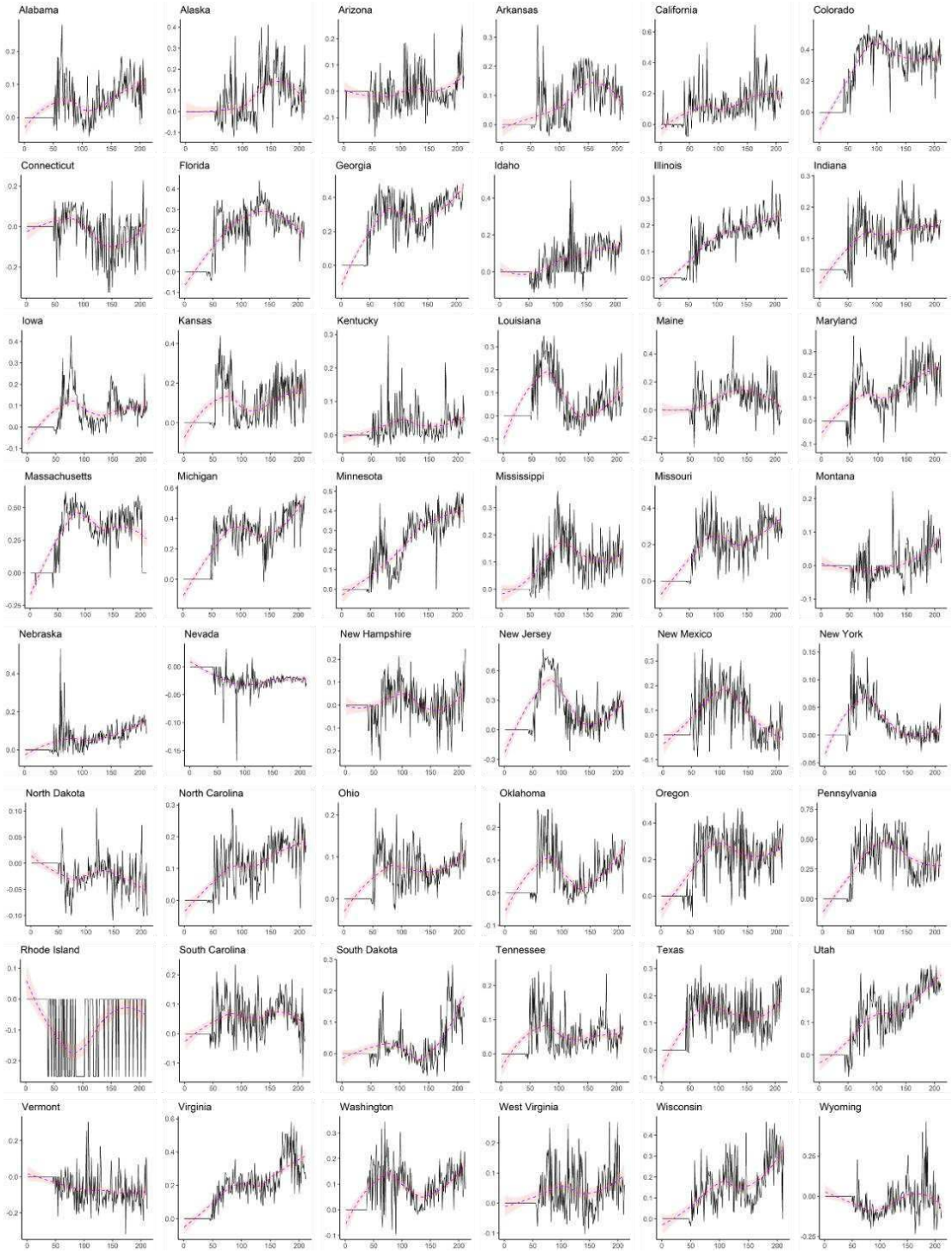


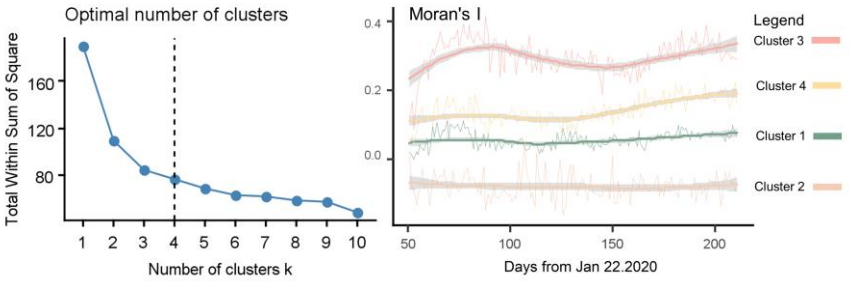
Figure 2 Time series Moran's *I* index of each state in the United States since Jan 22, 2020.

The x-axis and y-axis indicated the value of time and Moran's I value correspondingly, wherein, the red curve represented the fitted value with a 95% confidence interval. Obvious changes from around the 50th day can be observed, displaying several patterns of temporal changes in spatial effects such as reversed U-shape, N-shape.

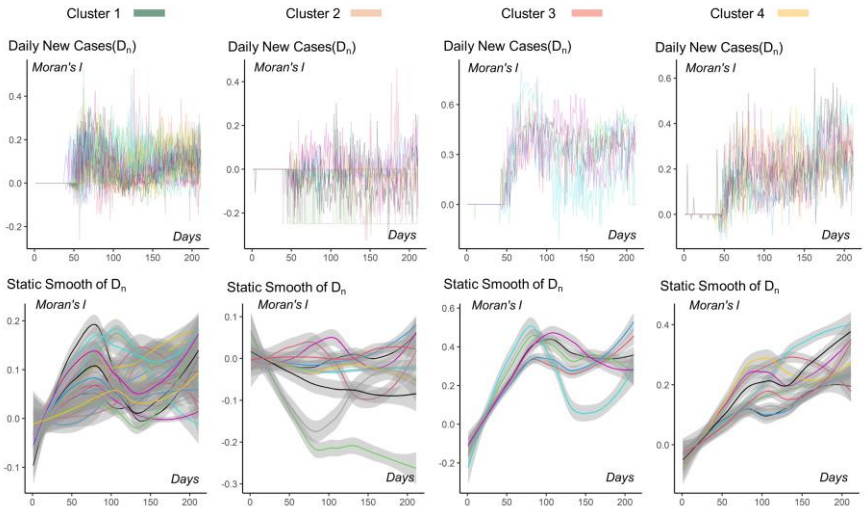
3.2 K-means Clustering of Time series Moran's *I*

The K-means clustering algorithm was further performed based on the time series Moran's *I* index. The results showed that the optimal group number was 4 (Figure 3a), and the four groups could be roughly defined as fluctuating growth, continuous growth, weak positive correlation, and weak negative correlation (Figure 3b). Figure 3c displays

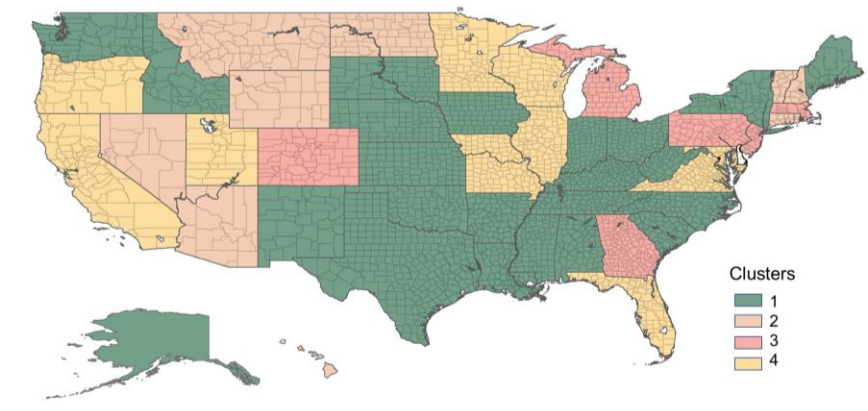
the Moran's *I* values of each group, wherein, the overall fluctuation ranges in cluster 1 is between -0.1-0.4, while the fitting curves and the 95% confidence interval were concentrated between -0.1 and 0.2, indicating a state of weakly positive spatial autocorrelated. Cluster 2 with a value ranging from -0.2 to 0.1, showed a feature of weakly negative spatial autocorrelated. Both Cluster 3 and Cluster 4 exhibited increasing trends, while the former was in a resurging status after the first wave, the latter was in a continuous increasing model with a relatively lower value of Moran's I. Figure 3d illustrating the clusters on the map, the spatial agglomeration could also be observed.



(a) optimal number of clusters (b) Time series Moran's I of 4 Clusters



(c) Time series Moran's I of states in four groups



(d) Spatial distribution of four groups

Figure 3 K-means Clustering of Time series Moran's *I* index.

The time series Moran's I value from the 50th day were used for K-means Clustering, as the values of most states are zero before the 50th day. (a) the AIC of clustering algorithm achieved the minimum value while the number

of clusters equals 4; (b) The four curves with grey ribbon represent the corresponding values and 95% CIs of the fitted line of four clusters; (c) The upper image was the original curves of Moran's I value, and the lower ones showed the fitted trends and 95% CIs. (d) Most of the states belonged to Cluster 1, while Cluster 2 was mainly located in the middle. Though the states in Cluster 3 were dispersed, Cluster 4 displayed a pattern of spatial aggregation.

3.3 SIRu integrated with Spatial Lag Model

The combined model of SLM-SIRu was tested with the spatial matrix and epidemic data of all the counties from Jan 22, 2020, to August 20, 2020. The results of all the states displayed high significance, verifying the feasibility of the SLM-SIRu model (Table 1). Wherein, the parameter λ ranged from -0.08 to 0.56, showing a normal distribution (Figure 4). The SLM-SIRu model also showed high R square values， most of which have a value larger than 0.5. In terms of fitness, the comparison between SIRu and SLM-SIRu indicated that the coefficient of spatial lag in SLM-SIRu improved the fitness of the original model of SIRu (Figure 5).

Table 1 Summary of Spatial Lag Coefficient in SLM-SIRu model

State	Spatial Lag Coefficient	Stand Error	t-value	p-value
Alabama	0.2297***	0.0084	27.2015	<0.001
Alaska	-0.081***	0.0126	-6.4374	<0.001
Arizona	0.09***	0.0124	7.2636	<0.001
Arkansas	0.2168***	0.0099	21.8995	<0.001
California	0.1472***	0.0098	14.9829	<0.001
Colorado	0.4275***	0.0083	51.6714	<0.001
Connecticut	0.3625***	0.0233	15.5821	<0.001
Delaware	0.278***	0.0310	8.9794	<0.001
Florida	0.2826***	0.0081	34.6982	<0.001
Georgia	0.2395***	0.0051	47.2483	<0.001
Hawaii	-0.0489**	0.0197	-2.4883	0.0128
Idaho	0.1117***	0.0088	12.7027	<0.001
Illinois	0.1661***	0.0066	25.0248	<0.001
Indiana	0.2052***	0.0084	24.3700	<0.001
Iowa	0.1312***	0.0087	15.0073	<0.001
Kansas	0.0773***	0.0081	9.5291	<0.001
Kentucky	0.1032***	0.0078	13.237	<0.001
Louisiana	0.3791***	0.0086	44.0066	<0.001
Maine	0.1497***	0.0214	7.004	<0.001
Maryland	0.1864***	0.0149	12.5354	<0.001
Massachusetts	0.5202***	0.0111	46.704	<0.001
Michigan	0.3771***	0.0075	50.2282	<0.001
Minnesota	0.2892***	0.0073	39.4947	<0.001
Mississippi	0.3353***	0.0081	41.2161	<0.001
Missouri	0.3031***	0.0065	46.2933	<0.001
Montana	0.0859***	0.0107	8.0393	<0.001
Nebraska	0.1004***	0.0085	11.7588	<0.001

Nevada	-0.0036**	0.0168	-0.2142*	0.0184
New Hampshire	0.1856***	0.0265	7.0142	<0.001
New Jersey	0.5621***	0.0114	49.3291	<0.001
New Mexico	0.2381***	0.0128	18.5421	<0.001
New York	0.2663***	0.0090	29.7591	<0.001
North Carolina	0.1598***	0.0071	22.449	<0.001
North Dakota	0.0348**	0.0112	3.1114**	0.0019
Ohio	0.0812***	0.0090	9.0392	<0.001
Oklahoma	0.1172***	0.0075	15.6807	<0.001
Oregon	0.1709***	0.0126	13.5983	<0.001
Pennsylvania	0.3031***	0.0088	34.2476	<0.001
Rhode Island	0.1400***	0.0371	3.7685	<0.001
South Carolina	0.2134***	0.0095	22.4236	<0.001
South Dakota	0.0560***	0.0117	4.8008	<0.001
Tennessee	0.1216***	0.0079	15.3701	<0.001
Texas	0.0376***	0.0052	7.2084	<0.001
Utah	0.0377***	0.0089	4.2494	<0.001
Vermont	0.1746***	0.0243	7.1958	<0.001
Virginia	0.2374***	0.0063	37.9368	<0.001
Washington	0.2351***	0.0126	18.6703	<0.001
West Virginia	0.1351***	0.0119	11.3082	<0.001
Wisconsin	0.1627***	0.0078	20.7673	<0.001
Wyoming	0.1288***	0.0188	6.8477	<0.001

Notes: ***Correlation is significant at the .001 level. ** Correlation is significant at the .01 level.

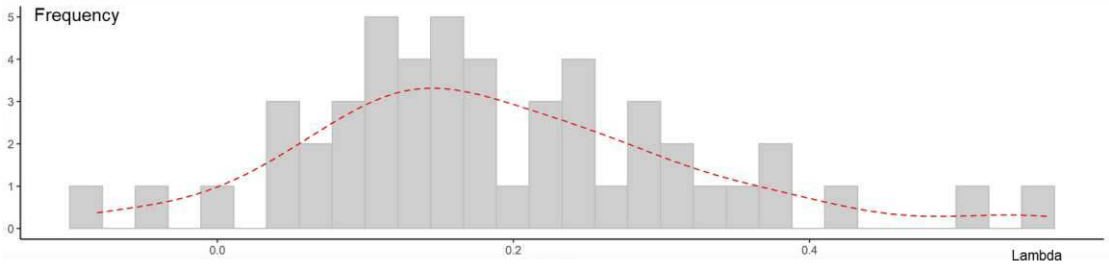


Figure 4 Histogram and Density plot of Spatial correlation coefficient λ.

The grey bars showed the frequency of λ values, the red line was the Probability Density Function (PDF) curves of λ values.

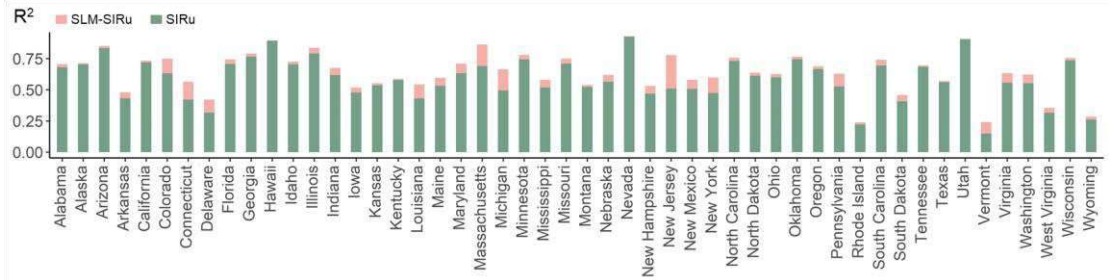


Figure 5 Fitness comparison of SLM-SIRu and SIRu.

The R^2 range of original SIRu was 0.1490 -0.9250(mean=0.5894), while that of SLM-SIRu was 0.2399 - 0.9303 (mean=0.6443), thus the percentage of R^2 improvement is 0.35% --62.72% with a mean of 11.54%.

Based on the mapping of spatial autocorrelation coefficients λ with four levels, it can be seen that only Alaska, Nevada, and Hawaii showed small negative values, while most of the central states have lower positive values, except the values of Colorado and New Mexico. The states in the southeast had similarly high values, wherein Louisiana was the largest one. In the north, Michigan, Massachusetts, Connecticut, and New Jersey also have relatively high spatial autocorrelation coefficients, indicating that the inter-county human flow in such states remains high.

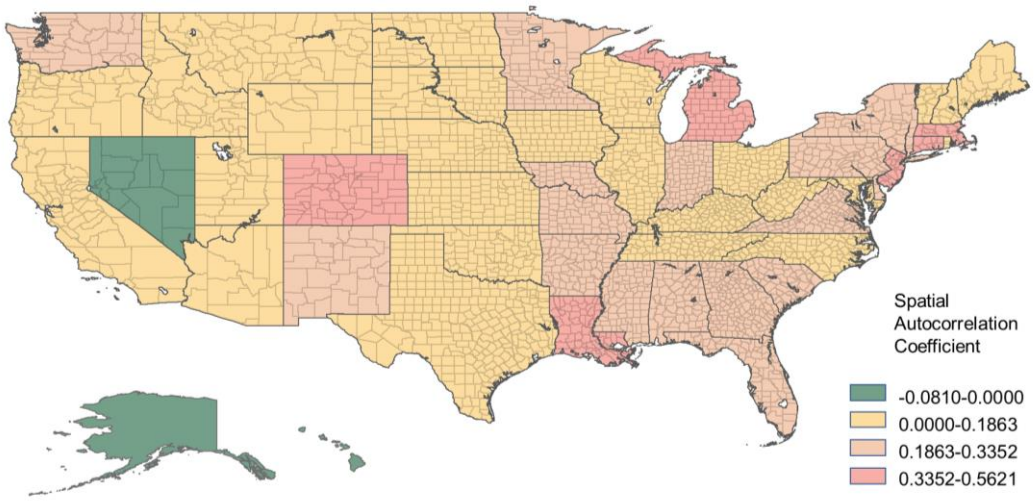
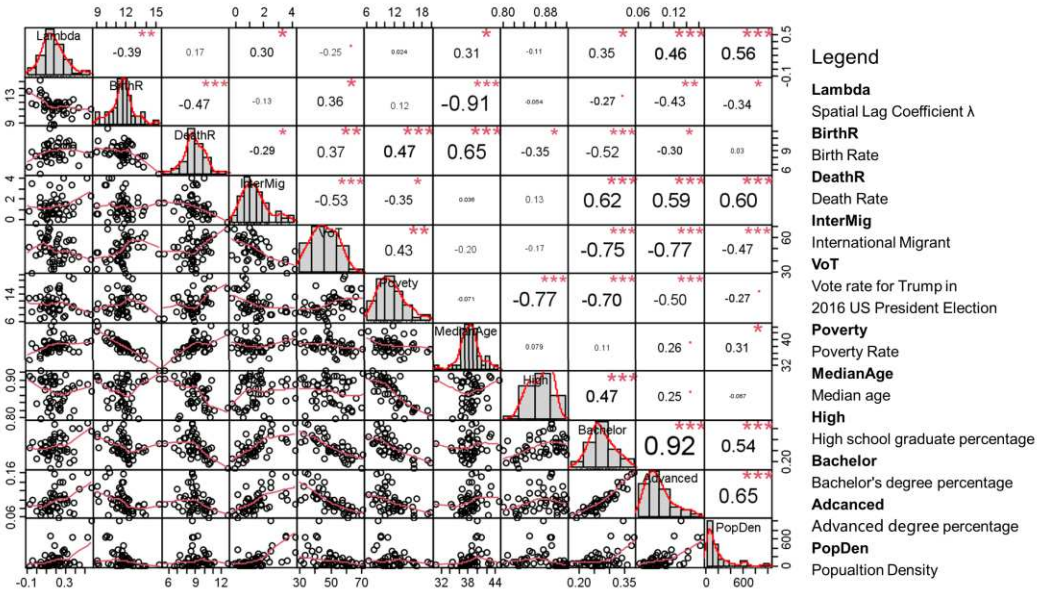


Figure 6 The choropleth map of spatial correlation coefficient λ.

The choropleth map of Spatial correlation coefficient λ used the Jerkens breakpoints with 4 levels.

3.4 Correlation and Regression with Socioeconomic Variables

To explore the correlation between the social, economic, and political factors and spatial correlation coefficient, the Spears test was applied. The result showed that there existed a significant negative relationship with the birth rate, and an obvious positive correlation with the proportion of international immigrants, median age, high education rate, and population density. In terms of the political factor, President Trump's support rate in the 2016 U.S. election showed a weak negative correlation (90% CI).



Notes: ***, **, * and. means that Correlation is significant at the 0.001, 0.01, 0.5,0.1 levels correspondingly.

Figure7 The correlation test between λ and Socioeconomic Variables.

The left part under the diagonal line is the scatter points plots, and the numbers are the correlation coefficients. The diagram along the diagonal line is the histograms.

A further Stepwise OLS review retained five variables, wherein only three of the variables were significant: population density, poverty rate, and bachelor degree rate. Among them, the population density and poverty rate are more significant, indicating that the intrastate population flow caused by population density and poverty is still dominant, and the increase of the highly educated population ratio would also increase the risk of inter-regional human flows.

Table 2 Summary of Stepwise OLS regression

	Estimate	Std. Error	t value	P-value	VIF
Intercept	0.0000	0.1093	0.000	1.0000	
BirthR	-0.2419	0.1214	-1.992	0.0525	1.208749
VoT	0.2669	0.1752	1.523	0.1349	2.518246
Poverty	0.4468**	0.1589	2.810	0.0073	2.072469
Bachelor	0.5706*	0.2292	2.489	0.0166	4.310079
PopDen	0.4089**	0.1371	2.983	0.0046	1.541248
Multiple R^2	0.4633				
Adjusted R^2	0.4023				
p-value	0.00003247				

4 Discussion

The initial objective of the project was to measure the inter-county spatiotemporal interaction effect of COVID-19 transmission in the United States and explore the correlation between spatiotemporal interaction coefficient and socioeconomic features. The results of this study indicated that the inter-county spatial effects in the states are changing with time, displaying four types of spatial correlation trends: continuous increasing,

fluctuating increasing, weak positive, and weak negative. Such clustering, though never been reported, could be explained by the heterogeneity in the social and spatial peripheries[37].

The fitnesses of SIRu models in all states have an average value above 0.75, indicating that the model can explain the epidemic dynamic of COVID-19 transmission in the United States. The SLM-SIRu model showed better fitness than the SIRu model with statistical significance, further verifying the hypothesis of spatial heterogeneity in epidemic dynamics[38]. The result indicated that the eastern states have a relatively high spatial interaction coefficient, showing a high possibility of inter-county flow.

In terms of socioeconomic features, the spatiotemporal interaction coefficients in each state were found positively correlated with the proportion of international immigrants, median age, the proportion of the highly educated population, and population density, but negatively correlated with the birth rate. The correlation in the two variables of median age and international immigrants ratio verified that the residents in inequitable living, working and environmental conditions may face a greater risk for COVID-19 infection[39]. What is interesting is that the Vote rate for Donald Trump in the 2016 U.S. presidential election showed a weak negative correlation, which implies that political factors may also some impact on the inter-county flow in the states[40].

The results of stepwise OLS regression suggested that poverty rate, population density, and the proportion of the highly educated population and are the three main positive correlated variables. Wherein, population density, and poverty rate have been considered highly correlated to COVID-19 transmission rate in previous studies[41, 42], while as our result implied that more inter-county flow may also occur in the states with higher population density or with low income in the United States. Such a result may support the study in Europe which proposed that high population density states appear to benefit more from their Shelter-in-Place Orders[43]. Of course, the cross-regional movement of highly educated people is also noteworthy, as these groups have better medical resources and lower infection rates[44], but their unrestricted movement may bring risks to other vulnerable groups.

The spatiotemporal dynamic model established in this research still has many deficiencies in terms of spatial weight calculation and model optimization. The COVID-19 transmission in each state not only occur in adjacent counties but also emerge among different states, which makes spatial weight calculation need more exploration with travel networks. Moreover, the relevance was explored by spatial effect measuring and OLS regression separately, which can be improved by other robust methods such as machine learning. The SIRu model, adopted to adjust the impact of data depressing, can also be further optimized in terms of accuracy.

5 Conclusion

This study proposes that inter-county movement within states has an impact on disease transmission, displaying obvious spatial heterogeneity which is potentially related to the social, economic, and political factors of each state. This result suggests that various state policies in the U.S. impose different impacts on COVID-19 transmission among counties. All states should provide more protection and support for the low-income population, high-density populated states need to strengthen regional mobility restrictions, and the highly educated group should reduce unnecessary regional movement and strengthen self-protection.

This study provides an integrated model to capture the spatiotemporal interaction effect, which may help assess the impact of cross-regional mobility on epidemic transmission and assist the government in optimize COVID-19 control policy.

Declarations

Ethical Approval and Consent to participate

Not applicable

Consent for publication

We confirmed that all authors have approved the manuscript for submission

Availability of supporting data

The data used in this paper is available in the Github repository by John Hopkins University https://github.com/CSSEGISandData/COVID-19

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

National Key Research and Development Project (2019YFB2101803)

National Natural Science Foundation of China (52078390)

Wuhan University Experiment Technology Project Funding

Authors' contributions

Lingbo Liu contributed to the conception of the study;

Lingbo Liu, Hao Wu performed the experiment;

Lingbo Liu, Tao Hu contributed significantly to analysis and manuscript preparation;

Lingbo Liu, Ru Wang performed the data analyses and wrote the manuscript;

Shuming Bao, Zhenghong Peng helped perform the analysis with constructive discussions.

Acknowledgments

The authors would like to acknowledge Xun Shi and other experts for their suggestions on the presentation of Data Statistical Analysis and Spatiotemporal Prediction Models of COVID-19 based on Workflow in the COVID-19 Data Analysis Webinar.

Figure legend

Figure 1 Research Workflow.

The research workflow explains the data, methods, and models used in the article. Moran's I and K-means clustering were introduced to capture the spatiotemporal feature of COVID-19daily new case changes in the US, then the spatial lag effect underlying the SIR model was further estimated by SLM-SIRu model and further used for correlation test with socio-economic variables.

Figure 2 Time series Moran's *I* index of each state in the United States since Jan 22, 2020.

The x-axis and y-axis indicated the value of time and Moran's I value correspondingly, wherein, the red curve represented the fitted value with a 95% confidence interval. Obvious changes from around the 50th day can be observed, displaying several patterns of temporal changes in spatial effects such as reversed U-shape, N-shape.

Figure 3 K-means Clustering of Time series Moran's *I* index.

The time series Moran's I value from the 50^{th} day were used for K-means Clustering, as the values of most states are zero before the 50^{th} day. (a) the AIC of clustering algorithm achieved the minimum value while the number of

clusters equals 4; (b) The four curves with grey ribbon represent the corresponding values and 95% CIs of the fitted line of four clusters; (c) The upper image was the original curves of Moran's I value, and the lower ones showed the fitted trends and 95% CIs. (d) Most of the states belonged to Cluster 1, while Cluster 2 was mainly located in the middle. Though the states in Cluster 3 were dispersed, Cluster 4 displayed a pattern of spatial aggregation.

Figure 4 Histogram and Density plot of Spatial correlation coefficient λ.

The grey bars showed the frequency of λ values, the red line was the Probability Density Function (PDF) curves of λ values.

Figure 5 Fitness comparison of SLM-SIRu and SIRu.

The R^2 range of original SIRu was 0.1490 -0.9250(mean=0.5894), while that of SLM-SIRu was 0.2399 - 0.9303 (mean=0.6443), thus the percentage of R^2 improvement is 0.35% --62.72% with a mean of 11.54%.

Figure 6 The choropleth map of spatial correlation coefficient λ.

The choropleth map of Spatial correlation coefficient λ used the Jerkens breakpoints with 4 levels.

Figure7 The correlation test between λ and Socioeconomic Variables.

The left part under the diagonal line is the scatter points plots, and the numbers are the correlation coefficients. The diagram along the diagonal line is the histograms.

Reference

1. Hu, T., et al., *Building an Open Resources Repository for COVID-19 Research.* Data and Information Management, 2020. **4**(3): p. 130.
2. Yang, C., et al., *Taking the pulse of COVID-19: a spatiotemporal perspective.* International Journal of Digital Earth, 2020: p. 1-26.
3. Loeffler-Wirth, H., M. Schmidt, and H. Binder, *Covid-19 Transmission Trajectories–Monitoring the Pandemic in the Worldwide Context.* Viruses, 2020. **12**(7): p. 777.
4. Walker, P.G.T., et al., *The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries.* Science, 2020. **369**(6502): p. 413-422.
5. Dowd, J.B., et al., *Demographic science aids in understanding the spread and fatality rates of COVID-19.* Proceedings of the National Academy of Sciences, 2020. **117**(18): p. 9696.
6. Merlo, J., et al., *A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena.* Journal of Epidemiology and Community Health, 2006. **60**(4): p. 290.
7. Qi, H., et al., *COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis.* Science of The Total Environment, 2020. **728**: p. 138778.
8. Wu, X., et al., *Natural and human environment interactively drive spread pattern of COVID-19: A city-level modeling study in China.* Science of The Total Environment, 2021. **756**: p. 143343.
9. Rader, B., et al., *Crowding and the shape of COVID-19 epidemics.* Nature Medicine, 2020. **26**(12): p. 1829-1834.
10. Roques, L., et al., *A parsimonious approach for spatial transmission and heterogeneity in the COVID-19 propagation.* Royal Society Open Science, 2020. **7**(12): p. 201382.
11. Viguerie, A., et al., *Simulating the spread of COVID-19 via a spatially-resolved susceptible–exposed–infected–*

recovered–deceased (SEIRD) model with heterogeneous diffusion. Applied Mathematics Letters, 2021. **111**: p. 106617.

12. Adekunle, I.A., et al., *Modelling spatial variations of coronavirus disease (COVID-19) in Africa.* Science of The Total Environment, 2020. **729**: p. 138998.
13. Xiong, C., et al., *Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections.* Proceedings of the National Academy of Sciences, 2020. **117**(44): p. 27087.
14. Aleta, A., et al., *Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19.* Nature Human Behaviour, 2020. **4**(9): p. 964-971.
15. Ascani, A., A. Faggian, and S. Montresor, *The geography of COVID-19 and the structure of local economies: The case of Italy.* Journal of Regional Science, 2020. **n/a**(n/a).
16. Hellewell, J., et al., *Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts.* The Lancet Global Health, 2020.
17. Abel, T. and D. Mcqueen, *The COVID-19 pandemic calls for spatial distancing and social closeness: not for social distancing!* International Journal of Public Health, 2020. **65**.
18. Lai, S., et al., *Effect of non-pharmaceutical interventions to contain COVID-19 in China.* Nature, 2020. **585**(7825): p. 410-413.
19. Kraemer, M.U.G., et al., *The effect of human mobility and control measures on the COVID-19 epidemic in China.* Science, 2020. **368**(6490): p. 493.
20. Flaxman, S., et al., *Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe.* Nature, 2020. **584**(7820): p. 257-261.
21. Bertuzzo, E., et al., *The geography of COVID-19 spread in Italy and implications for the relaxation of confinement measures.* Nature Communications, 2020. **11**(1): p. 4264.
22. Liu, F., et al., *Predicting and analyzing the COVID-19 epidemic in China: Based on SEIRD, LSTM and GWR models.* PLOS ONE, 2020. **15**(8): p. e0238280.
23. Das, A., et al., *Living environment matters: Unravelling the spatial clustering of COVID-19 hotspots in Kolkata megacity, India.* Sustainable Cities and Society, 2020: p. 102577.
24. Fingleton, B., D. Olner, and G. Pryce, *Estimating the local employment impacts of immigration: A dynamic spatial panel model.* Urban Studies, 2019. **57**(13): p. 2646-2662.
25. Sannigrahi, S., et al., *Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach.* Sustainable Cities and Society, 2020. **62**.
26. Mollalo, A., B. Vahedi, and K.M. Rivera, *GIS-based spatial modeling of COVID-19 incidence rate in the continental United States.* Science of The Total Environment, 2020. **728**: p. 138884.
27. Anderson, R.M., et al., *How will country-based mitigation measures influence the course of the COVID-19 epidemic?* The Lancet, 2020. **395**(10228): p. 931-934.
28. Lau, H., et al., *Internationally lost COVID-19 cases.* Journal of Microbiology, Immunology and Infection, 2020. **53**(3): p. 454-458.
29. Spychalski, P., A. Błażyńska-Spychalska, and J. Kobiela, *Estimating case fatality rates of COVID-19.* The Lancet. Infectious diseases, 2020. **20**(7): p. 774-775.
30. Wu, S.L., et al., *Substantial underestimation of SARS-CoV-2 infection in the United States.* Nature Communications, 2020. **11**(1): p. 4507.
31. Leon, D.A., et al., *COVID-19: a need for real-time monitoring of weekly excess deaths.* The Lancet, 2020. **395**(10234): p. e81.
32. Lipsitch, M., et al., *Potential Biases in Estimating Absolute and Relative Case-Fatality Risks during Outbreaks.* PLoS neglected tropical diseases, 2015. **9**(7): p. e0003846-e0003846.

33. Verity, R., et al., *Estimates of the severity of coronavirus disease 2019: a model-based analysis.* Lancet Infect Dis, 2020. **20**(6): p. 669-677.
34. Liu, Z., et al., *A COVID-19 epidemic model with latency period.* Infectious Disease Modelling, 2020. **5**: p. 323-337.
35. Cave, B., et al., *Applying an equity lens to urban policy measures for COVID-19 in four cities.* Cities & Health, 2020: p. 1-5.
36. Hamidi, S., S. Sabouri, and R. Ewing, *Does Density Aggravate the COVID-19 Pandemic?: Early Findings and Lessons for Planners.* Journal of the American Planning Association, 2020: p. 1-15.
37. Biglieri, S., L. De Vidovich, and R. Keil, *City as the core of contagion? Repositioning COVID-19 at the social and spatial periphery of urban society.* Cities & Health, 2020: p. 1-3.
38. Sun, F., et al., *A spatial analysis of the COVID-19 period prevalence in U.S. counties through June 28, 2020: where geography matters?* Annals of Epidemiology, 2020. **52**: p. 54-59.e1.
39. Cole, H.V.S., et al., *The COVID-19 pandemic: power and privilege, gentrification, and urban environmental justice in the global north.* Cities & Health, 2020: p. 1-5.
40. Yamey, G. and G. Gonsalves, *Donald Trump: a political determinant of covid-19.* BMJ, 2020. **369**: p. m1643.
41. White, E.R. and L. Hébert-Dufresne, *State-level variation of initial COVID-19 dynamics in the United States.* PLOS ONE, 2020. **15**(10): p. e0240648.
42. Berkowitz, R.L., et al., *Structurally vulnerable neighbourhood environments and racial/ethnic COVID-19 inequities.* Cities & Health, 2020: p. 1-4.
43. Dave, D., et al., *WHEN DO SHELTER-IN-PLACE ORDERS FIGHT COVID-19 BEST? POLICY HETEROGENEITY ACROSS STATES AND ADOPTION TIME.* Economic Inquiry, 2021. **59**(1): p. 29-52.
44. Cordes, J. and M.C. Castro, *Spatial analysis of COVID-19 clusters and contextual factors in New York City.* Spatial and Spatio-temporal Epidemiology, 2020. **34**: p. 100355.

Figures

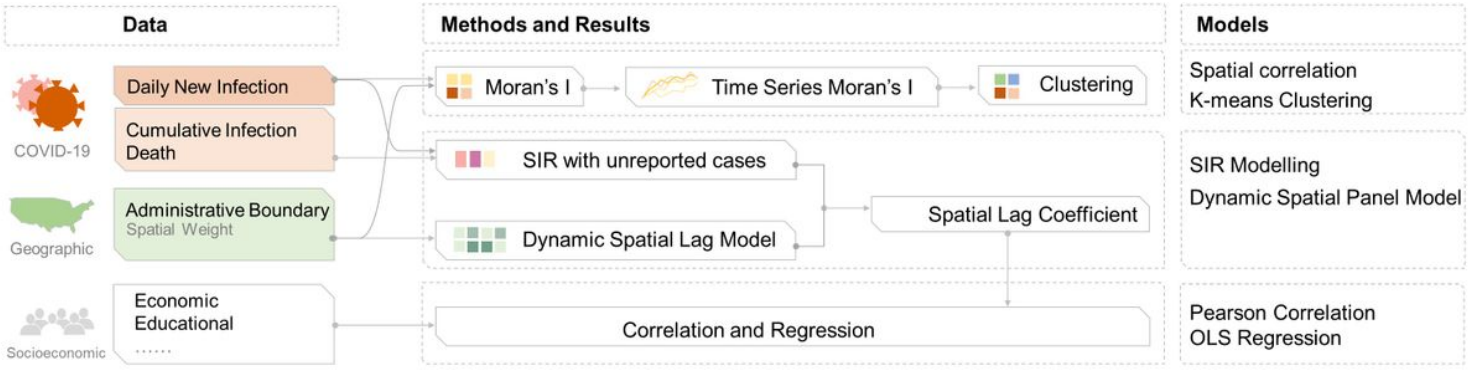


Figure 1

Research Workflow. The research workflow explains the data, methods, and models used in the article. Moran's I and K-means clustering were introduced to capture the spatiotemporal feature of COVID-19daily new case changes in the US, then the spatial lag effect underlying the SIR model was further estimated by SLM-SIRu model and further used for correlation test with socio-economic variables.

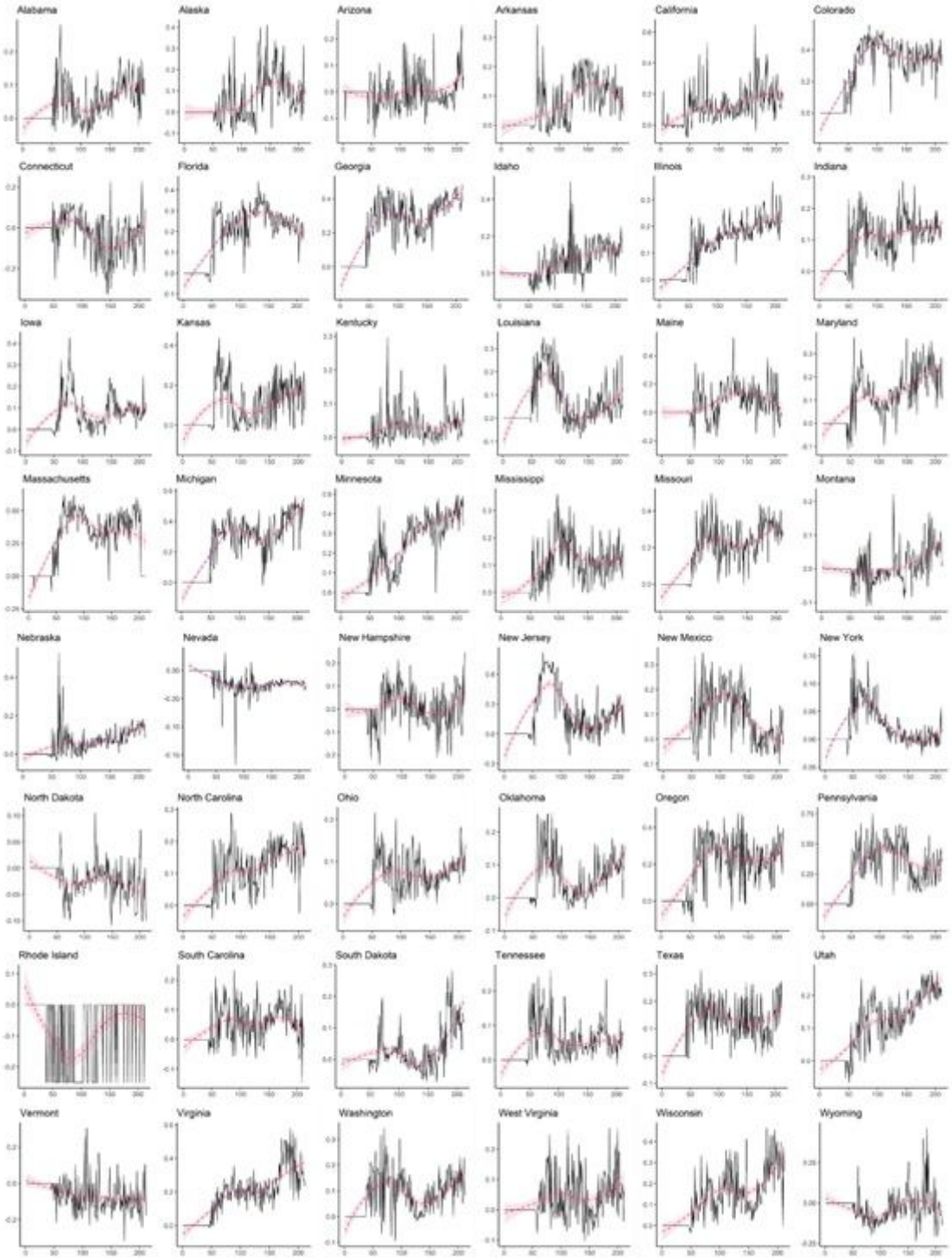


Figure 2

Time series Moran's I index of each state in the United States since Jan 22, 2020. The x-axis and y-axis indicated the value of time and Moran's I value correspondingly, wherein, the red curve represented the fitted value with a 95% confidence interval. Obvious changes from around the 50th day can be observed, displaying several patterns of temporal changes in spatial effects such as reversed U-shape, N-shape.

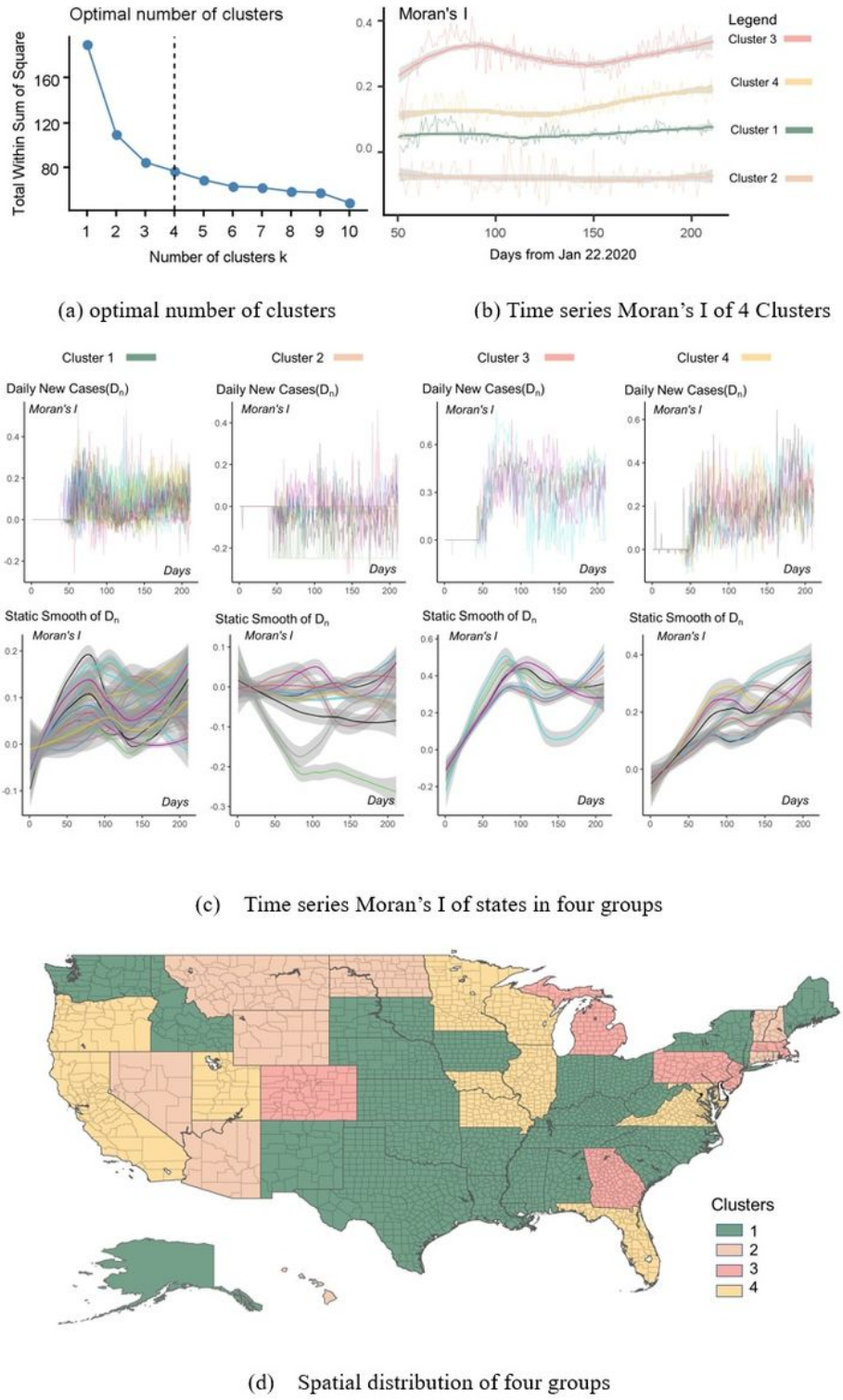


Figure 3

K-means Clustering of Time series Moran's I index. The time series Moran's I value from the 50th day were used for K-means Clustering, as the values of most states are zero before the 50th day. (a) the AIC of clustering algorithm achieved the minimum value while the number of clusters equals 4; (b) The four curves with grey ribbon represent the corresponding values and 95% CIs of the fitted line of four clusters; (c) The upper image was the original curves of Moran's I value, and the lower ones showed the fitted

trends and 95% CIs. (d) Most of the states belonged to Cluster 1, while Cluster 2 was mainly located in the middle. Though the states in Cluster 3 were dispersed, Cluster 4 displayed a pattern of spatial aggregation. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

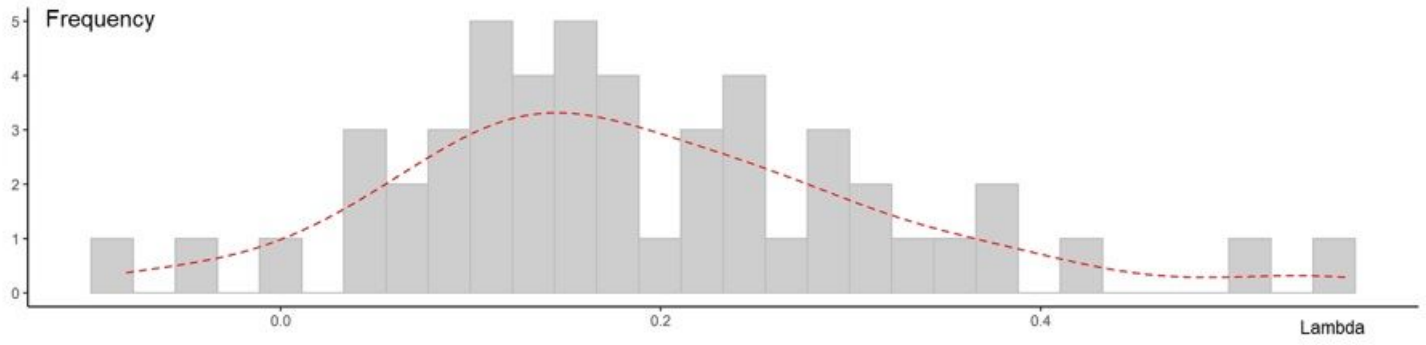


Figure 4

Histogram and Density plot of Spatial correlation coefficient λ. The grey bars showed the frequency of λ values, the red line was the Probability Density Function (PDF) curves of λ values.

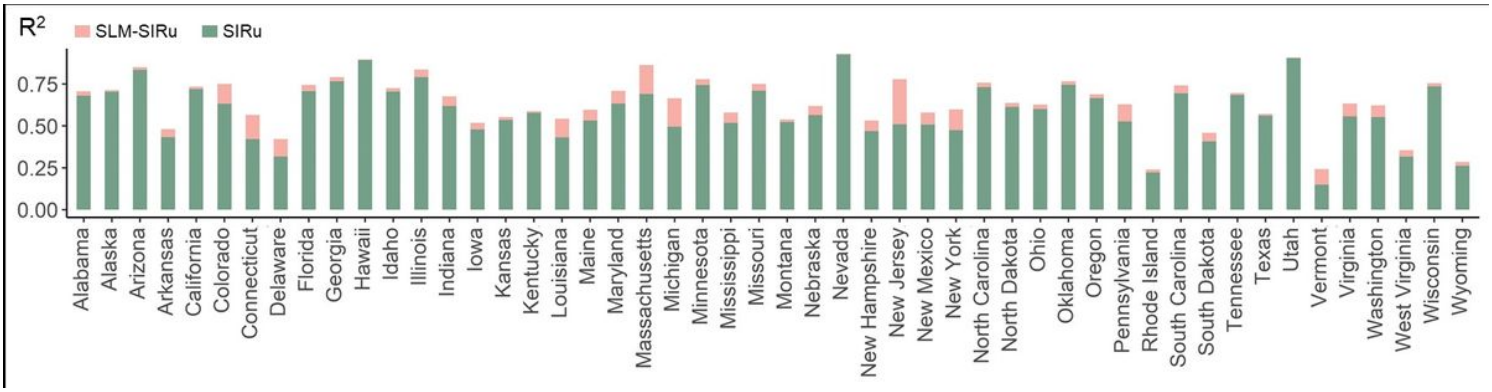


Figure 5

Fitness comparison of SLM-SIRu and SIRu. The R2 range of original SIRu was 0.1490 -0.9250(mean=0.5894), while that of SLM-SIRu was 0.2399 - 0.9303 (mean=0.6443), thus the percentage of R2 improvement is 0.35% –62.72% with a mean of 11.54%.

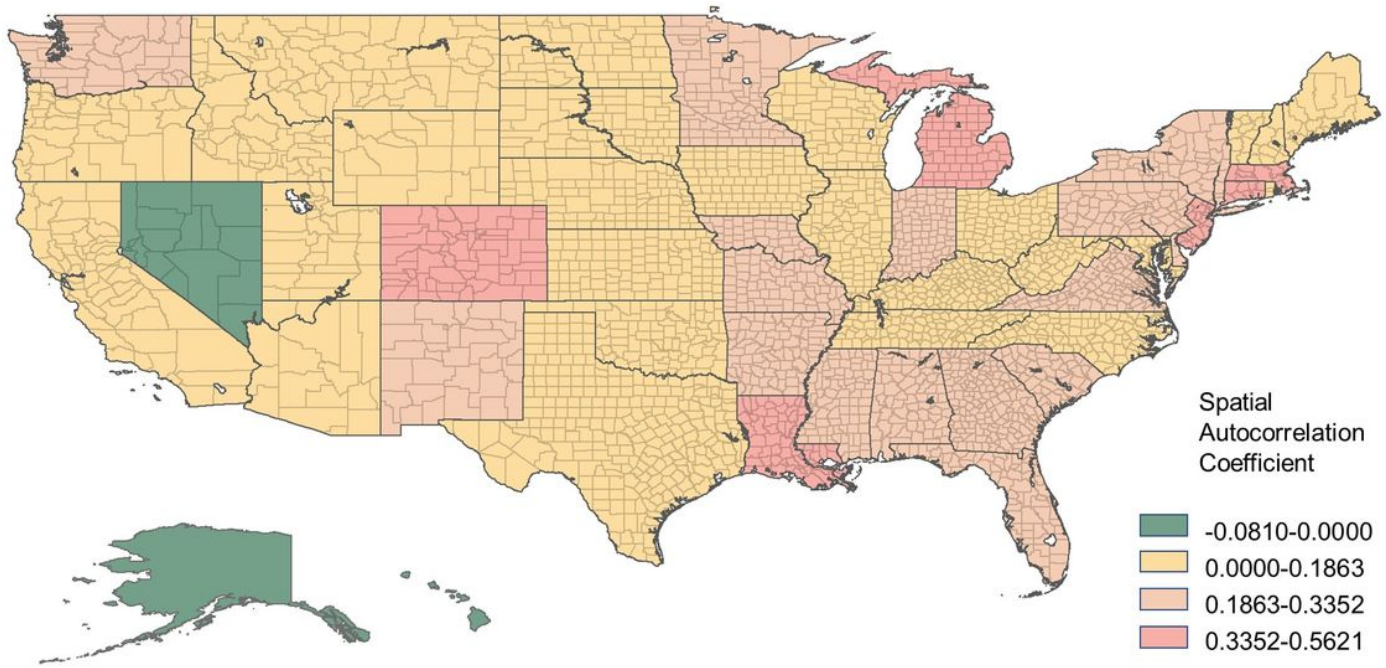


Figure 6

The choropleth map of spatial correlation coefficient λ. The choropleth map of Spatial correlation coefficient λ used the Jerkens breakpoints with 4 levels. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

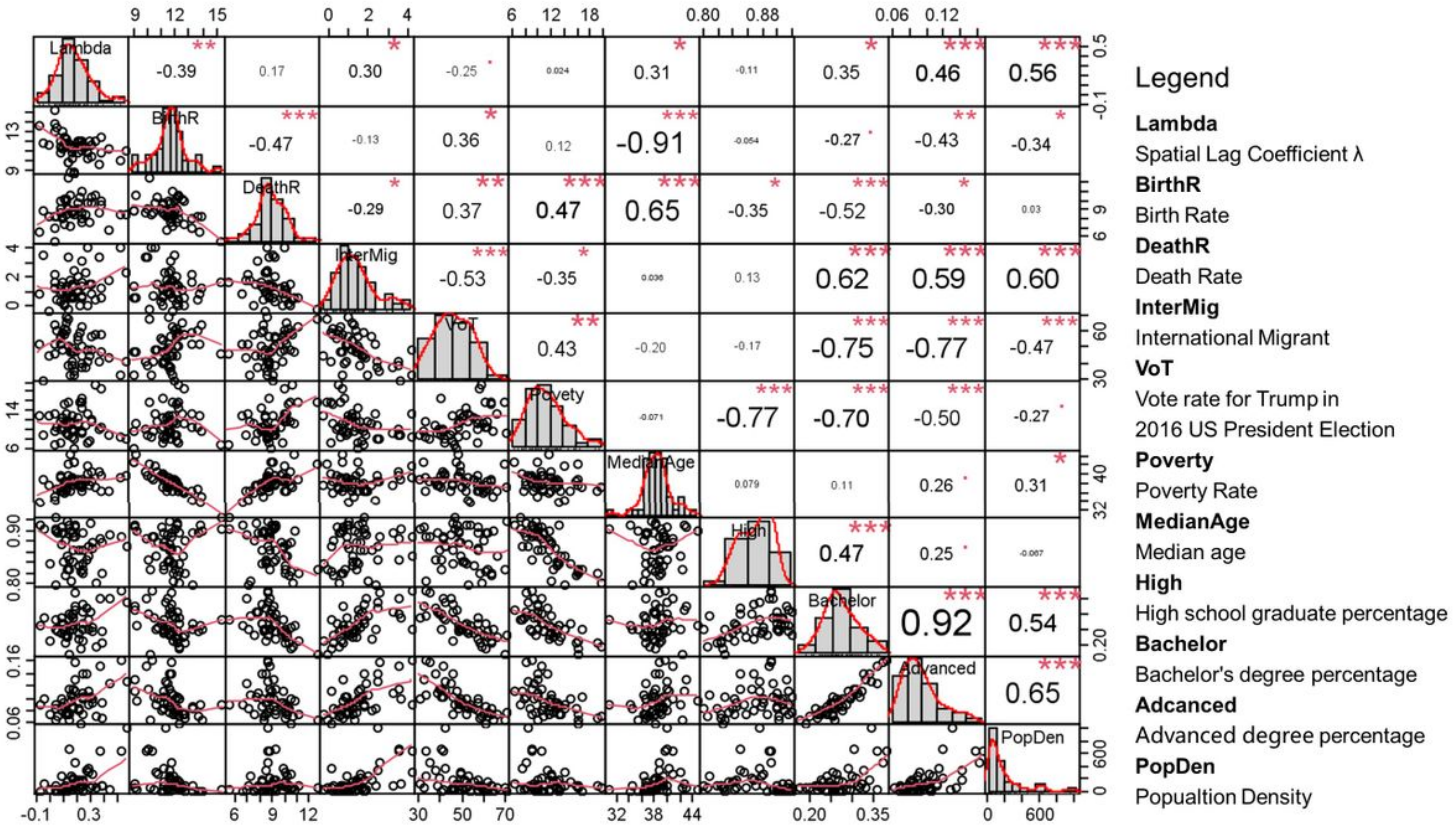


Figure 7

The correlation test between λ and Socioeconomic Variables. The left part under the diagonal line is the scatter points plots, and the numbers are the correlation coefficients. The diagram along the diagonal line is the histograms. Notes: ***, **, * and. means that Correlation is significant at the 0.001, 0.01, 0.5,0.1 levels correspondingly.