2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Statistical Analysis of Spatial Network Characteristics in Relation to COVID-19 Transmission Risks in US Counties

Siqi Zhang, Sihan Yang and Hui Yang*

***Abstract*—Since the pandemic of COVID-19 began in January 2020, the world has witnessed drastic social-economic changes. To harness the virus spread, several studies have been done to study contributing factors that are pertinent to COVID-19 transmission risks. However, little has been done to investigate how human activities on the spatial network are correlated to the virus transmission and spread. This paper performs a statistical analysis to examine interrelationships between spatial network characteristics and cumulative cases of COVID-19 in US counties. Specifically, both county-level transportation profiles (e.g., the total number of commute workers, route miles of freight railroad) and road network characteristics of US counties are considered. Then, the lasso regression model is utilized to identify a sparse set of significant variables that are sensitive to the response variable of COVID-19 cases. Finally, the fixed-effect model is built to capture the relationship between the selected set of predictors and the response variable. This work helps identify and determine salient features from spatial network characteristics and transportation profiles, thereby improving the understanding of COVID-19 spread dynamics. These significant variables can also be utilized to develop simulation models for the prediction of real-time positions of virus spread and the optimization of intervention strategies.**

***Index Terms*—Infectious disease, spatial network, transportation profiles, US county, COVID-19, virus spread**

I. Introduction

THe pandemic of COVID-19 poses great challenges to our society. Because of high virus infectivity and asymptomatic infection, it is often difficult to make quick responses and thereby control the transmission of COVID-19 in human populations. As of May 2, 2021, the US has reported more than 32.39 million infected cases and 576k deaths [1]. With rapid advances in epidemic surveillance systems, abundant data of infection are collected. The availability of these data provides unprecedented opportunities to investigate the relationships between different risk factors (e.g., age, commute settings) and epidemic characteristics.

Several works have been done to study different contributing factors that could affect the virus transmission and spread process in spatial regions. For example, Yang et. al [2] performed a statistical analysis to examine the relationship between a variety of factors (e.g., social-economic factors, healthy factors, demography factors) and COVID-19 infection. One significant variable from the study is the number of households with grandparents and grandchildren. Through an analysis of mobility data from 52 countries around the world, Nouvellet et al. [3] found that the reduction of human mobility significantly slows down the virus spread process. Pramanik et al. investigate the climatic influence on COVID-19 transmission risks in 228 cities across three climatic zones. They found that there are strong relationships between the average diurnal temperature range and the COVID-19 outbreak in tropical regions. Khanijahani [4] studied county-level disparities among ethnic and socioeconomic factors in confirmed COVID-10 cases and mortality rates in the United States. It is reported that the population of adults with less than a high school diploma had a high infection and mortality rate. However, very little has been done to investigate how human interactions in the spatial network are related to the extent of virus spreads in different US counties.

Indeed, human mobility patterns are primarily attributed to the structure of the spatial network [5]. In a small-scale spatial environment, individuals often follow the road network to visit a set of locations (e.g., workplaces, shopping malls, grocery stores). In a large spatial environment, people tend to travel through a spatial network of airports, railroads, and highways. As such, spatial network characteristics can also be potentially related to the COVID-19 spread in the spatial environment. Figure 1 illustrates the road network of Centre County, PA. Note that nodes are more densely scattered in the downtown area, which also consists of more human traffic flows. Because human interactions are frequent in this area, the virus tends to spread at a fast rate [6].

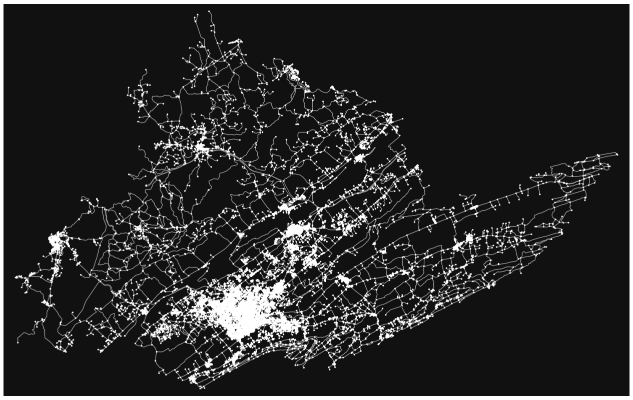

Fig. 1. The road network of Centre County, PA

In this paper, we perform a statistical analysis to investigate contributing factors from spatial network characteristics to COVID-19 transmission risks in US counties. First, relevant data are collected and categorized into two groups: trans-

*Corresponding author: Hui Yang (email:huy25@psu.edu)
The authors are with Complex Systems Monitoring, Modeling and Control Lab, The Pennsylvania State University, University Park, PA 16802 USA

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

portation profiles and road network characteristics. Then, we utilize the lasso regression model to shrink the number of predictors to a sparse set of predictors that are sensitive to the response variable of cumulative cases. Finally, we build a fixed-effect model to capture dynamics of COVID-19 spread in different US counties.

II. Data Integration

In this study, we used cumulative cases of COVID-19 at the county level as the dependent variable, which are readily available from the New York Times data repository (i.e., https://github.com/nytimes/covid-19-data). We then processed infection data from July 15, 2020 to April 15, 2021 for monthly cumulative cases of COVID-19. Figure 2(a) depicts the distribution of cumulative cases of April 15, 2021 in US counties. Note that the COVID-19 has almost spread to the whole US and only a few counties are the exception.

For the independent variables, a total of 33 predictors at the county level were extracted from the official website of the Bureau of Transportation Statistics of the U.S. Department of Transportation (i.e., https://www.bts.gov/ctp) and OSMnx supported by the OpenStreetMap platform [7]. Figure 2(b-c) provides the geographical distribution of some predictors (e.g., total street length and the number of business establishments). These predictors are categorized into two groups, namely transportation and road network as follows,

- **Transportation predictors**: We extracted 17 variables about U.S. transport infrastructure from Bureau of Transportation Statistics as follows: number of primary and commercial airports (x_1), number of non-commercial civil public-use airports and seaplane base (x_2), number of non-commercial other aerodromes (x_3), number of bridges (x_4), percent of poor condition bridges (x_5), number of business establishments (x_6), percent of resident workers who commute by transit (x_7), number of resident workers who work at home (x_8), number of workers from other counties who commute to work in the county (x_9), number of resident workers who commute to work in other counties (x_{10}), number of resident workers who commute within a county (x_{11}), number of resident workers (x_{12}), number of residents (x_{13}), number of total docks (x_{14}), route miles of freight railroad (x$_{15}$), percent of medium to fair condition bridges (x_{16}) and route miles of passenger railroad and rail transit (x_{17}).
- **Road network predictors**: In addition, road data were first downloaded from the OpenStreetMap and used to construct the spatial network of US counties. Then, 16 network measures are calculated as dependent variables: number of nodes(x_{18}), number of edges (x_19), average degree of nodes (x_{20}), number of intersections (x_{21}), average streets per node (x_{22}), total edge length (x_{23}), average edge length (x_{24}), total street length (x_{25}), average street length (x_{26}), number of street segments (x_{27}), node density in square kilometers (x_{28}), intersection density in square kilometers (x_{29}), edge density in square kilometers (x_{30}), street density in square kilometers (x_{31}), average circuity (x_{32}) and self-loop proportion (x_{33}).

III. Statistical Modeling

Regression models present a functional relationship between a multivariate set of predictors and responses. However, there is often multicollinearity among predictors that can cause the model to be sensitive and unstable to extraneous noises [8], [9]. Hence, the lasso regression model is utilized to perform the selection of significant variables from transportation and road network predictors [10]. For a nonnegative λ, it penalizes the sum of $L1$ norm of model parameters, which can be expressed as follows,

$$\underset{\beta_0,\boldsymbol{\beta}}{\operatorname{argmin}}\left(\frac{1}{N}\left(\sum_{i=1}^{N} y_i-\beta_0-\boldsymbol{X_i}\boldsymbol{\beta}\right)+\lambda\sum_{j=1}^{P}|\beta_j|\right) \tag{1}$$

where N is the number of observations, y_i is the cumulative cases of i-th observation, $\boldsymbol{X_i}$ is a vector of predictor values of i-th observation and P is the number of predictors. Note that λ controls the magnitude of the penalty. When λ is getting bigger, the model tends to penalize insignificant predictors, and therefore only significant predictors will be retained.

Because the control policy of COVID-19 may vary due to different state-level governments, the infection rate of COVID-19 is correlated to the state to which the county belongs. Hence, we consider the fixed-effect model to capture the relationship between the selected set of significant predictors and the response variable of cumulative cases in different US counties. Specifically, the fixed-effect model can be described as follows,

$$y_i|t=\beta_0+\sum_{j=1}^{P}\beta_j x_{ij}+\sum_{k=1}^{K}\gamma_k u_{ik}+\epsilon \tag{2}$$

where y_i is the total number of cumulative COVID-19 cases of county i, β_0 and β_j are estimated coefficients, x_{ij} is the value of j-th predictor of county i, γ_k is the fixed effect for state k, u_{ik} is the indicator variable of whether county i belongs to state k and $\epsilon \sim N(0,\sigma^2)$.

IV. Experimental Results

Figure 3(a) shows the Pearson correlation between 33 predictors and confirmed COVID-19 cases in US counties over 10 months. Overall, there are high correlations between COVID-19 infections and transportation, road network predictors. Four predictors in county-level transportation profiles correlate more than 0.9, namely the number of residents, the number of resident workers, the number of resident workers who commute within county, and the number of business establishments. This demonstrates that the more residents, resident workers, and commuting workers a county has, the more infection it will have. This is not surprising because the virus spreads at a rapid rate when humans make frequent contacts with each other during commuting and at

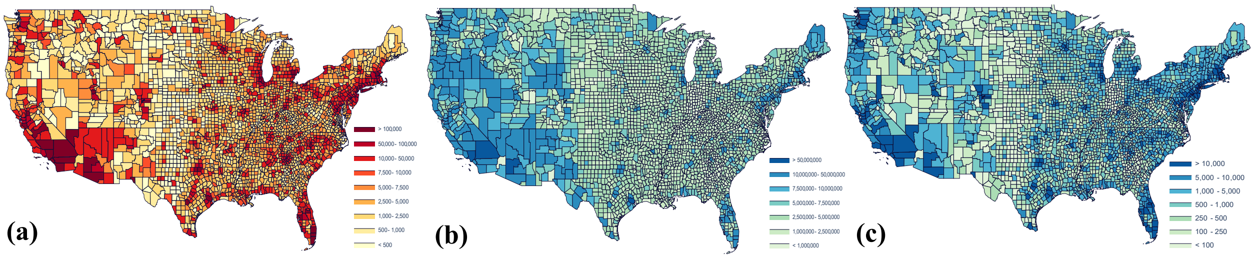

Fig. 2. The geographical distribution of (a) cumulative cases of 04/15/2021 (b) total street length (c) the number of business establishment for US counties

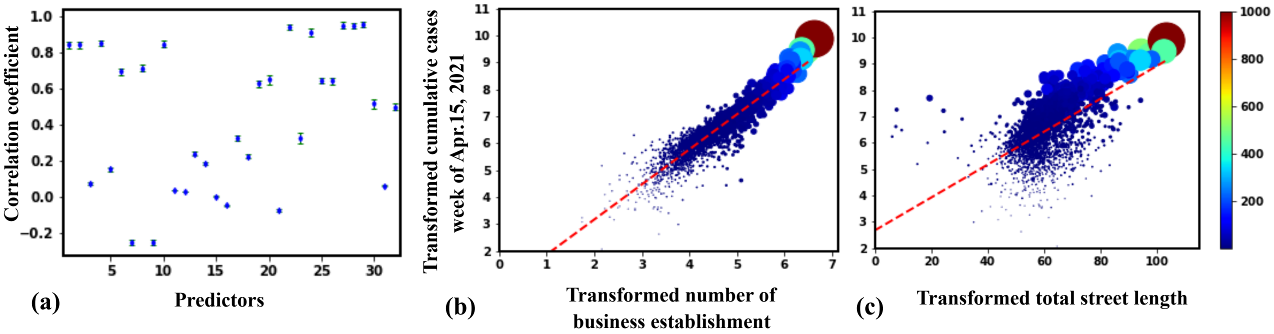

Fig. 3. (a) Pearson correlation between 33 predictors and confirmed COVID-19 cases in US counties. Scatter plots of monthly cumulative cases vs. (b) transformed number of business establishment (c) transformed total street length

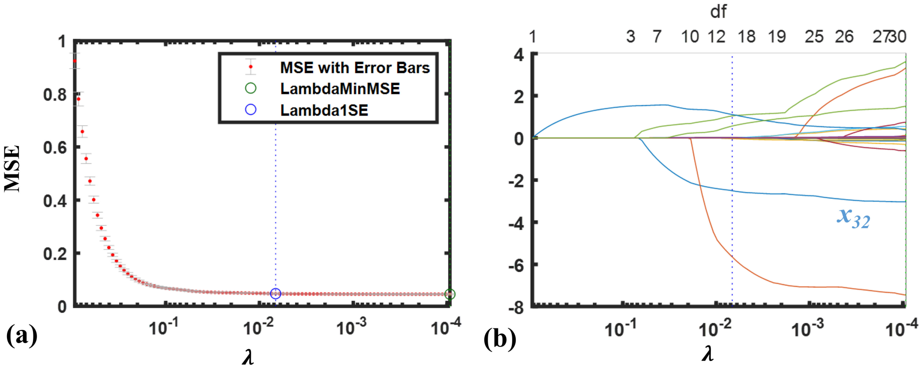

Fig. 4. (a) The variations of prediction errors vs. regularization parameter in Lasso regression with cumulative cases and 33 predictors. (b) The coefficient path for the Lasso regression

the workplace. Because business establishments are often visited by a large number of residents every day, infections are more likely to occur and thereby the virus will quickly spread to the entire spatial area. This can also be observed from the scatter plot in Figure 3(b). Network-based predictors tend to have a slightly smaller correlation with the cumulative cases at the county level. One of these predictors (i.e., total street length) has reported a correlation of 0.712, which can also be seen in the scatter plot of 3(c). When the total street length is large, people living in the spatial area tend to be more connected and concentrated. As such, individuals are more vulnerable and likely to transmit the virus.

The lasso experiment is performed on the response variable of cumulative cases and 33 predictors using 10-fold cross-validations. Figure 4(a) shows the variations of mean squared errors (MSE) at different choices of regularization parameters. The green circle and dotted line refer to λ with the minimum cross-validation error while the blue circle and dotted lines locate λ with minimum cross-validation error plus one standard deviation. Note that when the value of λ decreases, the number of predictors that enters the model also increases. Because MSE shows a trend of exponential decay with the regularization parameter λ, we choose the green dot as the optimal location of λ, which corresponds to

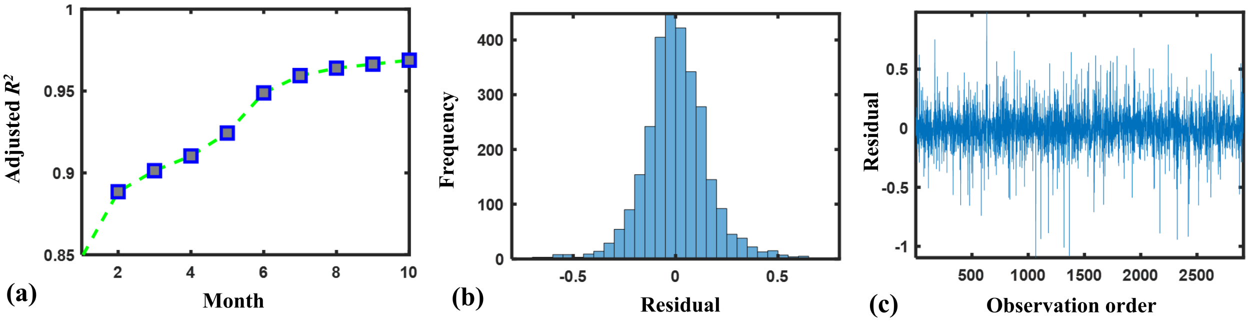

Fig. 5. (a) The variations of adjusted R^2 for the fixed-effect model with the response variable of cumulative cases (from July 2020 to April 2021). The residual diagnosis of the fixed-effect model includes (b) the histogram of residuals (c) residuals in the order of observations

the selection of 13 predictors as a sparse set of sensitive parameters to the response variable. The coefficient paths of 33 predictors when λ is reduced from 1 to 0.0001 are illustrated in Figure 4(b). One significant variable is the average circuity x_{32} that enters the model very early and then steadily negatively influence the response variable. Average circuity x_{32} describes the ratio between the length of network path and euclidean distance between two nodes (or locations) in the road network. The smaller the value of x_{32} is, the more efficient the road network. This indicates there is a negative relationship between the COVID-19 infection and the efficiency of the road network.

Figure 5(a) shows the variations of adjusted R^2 for the response variable of cumulative COVID-19 cases. When additional infection data are collected and added to the fixed-effect model, the adjusted R^2 increases from 0.85 to 0.96. Figure 5(b) and (c) provides an example of the residual diagnosis of the fixed effect model. The histogram of residuals show that the normality assumption is valid and there are no symmetric patterns in the residual plots.

V. Conclusions

The pandemic of COVID-19 not only impacts the health of our society but also brings great disruptions to the world. Research on the contributing factors to the COVID-19 transmission is critical to understand the spread dynamics. Although many factors (e.g, mobility, social-economic, demography factors) have been studied in previous studies, little has been done to investigate the interrelationship between spatial network characteristics and COVID-19 infections in US counties. Note that human interactions and dynamic movements are performed on the structure of spatial networks. This paper provides a statistical analysis to study this relationship between the characteristics of transportation networks and cumulative cases of COVID-19 in US counties. Experimental results show that network factors (e.g., average circuity, total street length of a spatial region) are highly correlated to the virus transmission process. These significant factors can be further fed into simulation models and/or health policy design to control the virus spread process.

Acknowledgement

The authors would like to acknowledge the NSF I/UCRC Center for Healthcare Organization Transformation (CHOT), NSF I/UCRC award IIP-1624727, and NSF RAPID grant IIP-2026875 for funding this research.

Disclaimer

This paper doesn't involve experimental procedures on either human subjects or animal models, but rather use the data available in the public domain. Thus, approval is not needed from the Institutional Review Board.

References

[1] "Coronavirus Disease 2019 (COVID-19) in the U.S. https://www.cdc.gov/coronavirus/2019-ncov/."

[2] H. Yang, S. Zhang, R. Liu, A. Krall, Y. Wang, M. Ventura, and C. Deflitch, "Epidemic informatics and control: A holistic approach from system informatics to epidemic response and risk management in public health," in *AI and Analytics for Public Health - Proceedings of the 2020 INFORMS International Conference on Service Science*. Springer, 2021, ch. 1, pp. 1–46.

[3] P. Nouvellet, S. Bhatia, A. Cori, K. E. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper *et al.*, "Reduction in mobility and covid-19 transmission," *Nature communications*, vol. 12, no. 1, pp. 1–9, 2021.

[4] A. Khanijahani, "Racial, ethnic, and socioeconomic disparities in confirmed covid-19 cases and deaths in the united states: a county-level analysis as of november 2020," *Ethnicity & health*, vol. 26, no. 1, pp. 22–35, 2021.

[5] B. Jiang, J. Yin, and S. Zhao, "Characterizing the human mobility pattern in a large street network," *Physical Review E*, vol. 80, no. 2, p. 021136, 2009.

[6] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. Du Plessis, N. R. Faria, R. Li, W. P. Hanage *et al.*, "The effect of human mobility and control measures on the covid-19 epidemic in china," *Science*, vol. 368, no. 6490, pp. 493–497, 2020.

[7] G. Boeing, "Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Computers, Environment and Urban Systems*, vol. 65, pp. 126–139, 2017.

[8] Y. Chen and H. Yang, "A novel information-theoretic approach for variable clustering and predictive modeling using dirichlet process mixtures," *Scientific reports*, vol. 6, no. 1, pp. 1–13, 2016.

[9] G. Liu and H. Yang, "Self-organizing network for variable clustering," *Annals of Operations Research*, vol. 263, no. 1, pp. 119–140, 2018.

[10] G. Liu, C. Kan, Y. Chen, and H. Yang, "Model-driven parametric monitoring of high-dimensional nonlinear functional profiles," in *2014 IEEE international conference on automation science and engineering (CASE)*. IEEE, 2014, pp. 722–727.