Original Paper

Quantifying the Regional Disproportionality of COVID-19 Spread: Modeling Study

Kenji Sasaki[1], PhD; Yoichi Ikeda[1], PhD; Takashi Nakano[1,2], PhD

[1]Center for Infectious Disease Education and Research, Osaka University, Suita, Osaka, Japan
[2]Research Center for Nuclear Physics, Osaka University, Ibaraki, Osaka, Japan

Corresponding Author:

Kenji Sasaki, PhD
Center for Infectious Disease Education and Research
Osaka University
Co-creation BLDG. D88-1, 2-1 Yamadaoka
Suita, Osaka, 565-0871
Japan
Phone: 81 50-5604-3730
Email: kenjis@cider.osaka-u.ac.jp

Abstract

Background: The COVID-19 pandemic has caused serious health, economic, and social consequences worldwide. Understanding how infectious diseases spread can help mitigate these impacts. The Theil index, a measure of inequality rooted in information theory, is useful for identifying geographic disproportionality in COVID-19 incidence across regions.

Objective: This study focused on capturing the degrees of regional disproportionality in incidence rates of infectious diseases over time. Using the Theil index, we aim to assess regional disproportionality in the spread of COVID-19 and detect epicenters where the number of infected individuals was disproportionately concentrated.

Methods: To quantify the degree of disproportionality in the incidence rates, we applied the Theil index to the publicly available data of daily confirmed COVID-19 cases in the United States over a 1100-day period. This index measures relative disproportionality by comparing daily regional case distributions with population proportions, thereby identifying regions where infections are disproportionately concentrated.

Results: Our analysis revealed a dynamic pattern of regional disproportionality in the confirmed cases by monitoring variations in regional contributions to the Theil index as the pandemic progressed. Over time, the index reflected a transition from localized outbreaks to widespread transmission, with high values corresponding to concentrated cases in some regions. We also found that the peaks in the Theil index often preceded surges in confirmed cases, suggesting its potential utility as an early warning signal.

Conclusions: This study demonstrated that the Theil index is one of the effective indices for quantifying regional disproportionality in COVID-19 incidence rates. Although the Theil index alone cannot fully capture all aspects of pandemic dynamics, it serves as a valuable tool when used alongside other indicators such as infection and hospitalization rates. This approach allows policy makers to monitor regional disproportionality efficiently, offering insights for early intervention and targeted resource allocation.

JMIR Form Res 2025;9:e59230; doi: 10.2196/59230

Keywords: infectious disease; COVID-19; epidemiology; public health; SARS-CoV-2; pandemic; inequality measure; information theory; Kullback-Leibler divergence

Introduction

The COVID-19 pandemic has caused serious health problems and has had major economic and social consequences worldwide. It has highlighted the need to understand regional disparities in infection rates to strengthen public health responses since infection dynamics are influenced by factors such as population density, socioeconomic conditions, and health care infrastructure [1,2]. Numerous indicators and models have been proposed to address the problem, and mechanisms for the spread of the infection and intervention measures to control the pandemic have been studied [3-9].

Several recent studies have investigated regional differences in COVID-19 prevalence [10-13]. Differences in the prevalence rates between regions highlight the need to understand regional inequalities in pandemic response strategies. Effectively addressing these disparities requires accurate quantification and understanding of regional disproportionalities in daily confirmed COVID-19 cases.

In the field of economics, various indicators have been developed to measure resource and income inequality, including an index proposed by Theil, which incorporated information theory [14]. Manz and Mansmann [15] have demonstrated the importance of using inequality indices for monitoring changes in geographic inequality; for instance, the Theil index was used to track geographic disproportionality over time during the COVID-19 pandemic, providing important insights for public health policy.

The aim of this paper is to quantify the interregional disproportionality in the number of confirmed cases using the Theil index, which mathematically corresponds to the Kullback-Leibler (KL) divergence in information theory [16]. The Theil index is an effective method of measuring the degree of disproportionality and objectively assessing biases in the interregional distribution of infected individuals.

Methods

Overview

We analyzed the time trends of daily COVID-19–confirmed cases in the United States over 1100 days since the first reported case on January 21, 2020 [17]. Data are taken from the COVID-19 data repository at the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [18]. US state population data were obtained from the US Census Bureau website [19]. Population changes due to migration, births, and deaths were not considered in the analysis.

The Disproportionality Measure: Theil Index

The Theil index is commonly applied in various fields including economics, sociology, and information theory. It quantifies the relative differences between various components of a dataset. In the context of regional analysis of the confirmed cases, the Theil index can be employed to evaluate the distribution of infected individuals across different regions. In this study, we use the Theil index to identify regions with disproportionate numbers of confirmed cases relative to their population size.

The discrete form of the Theil index is expressed as:

$$T = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} p_i ln \frac{p_i}{q_i}$$

where N is the total number of regions being considered and ln is the natural logarithm. The Theil index T is composed of a sum of t_i which is a partial contribution from region i. The discrete probability distribution, p_i, in region i is defined as the ratio of daily confirmed cases region i to the total confirmed cases across all regions for that day. Similarly, the population ratio, q_i, in region i is the ratio of the population in region i to the total population across all regions.

The Theil index, which is mathematically related to the KL divergence, is a nonsymmetric metric that measures the relative entropy or informational difference between two distributions. It is sensitive to the interregional variations in the distribution of the confirmed cases, with its maximum value attained when the confirmed cases are concentrated in areas with the smallest population proportion. Consequently, the index tends to exhibit higher values when a small number of regions account for a large share of the confirmed cases, and conversely, lower values when the confirmed cases are more evenly distributed across regions. Notably, it remains nonnegative and reaches a minimum value of 0 only when the two distributions are identical. Therefore, applying the Theil index to the time-series data of the confirmed cases, and monitoring changes in the index over time, we quantified the degree of spread of COVID-19 cases and assessed whether the confirmed cases were disproportionately concentrated relative to the regional population sizes over time.

Ethical Considerations

This study used publicly available, deidentified COVID-19 data from CSSE at Johns Hopkins University [18], and therefore, additional ethics approval and informed consent were not required. The aggregated data ensured privacy and confidentiality, and no direct human participants were involved; thus, no compensation was provided. No identifiable information appears in any images or materials.

Results

To address fluctuations in the Theil index caused by data aggregation inconsistencies during holidays across different regions, the 7-day average of confirmed COVID-19 cases was used instead of raw data.

Figure 1 illustrates a 2-axis graph showing the time trends of the Theil index (left axis) and the number of confirmed cases in a logarithmic scale (right axis, logarithmic scale). The horizontal axis represents the number of days elapsed (denoted by d in the text) since the date of the first reported case in the United States.

In Figure 1, there are eight notable surges of the confirmed cases, occurring at approximately d=80 (first), 180 (second), 350 (third), 450 (fourth), 580 (fifth), 720 (sixth), 900 (seventh), and 1080 (eighth), respectively. The presence of multiple peaks in the Theil index indicates that infected individuals were concentrated in specific regions during the period, and the degree of this concentration can be assessed by examining the numerical values. However, it is important to note that changes in the Theil index simply indicate the degree of regional disproportionality in the confirmed cases rather than absolute increases or decreases in the number of infected individuals. Therefore, this indicator is most effective when interpreted in conjunction with actual trends in the number of confirmed cases.

Before the first peak, the number of confirmed cases was quite low, and the Theil index fluctuated erratically. As *d* increased near the first peak, the Theil index gradually decreased, reaching a local minimum around *d*=120. This suggested that the initially localized epidemic began to spread throughout the US during the early stages of the global pandemic. Similar trends were observed during subsequent surges, such as slight a increase in the Theil index before the peak, followed by a decrease. This could be seen as a precursor to a surge in the number of infected individuals. This finding aligns with previous research by Ikeda, Sasaki, and Nakano [7].

The following examples provide interesting insights; when the Theil index value was high and the number of confirmed cases was low (*d*=60, 550, etc), it indicated that the infectious disease was localized and beginning to spread to various regions. Conversely, when the index was low and the number of confirmed cases was high (*d*=750, etc), it indicated that there was no obvious epicenter of the infectious disease, with the number of confirmed cases increasing relatively and evenly across different regions.

The contributions to the Theil index from each region (t_i), calculated from the number of cases on each date, were arranged in chronological order and visualized using a heatmap, as shown in Figure 2. Regions with a high proportion of confirmed cases are represented in red, while regions with a low proportion are colored blue. Notably, there are long intervals between the deep red patches in some regions such as California, Florida, and New York. Particularly, the periods of intense infection represented by these deep red patches were not repeated at short intervals. This phenomenon is of great importance in infectious disease management. Once a major epidemic in an area has subsided, the interval between subsequent outbreaks provides an opportunity to rebuild the health care infrastructure and implement preventive measures before the occurrence of the next epidemic.

Based on the observations from Figure 2, the epicenter of infectious diseases as indicated by the red patches alternates between New York, California, and Florida. This insight is crucial for understanding the underlying mechanisms of the spread of infectious diseases in the future. Furthermore, after *d*=750, both the red and blue colors fade over time, indicating the absence of a single epicenter, and a widespread outbreak of COVID-19. This pattern suggests the ineffectiveness of countermeasures against the spread of infectious diseases under these circumstances.

Figure 3A shows the contributions to the Theil index by region at *d*=60. The horizontal axis in the figure shows the state code (as listed in Multimedia Appendix 1). There is a significant contribution to the Theil index from New York State compared to the other regions. Figure 3B shows that at this point confirmed cases were highly localized in these regions.

There were relatively large negative contributions to the Theil index from California, Florida, and Texas, which were regions with high population ratios. It is interesting to note that there was little risk of infection in these regions at that point; however, the number of infected individuals rapidly increased following the concentration of confirmed cases in New York.

Figure 4 shows the contributions to the Theil index from each region at *d*=550 and 750. At *d*=550 shown in Figure 4A, the Theil index reaches a peak, and the trend of confirmed cases is increasing. This suggests that a new epidemic is emerging, mainly in Florida and Louisiana. However, their contributions are significantly smaller compared to New York at *d*=60, as seen in Figure 3. This indicates that regional disproportionality is much less pronounced than in the early stage of the COVID-19 pandemic. It is also interesting to look at data on *d*=750 as shown in Figure 4B, when confirmed cases in the United States are at their maximum. Although several regions show large contributions to the Theil index, the epicenter of COVID-19 is no longer obvious, suggesting that COVID-19 cases are uniformly distributed across the country.

Figure 1. Time trends of the Theil index on the left axis and the 7-day average number of the confirmed cases on the right axis on a logarithmic scale are shown in the red and blue curves, respectively. The horizontal axis is the number of days elapsed since January 21, 2020.

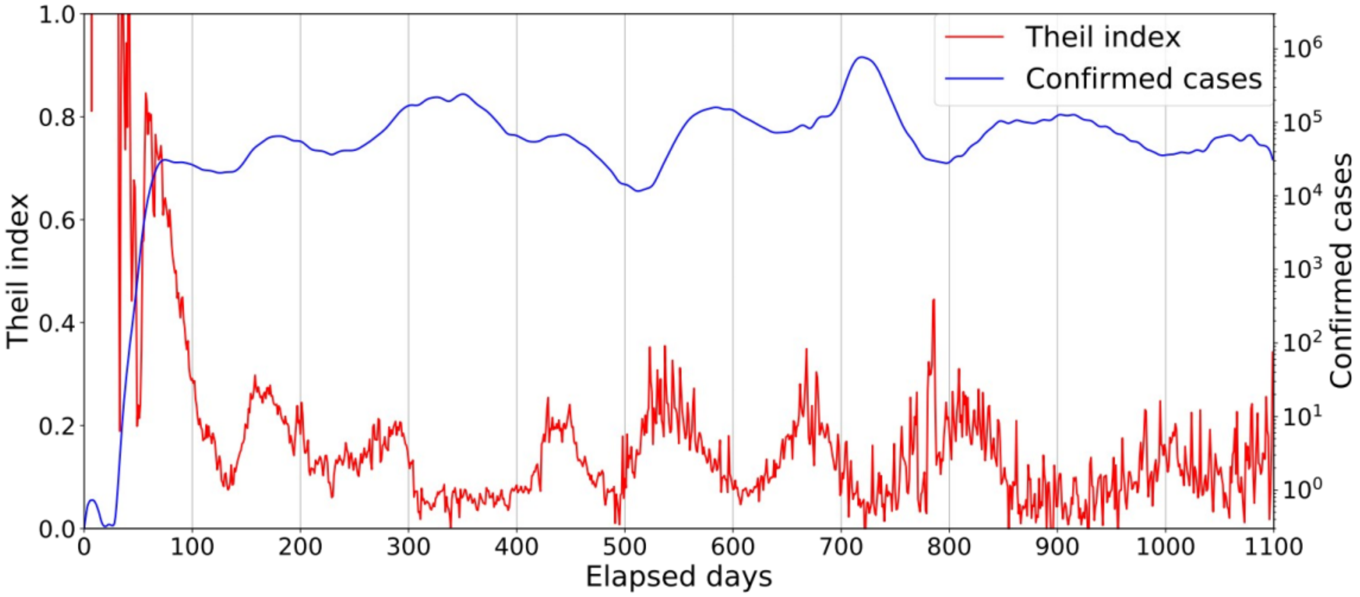


Figure 2. Partial contributions to the Theil index from each region, t_i, are displayed in a heatmap over time. The horizontal axis represents the number of days elapsed since January 21, 2020. The vertical axis shows the names of states in the United States. The positive (high concentration of incidences) and negative (low concentration) contributions to the Theil index correspond to deep red and blue colors, respectively.

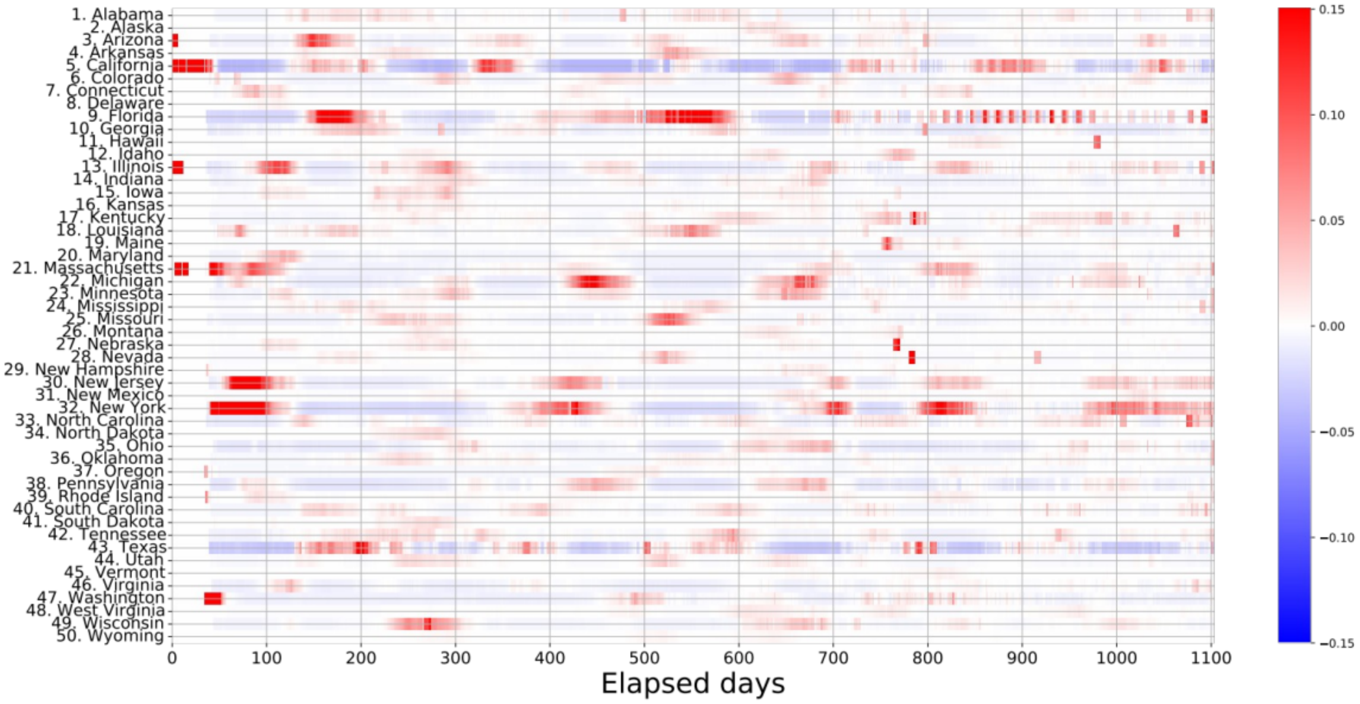


Figure 3. Partial contributions to the Theil index from each region, t_i, at d=60. The horizontal axis shows the state code given in Multimedia Appendix 1. (A) Comparison of the distribution of the confirmed cases and population. The vertical axis shows the ratio of a part to the whole region for populations and for the confirmed cases. (B) Contributions to the Theil index from each region. The vertical axis shows the strength of the contribution to the Theil index. The significantly high value of partial contribution to the Theil index is highlighted in orange.

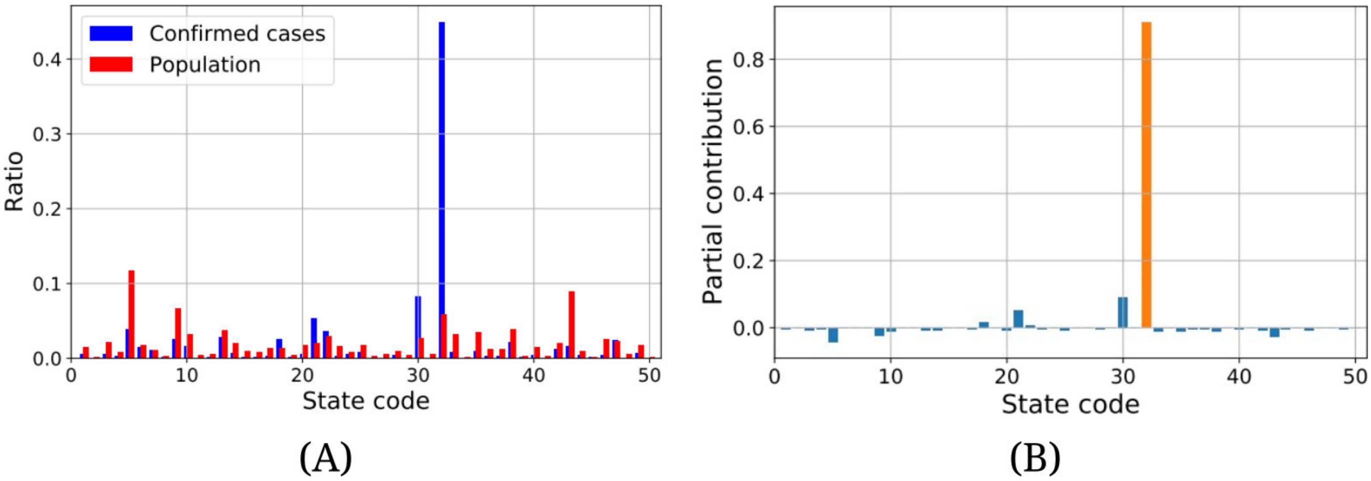
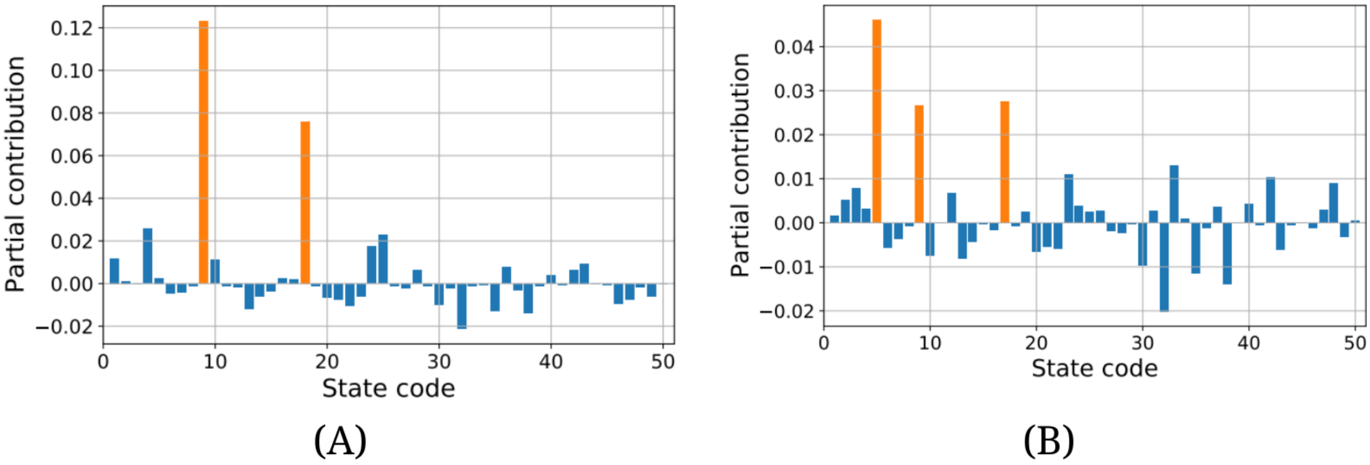


Figure 4. Partial contributions of the Theil index from each region, t_i at a specific date. The vertical axis shows the strength of contribution to the Theil index. The horizontal axis shows the state code given in Multimedia Appendix 1. (A) Contributions of the Theil index at d=550. (B) Contributions of the Theil index at d=750. The significantly high values of partial contribution to the Theil index are highlighted in orange.



Discussion

This study demonstrates the utility of the Theil index for quantifying regional disproportionalities in the distribution of COVID-19 cases. It offers an intuitive and efficient approach to identifying hotspots and monitoring the spread of infection. However, certain limitations may affect result interpretation.

The accuracy of the analysis depends on data quality; factors such as underreporting, delays in case confirmation, and regional differences in testing capacity may introduce biases into case counts. These issues could potentially impact the calculated Theil index and the assessment of regional disproportionalities.

Additionally, this study focuses primarily on confirmed cases rather than new infections, limiting its capacity to predict future spread. Therefore, the Theil index alone may not be sufficient for determining the timing and location of public health interventions, such as isolation measures. To support comprehensive policy-making, it should be used alongside other indicators, such as infection rates, hospitalization rates, and health care capacity.

Conventional spatiotemporal analysis methods [20,21] are widely used in epidemiology and public health to track infectious disease spread and visualize infection clusters over time in specific regions. These established tools effectively detect geographical clusters, identify areas with unusually high incidence, and reveal disease hotspots within defined spatial ranges.

In contrast, our method offers two distinct advantages. First, an increase in the Theil index acts as a precursor to a surge in the number of infected individuals. Second, it quantifies regional disproportionalities in incidence rates at any given time. Unlike conventional methods that emphasize physical distance and spatial proximity, our approach treats regions as discrete categories to calculate incidence rate disproportionalities. Although simple, this approach provides an intuitive way to identify epicenters at a lower computational cost compared to spatiotemporal scanning, enabling us to detect early surges in confirmed cases and pinpoint regions with concentrated infections.

For instance, the concentration of COVID-19 cases in New York at *d*=60 as shown in Figure 3A and B, cannot be overlooked when considering infection control. The lockdown was implemented in New York City [22] and coincided with a period when the contribution to the Theil index was concentrated in New York State. Although it is challenging to assess the direct impact of lockdown using the Theil index alone, the timing appears appropriate based on the pattern of concentration of confirmed cases.

Integrating our method with additional data sources, such as mobility patterns and health care capacity, will enhance pandemic response strategies, particularly for early intervention and efficient resource allocation.

In conclusion, this study demonstrates the application of the Theil index in quantifying regional disproportionalities in confirmed cases and monitoring their evolution over time. By analyzing confirmed case data in the United States, we have identified patterns of disproportionalities, specified epicenters, and characterized localized outbreaks.

Continued monitoring and analysis of regional differences in COVID-19 transmission remain essential, especially considering emerging variants and evolving public health responses. Our findings highlight the importance of understanding the regional dynamics of infected individuals for effective pandemic response interventions.

Incorporating the findings of this study will help policy makers refine strategies and address the diverse needs of different regions, ultimately increasing the effectiveness of pandemic response efforts and mitigating the impact of future health crises.

Lastly, the decomposability of the Theil index makes it possible to quantify and compare disproportionality in groups with specific characteristics, such as age, vaccination coverage, and health care accessibility. Identifying these disproportionalities will provide important insights for future pandemic responses.

Acknowledgments

We thank the members of the Division of Scientific Information and Public Policy (SiPP) at the Center for Infectious Disease Education and Research (CiDER) Osaka University for useful discussions. This research was supported by The Nippon Foundation—Osaka University Project for Infectious Disease Prevention.

Data Availability

Data were derived from public resources.

Conflicts of Interest

None declared.

Multimedia Appendix 1

US states codes.

[PNG File (Portable Network Graphics File), 58 KB-Multimedia Appendix 1]

References

1. Lopez L, Hart LH, Katz MH. Racial and ethnic health disparities related to COVID-19. JAMA. Feb 23, 2021;325(8):719-720. [doi: 10.1001/jama.2020.26443] [Medline: 33480972]
2. Adhikari S, Pantaleo NP, Feldman JM, Ogedegbe O, Thorpe L, Troxel AB. Assessment of community-level disparities in coronavirus disease 2019 (COVID-19) infections and deaths in large US metropolitan areas. JAMA Netw Open. Jul 1, 2020;3(7):e2016938. [doi: 10.1001/jamanetworkopen.2020.16938] [Medline: 32721027]
3. Perkins TA, España G. Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions. Bull Math Biol. Sep 4, 2020;82(9):118. [doi: 10.1007/s11538-020-00795-y] [Medline: 32888118]
4. Nakano T, Ikeda Y. Novel indicator to ascertain the status and trend of COVID-19 spread: modeling study. J Med Internet Res. Nov 30, 2020;22(11):e20144. [doi: 10.2196/20144] [Medline: 33180742]
5. Buchy P, Buisson Y, Cintra O, et al. COVID-19 pandemic: lessons learned from more than a century of pandemics and current vaccine development for pandemic control. Int J Infect Dis. Nov 2021;112:300-317. [doi: 10.1016/j.ijid.2021.09.045] [Medline: 34563707]
6. Stenseth NC, Dharmarajan G, Li R, Shi ZL, Yang R, Gao GF. Lessons learnt from the COVID-19 pandemic. Front Public Health. 2021;9:694705. [doi: 10.3389/fpubh.2021.694705] [Medline: 34409008]
7. Ikeda Y, Sasaki K, Nakano T. A new compartment model of COVID-19 transmission: the broken-link model. Int J Environ Res Public Health. Jun 3, 2022;19(11):6864. [doi: 10.3390/ijerph19116864] [Medline: 35682447]
8. Pandemic preparedness. Nature. Oct 26, 2022. URL: https://www.nature.com/collections/jaacfgeief [Accessed 2024-03-14]
9. Sasaki K, Ikeda Y, Nakano T. The effects of behavioral restrictions on the spread of COVID-19. Rep. 2022;5(4):37. [doi: 10.3390/reports5040037]

10. Jinjarak Y, Ahmed R, Nair-Desai S, Xin W, Aizenman J. Accounting for global COVID-19 diffusion patterns, January-April 2020. Econ Disaster Clim Chang. 2020;4(3):515-559. [doi: 10.1007/s41885-020-00071-2] [Medline: 32901228]
11. Mollalo A, Vahedi B, Rivera KM. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Sci Total Environ. Aug 1, 2020;728:138884. [doi: 10.1016/j.scitotenv.2020.138884] [Medline: 32335404]
12. Villanustre F, Chala A, Dev R, et al. Modeling and tracking Covid-19 cases using big data analytics on HPCC system platform. J Big Data. 2021;8(1):33. [doi: 10.1186/s40537-021-00423-z] [Medline: 33614394]
13. Yue H, Hu T. Geographical detector-based spatial modeling of the COVID-19 mortality rate in the continental United States. Int J Environ Res Public Health. Jun 25, 2021;18(13):6832. [doi: 10.3390/ijerph18136832] [Medline: 34202168]
14. Theil H. Economics and Information Theory. North-Holland Publishing Company: Amsterdam, North-Holland; 1967. [Accessed 2024-03-14] ISBN: 0444102825
15. Manz KM, Mansmann U. Inequality indices to monitor geographic differences in incidence, mortality and fatality rates over time during the COVID-19 pandemic. PLoS ONE. 2021;16(5):e0251366. [doi: 10.1371/journal.pone.0251366] [Medline: 33984055]
16. Kullback S, Leibler RA. On information and sufficiency. Ann Math Statist. Mar 1951;22(1):79-86. [doi: 10.1214/aoms/1177729694]
17. Holshue ML, DeBolt C, Lindquist S, et al. First case of 2019 novel coronavirus in the United States. N Engl J Med. Mar 5, 2020;382(10):929-936. [doi: 10.1056/NEJMoa2001191] [Medline: 32004427]
18. COVID-19 data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Github. URL: https://github.com/CSSEGISandData/COVID-19 [Accessed 2024-03-14]
19. National population totals and components of change: 2020-2023. US Census Bureau. URL: https://www.census.gov/data/datasets/time-series/demo/popest/2020s-national-total.html [Accessed 2024-03-14]
20. Kulldorff M, Heffernan R, Hartman J, Assunção R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS Med. Mar 2005;2(3):e59. [doi: 10.1371/journal.pmed.0020059] [Medline: 15719066]
21. Desjardins MR, Hohl A, Delmelle EM. Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: detecting and evaluating emerging clusters. Appl Geogr. May 2020;118:102202. [doi: 10.1016/j.apgeog.2020.102202] [Medline: 32287518]
22. Huang Y, Li R. The lockdown, mobility, and spatial health disparities in COVID-19 pandemic: a case study of New York City. Cities. Mar 2022;122:103549. [doi: 10.1016/j.cities.2021.103549] [Medline: 35125596]

Abbreviations

CSSE: Center for Systems Science and Engineering
KL: Kullback-Leibler

Edited by Amaryllis Mavragani; peer-reviewed by Alonso Inostrosa-Psijas, Zihao Guo; submitted 06.04.2024; final revised version received 04.11.2024; accepted 06.11.2024; published 03.01.2025

Please cite as:
Sasaki K, Ikeda Y, Nakano T
Quantifying the Regional Disproportionality of COVID-19 Spread: Modeling Study
JMIR Form Res 2025;9:e59230
URL: https://formative.jmir.org/2025/1/e59230
doi: 10.2196/59230