A Longitudinal and Geospatial Analysis of COVID-19 Tweets During the Early Outbreak Period in the United States

Raphael Cuomo
University of California, San Diego
Vidya Purushothaman
University of California, San Diego
Jiawei Li
S-3 Research LLC, San Diego
Mingxiang Cai
University of California, San Diego
Timothy Mackey (tmackey@ucsd.edu)
University of California, San Diego

Research Article

Keywords: Ecological Epidemiology, COVID-19, Infectious Diseases, Geospatial Analysis, Social Media

Posted Date: November 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-109161/v1

Version of Record: A version of this preprint was published at BMC Public Health on April 24th, 2021. See the published version at https://doi.org/10.1186/s12889-021-10827-4.

Abstract

Introduction: Local rates of COVID-19 cases and deaths may not accurately convey the variability in community-level concern about COVID-19 during the early outbreak period in the United States. Social media interaction may elucidate communication about COVID-19 in this critical period, during which communities may have formulated initial conceptions pertaining to the perceived gravity of the disease and potential behavioral strategies for prevention.

Methods: Scripts were written to obtain tweets related to COVID-19 from Twitter. Using manually-annotated tweets about symptom-related concerns from a prior study, a machine learning classifier was applied to obtain a subset of tweets about concerns relating to COVID-19. The longitudinal relationship between these social media posts and active COVID-19 cases was assessed using linear and exponential regression. Changes in the geospatial clustering of tweets was assessed for the top five most populous cities in the United States.

Results: Social media posts relating to COVID-19 concerns appeared more predictive of active COVID-19 cases as temporal distance increased. The distribution of tweets in New York City and Phoenix appeared concentrated in city centers, whereas tweets from other cities were more residential. Tweets from New York City became more highly concentrated, but the opposite trend was observed in tweets from Los Angeles.

Conclusion: Clustering of social media posts about COVID-19 revealed discrepancies across major US cities. General concern about the COVID-19 pandemic may moderate the relationship between behavioral/environmental factors and COVID-19 transmission. The degree and modality of this moderating effect may differ across US areas.

Background

In mid-March 2020, approximately 150,000 cases of coronavirus 2019 (COVID-19) had been confirmed globally, with only about 2,000 of these cases occurring in the United States [1]. Attention to the COVID-19 pandemic was likely to be widespread [2-4], though current data do not allow for a highly valid means of estimating the extent of concern. Furthermore, many who become infected do not exhibit symptoms or exhibit very mild symptoms, complicating the relationship between case counts and attention to the pandemic [5, 6]. Despite this, an asymptomatic or mildly symptomatic individual would still be able to transmit COVID-19 to individuals who may be at higher risk.

In the first half of March 2020, limited numbers of tests for COVID-19 led public health officials to suggest that only certain individuals need to seek confirmation of COVID-19 infection with a diagnostic test [7, 8]. As a result, during this time, it was recommended that testing be reserved for individuals suffering relatively extreme symptoms requiring hospitalization. Hence, in the early outbreak period, there may have been marked potential to convey inaccurate spatial variation for pandemic-related concern; limited numbers of diagnostic tests indicate that it may be worthwhile to assess distributions presented by non-

traditional sources of data. One approach to estimating this volume is by using “infoveillance” approaches, including using Internet and social media data to identify the distribution and determinants of disease-related concern [9-11].

Concordantly, this study aims to explore the use of geospatial and statistical methods to understand how social media data from the popular microblogging global platform Twitter may be leveraged to better estimate geographic distributions of attention to the COVID-19 pandemic during the early outbreak period in the United States.

Methods

Study methods broadly included Python scripting, mathematical transformations, regression analysis, and geospatial statistics. The distributions predominantly under scrutiny were spatial, and temporal fluctuations were also assessed. Analysis was conducted in ArcGIS version 10.6 and R version 3.6.0.

Data Collection

Scripts for data collection were written in the Python programming language to access the Twitter public API stream and to prospectively download publically-available posts from the United States between March 3rd and March 17th, inclusive. Keywords used to obtain tweets were intended to obtain a broad representation of conversations regarding COVID-19. These were “corona outbreak,” “corona,” “anticorona,” “coronavirus,” “Wuhan virus,” “COVID,” “Wuhan pneumonia,” and “pneumonia of unknown cause.” These keywords were chosen on the basis of structured manual searches conducted on Twitter that detected content related to the COVID-19 outbreak as posted by users, and they have been used to identify tweets pertaining to COVID-19 in prior studies [12, 13]. Approximately 60 million messages were collected during this timeframe, and 459,937 had available geospatial information in the metadata of collected messages. Geospatial information was in the form of latitude and longitude coordinates. The original source of this information was the user’s device, which transmits location information to Twitter if the user has given permission for this to occur. Coordinates are then made available to third parties via the Twitter API.

COVID-19 cases at the county and national levels were available from the 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository, actively maintained on GitHub by the Johns Hopkins University Center for Systems Science and Engineering, which collects case information reported from a variety of validated sources [14]. Cases were obtained for each day when posts were collected from Twitter. Active cases were used in regression modeling, computed by subtracting county-level recoveries and deaths from confirmed cases. Normalization of tweets for national and local analysis was done by dividing the number of tweets by the amount of people living in a given county or census tract. Population at these ecological units was available from the US Census Bureau.

Machine Learning

Assisted by the biterm model, a natural language processing algorithm, a subset of the 60 million-tweet corpus was hand coded for self-identification of symptom-related concerns in a separate study. Coding involved multiple rates with high inter-rater reliability; a more detailed description of these methods can be found in Mackey et al., 2020 [12]. These posts largely did not contain geospatial information but were also from the early outbreak period. The number of coded posts about symptoms from this prior study were obtained that were within the time period of this study. These were used, along with a matching number of posts coded as unrelated to symptoms, to create a machine learning classifier using a support-vector machine (SVM) algorithm. The SVM classifier was applied to the 459,937 posts in this study with geospatial coordinates to identify 249,778 posts whose contents are more consistent with concerns about COVID-19 symptoms and less consistent with news coverage, jokes, celebrity news, and other less-relevant post typologies. This subset of 249,778 geo-identifiable posts was used in all regression models computed in this study.

Longitudinal Analysis

National analysis involved scrutiny of the longitudinal relationship between tweets and cases at the county level. Bivariate regressions were conducted to investigate the strength of relationships between county-level tweets and county-level active cases. These models were computed to compare the distribution of tweets on the same day as the distribution of active cases, as well as for every combination of time-lagged tweets with active cases. R^2 values to assess the fit of linear relationships were compared to Nagelkerke's R^2 to assess the fit of exponential relationships for the set of same-day or time-lagged relationships. Nagelkereke's method provides a range from zero to one, as with the R^2 statistic for linear relationships, in computing a fit relative to a nested null model without predictors.

Geospatial Analysis

To obtain a sense of the county-level distribution of tweets with COVID-19 keywords within the United States, tweets were aggregated across the March 3-17 data collection period and divided by county-level population from the US Census Bureau.

Geospatial cluster analysis was conducted for the top five most populous cities in the United States. For cities bounded by the perimeter of a single county, the distribution across census tracts was analyzed at the county level; otherwise, the distribution across census tracts was analyzed at the city level. This strategy was undertaken in order to relay relevant distributions across space, particularly with cities contiguous with numerous other cities and towns. Unique coordinates from the overall study period were utilized for cluster analysis of all five areas, so as to prevent clustering statistics from being biased toward locations of individuals with extreme propensities for posting. For areas with sufficient sample size, additional analysis was done on the first day (March 3rd) and the last day (March 17th) of the study period, in order to relay change in the distribution of social media messages. The computational analysis itself involved calculation of the Getis Ord Gi* statistic for each census tract within the area. These *G*

statistics were used to obtain corresponding *z* scores, which were then visualized to relay high-value “hot” spots and low-value “cold” spots.

Results

Approximately 460,000 tweets included geospatial information originating from 11,022 unique coordinates. Tweets were available for 1,971 US counties (55% of US counties), with tweets per capita ranging from 0.0000076 per capita to 0.75 per capita. National analysis assessed active cases from March 3rd, which totaled 59 cases across 24 counties ranging from 1 to 14 cases per county; increasing on March 17th to 5,911 active cases across 523 counties ranging from 1 to 807 per county.

Longitudinal Analysis

Regression models comparing non-normalized tweets and active cases tended to exhibit better fit as the day of active case data was further from the day of tweet data (**Table 1**). This prediction was especially strong for exponential models, with the fit of active cases by COVID-related tweets reaching an R^2 of 0.76 for the exponential model with tweets from March 4th predicting active cases 13 days later. Average R^2 for same-day prediction was 0.16 for linear models and 0.15 for exponential models. Although only one comparison was available, R^2 for 14-day prediction was 0.49 for the linear model and 0.73 for the exponential model. The tweet covariate was not statistically significant when predicting cases on March 3rd, March 4th, or March 5th. Conversely, 96% of bivariate models for subsequent days exhibited *p* values under 0.05 for the tweet covariate.

Geospatial Analysis

Between March 3rd and March 17th, we collected 11,022 tweets with unique coordinates, with 3,842 (34.9%) of these represented on March 3rd and 11,022 (76.4%) represented on March 17th. Within the five most populous cities in the United States (or their respective encompassing counties), there were 95 unique coordinates from New York City, with 35 (36.8%) on March 3rd and 66 (69.5%) on March 17th; 178 unique coordinates from Los Angeles County, with 58 (32.3%) on March 3rd and 147 (82.5%) on March 17th; 86 from Cook County (i.e. Chicago), with 27 (31.4%) on March 3rd and 59 (68.6%) on March 17th; 81 from Houston, with 26 (32.1%) on March 3rd and 73 (90.1%) on March 17th; and 40 from Maricopa County (i.e. Phoenix) with 13 (32.5%) on March 3rd and 25 (62.5%) on March 17th. Therefore, across all city areas, approximately one-third of the number of locations interacting with the COVID-19 topic were represented in early March. The number of locations approximately doubled by mid-March, consistent with the national trend.

Cluster analysis was conducted for each city area. In New York City, a cluster was detected in Manhattan, which is the most densely populated area of the city. Conversely, in Los Angeles County, the densely populated downtown area was a cold spot, whereas the relatively residential areas of West Los Angeles and San Gabriel Valley were hot spots. The same pattern was observed for Cook County (i.e. Chicago) and Houston, where city centers were relatively less representative of social media interaction with

COVID-19. However, the distribution of tweets within Maricopa County (i.e. Phoenix) seemed more consistent with that of New York City, with relatively high representation in the densely-populated city center (**Figure 1**).

Analysis for different time points was possible for New York City and Los Angeles County. On March 3rd, the distribution of coordinates in New York City spanned across lower and central Manhattan, radiating across the East River into parts of Brooklyn and Queens. Despite more social media activity on March 17th, the distribution of coordinates in New York City became much more concentrated on the island of Manhattan. The opposite was trend was observed in Los Angeles County. On March 3rd, small clusters of tweets were detected from Los Angeles International Airport and some areas of San Gabriel Valley. On March 17th, these small clusters appeared to have grown to encompass most areas within the county's South Bay, Westside, San Gabriel Valley, and Southeast regions (**Figure 2**).

Discussion

This study revealed a number of quantitative aberrations, discrepancies, and findings that deem further study and could help in better assessing the epidemiological characteristics of the current COVID-19 outbreak using geolocated tweets as a proxy indicator for community attention to disease outbreaks, with possible insights related to disease transmission trends.

Analysis of county-level data in the United States also suggested a time lag between social media posts and predicted COVID-19 cases. In this study, posts exhibited much better fit of active cases as the time gap between posts and cases increased, especially in exponential models. The reason for this discrepancy may be that social media users are responding to COVID-19 news coverage and perceived risk, though the actual disease burden based on case counts may not be reported until later, which could also be impacted by the availability and speed of testing.

Analysis of the areas for the five most populated cities in the United States revealed some consistencies and some differences. All cities appeared to follow the national longitudinal trend, in which about one-third of locations interacting with the pandemic topic were doing so on March 3rd, which about doubled on March 17th. This result appears to indicate that, despite the total number of posts being correlated to local COVID-19 rates, the unique locations interacting with this issue appeared to be consistent with the evolving national concern about the pandemic.

Temporal differences were observed between New York City and Los Angeles County, wherein represented locations became more concentrated in New York City and less concentrated in Los Angeles County. As New York City exhibited extremely high COVID-19 rates in mid-March compared to the rest of the country, this result may partly be due to the dramatic effect observed on the normally-bustling area of Manhattan, as reflected in the following tweets from Manhattan on March 17th: "It was so eerie with empty streets in NYC Sunday like a science fiction movie #covid-19 #coronavirus," "Here's what Grand Central Station looks like at 6:30 PM as the mayor ponders whether to require New Yorkers to shelter-in-place. #coronavirus #covid_19 @ Grand Central Terminal," and "As seen in #GrandCentral today. Normally

BUSTLING throughout. So eerie. #COVID19 #coronavirus #Covid_19." Los Angeles County exhibited the opposite trend, whereas the distribution of tweets became more dispersed between March 3rd and March 17th. Though some tweets from Los Angeles on March 17th observed the newfound lack of traffic, tweets were generally dispersed across topics that included general precautions and store/event closures, thereby indicating that a heightened degree of awareness to the pandemic's impact reached new local communities between March 3rd and March 17th.

The cities of New York and Phoenix exhibited different clustering patterns than those for the areas of Los Angeles, Chicago, and Houston. In New York and Phoenix, clusters were generally from relatively densely populated city centers. However, in Los Angeles, Chicago, and Houston, clusters were mostly outside city centers. This difference suggests that, in some areas, people have tweeted their concerns and reactions primarily from home, whereas this was not the case for New York City and Phoenix. The similarity of the Phoenix area to New York City may be concerning, as New York City was thought to be, at this time, a national outlier with respect to COVID-19 burden. However, relatively few tweets were collected from the Phoenix area, so this similarity should be considered preliminary and should be further investigated.

Results from this study bear consistency with several published social media analysis on prior spread of infectious diseases. A 2016 study of Japanese tweets containing influenza symptoms found a time lag between the rate of tweets with forecasting words and the national influenza rate, as we found in this study [15]. A study of Korean tweets from 2016 found that tweets with keywords related to Middle East respiratory syndrome (MERS) were more predictive of the Korean quarantine rate as the time lag increased, but less predictive of laboratory-confirmed cases [16]. Finally, a 2010 study of English-language tweets about the H1N1 pandemic found that tweets which were automatically coded as indicative of personal disease experience, based on keywords, exhibited high correlation with personal disease experience after manual verification [17].

Throughout the COVID-19 crisis, maps have been popularly used to describe the extent and distribution of the pandemic [14, 18, 19]. However, these maps have focused on the disease itself, whereas social consequences of the disease (such as social media posts) may also provide useful insights warranting the production of maps [20, 21]. Furthermore, there exist powerful geospatial and statistical methods that can applied to these data. Though the data themselves are not direct records of the disease, careful scrutiny of indirect data manifestations of COVID-19 may be able to lend insights about the disease which otherwise are impossible to obtain.

Limitations

Findings from this study are subject to a number of important limitations. Importantly, a fraction of overall posts were geolocated, which raises the possibility of sampling bias with respect to the overall tweet corpus. Furthermore, some communities and their specific demographic features may have a greater propensity to post Twitter messages, regardless of conditions experienced in any of its users' communities. While it is possible that this error is approximately systemic, and thereby may not

appreciably contribute to the discovery of spurious relationships, little analysis has been done to verify whether the proportion of posts responding to local conditions is consistent across geospatial units. Similarly, we have considered the variation in COVID-19 cases to be reflective of true variation at an artificially deflated magnitude, due to insufficient testing. However, testing initiatives of local public health bodies may have appreciably varied during the study period, potentially resulting in erroneous variation, in addition to the suspected erroneous variation in magnitude.

This study is intended to be primarily hypothesis generating, and findings from this study should be further validated in more highly controlled settings. For example, a study in a manageable set of smaller communities should seek to determine whether variation in social media data is highly predictive of community caseloads that were obtained by especially thorough testing of those communities. Such a study may also seek to assess differences in the predictive power of social media messages at different intervals from the caseload prediction time point.

Conclusion

Results from this study suggest that a homogenous strategy of outbreak-related risk communication may not be efficacious. Across five major US cities, geospatial patterns of online postings about pandemic-related concern revealed apparent discrepancies, with suggested further discrepancies relating to how these clustering patterns change over time. The utility of online communication for measuring early concern about infectious disease outbreaks warrants further study, as does the potential moderating effect of concern on behavior-related prevention of transmission.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials:

The datasets used and/or analyzed during the current study are available from the corresponding author upon request.

Competing Interests:

TKM, JL, and MC are employees of the startup company S-3 Research LLC. S-3 Research is a startup funded and currently supported by the National Institutes of Health – National Institute on Drug Abuse through a Small Business Innovation and Research contract for opioid-related social media research and technology commercialization. Author reports no other conflict of interest associated with this manuscript and have not been asked by any organization to be named on or to submit this manuscript. RC and VP report no conflicts of interest or financial relationships associated with this manuscript.

Funding:

This research was funded by a grant from the California HIV/AIDS Research Program (Award #R00RG2400).

Author's contributions:

RC and TK conceptualized the study. MC conducted data collection. RC and VP analyzed the data. RC and TK worked on writing the manuscript. All authors approved the final manuscript.

Acknowledgements: Not applicable

References

1. Verhagen MD, Brazel DM, Dowd JB, Kashnitsky I, Mills M: **Mapping hospital demand: demographics, spatial variation, and the risk of "hospital deserts" during COVID-19 in England and Wales**. In.: OSF Preprints; 2020.
2. Lau H, Khosrawipour V, Kocbach P, Mikolajczyk A, Ichii H, Schubert J, Bania J, Khosrawipour T: **Internationally lost COVID-19 cases**. *Journal of Microbiology, Immunology and Infection* 2020.
3. Wang X, Ma Z, Ning Y, Chen C, Chen R, Chen Q, Zhang H, Li C, He Y, Wang T: **Estimating the case fatality ratio of the COVID-19 epidemic in China**. *medRxiv* 2020.
4. Zhuang Z, Cao P, Zhao S, Lou Y, Wang W, Yang S, Yang L, He D: **Estimation of local novel coronavirus (COVID-19) cases in Wuhan, China from off-site reported cases and population flow data from different sources**. *medRxiv* 2020.
5. Tian S, Hu N, Lou J, Chen K, Kang X, Xiang Z, Chen H, Wang D, Liu N, Liu D: **Characteristics of COVID-19 infection in Beijing**. *Journal of Infection* 2020.
6. Hu Z, Song C, Xu C, Jin G, Chen Y, Xu X, Ma H, Chen W, Lin Y, Zheng Y: **Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China**. *Science China Life Sciences* 2020:1-6.
7. World Health Organization: **Global surveillance for COVID-19 caused by human infection with COVID-19 virus: interim guidance, 20 March 2020**. In.: World Health Organization; 2020.
8. Odendaal WG: **A Method to Model Outbreaks of New Infectious Diseases with Pandemic Potential such as COVID-19**. *medRxiv* 2020.
9. Eysenbach G: **Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet**. *Journal of medical Internet research* 2009, **11**(1):e11.
10. Boulos MNK, Geraghty EM: **Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics**. In.: BioMed Central; 2020.

11. Zhou C, Su F, Pei T, Zhang A, Du Y, Luo B, Cao Z, Wang J, Yuan W, Zhu Y: **COVID-19: Challenges to GIS with Big Data**. *Geography and Sustainability* 2020.
12. Mackey TK, Purushothaman V, Li J, Shah NS, Nali M, Bardier C, Cuomo RE: **Machine Learning to Detect Self-Reporting of COVID-19 Symptoms, Testing Access and Recovery on Twitter**. *[Under Review]* 2020.
13. Cuomo RE, Purushothaman V, Li J, Cai M, Mackey Tk: **A Statistical and Geospatial Analysis of Global and National COVID-19-Related Tweets**. *[Under Review]* 2020.
14. Dong E, Du H, Gardner L: **An interactive web-based dashboard to track COVID-19 in real time**. *The Lancet Infectious Diseases* 2020.
15. Hayate I, Wakamiya S, Aramaki E: **Forecasting word model: Twitter-based influenza surveillance and prediction**. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers: 2016*; 2016: 76-86.
16. Shin S-Y, Seo D-W, An J, Kwak H, Kim S-H, Gwack J, Jo M-W: **High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea**. *Scientific reports* 2016, **6**:32920.
17. Chew C, Eysenbach G: **Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak**. *PloS one* 2010, **5**(11):e14118.
18. **Novel Coronavirus (COVID-19)** [https://www.healthmap.org/covid-19/]
19. **Novel Coronavirus Infection Map** [https://hgis.uw.edu/virus/]
20. Rossman H, Keshet A, Shilo S, Gavrieli A, Bauman T, Cohen O, Balicer R, Geiger B, Dor Y, Segal E: **A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys**. *medRxiv* 2020.
21. Qiu Y, Chen X, Shi W: **Impacts of social and economic factors on the transmission of coronavirus disease (COVID-19) in China**. *medRxiv* 2020.

Tables

Due to technical limitations, table 1 is only available as a download in the Supplemental Files section.

Figures

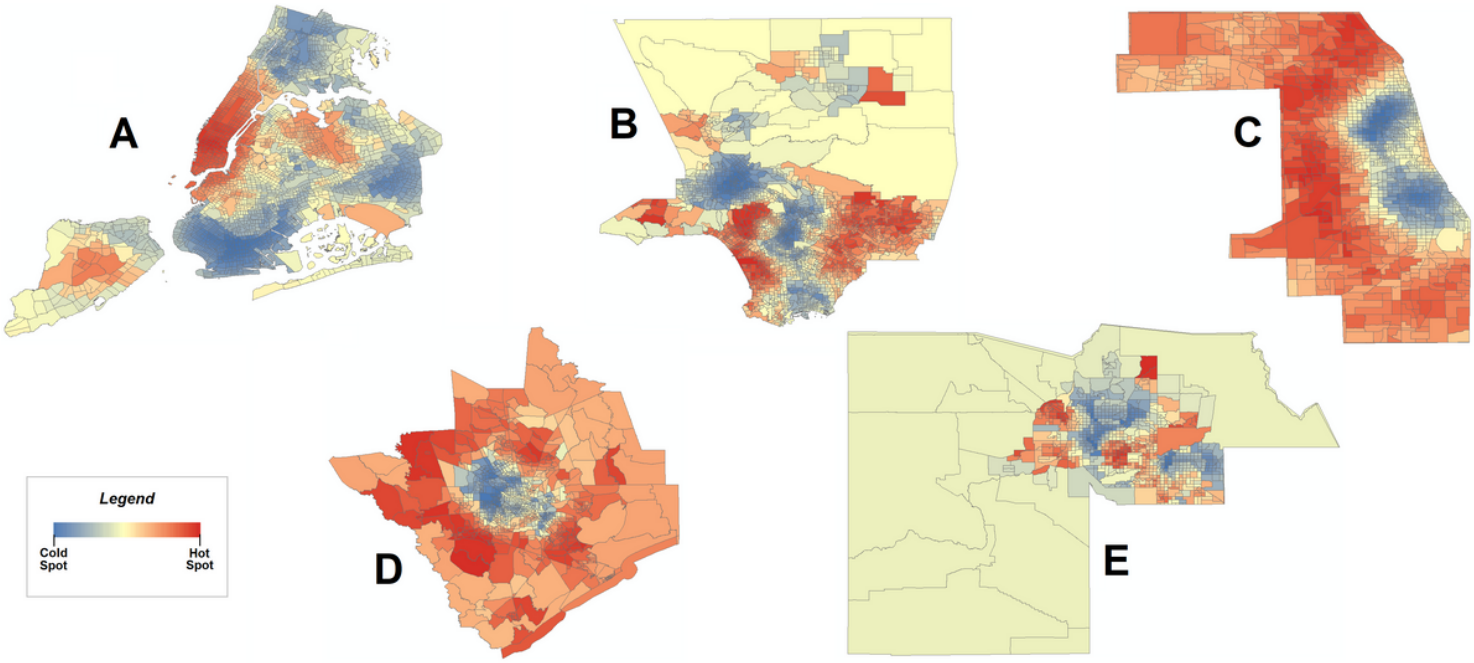

Figure 1

Z-scores for the Getis Ord Gi* statistic, indicating geospatial clustering of tweets about COVID-19 from (A) New York City, (B) Los Angeles County, (C) Cook County (i.e. Illinois), (D) Houston, and (E) Maricopa County (i.e. Phoenix) 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.

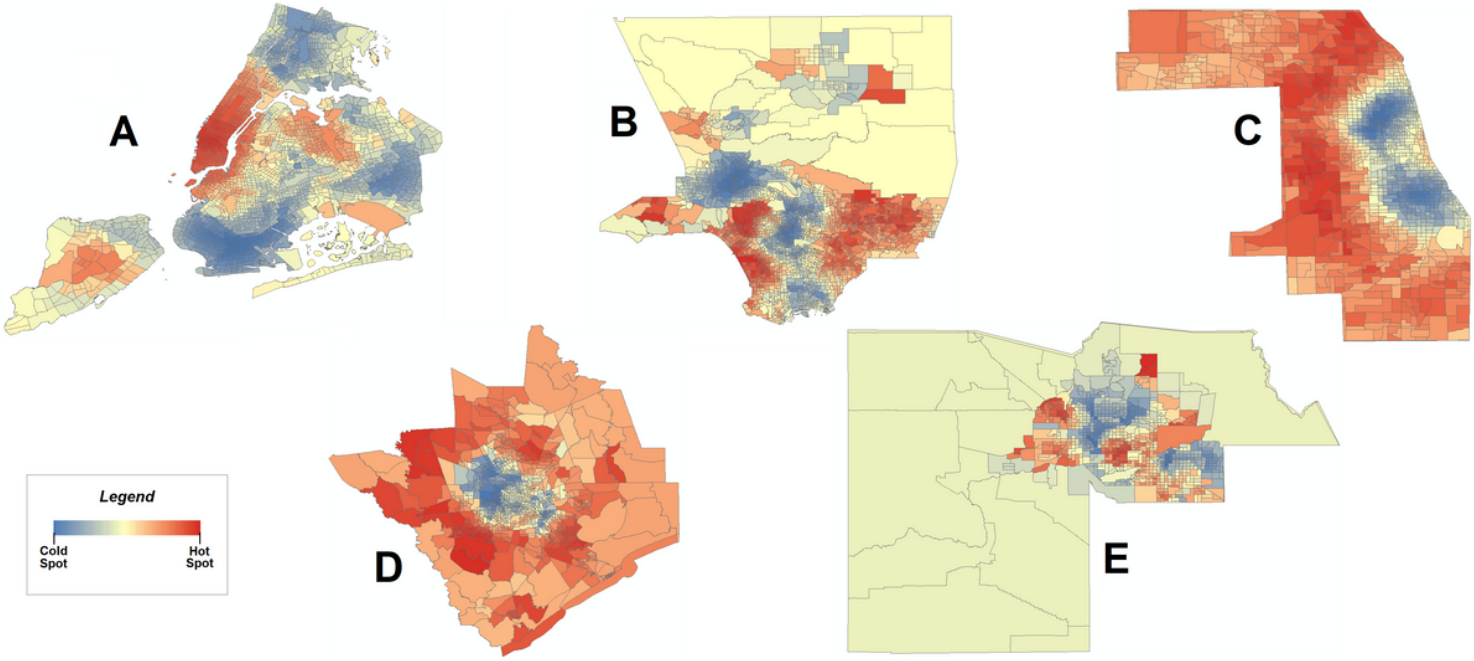

Figure 1

Z-scores for the Getis Ord Gi* statistic, indicating geospatial clustering of tweets about COVID-19 from (A) New York City, (B) Los Angeles County, (C) Cook County (i.e. Illinois), (D) Houston, and (E) Maricopa County (i.e. Phoenix) 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.

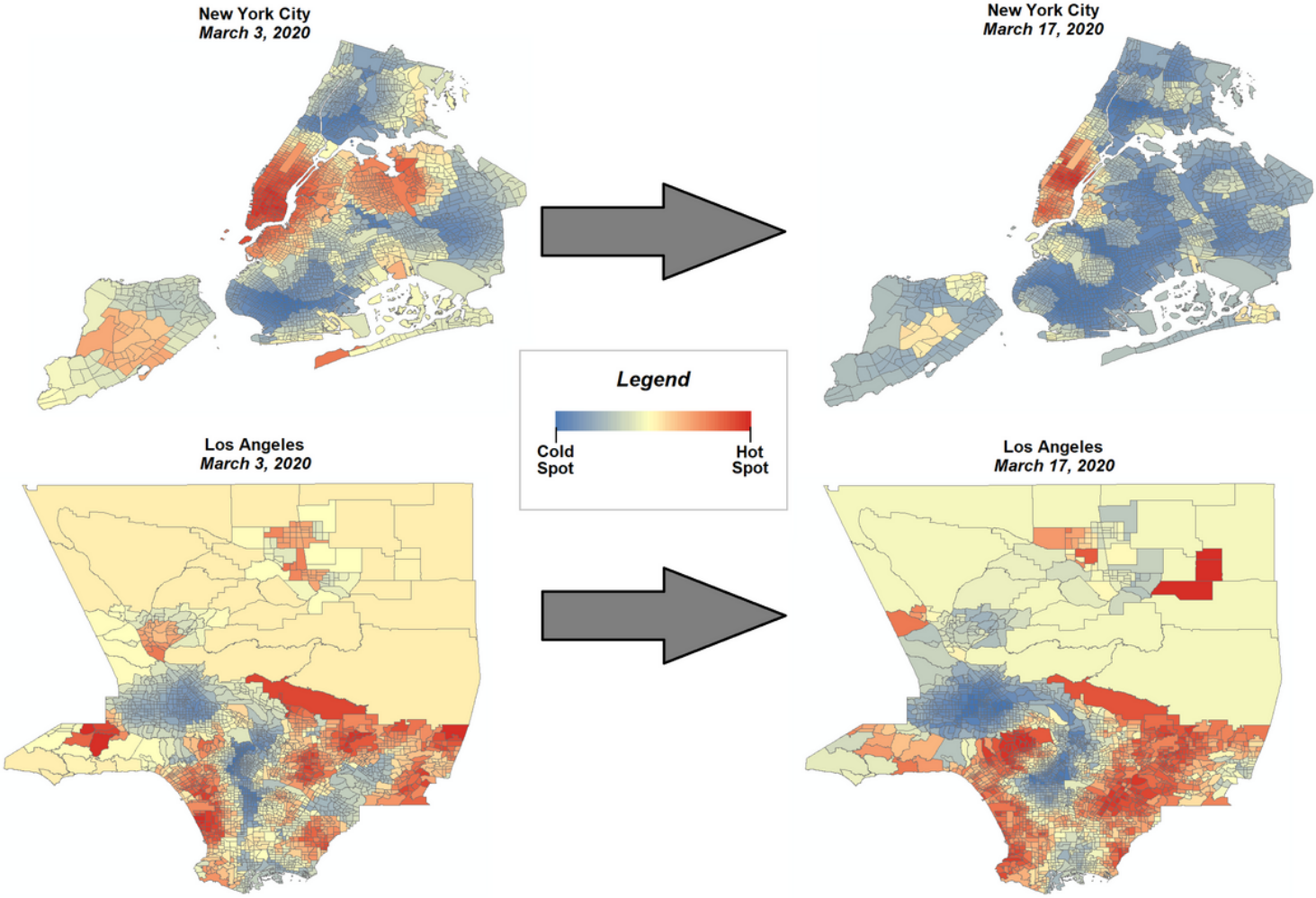

Figure 2

Changes in z-scores for the Getis Ord Gi* statistic, relaying clustering at the start of the study period (March 3rd) and the end of the study period (March 17th) for New York City and Los Angeles County 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.

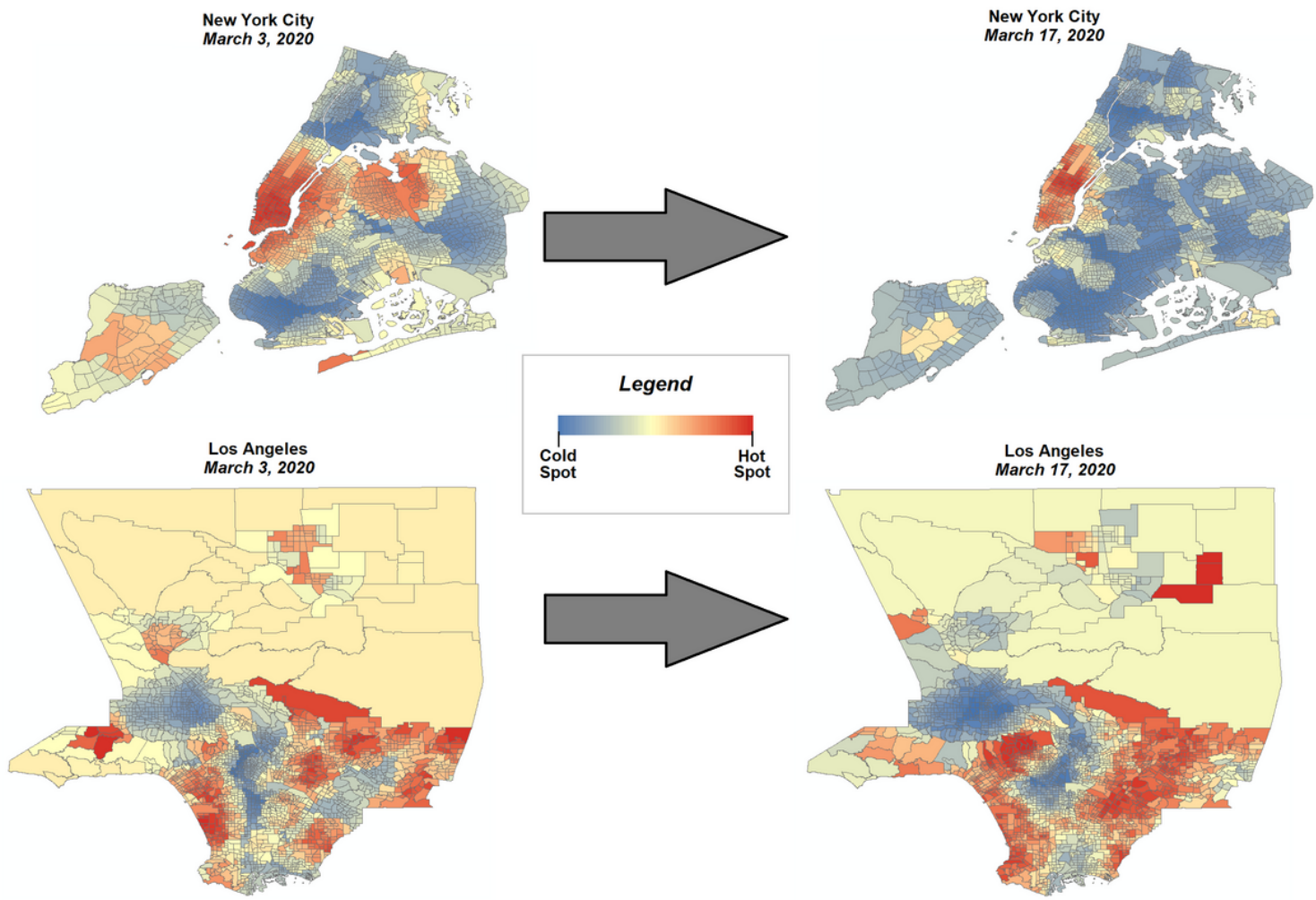

Figure 2

Changes in z-scores for the Getis Ord Gi* statistic, relaying clustering at the start of the study period (March 3rd) and the end of the study period (March 17th) for New York City and Los Angeles County 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.JPG
- Table1.JPG