On the predictability of COVID-19 in USA: A Google Trends analysis

Amaryllis Mavragani (amaryllis.mavragani1@stir.ac.uk)
Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling, FK9 4LA, Scotland, UK

Konstantinos Gillas
Department of Business Administration, University of Patras, Greece

Research Article

Keywords: big data, coronavirus, COVID-19, infodemiology, Google Trends, SARS-CoV-2, predictive analysis

Posted Date: May 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-27189/v1

Version of Record: A version of this preprint was published at Scientific Reports on November 26th, 2020. See the published version at https://doi.org/10.1038/s41598-020-77275-9.

On the predictability of COVID-19 in USA: A Google Trends analysis

Amaryllis Mavragani[1,*] and Konstantinos Gillas[2]
[1]*Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling, FK9 4LA, Scotland, UK;*
Tel.: +44 (0) 7523782711. Email: amaryllis.mavragani1@stir.ac.uk
[2]*Department of Business Administration, University of Patras, Greece; gillask@upatras.gr*

Abstract

During the difficult times that the world is facing due to the COVID-19 pandemic that has already had severe consequences in all aspects of our lives, it is imperative to explore novel approaches of monitoring and forecasting the regional outbreaks as they happen or even before they do. In this paper, the first approach of exploring the role of Google query data in the predictability of COVID-19 in the US at both national and state level is presented. The results indicate that Google Trends correlate with COVID-19 data, while the estimated models exhibit strong predictability of COVID-19. In line with previous work that has argued on the value of online real-time data in the monitoring and forecasting epidemics and outbreaks, it is evident that such infodemiology approaches can assist public health policy makers, in order to address the most crucial issue; that of flattening the curve, allocating health resources, and increasing the effectiveness and preparedness of the respective health care systems.

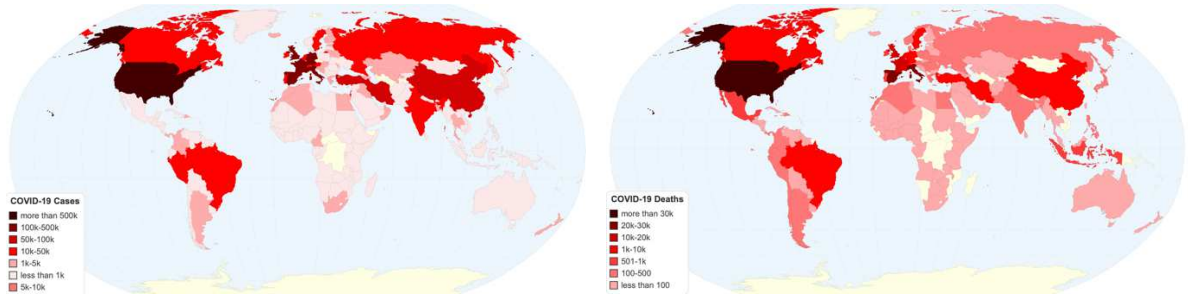
Keywords: big data**;** coronavirus; COVID-19; infodemiology; Google Trends; SARS-CoV-2; predictive analysis

Introduction

In December 2019, a novel coronavirus of unknown source was identified in a cluster of patients in the city of Wuhan, in Hubei, China [1]. The outbreak first came to international attention after WHO reporting of a cluster of pneumonia cases on Twitter on January 4th [2], followed by an official report on the 5th [3]. China reports its first COVID-19 related death on January 11th, while on the 13th, the first case outside China was identified [4]. On January 14th, the World Health Organization (WHO) tweeted that Chinese preliminary investigations reported that no human-to-human transmission had been identified [5]. However, the virus quickly spread to other Chinese regions and neighboring countries, while Wuhan, which was identified as the epicenter of the outbreak, was cut off by the authorities on January 23rd, 2020 [6]. On January 30th, WHO declared the epidemic as a public health emergency [1], and the disease caused by the virus, received its official naming, COVID-19, on February 11th [7].

The first serious COVID-19 outbreak in Europe was identified in northern Italy in February, with the country having its first death on the 21st [8]. The novel coronavirus was transmitted to all parts of Europe within the next few weeks, resulting in WHO declaring COVID-19 a pandemic on March 11th, 2020.

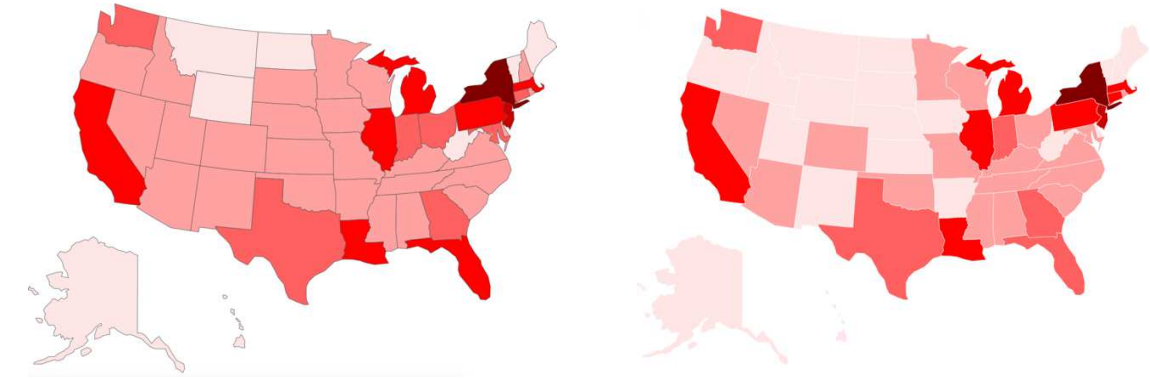
As of April 18th, 2020, 16:48 GMT [9], there have been 2,287,369 confirmed cases worldwide, with 157,468 confirmed deaths, and 585,838 recovered. The most affected countries with more than 100K cases (in absolute numbers, not divided by population) are: USA with 715,105 confirmed cases and 37,889 deaths; Spain with 191,726 confirmed cases and 20,043 deaths; Italy with 175,925 confirmed cases and 23,227 deaths; France with 147,969 confirmed cases and 18681 deaths; Germany with 142,614 confirmed cases and 4405 deaths; and the UK with 114,217 confirmed cases and 15,464 deaths, as depicted in Figure 1 that consists of the heat maps for the worldwide cases and deaths by country.



Cases **Deaths**

Figure 1. Geographical distribution of Worldwide COVID-19 cases and deaths as of April 18th.

As evident, Europe is severely hit by COVID-19; however, the spread of the disease now indicates that the center of the epidemic has moved to the US, which is the most affected country in terms of cases and deaths, with the state of New York counting more than 240K cases and 17K casualties. Figure 2 shows the distribution of the COVID-19 cases and deaths in the US by state, as of April 18th, 2020.



Cases **Deaths**

Figure 2. Geographical distribution of US COVID-19 cases and deaths as of April 18th.

Towards the direction of finding new methods and approaches for disease surveillance, it is crucial to make use of real time internet data. Infodemiology, i.e., information epidemiology, is a concept introduced by Gunther Eysenbach [10-11]. In the field of infodemiology, internet sources and data are employed in order to inform public health and policy [12-13], and are valuable for the monitoring and forecasting of outbreaks and epidemics [14], as for example Ebola [15], Zika [16], MERS [17], influenza [18], and measles [19-20].

During this pandemic, several approaches in using Web based data have been already published in this line of research. Google Trends, the most popular infodemiology source along with Twitter, has been widely used in health and medicine for the analysis and forecasting of diseases and epidemics [21]. As of April 20, 2020, already seven (7) papers on the topic of tracking and forecasting COVID-19 using Google Trends data have been published, according to PubMed (advanced search: covid AND google trends) [22], monitoring, analyzing, or forecasting COVID-19 in several regions like Taiwan [23], China [24-25], Europe [26-27], USA [27-28], Iran [27, 29]. Note that for Twitter publications related to the COVID-19 pandemic, eight papers (8) are online up to this point (PubMed advanced search: covid AND twitter [22]), published from March 13 to April 20, 2020 [30-37]. Table 1 consists of the systematic reporting of COVID-19 Google Trends studies, in the order of the reported publication date.

Table 1. Systematic reporting of publications in COVID-19 using Google Trends as of April 20th, 2020.

Authors	Date	Region	Objective	Publisher	Journal
Husnayain et al. [23]	March 12	Taiwan	Analyzing COVID-19 related searches	Elsevier	International Journal of Infectious Diseases
Li et al. [24]	March 25	China	Correlating Internet searches with COVID-19 cases	Eurosurveillance	Eurosurveillance
Mavragani [26]	April 2	Europe	Correlating Google Trends data with COVID-19 cases & deaths	JMIR	JMIR Public Health and Surveillance
Hong et al. [28]	April 7	USA	Relationship between telehealth searches and COVID-19	JMIR	JMIR Public Health and Surveillance
Walker et al. [27]	April 11	USA, Iran Europe	Exploring of the online activity related to loss of smell	Wiley	International Forum of Allergy & Rhinology
Ayyoubzadeh et al. [29]	April 14	Iran	Prediction of COVID-19 cases	JMIR	JMIR Public Health and Surveillance
Effenberger et al. [25]	April 16	China	Correlation between Google Trends data and COVID-19 cases	Elsevier	International Journal of Infectious Diseases

In this paper, USA Google Trends data on the topic of "Coronavirus (Virus)" are employed at both national and state level, in order to explore the relationship between COVID-19 data and the online interest on the virus. At first, the correlations between Google Trends and COVID-19 data are calculated, followed by exploring the role of Google Trends data in the predictability of COVID-19. To the best of our knowledge, this is the first attempt of this kind in the US.

The rest of the paper is structured as follows: the Methods section details the procedure of the data collection and the statistical analysis tools and methods, the Results section includes the nowcasting models at both national and state level, and the Discussion section consists of the main findings of this work, along with the limitations and future research suggestions.

Methods

Data from the Google Trends platform are retrieved in .csv [38]. Data are normalized over the selected period and Google Trends reports the adjustment procedure as follows: "*Search results are normalized to the time and location of a query by the following process: Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same search interest for a term don't always have the same total search volumes*" [39]. The methodology for the data collection is designed based on the Google Trends Methodology Framework in Infodemiology and Infoveillance [40]. Note that data may slightly vary based on the time of retrieval.

For the keyword selection, the online interest in all commonly used variations of referring to the virus are examined and compared, i.e., "Coronavirus (Virus)"; "COVID-19 (Search term)"; "SARS-COV-2 (Search term)"; "2019-nCoV (Search term)"; "Coronavirus (Search term)". Only "Coronavirus (Virus)" and "Coronavirus (Search term)" yield significantly high online interest, which is also quite expected. Between the two, i.e., the Topic (Virus) and the Search term, the "Coronavirus (Virus)" is selected for further analysis.

Data for the worldwide distribution of the COVID-19 cases and deaths are retrieved from Worldometer [9], and maps on COVID-19 cases and deaths are recreated by the authors using the free online tools Pixelmap [41] and Chartsbin [42]. Data for the US analysis on COVID-19 are retrieved by "The COVID Tracking Project", providing detailed structured data on COVID-19 cases and deaths nationally and at state level [43].

As Google Trends data are normalized the time frame for which search traffic data are retrieved should exactly match the period for which COVID-19 data are available. Therefore, the timeframes for which analysis is performed is different for the states, starting either on March 4th or on the date for which the first confirmed case is identified in each state, as shown in Table 2:

Table 2. Timeframes for which Google Trends data are retrieved, by state.

March 4th - April 15th	USA; Arizona; California; Florida; Georgia; Illinois; Massachusetts; New Hampshire; New York; North Carolina; Oregon; Texas; Washington; Wisconsin
March 5th - April 15th	Nevada; New Jersey; Tennessee
March 6th - April 15th	Colorado; Indiana; Maryland; Pennsylvania
March 7th - April 15th	Hawaii; Kentucky; Minnesota; Nebraska; Oklahoma; Rhode Island; South Carolina; Utah
March 8th - April 15th	Connecticut; District of Columbia; Kansas; Missouri; Vermont; Virginia
March 9th – April 15th	Iowa; Louisiana; Ohio
March 11th – April 15th	Delaware; Michigan; New Mexico; South Dakota
March 12th – April 15th	Arkansas; Maine; Mississippi; Montana; North Dakota; Wyoming
March 13th – April 15th	Alabama; Alaska
March 14th - April 15th	Idaho
March 18th – April 15th	West Virginia

Each variable used in this study is divided by its full-sample standard deviation, estimated or calculated based on the basic formula of standard deviation of a variable. By doing this, the inherent variability of each variable was moved, and thus all of them have a standard deviation equal to 1. This allows us to compare the strength of the impact of explanatory variables used on the dependent variable. The non-parametric [44] unit root test is also applied, in order to reveal whether or not both variables are stationary. The results suggest that both variables can be used directly without further transformation in the present analysis.

The first step towards exploring the role of Google Trends in the predictability of COVID, is to examine the relationship between Google Trends and COVID-19 incidence. To this direction, the Pearson correlation coefficients (*r*) between the ratio (COVID-19 Deaths)/(COVID-19 Cases) and Google Trends data are constructed. In particular, a minimum variance bias-corrected Pearson correlation coefficient [45-46] via a bootstrap simulation is applied, in order to deal with the limited number of observations, and thus, with the small sample estimation bias (also see [46]). The bias-corrected bootstrap coefficient $\tilde{\rho}^b$ for the Pearson correlation is given by:

$$\tilde{\rho}^b = B^{-1} \sum_{b=1}^{B} \tilde{\rho}_j^b(\rho)$$

where B corresponds to the length of the bootstrap samples; in this case set equal to 999.

Next, predictive analysis for USA and all US states (plus DC) is performed. The predictive model is a quantile regression, considered to be a robust regression analysis against the presence of outliers in the sample; introduced by Koenker and Bassett [47]. Building on the study implemented by Karlsson [46], a bias corrected via balanced bootstrapping quantile regression is employed. Such a model is the appropriate statistical approach to mitigate the small sample estimation bias and the present of outliers in the dataset, as it combines the advantages of bootstrap standard errors and the merits of quantile regression.

More specifically, let Y_t, where $t \in T$, be a time series representing the dependent variable, supposing a bivariate specification. A quantile regression estimates the impact of explanatory variable X_t, where $t \in T$, on the variable Y_t at different points of conditional q-quantile, where $q \in (0,1)$, of the conditional distribution. A value of q-quantile close to zero and a value of q-quantile close to one, represent the left (lower) and the right (upper) tail of the conditional distribution, respectively. The conditional quantile function is defined by:

$$Q_{Y|X}(q) = X'\beta_q$$

given the distribution of Y_t, the estimation of the conditional quantile functions β_q can be obtained by solving the following minimization problem:

$$\beta_q = \arg\min_{\beta \in \mathbb{R}^k} E\left(\rho_q(Y - X\beta)\right)$$

where $\rho_q(y) = y\left(q - 1_{\{y<0\}}\right)$ represents the loss function. By minimizing the sample analog $\{y_1, \ldots, y_n\}$ that corresponds to a q^{th} quantile sample, the estimator β_q takes the form:

$$\beta_q = \arg\min_{\beta \in \mathbb{R}^k} \sum_{t=1}^{n} \rho_q(Y_t - X_t'\beta) = \arg\min_{\beta \in \mathbb{R}^k} \left[q \sum_{Y_t \geq \beta X_t} |Y_t - \beta X_t| + (1-q) \sum_{Y_t < \beta X_t} |Y_t - \beta X_t| \right]$$

where βX_t is an approximation to the conditional q-quantile of the variable Y_t.

In our analysis, Y_t stands for the ratio (COVID-19 Deaths)/(COVID-19 Cases), X_{t-1} is the respective Google Trends value in lag order, and $t = 1, \ldots, T$, with T being the respective number of observations. A linear trend is also used.

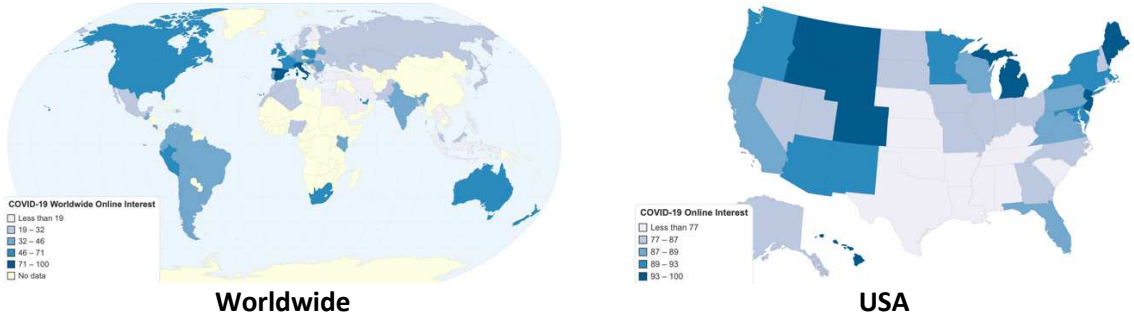
Finally, the bias corrected parameter estimate is estimated as:

$$\tilde{\beta}_i^b(q) = \hat{\beta}_i(q) - \widehat{bias}\left(\hat{\beta}_i(q)\right)$$

where the $\widehat{bias}\left(\hat{\beta}_i(q)\right)$ is given by $B^{-1}\sum_{b=1}^{B}\hat{\beta}_i^*(q) - \hat{\beta}_i(q)$ and $q \in (0,1)$ stands for the quantile considered; in this case set equal to 0.5 (median). A median regression is considered as more robust to outliers than, for example, least squares regression, and it also avoids assumptions about the error parametric distribution (see [48]).

Results

In Figure 3, the worldwide and US online interest in terms of Google queries in the "Coronavirus (Virus)" Topic from January 22nd to April 15th, 2020, is depicted, showing that said topic is very popular, and especially in Europe and in North America, where, in the US, the interest is significantly high -i.e. above 70- for all US states.



Worldwide **USA**

Figure 3. Heat maps of the worldwide and US online interest in "Coronavirus (Virus)".

Following, the correlations between Google Trends and COVID19 data are calculated. Table 3 consists of the Pearson correlation analysis, while Figure 4 depicts the Pearson correlations heat map in the US. As evident, statistically significant correlations are observed in USA and in the states of Alabama, Arkansas, California, Colorado, DC, Florida, Georgia, Illinois, Kentucky, Massachusetts, Minnesota, Nebraska, Nevada, New Hampshire, New York, North Carolina, Oregon, Pennsylvania, South Dakota, Tennessee, Vermont, Virginia, Washington, Wisconsin, and Wyoming.

Table 3. Pearson correlations by state.

State	Pearson Correlation	Standard Error	Wald Test (r=0)	p-value	State	Pearson Correlation	Standard Error	Wald Test (r=0)	p-value
USA	-0.7054***	(0.0536)	[13.1672]	<.0001	**Missouri**	-0.2627	(0.1608)	[1.6333]	0.1024
Alabama	-0.6896***	(0.0748)	[9.2185]	<.0001	**Montana**	-0.063	(0.1727)	[0.3651]	0.7151
Alaska	-0.1162	(0.1276)	[0.9107]	0.3625	**Nebraska**	-0.2763*	(0.1503)	[1.8381]	0.0661
Arizona	-0.313*	(0.1292)	[2.4225]	0.0154	**Nevada**	-0.3452**	(0.1519)	[2.273]	0.0230
Arkansas	0.4282***	(0.1105)	[3.8742]	0.0001	**New Hampshire**	-0.406***	(0.1432)	[2.8349]	0.0046
California	-0.4123***	(0.1300)	[3.1711]	0.0015	**New Jersey**	-0.065	(0.2013)	[0.3227]	0.7469
Colorado	0.435**	(0.1761)	[2.4694]	0.0135	**New Mexico**	-0.1474	(0.1367)	[1.0783]	0.2809
Connecticut	-0.1266	(0.1895)	[0.668]	0.5041	**New York**	-0.5925***	(0.0790)	[7.5016]	<.0001
Delaware	0.182	(0.2004)	[0.908]	0.3639	**North Carolina**	-0.3172**	(0.1561)	[2.032]	0.0421
DC	-0.3464**	(0.1632)	[2.1219]	0.0338	**North Dakota**	0.2567	(0.1705)	[1.5056]	0.1322
Florida	-0.3171**	(0.1559)	[2.034]	0.0420	**Ohio**	-0.1645	(0.1979)	[0.8311]	0.4059
Georgia	-0.3467**	(0.1462)	[2.3708]	0.0178	**Oklahoma**	-0.1703	(0.1713)	[0.9944]	0.3200
Hawaii	-0.1591	(0.1692)	[0.9405]	0.3470	**Oregon**	0.4605***	(0.1432)	[3.2154]	0.0013
Idaho	0.0614	(0.1436)	[0.4276]	0.6689	**Pennsylvania**	-0.3645**	(0.1446)	[2.5218]	0.0117
Illinois	0.2501*	(0.1512)	[1.6541]	0.0981	**Rhode Island**	-0.0366	(0.1805)	[0.2031]	0.8391
Indiana	0.0162	(0.1884)	[0.086]	0.9314	**South Carolina**	-0.2094	(0.1400)	[1.4958]	0.1347
Iowa	-0.2172	(0.1539)	[1.4112]	0.1582	**South Dakota**	0.3518*	(0.1920)	[1.8323]	0.0669
Kansas	0.1141	(0.1748)	[0.6531]	0.5137	**Tennessee**	-0.3878***	(0.1495)	[2.5937]	0.0095
Kentucky	-0.2789*	(0.1663)	[1.677]	0.0935	**Texas**	0.0223	(0.1931)	[0.1157]	0.9079
Louisiana	-0.2422	(0.1713)	[1.4141]	0.1573	**Utah**	-0.2135	(0.1448)	[1.4749]	0.1402
Maine	-0.1811	(0.1387)	[1.3062]	0.1915	**Vermont**	-0.3255**	(0.1549)	[2.1007]	0.0357
Maryland	-0.0385	(0.2045)	[0.1884]	0.8505	**Virginia**	-0.286**	(0.1414)	[2.0228]	0.0431
Massachusetts	-0.4285***	(0.1421)	[3.0152]	0.0026	**Washington**	-0.5805***	(0.0835)	[6.9492]	<.0001
Michigan	-0.1045	(0.1757)	[0.5949]	0.5519	**West Virginia**	0.0033	(0.0426)	[0.0781]	0.9378
Minnesota	-0.3513**	(0.1550)	[2.2657]	0.0235	**Wisconsin**	-0.3972***	(0.1285)	[3.09]	0.002
Mississippi	0.308	(0.1975)	[1.5599]	0.1188	**Wyoming**	0.396**	(0.1840)	[2.1524]	0.0314

*p<0.1; **p<0.05; ***p<0.01

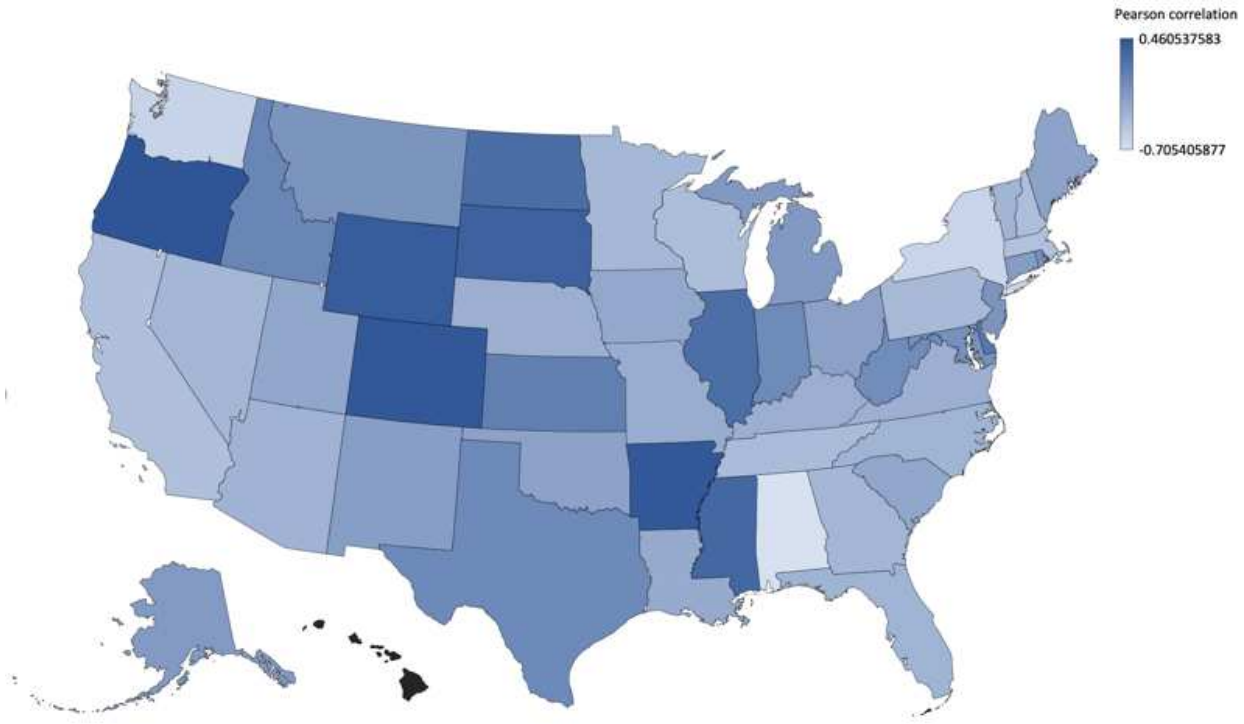


Figure 4. Heat map of the Pearson correlations by state.

Proceeding with the results of the predictive analysis, Table 4 consists of the estimated models for the US and for each US state (plus DC), and Figure 5 depicts the heat map for $\boldsymbol{\beta_1}$ by state. Due to low number of observations, the states of Maine, Montana, North Dakota, West Virginia, and Wyoming were not included in the predictive analysis results, however included in the heat map for uniformity reasons. As is evident, the estimated Google Trends models exhibit strong COVID-19 predictability.

Table 4. Predictive analysis by state.

	$\boldsymbol{\beta_0}$			$\boldsymbol{\beta_1}$			$\boldsymbol{\beta_2}$		
USA	-0.0509	-(0.4339)	-[0.1172]	-0.7506	-(0.2197)	-[3.4173]	-0.0014	-(0.0169)	-[0.0831]
AL	0.8944	(0.2176)	[4.1099]	-0.5961	(0.1160)	-[5.1383]	-0.0413	(0.0070)	-[5.8850]
AK	-1.4528	(0.2003)	-[7.2539]	-0.2449	(0.1006)	-[2.4341]	0.0663	(0.0087)	[7.6030]
AZ	-1.4183	(0.1309)	-[10.8362]	-0.2429	(0.0817)	-[2.9745]	0.0637	(0.0049)	[12.8777]
AR	-0.2565	(0.4658)	-[0.5507]	0.2785	(0.2531)	[1.1004]	0.0023	(0.0124)	[0.1825]
CA	-1.4274	(0.0936)	-[15.2521]	-0.1634	(0.0539)	-[3.0325]	0.0642	(0.0046)	[13.8481]
CO	-0.9688	(0.1916)	-[5.0561]	0.3007	(0.2587)	[1.1623]	0.0290	(0.0074)	[3.9132]
CT	-1.7866	(0.0654)	-[27.3353]	-0.1645	(0.0470)	-[3.4989]	0.0782	(0.0026)	[30.6221]
DE	-2.0415	(0.4639)	-[4.4003]	-0.2687	(0.2446)	-[1.0987]	0.0715	(0.0110)	[6.4873]
DC	-1.3077	(0.1980)	-[6.6064]	-0.1548	(0.0849)	-[1.8228]	0.0578	(0.0094)	[6.1513]
FL	-1.5483	(0.0766)	-[20.2209]	-0.2128	(0.0431)	-[4.9412]	0.0715	(0.0024)	[29.3170]
GA	-1.5727	(0.0808)	-[19.4690]	-0.2047	(0.0570)	-[3.5898]	0.0721	(0.0042)	[17.2658]
HI	-1.6732	(0.0873)	-[19.1647]	-0.2083	(0.0470)	-[4.4343]	0.0758	(0.0041)	[18.3027]
ID	-1.8929	(0.1465)	-[12.9167]	-0.2686	(0.0663)	-[4.0507]	0.0866	(0.0067)	[12.8631]
IL	-1.4466	(0.1404)	-[10.3063]	0.3943	(0.0707)	[5.5764]	0.0680	(0.0056)	[12.2022]
IN	-1.4674	(0.2157)	-[6.8020]	0.0977	(0.1624)	[0.6018]	0.0693	(0.0065)	[10.7392]
IA	-1.5912	(0.1402)	-[11.3507]	-0.2957	(0.0733)	-[4.0346]	0.0732	(0.0042)	[17.3342]
KS	-1.5579	(0.2298)	-[6.7799]	0.0463	(0.1101)	[0.4204]	0.0635	(0.0106)	[5.9774]
KY	-1.5530	(0.1396)	-[11.1222]	-0.2415	(0.0599)	-[4.0291]	0.0719	(0.0062)	[11.5292]
LA	-1.6432	(0.0602)	-[27.2763]	-0.2050	(0.0357)	-[5.7381]	0.0751	(0.0026)	[28.6534]
MD	-1.1066	(0.2339)	-[4.7306]	0.1135	(0.1008)	[1.1255]	0.0550	(0.0088)	[6.2834]
MA	-1.6424	(0.0771)	-[21.3061]	-0.1757	(0.0538)	-[3.2668]	0.0742	(0.0034)	[21.8651]
MI	-1.7657	(0.0813)	-[21.7133]	-0.1884	(0.0406)	-[4.6375]	0.0800	(0.0032)	[25.2349]
MN	-1.6085	(0.0773)	-[20.7963]	-0.2344	(0.0521)	-[4.4970]	0.0728	(0.0027)	[26.9966]
MS	-1.3047	(0.2959)	-[4.4088]	0.1773	(0.1600)	[1.1086]	0.0570	(0.0082)	[6.9200]
MO	-1.5382	(0.0883)	-[17.4271]	-0.2326	(0.0478)	-[4.8610]	0.0718	(0.0051)	[14.0987]
NE	-1.4875	(0.1909)	-[7.7908]	-0.2192	(0.0746)	-[2.9375]	0.0717	(0.0063)	[11.3935]
NV	-1.6778	(0.0862)	-[19.4683]	-0.1872	(0.0348)	-[5.3846]	0.0763	(0.0037)	[20.4946]
NH	-1.6586	(0.0723)	-[22.9526]	-0.1515	(0.0365)	-[4.1562]	0.0741	(0.0025)	[30.0037]
NJ	-1.8518	(0.2428)	-[7.6277]	-0.2395	(0.2427)	-[0.9867]	0.0688	(0.0060)	[11.3949]
NM	-1.2414	(0.1640)	-[7.5679]	-0.1188	(0.0803)	-[1.4805]	0.0593	(0.0066)	[8.9371]
NY	-1.2201	(0.0468)	-[26.0596]	-0.1482	(0.0562)	-[2.6358]	0.0482	(0.0043)	[11.2916]
NC	-1.6575	(0.0953)	-[17.3914]	-0.1613	(0.0476)	-[3.3848]	0.0722	(0.0038)	[18.8471]
OH	-1.8408	(0.1464)	-[12.5751]	-0.1758	(0.0750)	-[2.3436]	0.0790	(0.0048)	[16.3817]
OK	-1.7038	(0.0544)	-[31.2986]	-0.2463	(0.0318)	-[7.7497]	0.0767	(0.0026)	[29.5090]
OR	-0.7953	(0.2019)	-[3.9392]	0.4395	(0.1362)	[3.2257]	0.0293	(0.0069)	[4.2697]
PA	-1.3917	(0.1279)	-[10.8769]	-0.1845	(0.0758)	-[2.4348]	0.0716	(0.0041)	[17.5561]
RI	-1.4924	(0.0752)	-[19.8418]	-0.1461	(0.0408)	-[3.5844]	0.0588	(0.0049)	[12.1036]
SC	-1.2889	(0.0941)	-[13.7030]	-0.1816	(0.0513)	-[3.5395]	0.0520	(0.0069)	[7.5216]
SD	-1.1230	(0.2939)	-[3.8212]	0.2815	(0.1388)	[2.0277]	0.0537	(0.0084)	[6.4280]
TN	-1.5098	(0.0658)	-[22.9294]	-0.2157	(0.0524)	-[4.1179]	0.0676	(0.0020)	[33.1730]
TX	-1.4766	(0.3041)	-[4.8557]	0.2749	(0.1903)	[1.4442]	0.0660	(0.0077)	[8.5342]
UT	-1.4381	(0.1399)	-[10.2768]	-0.1586	(0.0723)	-[2.1944]	0.0720	(0.0069)	[10.3640]
VT	-1.5359	(0.1854)	-[8.2848]	-0.2499	(0.0848)	-[2.9476]	0.0770	(0.0081)	[9.5352]
VA	-1.5878	(0.2504)	-[6.3400]	-0.3147	(0.1021)	-[3.0837]	0.0767	(0.0106)	[7.2484]
WA	-1.3476	(0.1540)	-[8.7488]	-0.2236	(0.1007)	-[2.2212]	0.0660	(0.0101)	[6.5118]
WI	-1.3407	(0.0992)	-[13.5142]	-0.2143	(0.0698)	-[3.0711]	0.0618	(0.0053)	[11.6287]

Parenthesis reports the standard errors; t-statistics are given in brackets.

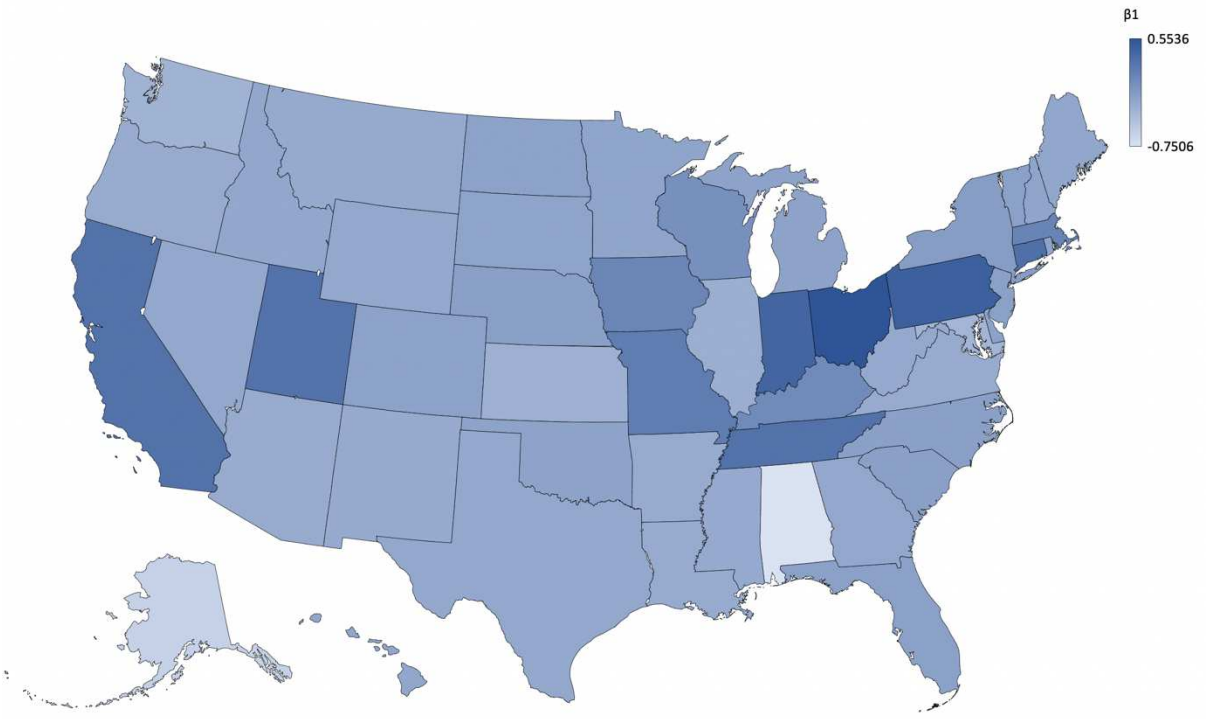


Figure 5. Heat map of the predictive analysis models' statistical significance.

Discussion

In light of the COVID-19 pandemic and towards finding new ways of forecasting the disease spreading, in this study Google Trends data on the "Coronavirus (Virus)" Topic were used in order to explore the predictability of COVID-19 in the US. At first, statistically significant correlations were observed for the US and several US states -as shown in a more elaborate depiction of a spider web chart of said correlations in Figure 6-, which is in line with previous studies that have suggested that correlations are observed between Google Trends and COVID-19 data.

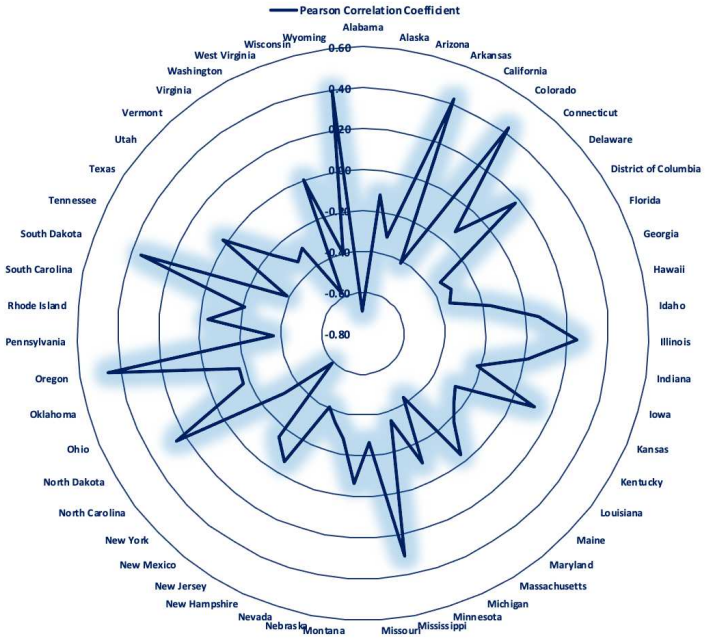


Figure 6. Radar chart of the Pearson correlations coefficients by state.

Figure 7 consists of the graph of the COVID-19 Deaths/Cases ratio and the respective Google Trends normalized data in the US from March 4th to April 15th, 2020. For graph consistency purposes, the COVID-19 Deaths/Cases ratio is normalized on a 0-100 scale. As depicted in the graph and also confirmed by the predictive analysis, it is evident that the two variables are not linearly dependent, rather than have an inversely proportional relationship, meaning that as COVID-19 progresses, the online interest decreases.

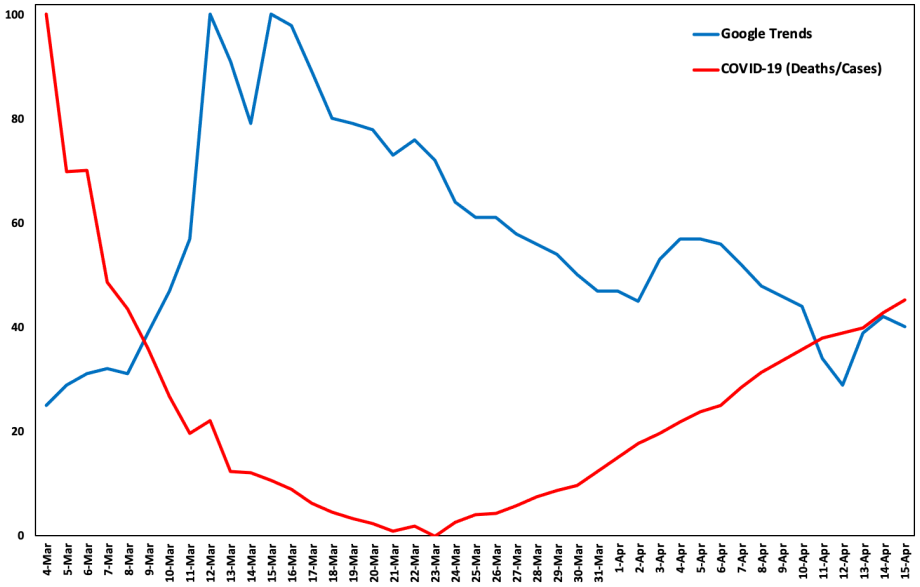


Figure 7. COVID-19 and Google Trends data from March 4th to April 15th in the US.

In sense and from a behavioral point of view, this can be explained as follows: The interest started increasing at first and reached a peak as the confirmed cases reached a high number and as deaths rates started exhibiting the real threat of this pandemic, while after a while the interest has an inverse course, which could also be indicating that the public can be overwhelmed by all this information overload and turns to decreased information intake. The spike in Google queries and the decline in the ratio of COVID-19 Deaths/Cases, could be due to the spreading of the virus over these days and the "delay" in deaths, i.e., cases increasing while total number of deaths has not started significantly increasing yet.

The latter is in line with the recent publication of Mavragani [26], that suggested that, though significant correlations between COVID-19 and Google data are observed, they tend to decrease both in strength and significance as time moves forward in regions that have been affected by COVID-19, because the interest decreases. This counter-intuitively happens before the cases' and deaths' curves start exhibiting a downward trend, i.e. when a region is being heavily affected, independently of having or not reached its peak yet. However, it would be interesting to explore the relationship from this point onwards, since, as shown in the graph, the lines meet, which could indicate a future change in the relationship dynamics when deaths peak at a later point and also when they start their downward course.

This study has limitations. At first, only data from Google Trends were considered. Though this is the most popular search engine, some data on the topic of Coronavirus from other search engines were not included in this analysis. Second, data at this point are very limited, thus the results are based on fewer observations. Third, the 51 states exhibit diversity in terms of confirmed cases and deaths, thus any conclusions drawn from this analysis refer to each case individually. Despite the known limitations of online search traffic data though, using Infodemiology metrics in informing public health and policy in general and for the monitoring of outbreaks and epidemics in specific, has received wide attention recently, with several successful attempts of forecasting disease spreading.

Towards exploring the dynamic of finding determinants of COVID-19, the predictive analysis in this study gives insight on how online search traffic data can play a significant part in forming public health policies, especially in times of epidemics and outbreaks, when real-time data are essential. With the COVID-19 pandemic, the world is in uncharted territory, in scientific, financial, and social terms. This calls for immediate action and open research and data, and the term "multidisciplinary" has never before been more important. To this direction, the role of big data in providing "*opportunities for performing modeling studies of viral activity and for guiding individual country healthcare policymakers to enhance preparation for the outbreak*" has been acknowledged [49], and current research on the subject should focus on both exploring the role of more infodemiology variables as well as combine infodemiology with traditional sources, in order to explore the full potential of what online, real-time data have to offer to disease surveillance.

References

1. World Health Organization. WHO Timeline – COVID-19. (https://www.who.int/news-room/detail/08-04-2020-who-timeline---covid-19; April 19 2020).
2. World Health Organization. Twitter account. (https://twitter.com/WHO/status/1213523866703814656?s=20; April 21 2020.
3. World Health Organization. Pneumonia of unknown cause. (https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/; April 21 2020).
4. Business Insider. A comprehensive timeline of the new coronavirus pandemic, from China's first COVID-19 case to the present. (https://www.businessinsider.com/coronavirus-pandemic-timeline-history-major-events-2020-3; April 21 2020).
5. World Health Organization. Twitter account. (https://twitter.com/who/status/1217043229427761152?lang=en; April 21 2020).
6. New York Times. Wuhan, Center of Coronavirus Outbreak, Is Being Cut Off by Chinese Authorities. URL: https://www.nytimes.com/2020/01/22/world/asia/china-coronavirus-travel.html Accessed April 21 2020.
7. BBC News. Coronavirus disease named Covid-19. (https://www.bbc.com/news/world-asia-china-51466362; April 21 2020).
8. Wolrdometer. COVID coronavirus Outbreak. Italy (https://www.worldometers.info/coronavirus/country/italy/; April 19 2020).
9. Wolrdometer. COVID coronavirus Outbreak. (https://www.worldometers.info/coronavirus/; April 19 2020)
10. 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. *J Med Internet Res.* **11**(1), e11 (2009).
11. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *Am J Prev Med*. **40**(5 Suppl 2), S154-8 (2011).
12. Mavragani A. Infodemiology and Infoveillance: A Scoping Review. *J Med Internet Res* (2020) In press.
13. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation. J Med Internet Res. **15**(7), e147 (2013).
14. Eysenbach G. SARS and Population Health Technology. *J Med Internet Res*. **5**(2), e14 (2003).
15. van Lent LG, Sungur H, Kunneman FA, van de Velde B, Das E. Too Far to Care? Measuring Public Attention and Fear for Ebola Using Twitter. *J Med Internet Res*. **19**(6), e193 (2017).

16. Farhadloo M, Winneg K, Chan MS, Hall JK, Albarracin D. Associations of Topics of Discussion on Twitter With Survey Measures of Attitudes, Knowledge, and Behaviors Related to Zika: Probabilistic Study in the United States. *JMIR Public Health Surveill*. **4**(1), e16 (2018).
17. Poletto C, Boëlle P, Colizza V. Risk of MERS importation and onward transmission: a systematic review and analysis of cases reported to WHO. *BMC Infect Dis*. **16**(1), 448 (2016).
18. Samaras L, García-Barriocanal E, Sicilia MA. comparing Social media and Google to detect and predict severe epidemics. *Sci Rep.* **10**, 4747 (2020).
19. Mavragani A, Ochoa G. The Internet and the Anti-Vaccine Movement: Tracking the 2017 EU Measles Outbreak. *Big Data Cog Comp*. **2**(1), 1 (2018).
20. Du J, Tang L, Xiang Y, Zhi D, Xu J, Song HY, Tao C. Public Perception Analysis of Tweets During the 2015 Measles Outbreak: Comparative Study Using Convolutional Neural Network Models. *J Med Internet Res*. **20**(7), e236 (2018).
21. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review. *J Med Internet Res.* **20**(11):e270 (2018).
22. PubMed search: Google Trends & COVID. (https://www.ncbi.nlm.nih.gov/pubmed/ April 20, 2020).
23. Husnayain A, Fuad A, Su EC. Applications of google search trends for risk communication in infectious disease management: A case study of COVID-19 outbreak in Taiwan. *Int J Infect Dis*. pii: S1201-9712(20)30140-5 (2020).
24. Li C, Chen LJ, Chen X, Zhang M, Pang CP, Chen H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Euro Surveill*. **25**(10) (2020).
25. Effenberger M, Kronbichler A, Shin JI, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with Internet Search Volumes: A Google Trends(TM) Analysis. Int J Infect Dis. 2020 Apr 16. pii: S1201-9712(20)30249-6.
26. Mavragani A. Tracking COVID-19 in Europe: Infodemiology Approach. *JMIR Public Health Surveill*. **6**(2), e18941 (2020).
27. Walker A, Hopkins C, Surda P. The use of google trends to investigate the loss of smell related searches during COVID-19 outbreak. *Int Forum Allergy Rhinol*. In press (2020).
28. Hong YR, Lawrence J, Williams D Jr, Mainous Iii A. Population-Level Interest and Telehealth Capacity of US Hospitals in Response to COVID-19: Cross-Sectional Analysis of Google Search and National Hospital Survey Data. *JMIR Public Health Surveill*. **6**(2), e18961 (2020).
29. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR Public Health Surveill*. **6**(2):e18828 (2020).
30. Rufai SR, Bunce C. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *J Public Health (Oxf)*. pii: fdaa049 (2020).
31. Kouzy R, Abi Jaoude J, Kraitem A, El Alam MB, Karam B, Adib E, Zarka J, Traboulsi C, Akl EW, Baddour K. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*. **12**(3), e7255 (2020).
32. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: A surveillance study. *J Med Internet Res*. (2020)
33. Dost B, Koksal E, Terzi Ö, Bilgin S, Ustun YB, Arslan HN. Attitudes of Anesthesiology Specialists and Residents Toward Patients Infected with the Novel Coronavirus (COVID-19): A National Survey Study. *Surg Infect (Larchmt)*. (2020).
34. Simcock R, Thomas TV, Estes C, Filippi AR, Katz MA, Pereira IJ, Saeed H. COVID-19: Global radiation oncology's targeted response for pandemic preparedness. *Clin Transl Radiat Oncol.* **22**, 55-68 (2020).
35. Kim B. Effects of Social Grooming on Incivility in COVID-19. *Cyberpsychol Behav Soc Netw*. (2020).

36. Rosenberg H, Syed S, Rezaie S. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *CJEM*. **6**, 1-4 (2020).
37. Chan AKM, Nickson CP, Rudolph JW, Lee A, Joynt GM. Social media for rapid knowledge dissemination: early experience from the COVID-19 pandemic. *Anaesthesia*. (2020)
38. Google Trends Explore. (https://trends.google.com/trends/explore; April 18, 2020).
39. Google Trends Help. (https://support.google.com/trends/answer/4365533?hl=en; April 18, 2020).
40. Mavragani A, Ochoa G. Google Trends in Infodemiology and Infoveillance: Methodology Framework. JMIR Public Health Surveill 2019;5(2):e13439. PMID: 31144671
41. PixelMap. (https://pixelmap.amcharts.com; April 20, 2020).
42. ChartsBin. (http://chartsbin.com; April 20, 2020).
43. The COVID Tracking Project. (https://covidtracking.com; April 15, 2020).
44. Phillips PCB, Perron P. Testing for a unit root in time series regression. *Biometrica*. **75** (2), 335–346 (1988).
45. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*. **1**(1), 54–75 (1986).
46. Karlsson A. Bootstrap methods for bias correction and confidence interval estimation for nonlinear quantile regression of longitudinal data. *J Stat Comput Sim*. **79**, 10, 1205-1218 (2009).
47. Koenker R, Bassett G. Regression quantiles. *Econometrica*. **46**(1), 33–50 (1978).
48. Chernozhukov V, Hansen C, Jansson M. Finite sample inference for quantile regression models. *J Econom*. **152**, 93–103 (2009).
49. Shu Wei Ting D, Carin L, Dzau V, Wong TY. Digital Technology and COVID-19. Nat Med. **26**, 459-461 (2020).

Author Contributions

A.M. conceived the idea; A.M. and K.G. designed the methodology; A.M. performed the data collection and analysis; K.G. performed the statistical analysis; A.M. and K.G. interpreted the data; A.M. wrote the paper. Both authors reviewed the manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability

All data used in the analysis are open and publicly available in the cited sources.

Funding Disclosure

No external funding was received for the implementation of this study.

Figures

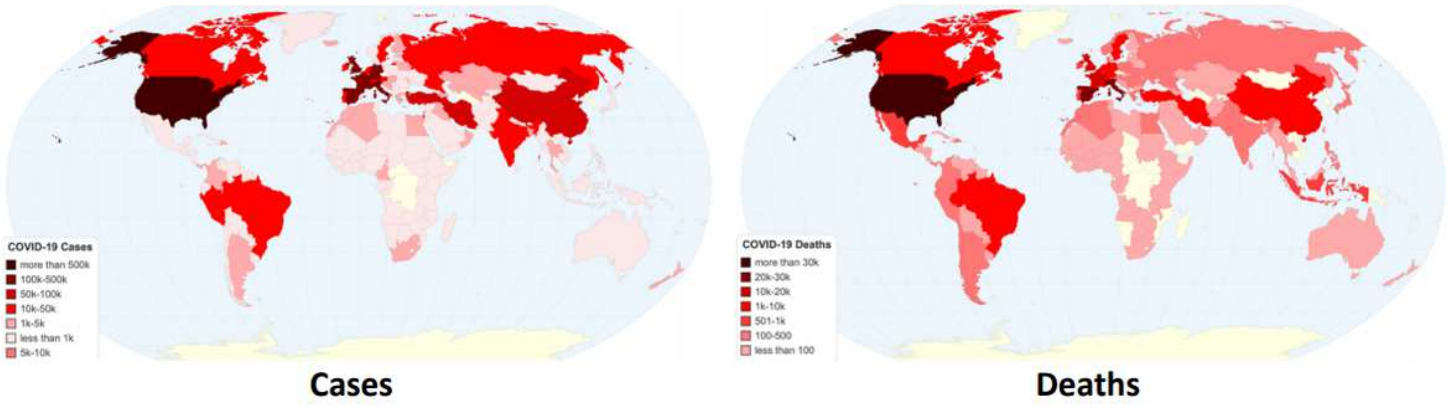


Figure 1

Geographical distribution of Worldwide COVID-19 cases and deaths as of April 18th. 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.

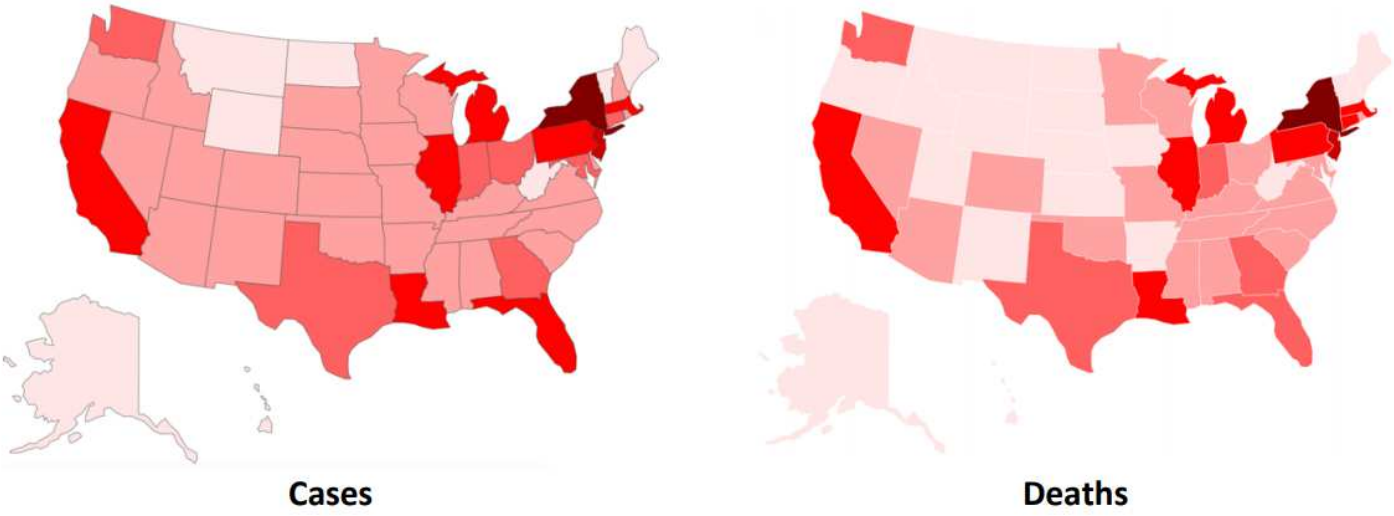


Figure 2

Geographical distribution of US COVID-19 cases and deaths as of April 18th

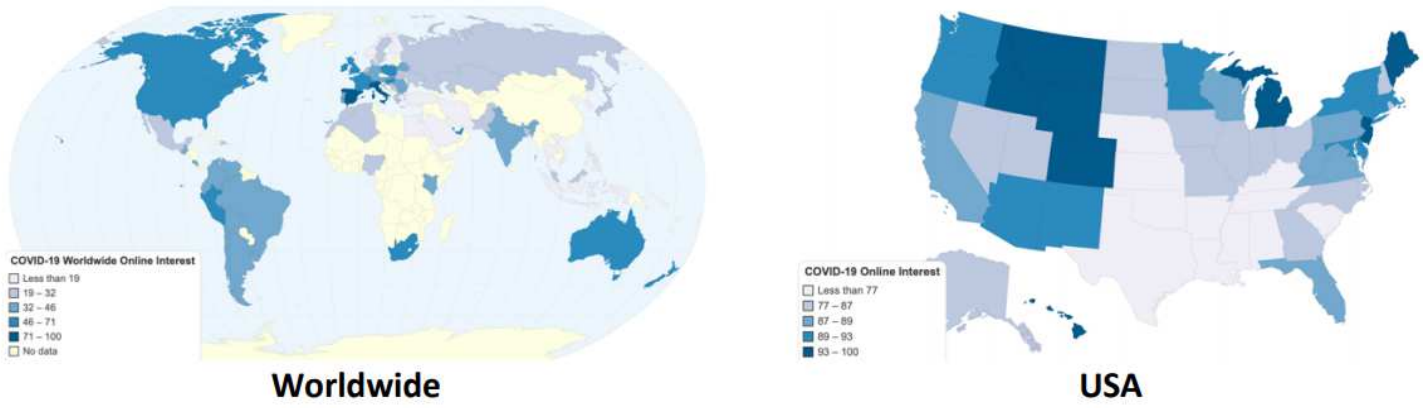


Figure 3

Heat maps of the worldwide and US online interest in “Coronavirus (Virus)”. 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.

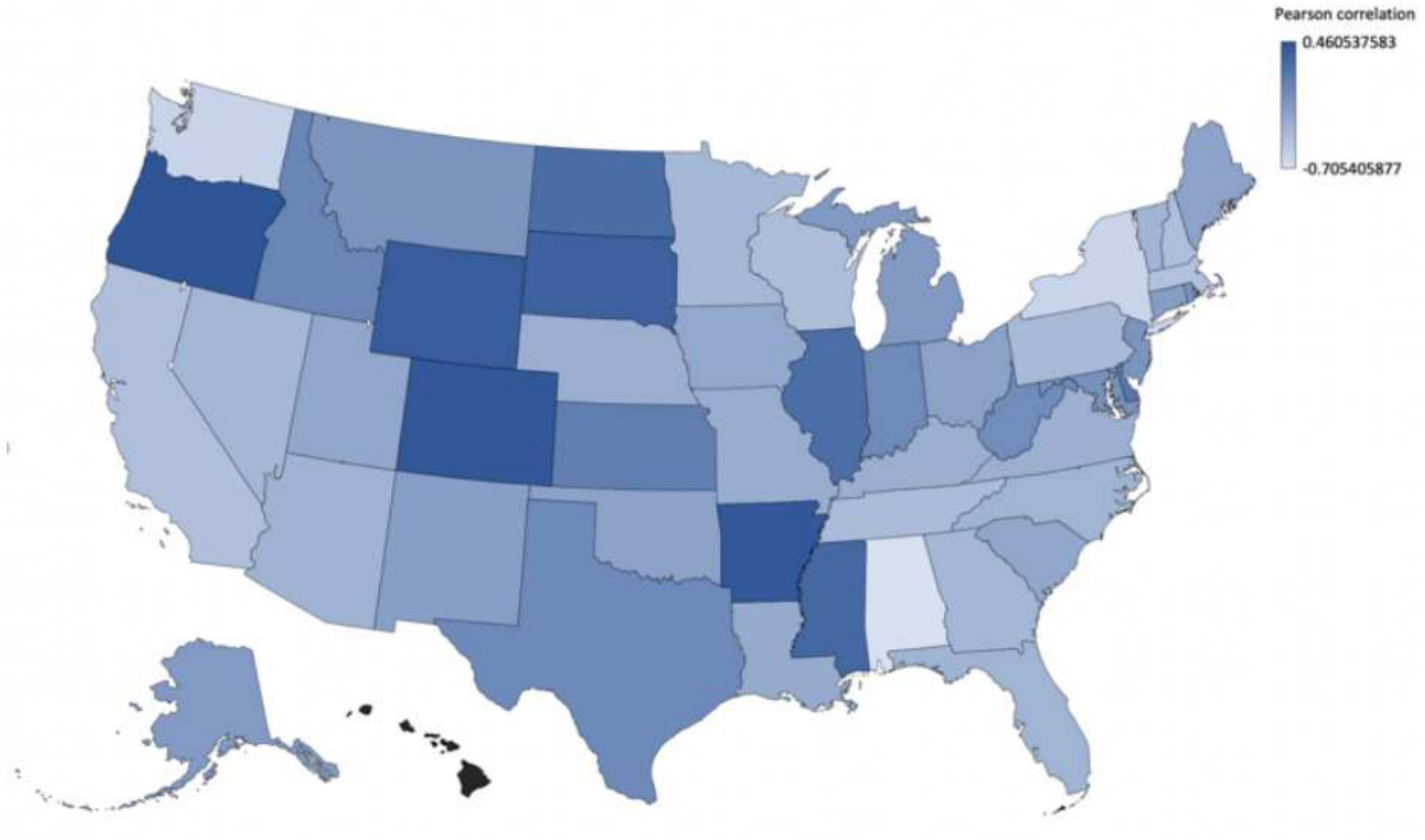


Figure 4

Heat map of the Pearson correlations by state.

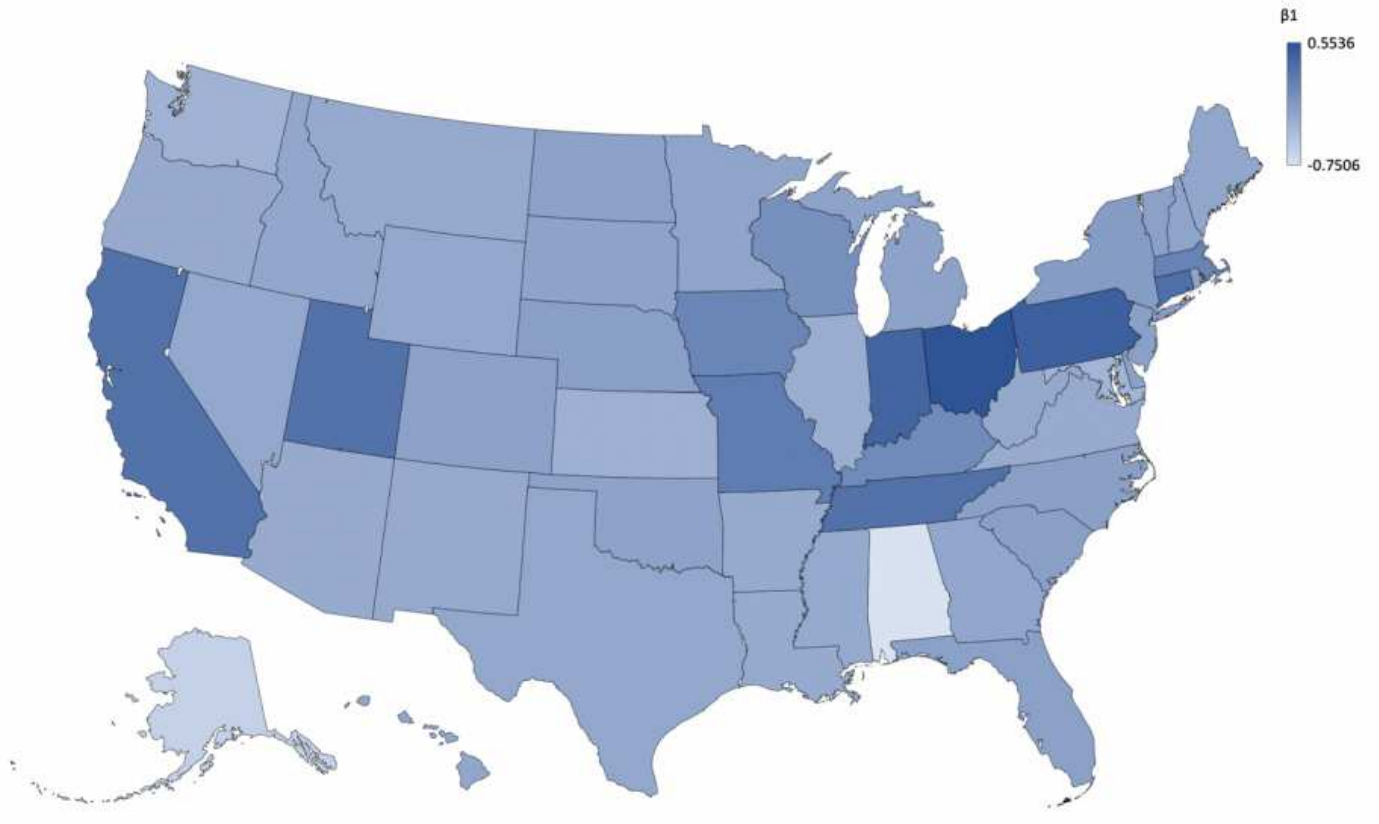


Figure 5

Heat map of the predictive analysis models' statistical significance.

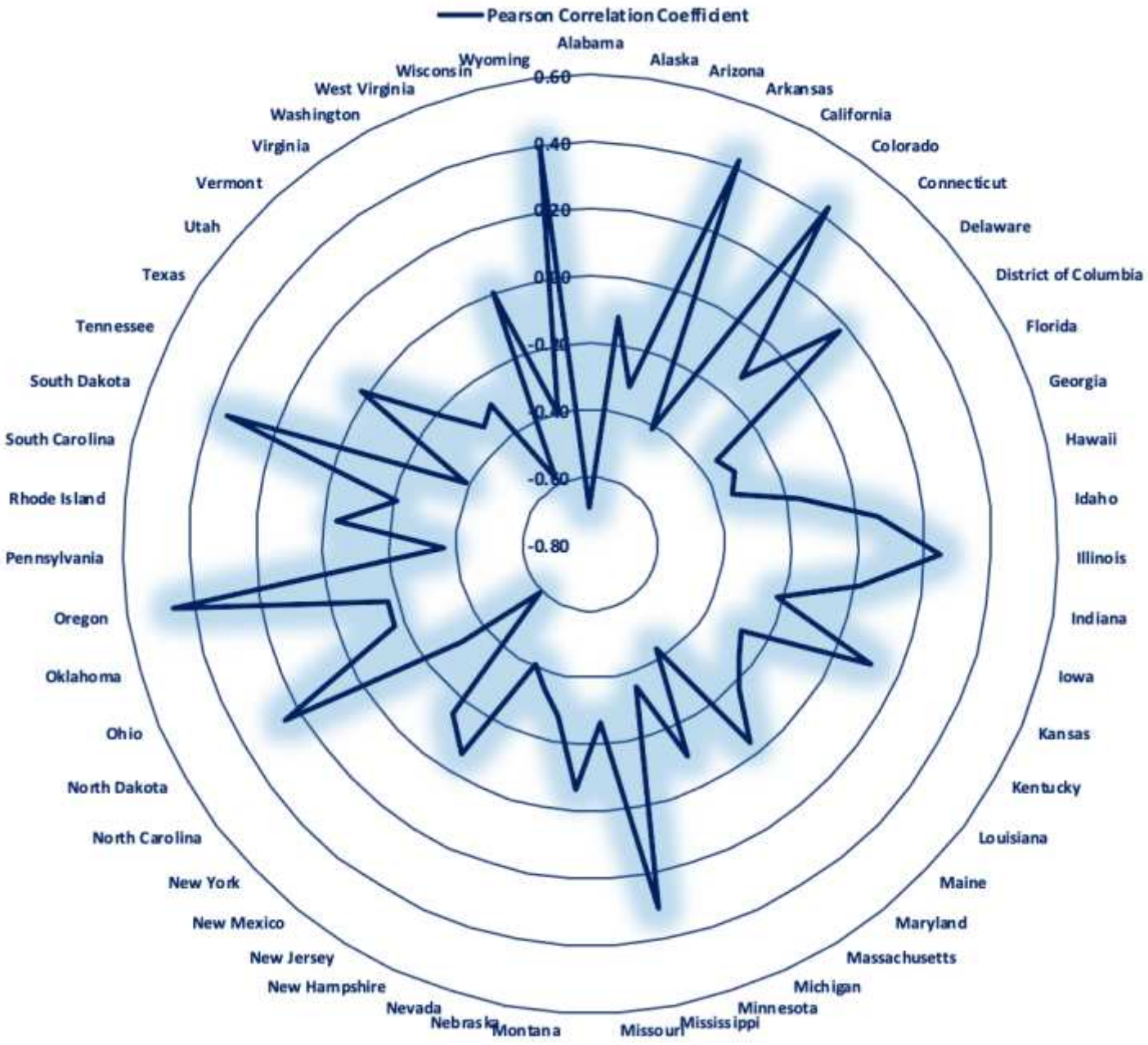


Figure 6

Radar chart of the Pearson correlations coefficients by state.

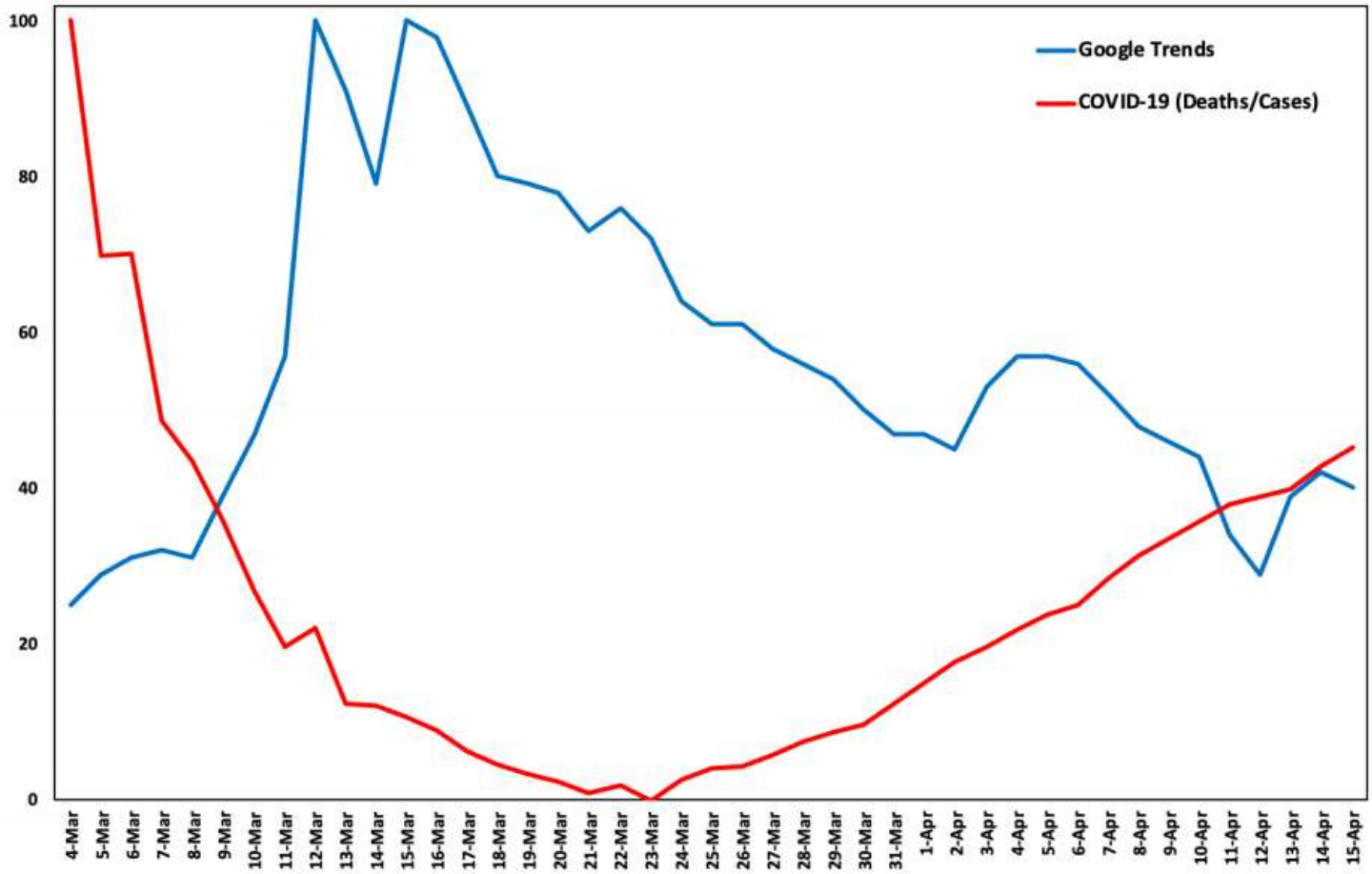


Figure 7

COVID-19 and Google Trends data from March 4th to April 15th in the US.