Title:	COVID-19: Spatiotemporal social data analytics and machine learning for pandemic exploration and forecasting
Author(s):	Vesselinov, Velimir Valentinov Middleton, Richard Stephen (Richard) Talsma, Carl James
Intended for:	Report
Issued:	2021-04-05

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

COVID-19: Spatiotemporal social data analytics and machine learning for pandemic exploration and forecasting

Description: This task focused on developing a preliminary approach to use machine learning (ML) to explore the relationship between county-level societal variables and COVID-19 parameters, including COVID-19 cases rates and counts and COVID-19 death rates and counts. The objective was to develop and test a prototype approach for linking COVID-19 and county-level data. The task focused on enhancing and applying existing LANL ML techniques to COVID-19. Our novel ML methods have been a subject of a recently approved U.S. patent. The codes based on these methods are already open-source released (http://tensors.lanl.gov). Our ML tools (NMFk/NTFk) are applied to extract hidden features (signals, waves) in the analyzed datasets and automatically identify their optimal number. The features are extracted by identifying counties that have similarities between the county-level societal variables and the COVID-19 parameters. These demonstration analyses will facilitate the ongoing pandemic simulations and predictions performed by Los Alamos other institutions, as well as lay the groundwork for future work.

COVID-19 datasets: We collected publicly-available county-level COVID-19 data for COVID-19 cases (rates and counts) and deaths (rates and counts) from March 1st 2020 until September 2020. For the ML analysis, we used the transient COVID-19 data (i.e., changing COVID-19 cases and rates over time) as well as multiple static COVID-19 snapshots.

Static county-level societal datasets: We collected static county-level data from 36 different static datasets, summing to a total of 924 variables that we tested using our ML approach. Each dataset and variable covered up to 3,142 US counties, though some variables had incomplete coverage. Examples of the datasets and representative variables included are:

- **Demographics:** Population count/density, race/ethnicity, age.
- **Census:** Households, housing type, housing occupancy, farm size, urban-rural metrics, largest city, poverty, citizenship, family size, vehicles, telephones, computers, etc.
- **Socioeconomics:** Unemployment/employment, poverty, social vulnerability, etc.
- **Crime:** Homicide, robbery, assault, burglary, incarceration trends, etc.
- **Education:** Education attainment, graduate degrees, etc.
- **Religion:** Multiple denominations, religious adherence, etc.
- **Personal:** Marriage/divorce rate, etc.
- **Opinion/politics:** 2016 election results, climate change, voter participation, implicit bias, etc.
- **Spending & revenues:** Federal spending (e.g., Medicare, social security, DOD), federal taxes, state taxes, etc.
- **Health:** Life expectancy, health activity, obesity, suicide, infant mortality, STDs, pregnancy, alcohol consumption, mental health, food insecurity, diabetes, overdoses, insufficient sleep, cancer, respiratory disease, mortality risk, Medicare rates, hypertension, smoking, etc.
- **Policy:** state of emergency dates, nursing home visits, non-essential business closures, masks in public,

Transient county-level societal datasets: Examples of the datasets and representative variables included are:

- **Mobility:** Movement by sector (residential, workplace, retail, grocery, etc.); median and max distance mobility; mobility for walking, driving, or public transit, etc.
- **Health:** the number of COVID related doctor's visits from outpatient visit data, etc..

Results: Below is a representative ML run that identified four different signals (or clusters) based on analyzing all 934 county-level variables and COVID-19 data representing changes in the death/case rates/counts from May, June, July, and August. The four signals are illustrated below.

Signal 1: Counties dominating Signal 1 are presented in Figure 1. The counties with high importance for the characterization of this signal have estimated weights close to 1 (red). The counties with low importance have weights close to 0 (green).

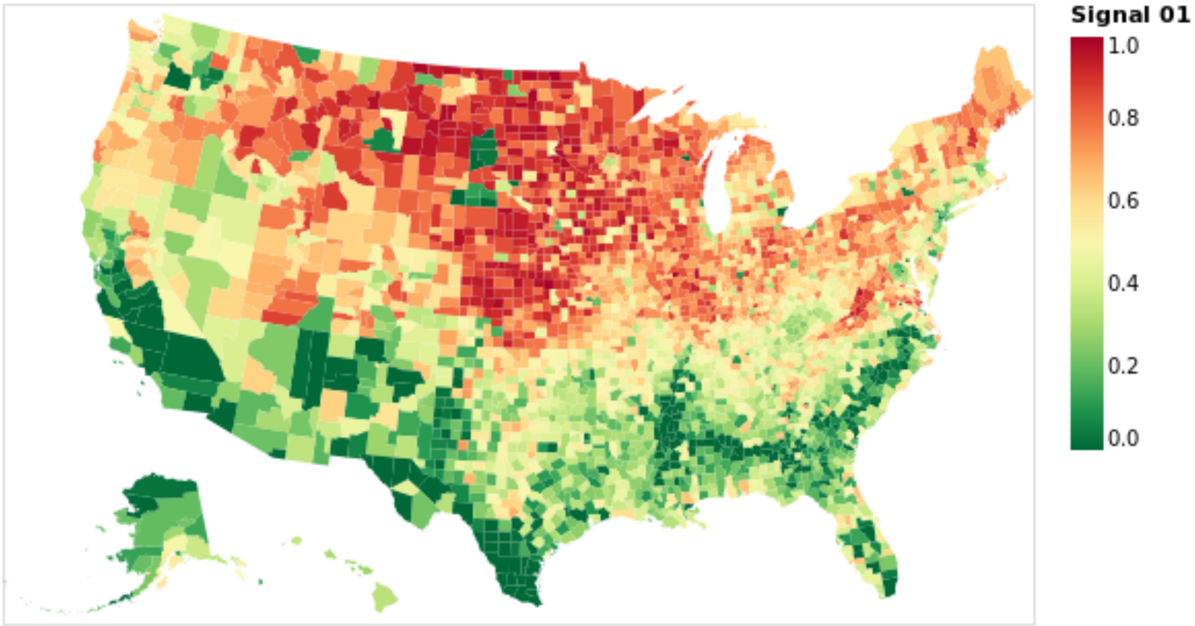


Figure 1: *Signal 1 - White, High Poverty, High Unemployment, Low COVID variables.*

Most influential county-level variables: Socioeconomic status, social vulnerability, unemployment insurance, disability, firearm businesses open.

Top COVID-19 variables: August COVID-19 case and death rate, worst COVID week (cases and deaths) up to August.

Interpretation: Signal 1 shows U.S. counties which show lower than expected COVID-19 spread. These counties can be characterized by being largely non-Hispanic white, exhibiting high unemployment, low percentile of income, and high economic social vulnerability as defined by the CDC.

Signal 2: Counties dominating Signal 2 are mapped in Figure 2.

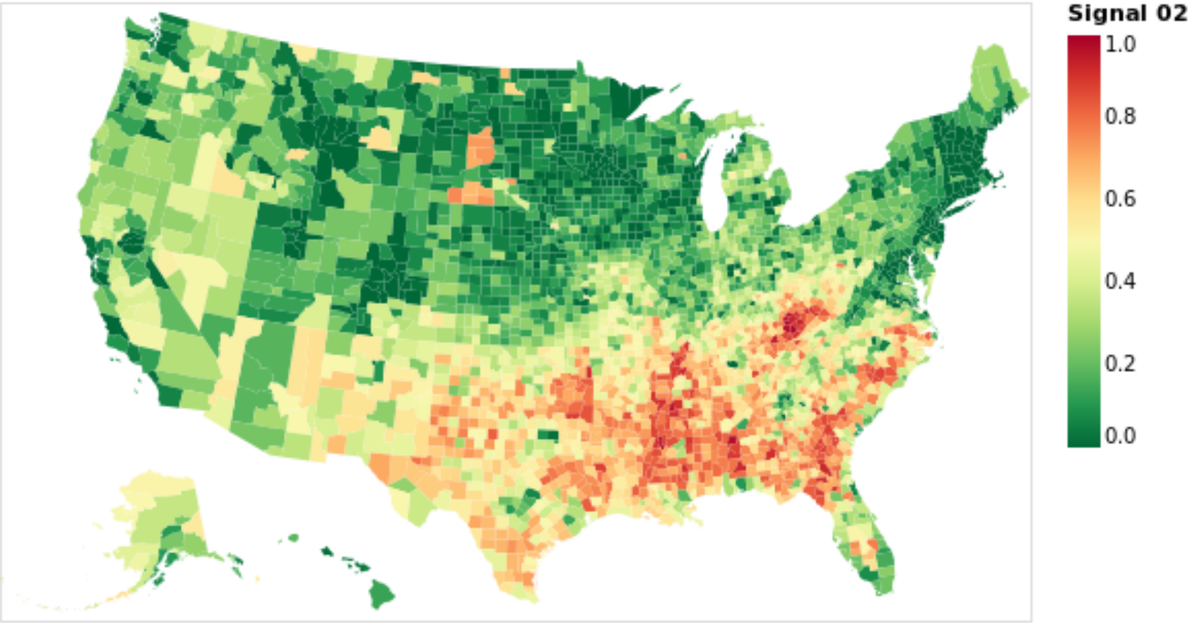


***Figure 2:** Signal 2 - High Poverty, Household Composition and Disability, High COVID Case Rate (08-31-2020, second wave).*

Most influential county-level variables: Alcohol/liquor stores being open, social vulnerability, unemployment.

Top COVID-19 variables: August COVID-19 case rate.

Interpretation: Signal 2, similarly to signal 1 shows areas of high unemployment and low economic status. However, August COVID-19 case rates are weighted highly in signal 2 (0.92), while signal 1 exhibited very low weight for COVID-19 related attributes. Additionally, signal 2 was represented by high percentile of population with a disability and a high percentile of population living in mobile homes. Alcohol and liquor stores being open was another highly-weighted attribute for signal 2, perhaps reflecting both high social vulnerability and limited policy response.

Signal 3: Counties dominating Signal 3 are mapped in Figure 3.

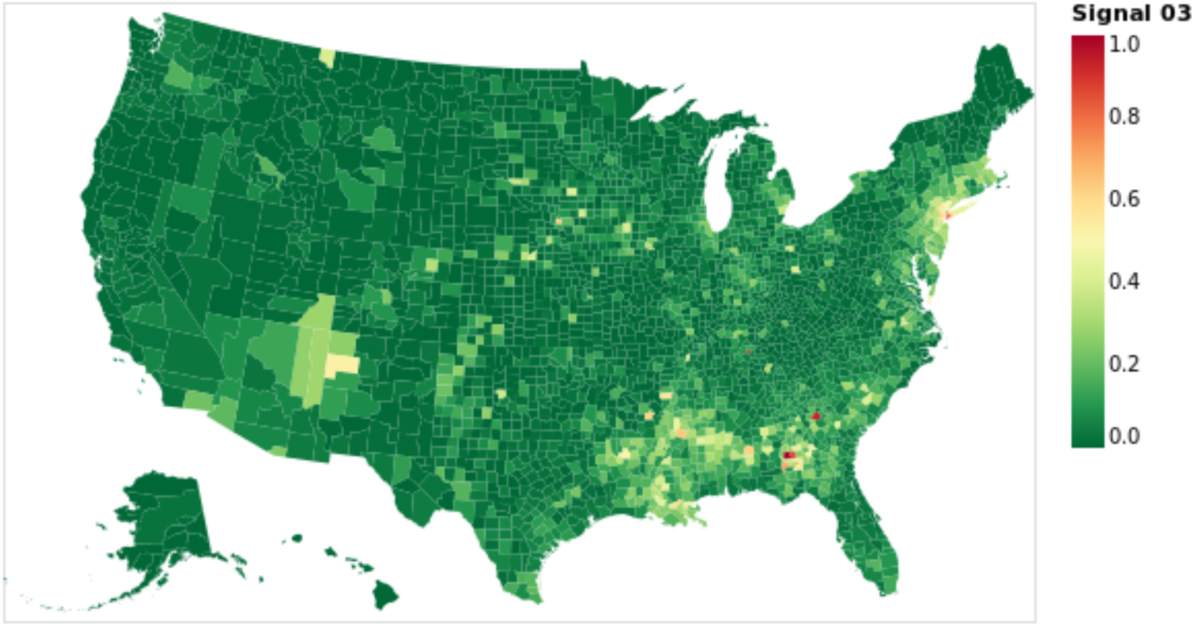


Figure 3: *Signal 3 - High COVID Death Rates.*

Most influential county-level variables: All COVID-19 related.

Top COVID-19 variables: COVID-19 death rates (multiple months).

Interpretation: Signal 3 is dominated by multiple COVID-19 parameters including COVID-19 case rates, death rates, and worst weekly rates for cases and deaths. The death rate is particularly important in signal 3, which shows counties where COVID-19 outbreaks have been particularly devastating and overwhelming. New York City, areas of the South and Navajo Nation regions are represented prominently by signal 3.

Signal 4: Counties dominating Signal 4 are mapped in Figure 4.

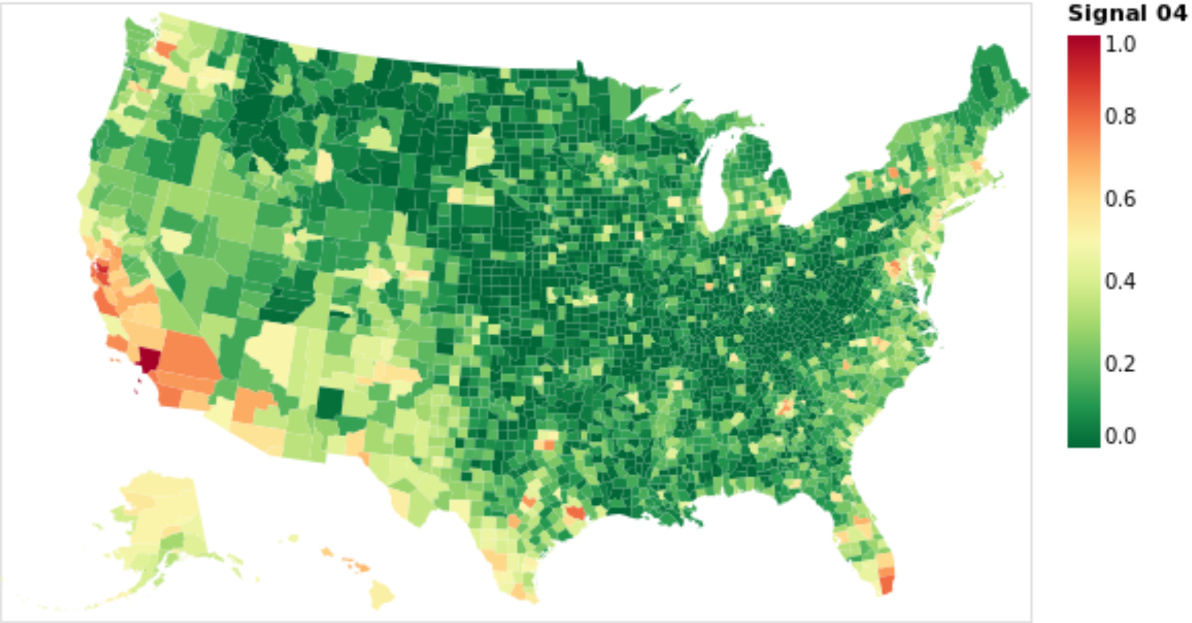


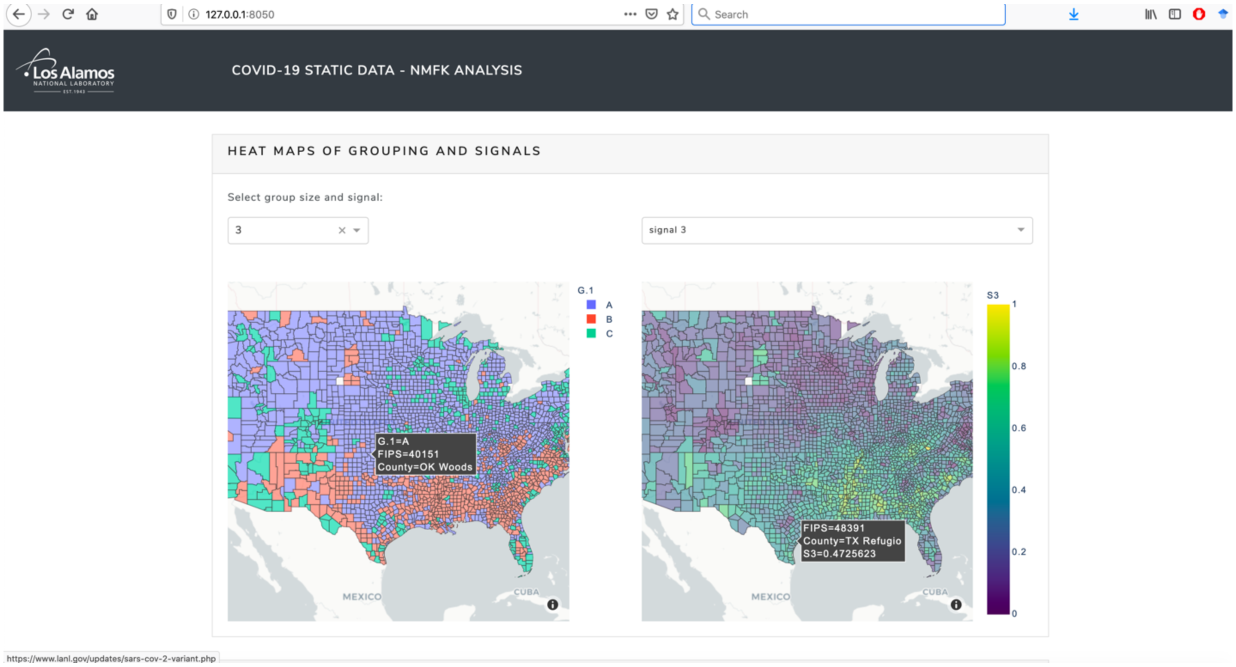
Figure 4: *Signal 4 - High Minority, non-English speakers, moderate COVID case rate.*

Most influential county-level variables: Minority status, access to exercise opportunities, ethnicity, climate change opinion, language.

Top COVID-19 variables: August and July COVID-19 case rates.

Interpretation: Signal 4 presents counties which have high COVID-19 case rates but low death rates. These counties have large minority and non-English speaking populations and rank high in the percentage of dense housing structures (10 or more units). Assessing which counties are represented by signal 4, it is evident that many larger urban counties are well represented. Interestingly, the percent of population believing that climate change is “happening” weighted significantly in signal 4. While the august COVID-19 case rate weighted moderately in signal 4 (0.58), it is not as highly weighted as is signal 2 (0.92).

Web-based visualization: We have developed a web-based visualization toolkit (Figure 5) that will allow users to dive deeper into the ML results. Users can dynamically change the group size and pick a particular signal to investigate the signal strength across the various counties. Furthermore, the toolkit also shows the top attributes contributing to the chosen signal. This allows one to understand the key attributes contributing to the spread of COVID-19. The toolkit is uses React.js in the backend and primarily uses the python packages pandas and numpy for query and the package plot.ly Dash for visualization components. The web front-end uses bootstrap and one can easily modify the front-end style with just a minor change in a python call. We plan on adding timeline and functionality to cluster keywords, that will allow visualization of transient ML outputs. This visualization toolkit can easily hosted on a LANL web server to either face internally or externally.



COVID-19 STATIC DATA - NMFK ANALYSIS

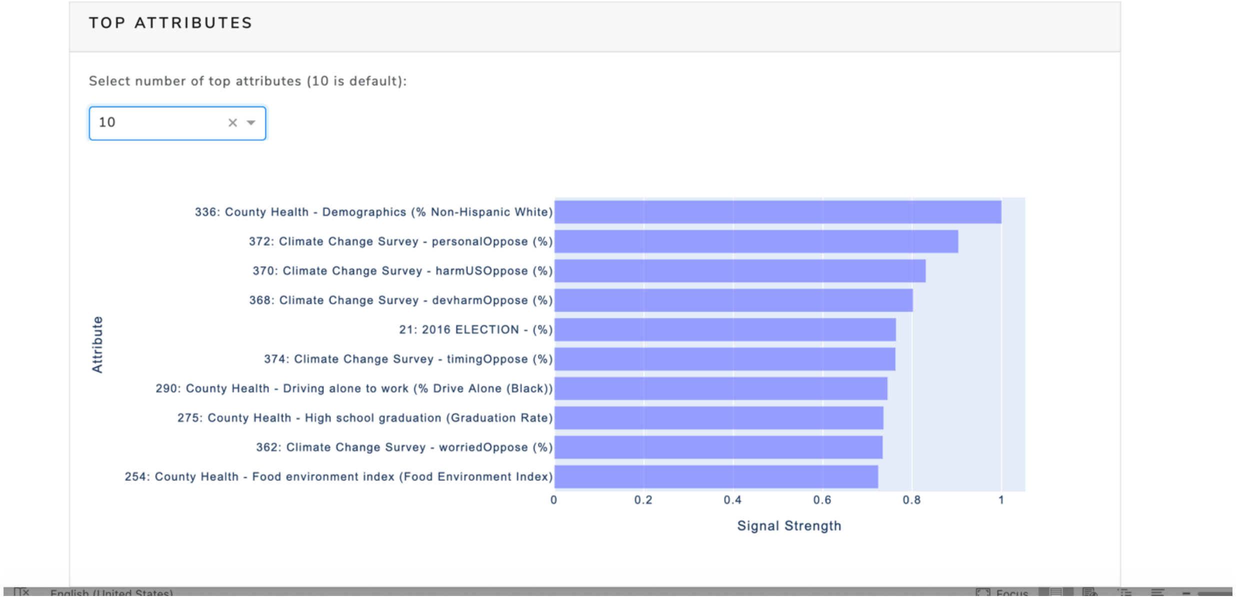
HEAT MAPS OF GROUPING AND SIGNALS

Select group size and signal:

3

signal 3

Figure 5: *Signal 4 - High Minority, non-English speakers, moderate COVID case rate.*



TOP ATTRIBUTES

Select number of top attributes (10 is default):

10

336: County Health - Demographics (% Non-Hispanic White)
372: Climate Change Survey - personalOppose (%)
370: Climate Change Survey - harmUSOppose (%)
368: Climate Change Survey - devharmOppose (%)
21: 2016 ELECTION - (%)
374: Climate Change Survey - timingOppose (%)
290: County Health - Driving alone to work (% Drive Alone (Black))
275: County Health - High school graduation (Graduation Rate)
362: Climate Change Survey - worriedOppose (%)
254: County Health - Food environment index (Food Environment Index)
Attribute
0
0.2
0.4
0.6
0.8
1
Signal Strength
English (United States)
Focus

Figure 5: Web-based visualization toolkit for the processed ML data

Conclusions: Our work has demonstrated a successful application of unsupervised ML methods for characterization of the COVID-19 pandemic. The temporal dynamics of COVID-19 cases and deaths are varying throughout the USA. It is not clear what causes these differences. Our work is a first attempt to explain these differences by relating the COVID-19 parameters to various county-level societal variables. Future work may involve incorporation of addition data such as mobility and hospitalization rates which can further improve our analyses.