ORNL/TM-2021/2329

COVID-19 Data Curation Effort: An Initial Analysis of the Data

Jesse Piburn
Robert Stewart
Jason Kaufman
Alex Sorokine
Eric Axley

September 2020

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

National Security Emerging Technologies Division

DOE COVID-19 Data Curation Effort: An Initial Analysis of the Data

Jesse Piburn, Robert Stewart, Jason Kaufman, Alex Sorokine, Eric Axley

Date Published:

September 2020

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

FIGURES

1. INTRODUCTION

During the COVID-19 pandemic of 2020, major case reporting outlets quickly coalesced around two or three primary vendors. Johns Hopkins University and *The New York Times* were among the more prominent, and all were of great value to the nation, particularly during the uncertain early stages of the pandemic. They primarily focused on three major attributes: number of new cases, deaths, and recovery. Recognizing that many states were reporting very detailed data sets (e.g., hospital beds) at a county level or finer, the ORNL Pandemic Modeling team embarked on a major data curation effort from March to June 2020 for the purpose of capturing this wealth of detailed data. The challenge of curating this data was daunting. The number of attributes reported by the states grew on almost on a weekly basis. States were routinely shifting their web tool strategies away from easily parsable HTML-based formatting to new Tableau and ArcGIS content. This growth in the sheer number of attributes, combined with the unpredictable shifts in data format, meant an aggressive and agile combination of automated scripting and manual scraping was required to capture new daily streams. Further, the team had to scale up staff and widen its approach for capture and storage. As a result, the team collected more than 11 million data points.

Following the close of this data collection effort on June 30th, 2020, the team embarked on a major effort to appraise what had been collected, including an inventory list, spatial completeness, temporal completeness, scale and geographic characteristics, and a determination. A report on this matter was submitted on September 15th, 2020, titled "DOE COVID-19 Data Curation Effort: Overview of Data Collection Coverage". Over 2000 unique attributes had been netted over a wide range of spatial scales, including state, county, zip codes, health regions, and census blocks. Over 11 million individual data points were collected across these attributes, and spatial coverage (in total) included all 50 states and multiple territories. What became apparent in the process is that in the absence of any data standards, many states reported a wide variety of unique attributes that were not always compatible with attributes reported in other states. As time continued, states began adding new attributes and offering finer grain detail in some older attributes. This meant that not all data streams existed for the entire time period; in fact, the number tended to increase dramatically towards the end. Often, states would begin an attribute series and then stop altogether. These highly variable and uncertain conditions illuminated the need for *harmonization approaches* that would reconcile and conflate changing attribute names and detail over time. For example, grouping racial data reported as either Black or African American, depending on the state, into a single harmonized attribute. These choices would make a *within-state* analysis possible during the time period and lead to potential between-state analytics later on. This was almost entirely a manual decision process, requiring some subjective decision-making at times, to prevent a fragmented, short-lived collection of time series fragments that would offer few insights into trends, patterns, and correlates.

This report imports harmonized data for state and county into the World Spatio-Temporal Analytics and Mapping Project (WSTAMP). WSTAMP is a major space-time analysis and visualization tool developed at ORNL for the National Geospatial-Intelligence Agency specifically for this kind of exploratory analysis. WSTAMP offers a rich analytical and graphical environment consisting of a wide range of analytics. These include time series plots, statistical summaries, data mining techniques, trend and pattern detection, and hypothesis generation.

The use of WSTAMP provided a first deep look at the data and allowed an uncovering of a number of critical findings not possible otherwise. These included QA/QC issues that require further attention, including mitigation of some point-wise curation errors and a need to review a select number of harmonization decisions. A first detailed look at attribute trends revealed that many were agreeably

continuous while some were fragmented and short lived. For the most complete attributes, a series of data mining exercises were conducted that surfaced the most common multi-variate trends, revealing the presence of correlation among many of the data.

The inquiry was organized around the principles of *within-state* and *between-state* analysis. In within-state analysis, attributes are analyzed for a specific state either at the county or state level. In between-state analysis, state level attributes that are reasonably represented across the nation are analyzed to reveal correlates, trends, and patterns in the analysis. This report focuses on two within-state exemplars: Ohio and Florida. Ohio's state level data was among the most consistent and reliable in the nation. Florida's county level data were similarly stable and reliable. The aim in these case studies is to demonstrate the kinds of analytics that are possible (particularly in WSTAMP) and explore how space-time data in this context can be analyzed and interpreted within a strictly exploratory aim.

This report begins with an overview of the data and a reminder of the conditions of curation and why harmonization techniques played such a significant role. Next, an overview of WSTAMP methods is presented, followed by results for within and between state analysis. Finally, a summary of the findings and proposed next steps is presented.

2. DATA

2.1 CURATION

Curating data from 50+ state level and equivalent webpages was extraordinarily challenging. As the pandemic unfolded across the nation from March onward, states responded with demands for information, using a wide variety of uncoordinated approaches. States began with data dissemination solutions they had in place at the start of the outbreak but transitioned to newer platforms (e.g., Tableau), making automated curation through scripts difficult to build and maintain. This, in combination with a substantial growth in the number and granularity of attributes reported by the state, meant that data curation was necessarily a combination of manual extraction and script automation. From day to day, scripts would fail due to state resource changes, requiring an agile move to manual curation and then back to scripting when updates were available. In addition, states presented widely varying reporting cadences, and many would periodically stop reporting and then resume again. In many cases, states did not maintain any kind of historical record on their websites. Instead, each day was only a snapshot, and any data missed during a 24-hour period was permanently lost. Attribute name changes, new data hierarchies, and platform shifts mean a constant effort was required to sustain curation during the initial wave of cases across the nation. Additionally, the team had to scale up staff and widen its approach for capture and storage.

Every effort was made to ensure quality during this exceptional curation environment. Unfortunately, the unrelenting operational cadence of the effort, the periodic shortage of manpower, and simply unprecedented circumstances surrounding a global pandemic meant that development and adherence to formal QA/QC procedures were simply not possible. During curation only, end-of-day cursory reviews were possible and indeed they did catch many problems. It was not until Phase 2 that attention could be fully given to the QA/QC challenge. The previous report "DOE COVID-19 Data Curation Effort: Overview of Data Collection Coverage" covers the first high level appraisal of the data, including spatial and temporal coverages and the pressing need for harmonization of the data. It was in preparation of this report and the consequential harmonization efforts that led to a substantial improvement in the COVID data cube and positioned it for ingestion into a tool such as WSTAMP. The discussion turns first to the issue of harmonization before returning to the challenge of QA/QC.

2.2 HARMONIZATION

The lack of any nationally coordinated data strategy meant that each state forged its own path for informing the public about COVID-19. Individual state strategies also evolved over time, with attribute name changes, addition of new attributes, and greater detail in those attributes. For example, early on, a state might have reported only cases and deaths. This would then expand to cases and deaths by gender and then cases and deaths by age and gender. This continual expansion constantly created new temporal trends that had not previously existed. While the proliferation of more precise data is very useful to help better understand the effects of the pandemic, the constant shifting of variables created challenges for consistent attributes across the dataset, even within a single state. This gives the appearance that one attribute may end while another begins immediately following it. When considering these dynamics across 50 states, the situation is substantially worse. Many attributes exist only within a given state, making comparison to other states difficult. Consider the issue of reporting cases and deaths. While every state reported cases and deaths at the state level, some reported by age brackets with widely varying time ranges. For example, Idaho reported as few as 3 age brackets while Montana used as many as 20. While geographically adjacent, it is impossible to compare these age brackets; Idaho separated ages into 0-17, 18-49, and 50+, whereas Montana used 5-year intervals. These age intervals do not cause any data issue looking within state for either Montana or Idaho, but they create large challenges when looking between-state for the region.

In order to maximize the value of these data, *harmonization strategies* addressed each opportunity for relinking, conflating, and restoring time series broken apart by name changes or finer scale attribution. Harmonization changes can be both major and minor. A good example of a minor harmonization change was combining the racial attributes ‘Black’ and ‘African American’ into ‘African American or Black.’ A much larger harmonization of racial data was the grouping of all racial and ethnic definitions into a single category, for example ‘Cases by Ethnicity and/or Race’. This significant change was necessary because a minority of states reported data in this format originally; it made more sense to create one inclusive attribute usable across all states rather than several related attributes that could only be used over some states. This ‘parent’ attribute then has the child attributes of various racial and ethnic characteristics, such as ‘African American or Black’. This harmonization then allows data reported with minor name differences to be comparable between states much easier than the original data. It also allows for greater temporal comparisons not only between-state but also within-state, if the state changed their reporting standards during the data collection period. It is important to note that attribute meanings also varied slightly from state to state. At this point in the investigation, there is no clear approach for dealing with these variations; therefore, this issue is postponed. In the end, the harmonization reduced almost 19,000 attributes into approximately 1900. Figure 1 (from the previous report, “DOE COVID-19 Data Curation Effort: Overview of Data Collection Coverage”) presents harmonized attributes for US county data.

Figure 1. Harmonized attributes by total US county coverage

It is critical to observe that after harmonization of data, there is still fragmented coverage both temporally and spatially for many of the data. Sporadic missing data are not usually problematic (for example Hospital cases by Age Range); however, large gaps in data or late startups (like Per Capita Population Tested) have limited use outside of the counties they are located within.

2.3 QA/QC IN WSTAMP

As mentioned earlier in the curation section, every effort was made to maintain quality assurance during this exceptional curation period; however, most reviews were limited to end-of-day cursory checks for obvious omissions and mistakes. Quality assurance was further assessed in our previous report by examining coverage and completeness, along with a range of harmonization choices that restored attribute

fidelity under a variety of conditions. In this report, WSTAMP offers an excellent opportunity to examine QA/QC in greater detail, in a rich visual analytical environment.

Analytics, like simple time series plots, time series with outliers, and find anomalous trends, easily reveal suspect data points that are difficult to see in a database or tabular format. In Figure 2, two likely errors stand out with single days far exceeding any other collection days. Cases by Age Range was also a cumulative value; it should steadily have increased each day instead of increasing or decreasing as cases waxed or waned. These are not the only errors of this type; a number of other smaller peaks are evident in the data. These are harder to identify with standard QA/QC efforts; they are not outliers in regard to the entire set of values but do stand out as anomalous for that state and date. This WSTAMP visualization shows these anomalies easily.

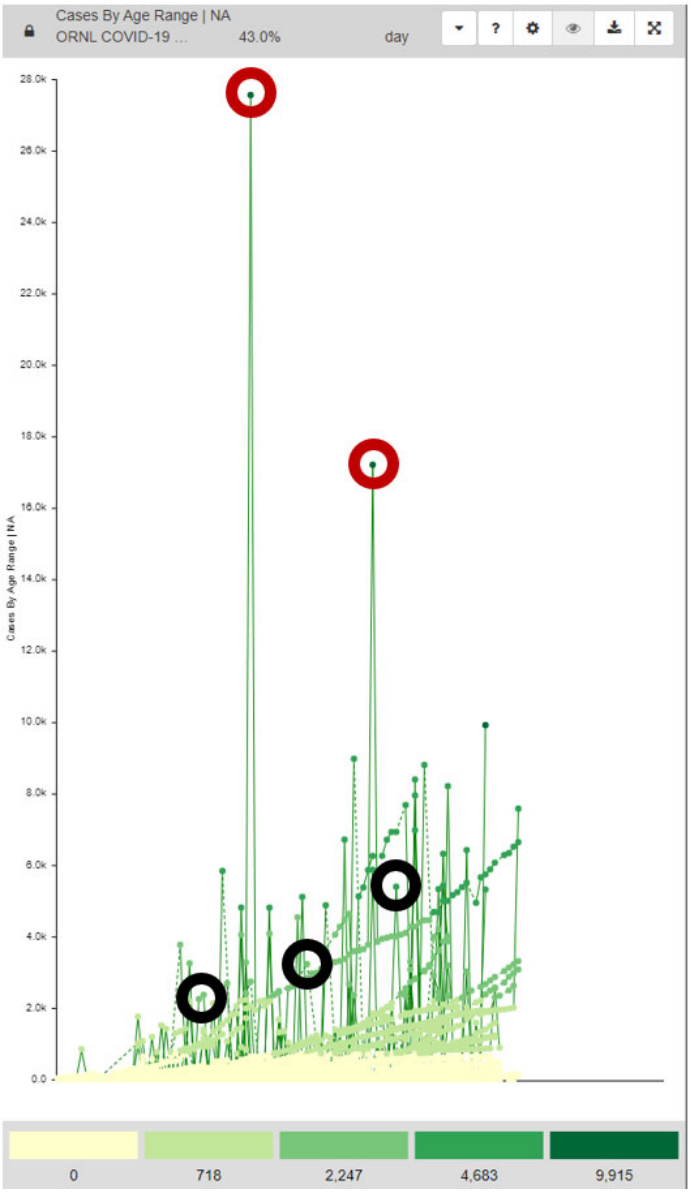

Figure 2. Cases by Age Range over the collection period (note the obvious collection errors in red and the less obvious in black).

These anomalies cannot be easily dismissed as false, however. States were under enormous pressure to deliver data in days that would have normally been delivered in months, which caused data quality errors. This often resulted in erroneous data reported by the states, only to be retracted the next day. This is most evident in deaths for small counties, where deaths of -1 were commonly reported on days after a new death. Rather than people brought back to life, these are simply instances of data quality failures; however, the data collection environment necessitated capturing all of these instances without editing the previous data. This was simply not possible within the workload of the project. These small scale anomalies are present throughout the data and are not mistakes but examples of poor data reporting by the source and are an example of the fast paced environment that defined curation efforts.

While the harmonization effort turns the fractured COVID-19 data collection environment into a uniform dataset for further analysis, further QAQC of the harmonization is likely needed for future work. In Figure 3, what is immediately obvious is an alternating set of hospitalized cases points, one slowly increasing and one quickly increasing. This is unlikely to be a data collection error; the pattern is far too regular and common. What is most likely is that the harmonization effort grouped two data points together that perhaps should not have been.

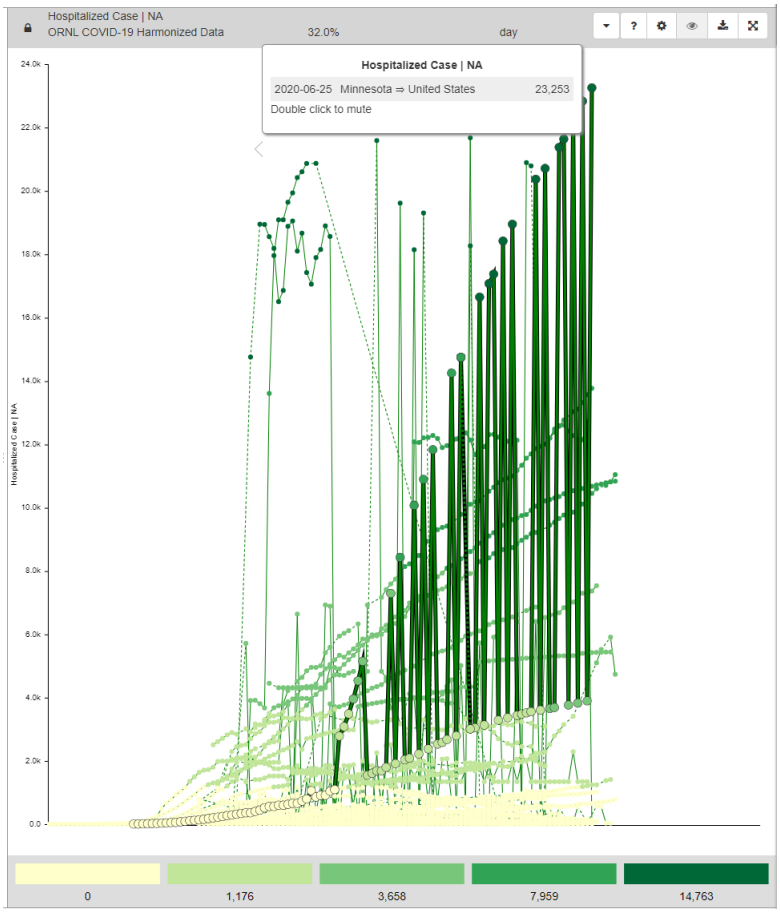

Figure 3. Hospitalized Cases over the collection period. Highlighted are hospitalized cases for Minnesota, which clearly oscillate between two unharmonized variables.

In the final assessment, two QA/QC action items are identified using WSTAMP that were not clear before: 1) revisitation of possible data entry errors like those seen in Figure 2 and 2) a closer look at harmonization choices particularly those that produce oscillatory behaviors such as those seen in Figure 3.

Fortunately, the number of obvious data entry errors is quite limited in the study, and WSTAMP analytics, particularly those involved in data mining, were fairly resilient to oscillatory artifacts emerging from harmonization. The next section discusses use cases and the supporting data for exploring within-state and between-state analysis.

2.4 USE CASE DATA

This report focuses on three case studies: 1) a county level within-state analysis of Florida, 2) a state level within-state analysis of Ohio, and 3) a between-state analysis of all 50 U.S. states and the District of Columbia. In the case of within-state analysis, Ohio and Florida make excellent exemplar studies owing to their high number of well populated consistent attributes, respectively, at the state and county level. To select attributes for each use cases study, the time period was set from April 1 to June 30, where the richest data collection occurred. Additionally, to limit the analysis from limitations caused by small data sets, only attributes with a completeness percentage above 20% were selected. Table 1 details attributes used in Florida for the Within-State County level analysis, Table 2 in Ohio for the Within-State State Level analysis, and Table 3 for the Between-State Analysis.

Table 1. Florida Within-State County Level Variables.

Attribute	Completeness Percentage
Cases by Age Range	68
Cases by Ethnicity and/or Race | African American or Black	44
Cases by Ethnicity and/or Race | Hispanic/Latinx	44
Cases by Ethnicity and/or Race | Non-Hispanic/Latinx	44
Cases by Ethnicity and/or Race | Other	44
Cases by Ethnicity and/or Race | Unknown	44
Cases by Ethnicity and/or Race | White or Caucasian	44
Cases by Gender | Female	59
Cases by Gender | Male	59
Cases by Gender | Unknown	59
Cases	70
Cases | State Residents	60
Deaths	68
Hospitalized Case	68
Number of Negative Tests	68
Number of People Tested	68

Table 2. Ohio Within-State State Level Variables

Attribute	Completeness Percentage
Cases by Age Range | Lower Bound of Estimate	62
Cases by Age Range	86
Cases by Age Range | Upper Bound of Estimate	62
Cases by Age Range, And by Gender | Female	74
Cases by Age Range, And by Gender | Male	74
Cases by Age Range, And by Gender | Unknown	74
Cases by Ethnicity and/or Race | African American or Black	63
Cases by Ethnicity and/or Race | American Indian or Alaskan Native	63
Cases by Ethnicity and/or Race | Asian	63
Cases by Ethnicity and/or Race | Hispanic/Latinx	62
Cases by Ethnicity and/or Race | Hawaiian or Pacific Islander	63
Cases by Ethnicity and/or Race | Non-Hispanic/Latinx	60
Cases by Ethnicity and/or Race | Other	63

Cases by Ethnicity and/or Race | Refused to Answer	44
Cases by Ethnicity and/or Race | Two or More Races	63
Cases by Ethnicity and/or Race | Unknown	63
Cases by Ethnicity and/or Race | White or Caucasian	63
Cases by Gender | Female	80
Cases by Gender | Male	80
Cases by Gender | Unknown	79
Cases	51
Counties with a Case	62
Counties with a Death	58
Counties with a Hospitalized Case	63
Deaths by Age Range	86
Deaths by Age Range, And by Gender | Female	74
Deaths by Age Range, And by Gender | Male	74
Deaths by Age Range, And by Gender | Unknown	74
Deaths by Ethnicity and/or Race | African American or Black	63
Deaths by Ethnicity and/or Race | American Indian or Alaskan Native	23
Deaths by Ethnicity and/or Race | Asian	59
Deaths by Ethnicity and/or Race | Hispanic/Latinx	62
Deaths by Ethnicity and/or Race	57
Deaths by Ethnicity and/or Race | Non-Hispanic/Latinx	62
Deaths by Ethnicity and/or Race | Other	63
Deaths by Ethnicity and/or Race | Refused to Answer	33
Deaths by Ethnicity and/or Race | Two or More Races	63
Deaths by Ethnicity and/or Race | Unknown	63
Deaths by Ethnicity and/or Race | White or Caucasian	63
Deaths by Gender | Female	85
Deaths by Gender | Male	68
Deaths by Gender | Unknown	49
Deaths with Pending Tests	43
Deaths	54
Hospitalized Case | In ICU	52
Hospitalized Case	52
Hospitalized Cases by Age Range	86
Hospitalized Cases by Age Range, And by Gender | Female	74
Hospitalized Cases by Age Range, And by Gender | Male	74
Hospitalized Cases by Age Range, And by Gender | Unknown	74
Hospitalized Cases by Ethnicity and/or Race | African American or Black	63
Hospitalized Cases by Ethnicity and/or Race | American Indian or Alaskan Native	63
Hospitalized Cases by Ethnicity and/or Race | Asian	63
Hospitalized Cases by Ethnicity and/or Race | Hispanic/Latinx	62
Hospitalized Cases by Ethnicity and/or Race | Hawaiian or Pacific Islander	52
Hospitalized Cases by Ethnicity and/or Race | Non-Hispanic/Latinx	62
Hospitalized Cases by Ethnicity and/or Race | Other	63
Hospitalized Cases by Ethnicity and/or Race | Refused to Answer	63
Hospitalized Cases by Ethnicity and/or Race | Two or More Races	63
Hospitalized Cases by Ethnicity and/or Race | Unknown	63
Hospitalized Cases by Ethnicity and/or Race | White or Caucasian	63
Hospitalized Cases by Gender | Female	74
Number of People Tested	51
Pending Tests	43
Probable Cases by Ethnicity and/or Race | Two or More Races	38

Table 3. US Between State Level Variables

Attribute	Completeness Percentage
Cases	71
Deaths	66
Cases by Age Range | 80 - 89	65
Cases by Gender | Female	59
Cases by Gender | Male	59
Hospitalized Case	55
Number of People Tested	48
Deaths by Age Range | 70 -79	44
Cases by Gender | Unknown	40
Cases by Ethnicity and/or Race | African American or Black	33
Cases by Ethnicity and/or Race | White or Caucasian	33
Cases by Ethnicity and/or Race | Hispanic/Latinx	32
Cases by Ethnicity and/or Race | Unknown	28
Cases by Ethnicity and/or Race | Other	28
Recovered Cases	26
Deaths by Gender | Male	24
Cases by Ethnicity and/or Race | Asian	24
Deaths by Gender | Female	24
Number of Negative Tests	23
Deaths by Ethnicity and/or Race | African American or Black	23
Deaths by Ethnicity and/or Race | Hispanic/Latinx	22
Hospitalized Case | In ICU	21

3. METHODS

In order to expedite a deeper analysis of the COVID cube data, harmonized attributes at county and state level were ingested into WSTAMP. WSTAMP is a rich visual and analytical environment developed at ORNL by NGA to analyze high dimensional (high number of attributes) spatio-temporal time series data. WSTAMP methods include a wide range of analytics, and development of the capability can be traced in Piburn et al. (2017a and 2017b) and Stewart et al. (2015). The public version of the tool, which includes data from the World Bank, the World Health Organization, and many other vendors, can be accessed at wstamp.ornl.gov. This section highlights analytics that are applied to the COVID cube data to uncover remaining QA/QC challenges, patterns, and correlates.

Time Series Analytic: The Time Series WSTAMP analytic allows visualization of how the values of the selected attribute change from one time observation to the next for the entire time selection. All available selected locations are displayed as individual lines. The X-axis is the time selection, and the Y-axis are the values of the selected attribute.

Time Series With Outliers Analytic: The Time Series with Outliers WSTAMP analytic displays the same data as the Show Time Series analytic in the form of a box plot, allowing the user to quickly see which observations in each time step are considered outliers. Values are displayed for the selected attribute, locations, and time range. Outliers are determined through a calculated quartile range of that time interval, depending on the selection. The interquartile range is calculated, then multiplied by 1.5, added to the 3rd quartile, and subtracted from the 1st quartile. Outliers identified in this graph will also be shown in the Find Unusual Trends analytic. Values within the interquartile range will be shown with a solid blue bar. Values outside of that, but not outliers, are shown with a light blue bar. Outliers are shown as white circles.

Net Change Over Time Analytic: The Show Net Change Over Time WSTAMP analytic assesses the amount of change from the first and last observation within the time selection across locations or attributes. This difference can be shown by location or by attribute.

Cluster by Trend Analytic: The Cluster by Trend WSTAMP analytic uses time series data mining algorithms to organize all the currently selected locations and attributes into clusters of how similar their behavior evolves over time. This analytic then displays a characteristic trend, representing the average for each of the number of clusters that were selected, allowing simultaneous organization of any number of attributes into a few manageable categories. The calculated trend groupings are displayed on the graph. Each similar trend can be clicked to take a deeper dive and see the individual trends that were clustered into that signature. The colors for each location on the map correspond to that specific trend, providing a map of behaviors over time and displaying which areas are spatial outliers (e.g., isolated) in their behavior.

Categorize Trends Analytic: The Categorize Trends WSTAMP analytic can be used to view all the Trend Taxonomies for any attribute for the selected locations. All locations will be placed into the Taxonomy that most closely matches its behavior, compared to the most recent value on the Y-axis, and the map will be colored to match the corresponding Trend Taxonomy for spatial analysis.

WSTAMP has 10 Trends in this analytic:

- Up: Values that are steadily increasing over the time selection
- Down: Values that are steadily decreasing over the time selection
- Smile: Values that are initially decreasing but with an upturn towards the end of the time selection
- Frown: Values that are initially increasing but with a downturn towards the end of the time selection
- Wave: Values that are showing erratic behavior over the time selection
- Constant: Values that are showing little to no change over the time selection
- Spike Up: Values that are showing little to no overall change except for a sudden rise during a limited time period in the selection, which returns to the previous values
- Spike Down: Values that are showing little to no overall change except for a sudden drop during a limited time period in the selection, which returns to the previous values
- Step Up: Values that are initially steady, stepping upward in value and holding at the new value for the rest of the time period
- Step Down: Values that are initially steady, stepping downward in value and holding at the new value for the rest of the time period

Find Unusual Trends Analytic: The Find Unusual Trends analytic can be used to find two different types of anomalies in WSTAMP data: Trend Anomalies, which are unlike other selected locations and attributes, and Magnitude Anomalies, which are values outside the calculated interquartile range, same as in Show Time Series With Outliers. This analytic is an exceptionally powerful tool that finds attributes in a large collection that are anomalies for individual countries. The chart will show trends and/or magnitude

values that have been identified as an anomaly, along with the last value of any anomalous attribute identified.

Find Trends Similar to Mine Analytic: The Find Trends Similar to Mine analytic uses a non-linear time series data mining algorithm to organize spatiotemporal trends in order of how similar their behavior evolves over time to the currently locked target location and locked attribute pair. Much like a distance measurement, the values indicate how close the trends are to one another in their temporal evolution and are arranged from most similar (low values) to least similar (high values).

Calculate Attribute Predictability Analytic: Calculate Attribute Predictability is a spatio-temporal data mining algorithm that measures an attribute's overall stability for a geographic entity. By integrating approximate entropy values over the entire time series, this analytic incorporates both how widely varying the values are and how predictable that variance is from one observation to the next. Larger resulting values indicate a more chaotic or less stable time series, while a lower value indicates a more stable trend. The units are based on the selected attribute's units and are a unit-entropy measure.

4. RESULTS

4.1 WITHIN-STATE COUNTY LEVEL ANALYSIS

Florida provided a good example as a dataset with diverse county level data by age and consistent data reporting. Sixteen county level attributes were chosen (Table 1) for this study making it possible to conduct both individual attribute studies as well as multivariate analysis. In this study we began with a plot of the time series and time series with outliers analytic to become familiar with the data and its possible limitations. Rather than produce 16 graphs for simple time series and 16 more for time series with outliers we showcase only a couple to demonstrate the kind of discoveries possible with WSTAMP analytics

Figure 4 shows the number of people who are no longer being monitored – defined by the state of Florida to no longer be a contagious case. Two things emerge quickly. First Lee and Collier county are clear outliers. Secondly the number of people meeting this definition significantly leveled off towards the month of June and for many examples appears to no longer be increasing. This can easily generate a pair of hypotheses for further analysis: was the outbreak largely contained in most counties by June, or did the state of Florida change the definition and stop tracking these cases.

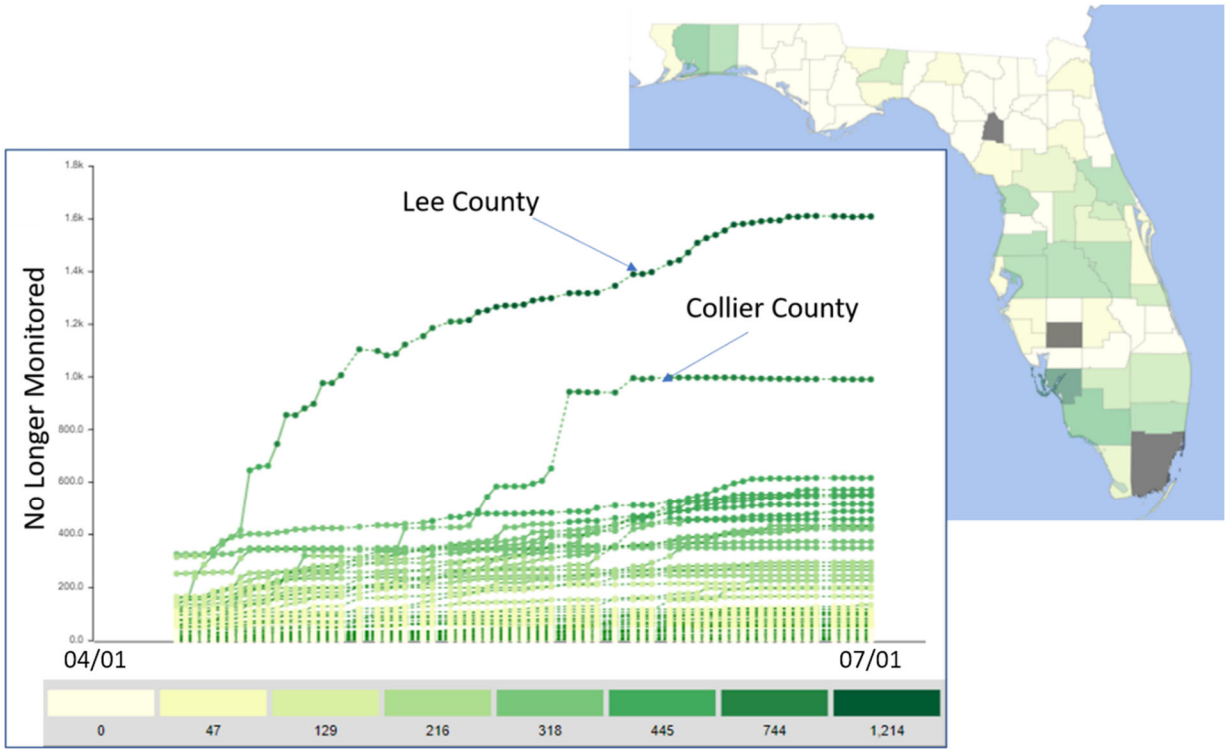

Figure 4. Persons no longer monitored in Florida counties over time

If we look at the attribute New Cases, showing the number of new infections yesterday in Florida using the Time Series With Outliers tool, it suggests that the second definition is more likely to be true, as seen in Figure 5. If the outbreak was largely contained by June in Florida, then the Time Series With Outliers would show that the size of the box plot was decreasing towards the end of the outbreak, with few high outliers outside of the box plot. Instead, the opposite is true; the highest outliers for new cases are at the very end of the study period and the box plot steady increases in size during the last quarter of the chart. This is opposed to what is shown in Figure 4, which suggests a largely contained outbreak, showing that despite what appears to be solid data, the a plausible explanation is that Florida simply stopped updating this data point for most counties.

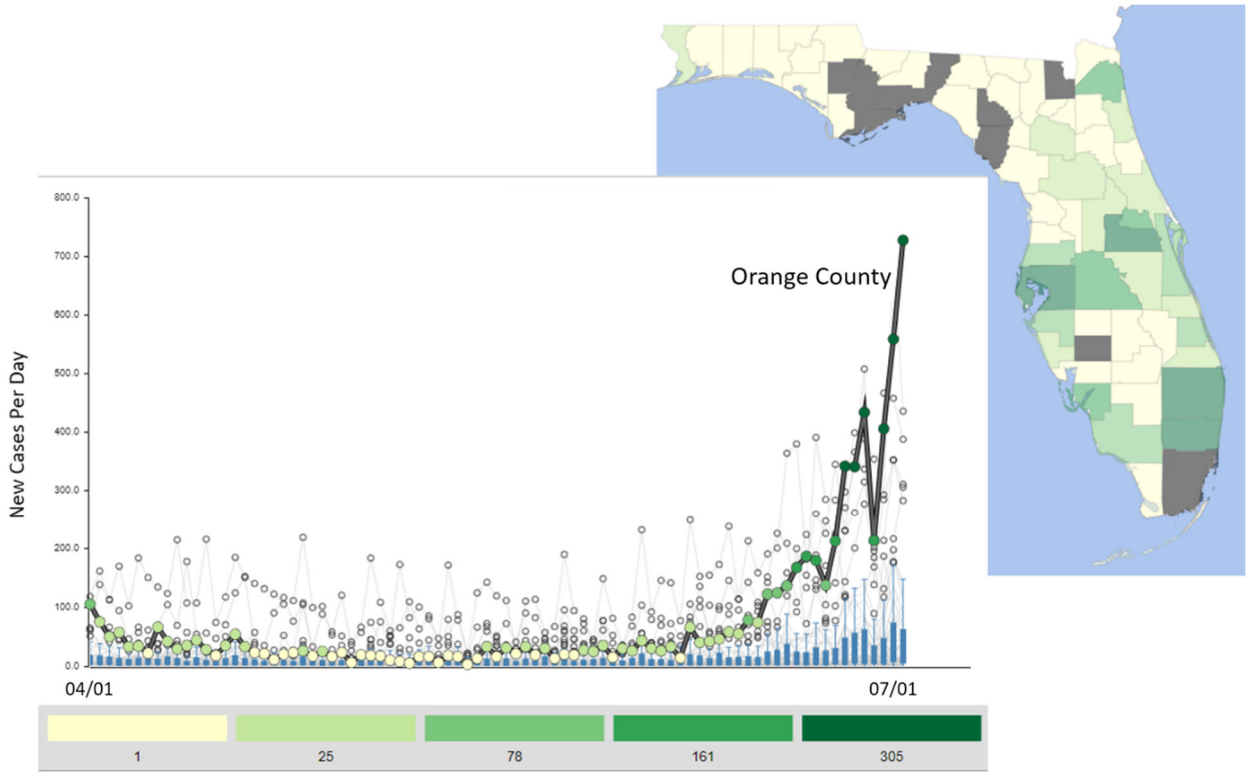

Figure 5. New Cases Per Day using the Time Series With Outliers. The map shows new case rates on June 23, showing new infections are largely concentrated in Central and South Florida

The challenge with these charts is that it's hard to see general trends in across the state as many are buried in the low value areas and the data is fairly noisy. The Cluster by Trend Analytic is used to sort these noisy trends into similar clusters where they can be more reasonably assessed for pattern and covariance structures. This tool takes the data from each county in Florida and all 35 attributes for each county – more than 2000 trend lines total – and groups them the 6 trend line groups in Figure 6. Additionally, the trend each location correlates to can be assigned on the map. In Figure 6, the map shows which counties have the trend line for New Cases by Day. For most of Florida, New Cases by Day falls into three trends. Trend 1 is characterized by steady growth throughout, covering the southern half of the Florida Panhandle. Trend 5 is characterized by explosive growth towards the end of the study period after previous low growth, and is generally found in the northern half of the Florida peninsula and parts of the Florida Panhandle. A third Trend, Trend 3, is characterized by slow and steady growth with a spike around the middle of the study period, and is most commonly seen in parts of the Panhandle not in Trend 5. Overall then, we can see that case growth in Florida began early in the South part of the state and never really was brought under control, most of the Panhandle never experienced explosive growth, and the central part of the state only saw high growth rates late in the study period.

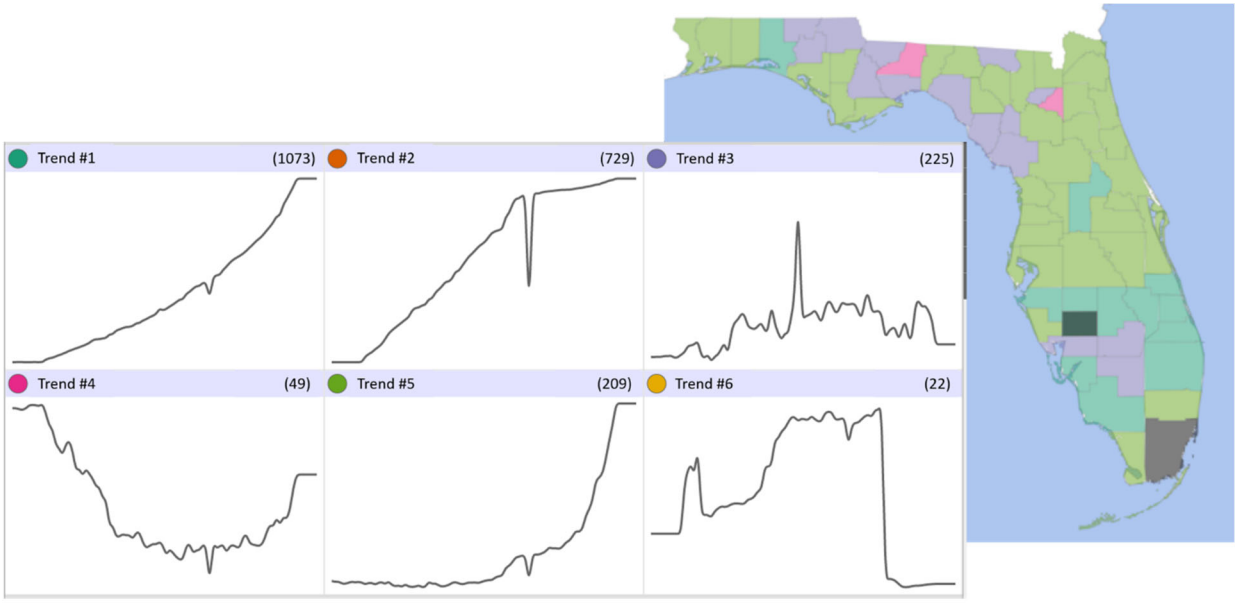

Figure 6. New Cases by Day trends by county are shown in the map based on which trend they are classified into by the Cluster by Trend Analytic.

The analytic Categorize Trends bins trend lines into 10 target groups. Using the attribute Deaths, this Figure 7 shows that Categorize Trends clusters county trends into 4 trends. These 4 trends are not unexpected; Deaths is a cumulative metric so decreasing deaths should not be found in any of the counties. Most counties fell into two trends, the Up and Step Up trend lines. WSTAMP allows further exploration of the strength of these memberships. For example Suwanee county is classifed as a "step up" but also scores very high for linear trend up as indicated by the bar charts in the hover over.

The county groupings based on Death in the map also largely match those found in Figure 6. Counties where deaths constantly increased were in the Peninsula, and counties where deaths peaked and leveled off were more likely to be in the Panhandle. Several counties, also mostly in the Panhandle, showed a flat rate of deaths throughout the pandemic. The one county that shows a spike trend – Monroe County - is an example of a data quality error discussed Section 2.3, where a single death was retracted the following day. Together these suggest that the Florida Peninsula was more impacted by COVID than the Florida Panhandle.

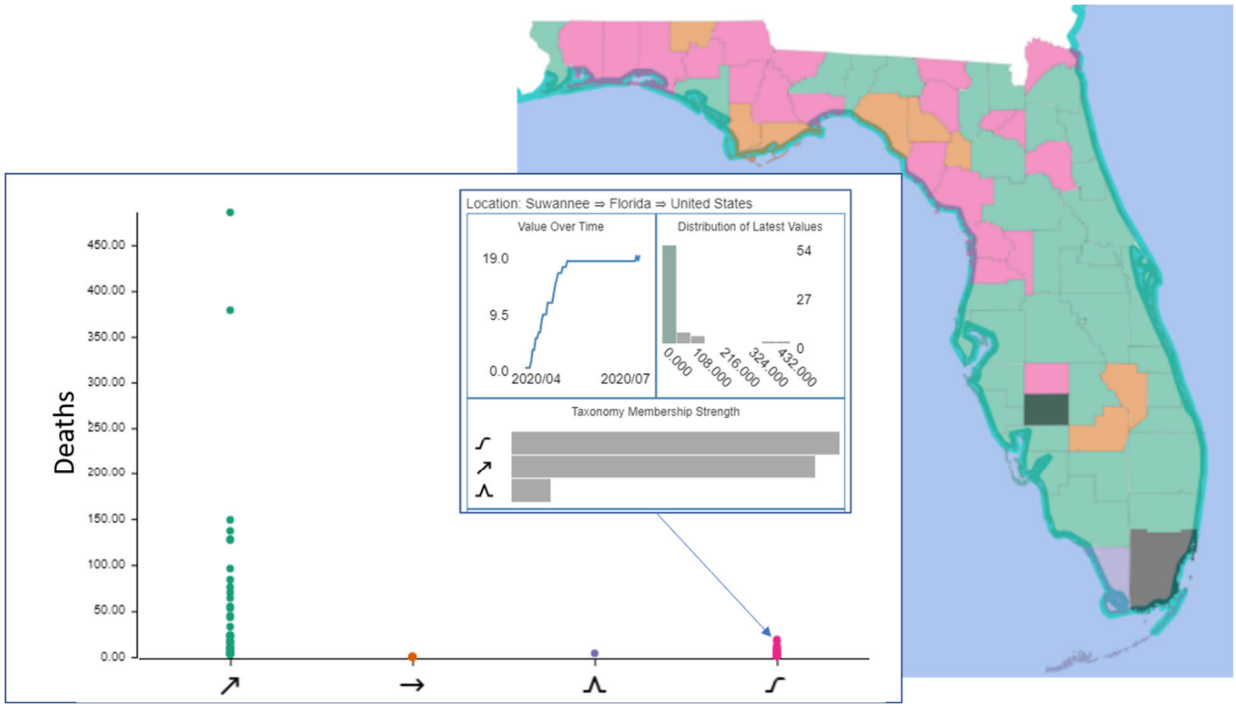

Figure 7. Deaths in the Categorize Trends WSTAMP Analytic. Counties on in the map are colored according to the trend they largely belong to on the right.

The WSTAMP analytic Find Unusual Trends allows analysts to quickly surface likely anomalous values (high or low) and behaviors. Figure 8 shows the results for set of 35 attributes across all Florida counties. WSTAMP bins the outcomes by geography or by attribute and maps the each county by the number of detected anomalies. For three counties – Broward, Hillsborough, and Palm Beach - there were outliers (primarily high outliers) present in every attribute. Across the state as the whole, the trend was that the highest number of unusual trends were in the southern part of the state decreasing northwards throughout state. The exception to this is Duval County in the Northwest corner of the state, which had outliers present in almost every attribute. As with other analytics, this suggests southern Florida experienced a higher proportion of anomalous trends as compared to the rest of the state.

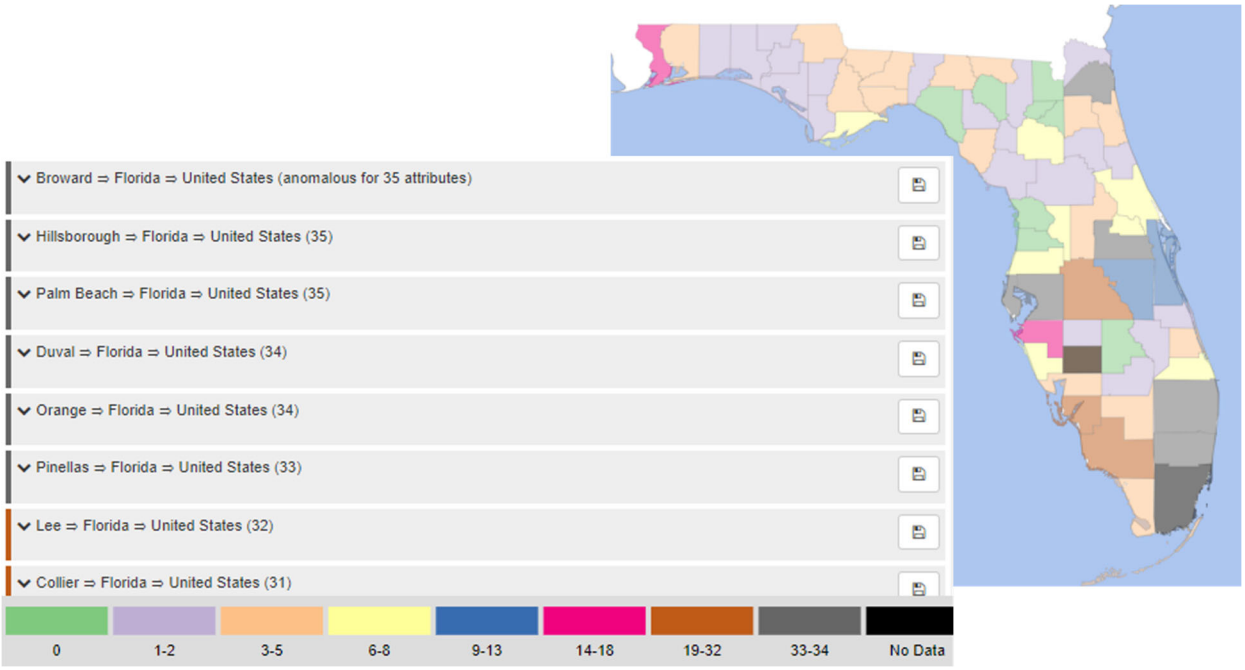

Figure 8. The Find Unusual Trends analytic uncovers likely Florida county anomalies based on magnitude or behavior.

4.2 WITHIN-STATE STATE LEVEL ANALYSIS

Based on the data availability and collection, Ohio provides a good test case for consistent data published at the state level across the pandemic. These attributes are detailed in Table 2, and fall broadly into the categories of cases, deaths, and hospitalizations. Using WSTAMP, we can easily explore each attribute one at a time or consider all 65 at once in a multi-variate analysis.

Quickly applying Cluster by Trend analytic to all 65 Ohio attributes, 4 iconic trends emerge (Figure 9). Of the 65 total attributes, 56 fell in Trend 1 and the remaining 9 attributes are split evenly between the other 3 trends. While most attributes fall into Trend 1, those that fall into trends two-four are also interesting. These trends are largely composed of attributes with limited data in the dataset or small population sizes in Ohio, such as Deaths by Gender where the gender is unknown, or Deaths by Ethnicity and/or Race for American Indians. The attributes within each trend are detailed in Table Alpha Omega and suggest groups of strongly correlated data for deeper statistical analysis. This is also an opportunity for hypothesis generation – what are these two attributes so well correlated? Or why are these two not well correlated? From these new deeper inquiries are possible.

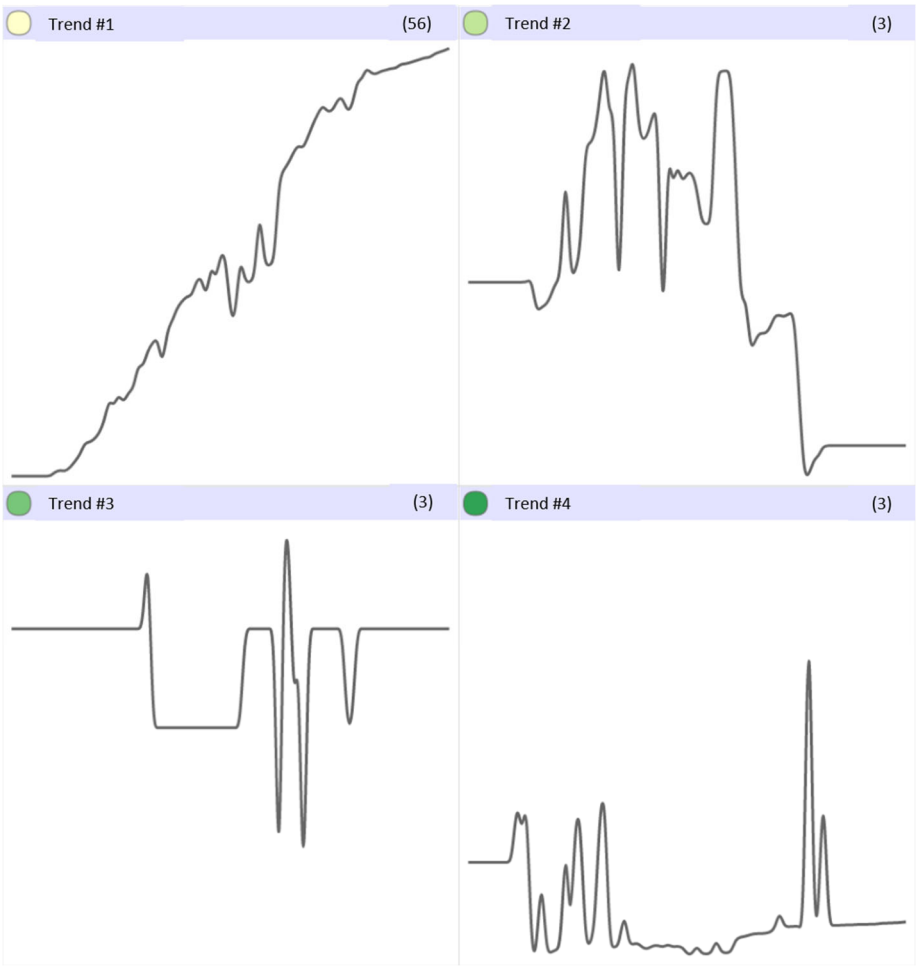

Figure 9. Cluster by Trend Analysis of Ohio Within-State Data

Table 4. Attribute trend groupings

Attribute	Associated Trend
Cases By Age Range | Lower Bound of Estimate	1
Cases By Age Range | NA	1
Cases By Age Range | Upper Bound of Estimate	1
Cases By Age Range, And By Gender | Female	1
Cases By Age Range, And By Gender | Male	1
Cases By Age Range, And By Gender | Unknown	1
Cases By Ethnicity and/or Race | African American or Black	1
Cases By Ethnicity and/or Race | American Indian or Alaskan Native	1
Cases By Ethnicity and/or Race | Asian	1
Cases By Ethnicity and/or Race | Hispanic/Latinx	1
Cases By Ethnicity and/or Race | Native Hawaiian or Pacific Islander	1
Cases By Ethnicity and/or Race | Non Hispanic/Latinx	1
Cases By Ethnicity and/or Race | Other	1

Cases By Ethnicity and/or Race | Two or More Races	1
Cases By Ethnicity and/or Race | Unknown	1
Cases By Ethnicity and/or Race | White or Caucasian	1
Cases By Gender | Female	1
Cases By Gender | Male	1
Cases By Gender | Unknown	1
Cases	1
Counties with a Case	1
Counties with a Death	1
Counties with a Hospitalized Case	1
Deaths By Age Range, And By Gender | Female	1
Deaths By Age Range, And By Gender | Male	1
Deaths By Age Range, And By Gender | Unknown	1
Deaths By Ethnicity and/or Race | African American or Black	1
Deaths By Ethnicity and/or Race | Asian	1
Deaths By Ethnicity and/or Race | Hispanic/Latinx	1
Deaths By Ethnicity and/or Race | Non Hispanic/Latinx	1
Deaths By Ethnicity and/or Race | Other	1
Deaths By Ethnicity and/or Race | White or Caucasian	1
Deaths By Gender | Female	1
Deaths By Gender | Male	1
Deaths With Pending Tests	1
Deaths	1
Hospitalized Case | In ICU	1
Hospitalized Case	1
Hospitalized Cases By Age Range, And By Gender | Female	1
Hospitalized Cases By Age Range, And By Gender | Male	1
Hospitalized Cases By Age Range, And By Gender | Unknown	1
Hospitalized Cases By Ethnicity and/or Race | African American or Black	1
Hospitalized Cases By Ethnicity and/or Race | American Indian or Alaskan Native	1
Hospitalized Cases By Ethnicity and/or Race | Asian	1
Hospitalized Cases By Ethnicity and/or Race | Hispanic/Latinx	1
Hospitalized Cases By Ethnicity and/or Racetive Hawaiian or Pacific Islander	1
Hospitalized Cases By Ethnicity and/or Race | Non Hispanic/Latinx	1
Hospitalized Cases By Ethnicity and/or Race | Other	1
Hospitalized Cases By Ethnicity and/or Race | Refused to Answer	1
Hospitalized Cases By Ethnicity and/or Race | Two or More Races	1
Hospitalized Cases By Ethnicity and/or Race | Unknown	1
Hospitalized Cases By Ethnicity and/or Race | White or Caucasian	1
Hospitalized Cases By Gender | Female	1
Number of People Tested	1
Pending Tests	1

Probable Cases By Ethnicity and/or Race | Two or More Races	1
Deaths By Ethnicity and/or Race	2
Deaths By Ethnicity and/or Race | Refused to Answer	2
Deaths By Ethnicity and/or Race | Unknown	2
Cases By Ethnicity and/or Race | Refused to Answer	3
Deaths By Ethnicity and/or Race | American Indian or Alaskan Native	3
Deaths By Gender | Unknown	3
Deaths By Age Range	4
Deaths By Ethnicity and/or Race | Two or More Races	4
Hospitalized Cases By Age Range	4

In Figure 10, we see the specific attribute trends from Trend 1 represented as thin gray lines and the average of these represented as a blue line. From this we see once again that Cluster by Trend is robust to noisy data and captures a primary behavior (up) in the set of trends belonging to this group. The brief oscillation in May on this graph stands out and perhaps is due to an interruption of data collection across the attributes rather than represent any slowing of the outbreak at that time. More investigation will be necessary.

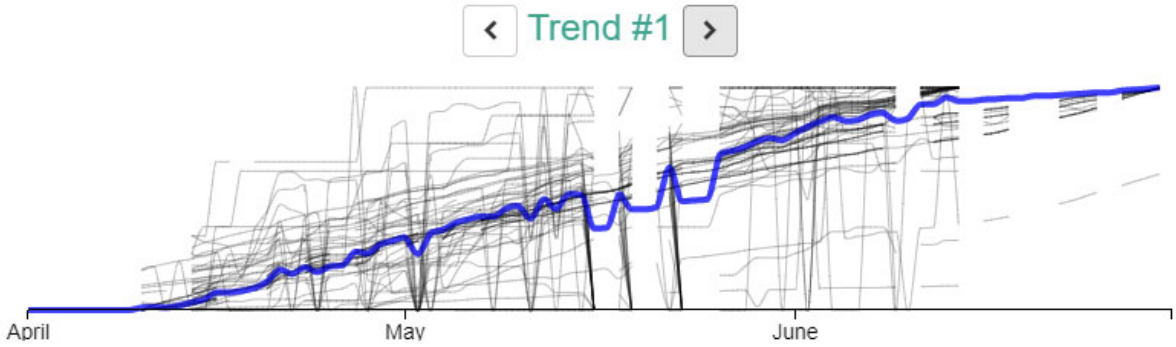

Figure 10. The 56 attribute noisy trends in Trend 1 show a general upward climb

The WSTAMP analytic "Find Trends Similar to Mine" scores each trend by how similar it is to a target attribute of interest. The smaller the similarity score the more similar the trends. In Figure 11, we can use the attribute Hospitalized Case as the match trend line and compare it against the other 64 selected trends in Ohio. As you can see from the graph, Hospitalized Cases trended most similarly to Cases overall, followed by Hospitalized Cases in the ICU and Deaths. Looking further down the graph we see that attributes like "Deaths by Ethnicity and/or Race" are significantly different than Hospitalized cases. This analysis allows one to examine possible correlates to a specific attribute and create new hypothesis about those relationships.

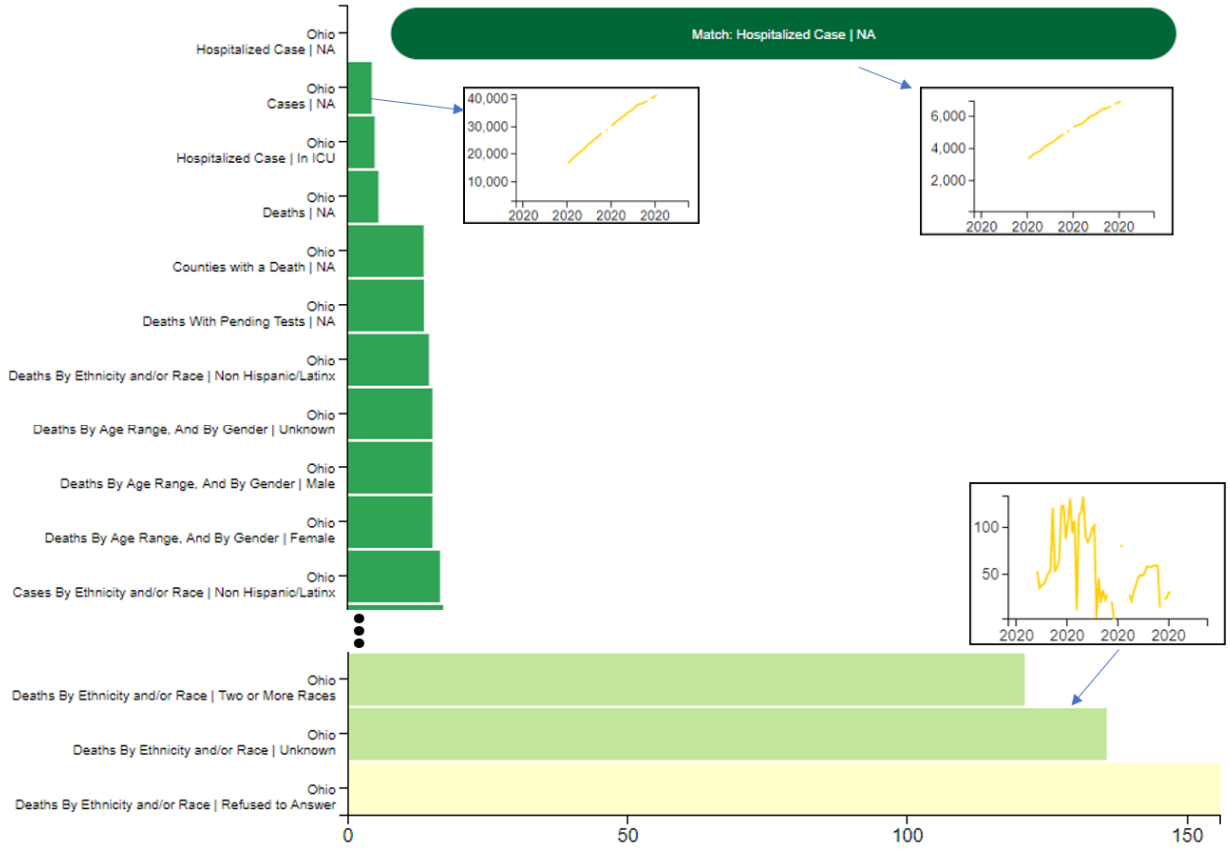

Figure 11. The Find Trends Similar to Mine analytic, abbreviated for readability.

4.3 BETWEEN STATE ANALYSIS

In this use case, our motivating interest is to determine what likely ST correlates exist in the selected attributes over 50 states. As with previous use cases, we begin with an examination of trends using simple time series plots and summary statistics. This cursory analysis revealed occasional QA/QC problems with the data. For example, when looking at the raw trends of state level Total Cases, a few outliers dominate the time series plot (Figure 12). Total cases on May 3rd is particularly noticeable for California. This is likely a parsing error that would need to be either removed or corrected. Other states cause issues with sporadic outliers including New York and Oklahoma.

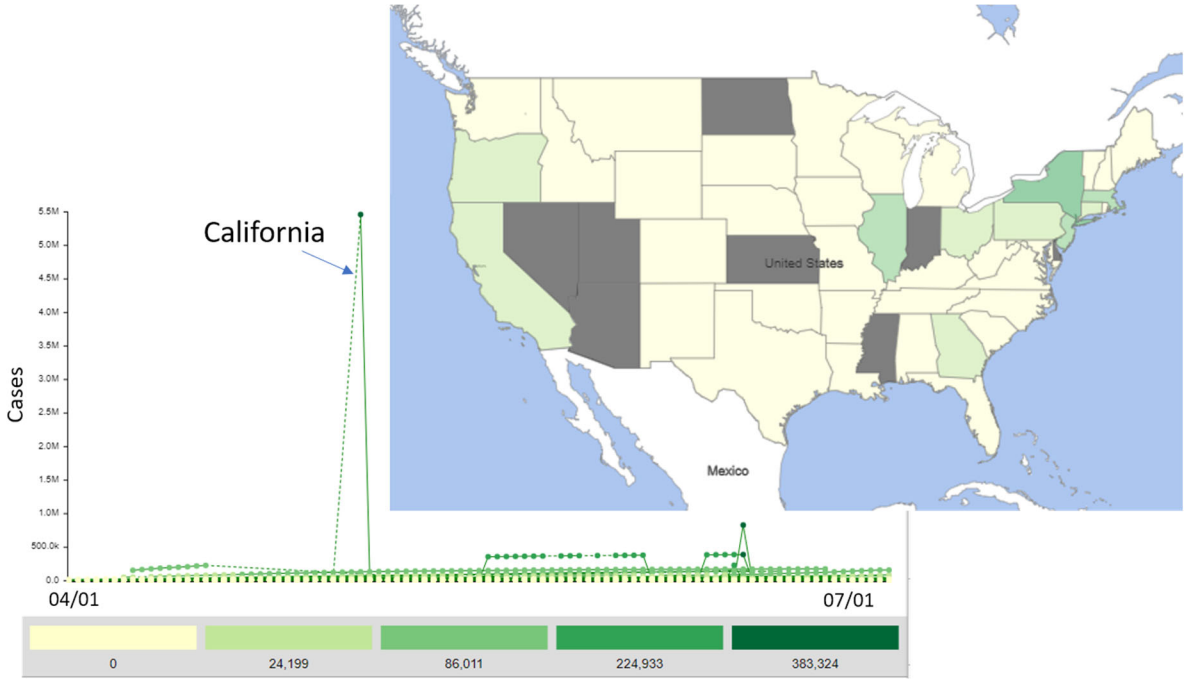

Figure 12. Time Series of State Level Cases, including outliers

When the states with the most egregious outliers (California, New York, Oklahoma, West Virginia) are temporally removed, more cogent trends emerge. Data consistency differences between states are evident. Some states like New Jersey, the upward and then flattening highest valued trend in the chart, provided consistent data access and measures and their trends can be reconstructed with relative confidence. A hand full of states, however, show the same spikiness that the previously removed states demonstrated, albeit with less extreme outliers. More investigation into the curation of those data is required.

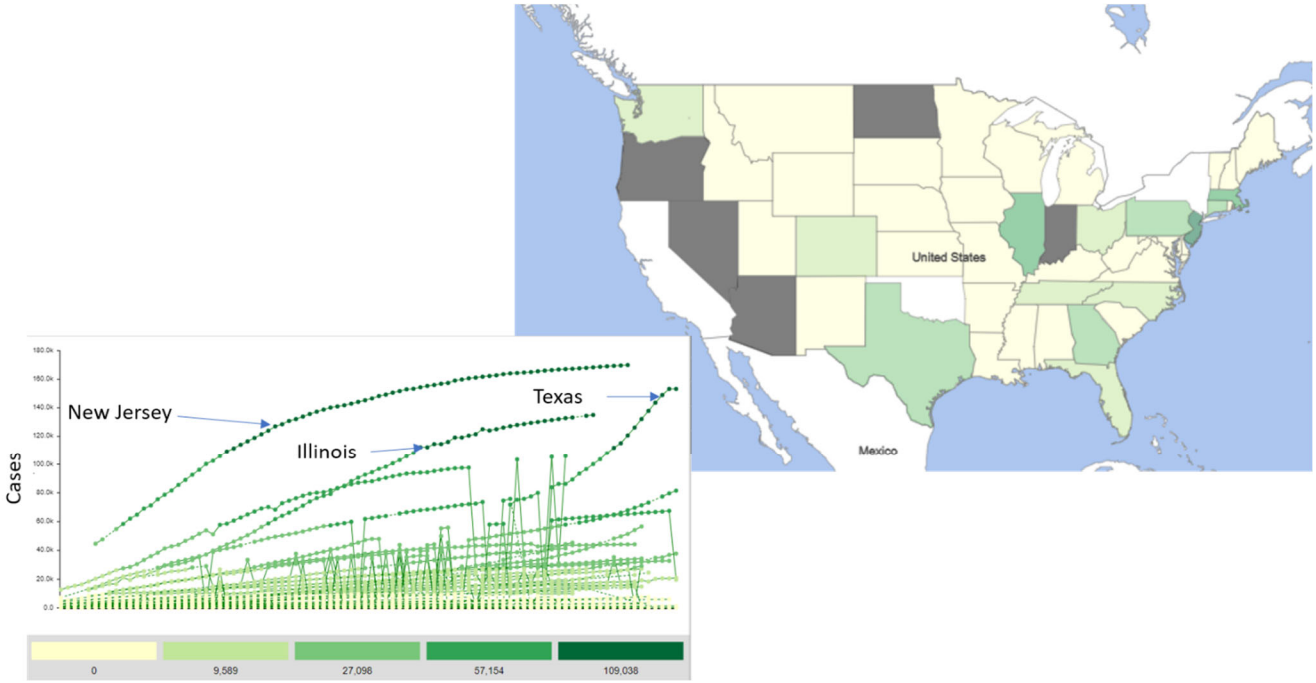

Figure 13 – Time Series of State Level Cases, with Largest Outliers Removed

When switching the chart to view the trends as boxplots, a steady increase in variance across states becomes clear. This overall trend is indicated by the increasing heights of the blue boxes representing the 25th (lower) and 75th (upper) percentiles.

.

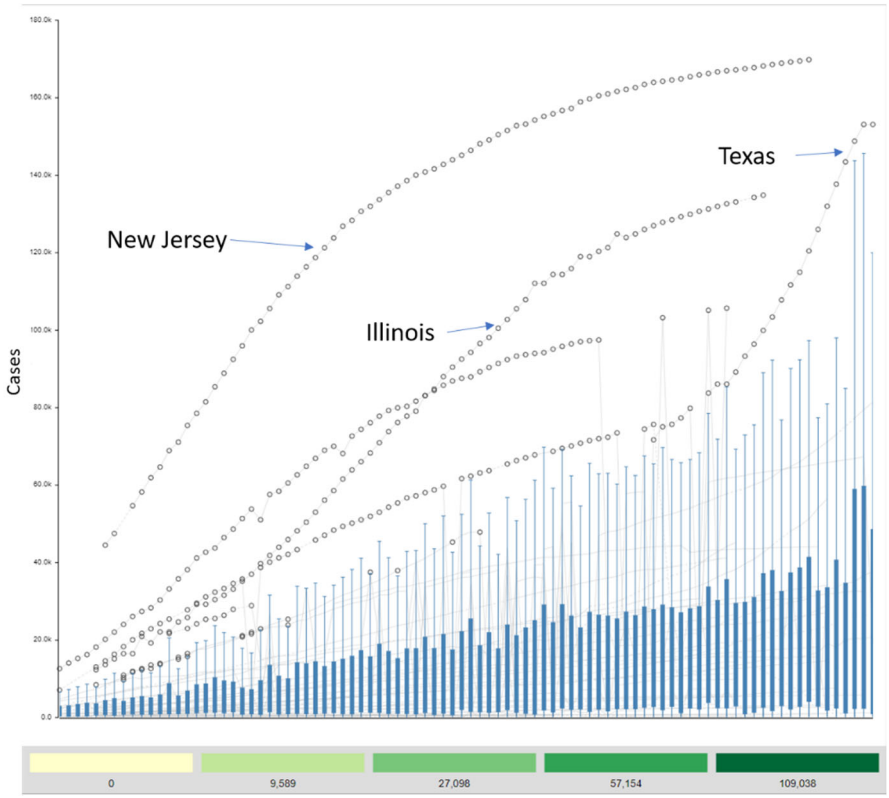

Figure 14 – Boxplots of State Level Cases, with Largest Outliers Removed

Based on the two figures above, the dominate behavior in total cases across states is generally increasing. This masks that some trends are which are concave or convex, many too entangled to be clear. Having a large picture understanding of the general kinds temporal behaviors in a dataset is very helpful when looking at more than a couple dozen trends and essential when there are hundreds or thousands as this dataset contains.

Applying Cluster by Trend Analytic to State Level Cases, 6 characteristic primary behaviors appear. Trends 1, 3, and 4 all show various evolutions of an upward trend. Trends 2 and 5 capture the more chaotic behavior we saw previously, while Trend 6 is only one state, Wisconsin. Whose trend is so distinctive that it was placed in its own group.

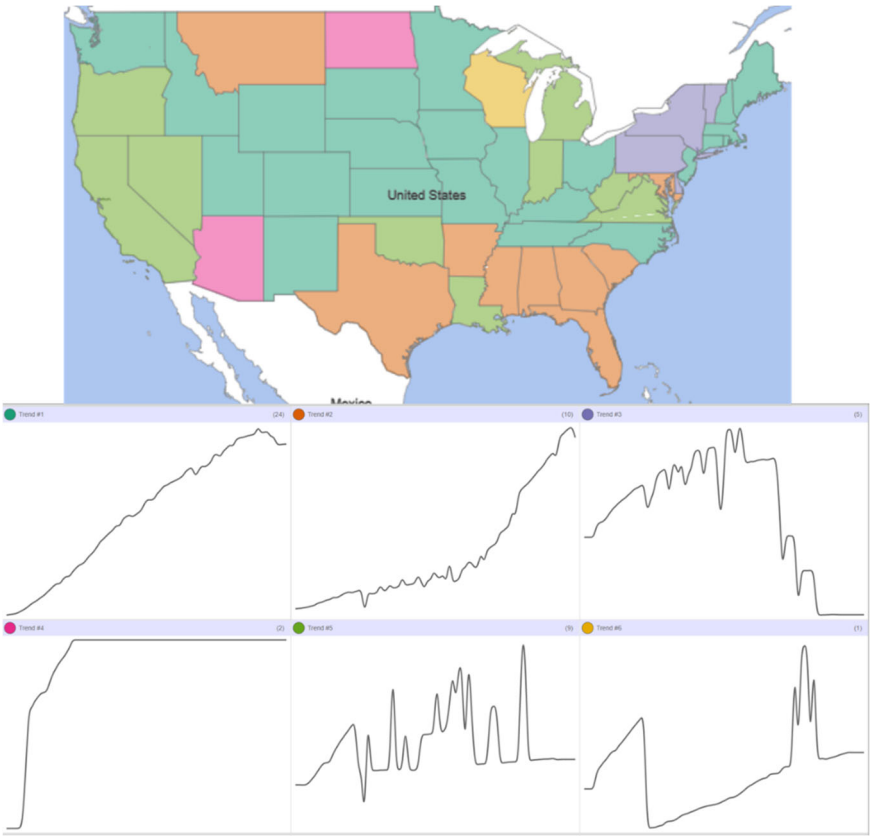

Figure 15 – The 6 Most Characteristic Behaviors of State Level Case Trends

This analysis technique is not limited to grouping only one attribute at a time. We can include several attributes at once and quickly find out the most characteristic behaviors across all locations and attributes. This allows us to quickly see if some attributes tend to co-evolve with one another.

Turning to the primary question of detecting correlates in these limited, noisy data we rely on Cluster by Trend to group approximately 876 state-attribute pairs into 6 manageable groups (Figure 16). Primary behaviors include decreasing (Trend 1, Trend 5) , increasing (Trend 3 and Trend 4), and unimodal (Trend 2 and Trend 6).

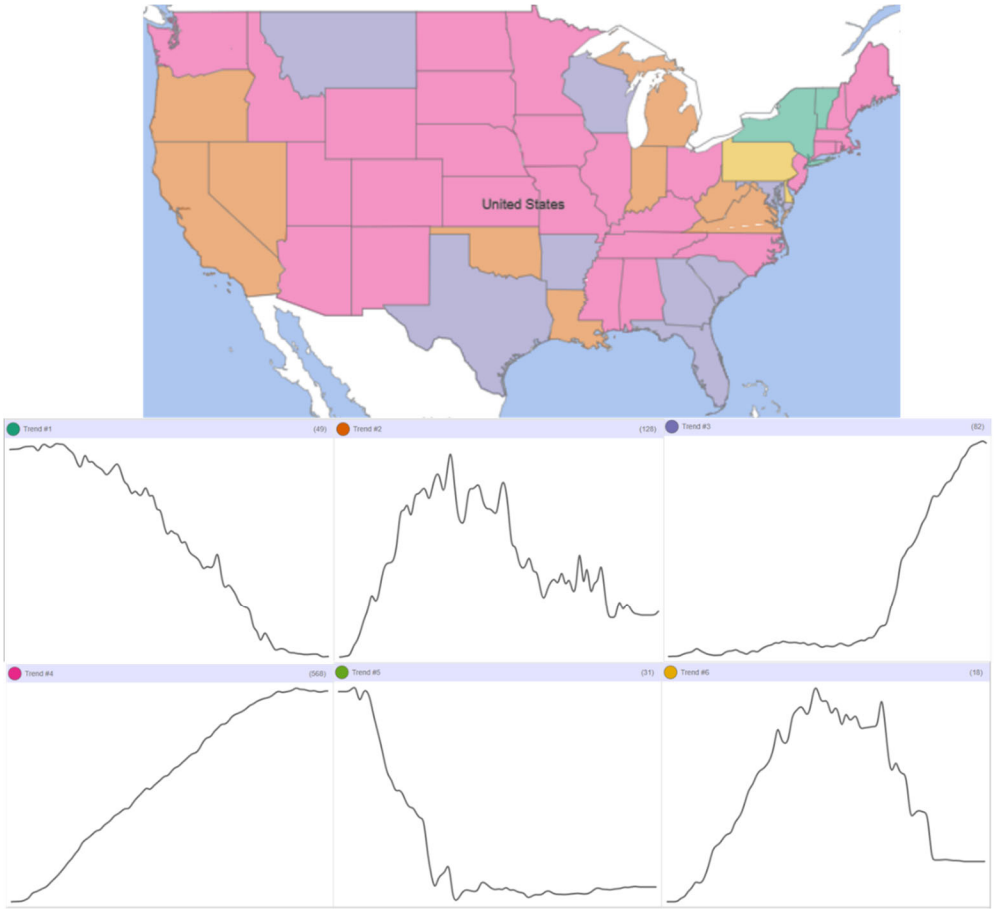

Figure 16 – The 6 Most Characteristic Behaviors Across All State Level Trends

Using WSTAMP it is possible to see the spatial distribution of trends for any specific attribute. In Figure 16, the US map shows the distribution of Deaths trends. Most states are increasing (pink/Trend 4). A handful (e.g. Alaska, Nevada, Illinois, Pennsylvania, Virginia, and Michigan) which surged and then decline (yellow/Trend 6). Standouts include South Carolina which is initially flat and then rapidly increasing (purple/Trend 3).

Post processing these Cluster By Trend results into a tabulation of concurrent trend memberships by attribute we can begin to surface those attributes that are most commonly colocated together in a trend and better understand likely correlates that exist across attributes. In Figure 17 we see that temporal patterns for Cases by Ethnicity and/or Race | Asian only have concurrent trend category memberships with Deaths By Age – Range in only 8 US States. The correlation here can be expected to be low. However, trends for number of people tested and deaths is concurrent in 36 states. Using this table analysts can see likely correlates at national scale based on trend membership categories. This signals further investigation into the specific trends or trend combinations.

	Cases	Cases By Age Range	Cases By Ethnicity and/or Race | African American or Black	Cases By Ethnicity and/or Race | Asian	Cases By Ethnicity and/or Race | Hispanic/Latinx	Cases By Ethnicity and/or Race | Other	Cases By Ethnicity and/or Race | Unknown	Cases By Ethnicity and/or Race | White or Caucasian	Cases By Gender | Female	Cases By Gender | Male	Cases By Gender | Unknown	Deaths	Deaths By Age Range	Deaths By Ethnicity and/or Race | African American or Black	Deaths By Ethnicity and/or Race | Hispanic/Latinx	Deaths By Gender | Female	Deaths By Gender | Male	Hospitalized Case	Hospitalized Case | In ICU	Number of Negative Tests	Number of People Tested	Recovered Cases
Cases	51	21	15	11	22	17	16	15	27	25	15	33	12	8	12	19	13	18	10	19	25	13
Cases By Age Range	21	50	15	13	23	19	15	19	26	27	14	27	9	10	14	20	12	16	7	19	28	16
Cases By Ethnicity and/or Race | African American or Black	15	15	39	21	24	27	17	28	24	23	13	19	11	21	16	19	17	19	13	12	19	10
Cases By Ethnicity and/or Race | Asian	11	13	21	29	20	22	16	23	17	18	11	16	8	15	15	14	10	13	7	11	16	7
Cases By Ethnicity and/or Race | Hispanic/Latinx	22	23	24	20	40	27	24	26	27	28	13	28	9	16	21	22	14	17	14	18	28	13
Cases By Ethnicity and/or Race | Other	17	19	27	22	27	37	20	28	25	24	16	23	8	16	17	20	14	18	11	15	23	14
Cases By Ethnicity and/or Race | Unknown	16	15	17	16	24	20	36	21	21	20	11	20	8	13	17	19	12	15	10	15	20	10
Cases By Ethnicity and/or Race | White or Caucasian	15	19	28	23	26	28	21	39	28	28	14	23	7	19	17	17	13	16	10	13	24	10
Cases By Gender | Female	27	26	24	17	27	25	21	28	49	42	19	27	11	16	20	24	21	19	11	19	28	12
Cases By Gender | Male	25	27	23	18	28	24	20	28	42	49	20	27	10	14	20	24	19	18	11	19	28	11
Cases By Gender | Unknown	15	14	13	11	13	16	11	14	19	20	40	16	10	10	12	13	13	12	8	6	16	6
Deaths	33	27	19	16	28	23	20	23	27	27	16	51	10	11	17	25	17	20	9	23	36	19
Deaths By Age Range	12	9	11	8	9	8	8	7	11	10	10	10	42	11	8	8	7	8	9	3	7	5
Deaths By Ethnicity and/or Race | African American or Black	8	10	21	15	16	16	13	19	16	14	10	11	11	33	16	14	13	14	9	8	11	6
Deaths By Ethnicity and/or Race | Hispanic/Latinx	12	14	16	15	21	17	17	17	20	20	12	17	8	16	34	18	14	10	6	12	18	6
Deaths By Gender | Female	19	20	19	14	22	20	19	17	24	24	13	25	8	14	18	36	26	19	9	18	24	11
Deaths By Gender | Male	13	12	17	10	14	14	12	13	21	19	13	17	7	13	14	26	35	18	7	12	15	7
Hospitalized Case	18	16	19	13	17	18	15	16	19	18	12	20	8	14	10	19	18	48	12	15	19	12
Hospitalized Case | In ICU	10	7	13	7	14	11	10	10	11	11	8	9	9	9	6	9	7	12	28	4	11	5
Number of Negative Tests	19	19	12	11	18	15	15	13	19	19	6	23	3	8	12	18	12	15	4	31	23	15
Number of People Tested	25	28	19	16	28	23	20	24	28	28	16	36	7	11	18	24	15	19	11	23	49	18
Recovered Cases	13	16	10	7	13	14	10	10	12	11	6	19	5	6	6	11	7	12	5	15	18	30

Figure 17. Cross tabulation of concurrent trends memberships by attribute.

5. SUMMARY

This report continues the examination of space-time public health data curated during the opening months of the COVID-19 pandemic. The report focuses on examination of the data from a strictly data science perspective, using the WSTAMP tool to assess the data for quality, trends, patterns, and anomalies among presently harmonized attributes. The study explored the data across three scales of analysis: within-state county level, within-state state level, and between-state. Florida and Ohio, respectively, served as exemplar use cases for within-state county and within-state state level analyses due to their relatively rich and consistent set of attributes at those scales. These two examinations shed light on how a state level analysis would be conducted and what type of correlates can emerge from the analysis for further investigation. Similarly, the national between-state analysis showed a range of results for trend and anomalies. The following are key findings of the report.

- Lack of national strategy for reporting public health information led to a significant effort in harmonizing, cleaning, and relinking data as the pandemic surged. A national data strategy for pandemic data would significantly improve the ability to track, infer, and learn from the spread of the disease.

- Harmonized data, as a mitigation for absent data standards, allowed for within-state and between-state analytics and will play a major role in any future comparative studies using this data.

- Two opportunities for QA/QC improvements are: a) investigation and possible correction of a limited number of suspect point data collects and b) a reassessment of a handful of harmonization decisions that led to unusual temporal artifacts.

- Twenty two harmonized attributes had sufficient space-time completeness to support between-state covariate analysis on a national scale.

- Cross tabulation of attribute trend behaviors indicate a range of likely correlations among harmonized variables on a national scale. These may serve well in future inference or explanatory analytics but more statistical evaluation will be needed.

- The number of harmonized attributes at the state and county level for within-state analysis will vary from state to state, but most states have numerous useable attributes.

- In Florida, the "no longer monitored" attribute suggests that the situation was improving in June, but the "new cases" attribute shows that the situation was deteriorating at that time, suggesting the data was no longer being reported correctly.

- Trend data across the state of Florida shows that southern Florida was disproportionally impacted by COVID-19, and the Florida panhandle was less impacted.

- Four to Six iconic trends were adequate to sort very noisy trend data into useable categorizes at all three scales. These trends were mostly variations of "trending up" although some included "smile" or down patterns as well.

Recommended next steps are:

- Investigate potential data errors and problematic harmonization.

- Some WSTAMP capabilities were designed for annual data (the majority of WSTAMP core database) and could not be applied to daily data directly. Examples include the completeness analytic and a variety of GUI interaction, such as time bar steps.

- Provide a quality score for each attribute that indicated its overall completeness and coverage.

- Create a data portal for wider data access by scientists and public health officials for evaluation against a particular set of questions they may pose.

6. REFERENCES

1. Piburn, J, **R.N. Stewart**, A Myers, A Sorokine, D Axley, D Anderson, J Burdette, C Biddle, A Hohl, R Eberle, J Kaufman, and A Morton (2017a), *The World Spatiotemporal Analytics and Mapping Project (WSTAMP): Further Progress in Discovering, Exploring, and Mapping Spatiotemporal Patterns Across the World's Largest Open Source Data Sets*, ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, Vol 4, pp 199-205.

2. Piburn, J, A Morton, and **R.N. Stewart** (2017b). *Attribute Portfolio Distance: A Dynamic Time Warping based approach to comparing and detecting common spatiotemporal patterns among multi-attribute data portfolios*. Advances in Geocomputation: Geocomputation 2015--The 13th International Conference. D. A. Griffith, Y. Chun and D. J. Dean. Cham, Springer International Publishing: 197-205.

3. **Stewart, R.N.**, J Piburn, A Sorokine, A Myers, and D White (2015) *World Spatiotemporal Analytics and Mapping Project (WSTAMP): Discovering, Exploring, and Mapping Spatiotemporal Patterns across the World's Largest Open Source Geographic Data Sets*, ISPRS Annals of Photogrammetry, Remote Sensing, and Spatial Information Sciences. Volume II-4W2.