Big Data
Volume 10, Number 2, 2022

DOI: 10.1089/big.2021.0023

ORIGINAL ARTICLE

Visual Analytics for the Coronavirus COVID-19 Pandemic

Christopher G. Healey,[1,2,*] Susan J. Simmons,[2] Chandra Manivannan,[1] and Yoonchul Ro[1]

Abstract

The coronavirus disease COVID-19 was first reported in Wuhan, China, on December 31, 2019. The disease has since spread throughout the world, affecting 227.2 million individuals and resulting in 4,672,629 deaths as of September 9, 2021, according to the Johns Hopkins University Center for Systems Science and Engineering. Numerous sources track and report information on the disease, including Johns Hopkins itself, with its well-known Novel Coronavirus Dashboard. We were also interested in providing information on the pandemic. However, rather than duplicating existing resources, we focused on integrating sophisticated data analytics and visualization for region-to-region comparison, trend prediction, and testing and vaccination analysis. Our high-level goal is to provide visualizations of predictive analytics that offer policymakers and the general public insight into the current pandemic state and how it may progress into the future. Data are visualized using a web-based jQuery+Tableau dashboard. The dashboard allows both novice viewers and domain experts to gain useful insights into COVID-19's current and predicted future state for different countries and regions of interest throughout the world.

Keywords: coronavirus; COVID-19; data analytics; visualization

Introduction

At the end of 2019, China alerted the World Health Organization (WHO) of an outbreak of a novel strain of coronavirus in the city of Wuhan.[1] The WHO published information on the outbreak on January 5. Although we use the common name COVID-19 here, the International Committee on Taxonomy of Viruses classified the disease as "severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)" on February 11, the same day the WHO announced COVID-19 as the name of the new disease. On January 13, the virus first appeared outside China, in Thailand.

China shared the genetic sequence of COVID-19 on February 12. By March 11, the virus had spread to 114 countries and caused 118,000 confirmed cases. At this point, the WHO declared COVID-19 a global pandemic. The first case of COVID-19 within the United States was reported in Snohomish County, Washington, on January 20, 2020.[2] As of September 9, 2021, the United States reported ~41.8 million confirmed cases of COVID-19, resulting in 670,128 deaths. Worldwide, COVID-19 has sickened 227.2 million individuals, with 4,672,629 succumbing to the disease.

As COVID-19 expanded, numerous agencies began publishing case counts both in text and visual format to track the spread of the disease. The Johns Hopkins University (JHU) Novel Coronavirus (COVID-19) Cases Data dashboard[3] provides up-to-date statistics on confirmed, recovered, and fatality cases, as well as detailed information on location and other pertinent facts. We use a modified version of JHUs data from data.world[4] as one source for our dashboard.

Another similar effort is CovidNet, a real-time COVID tracking project that presents information on both temporal and geographic trends.[5] CovidNet focuses on reliable timely data sources and provides a number of visualizations, including temporal epidemic curves and cross-region comparison based on absolute case numbers.

[1]*Department of Computer Science, North Carolina State University, Raleigh, North Carolina, USA.*
[2]*Institute for Advanced Analytics, North Carolina State University, Raleigh, North Carolina, USA.*

*Address correspondence to: *Christopher G. Healey, Department of Computer Science, North Carolina State University, Raleigh, 890 Oval Drive, Raleigh, NC 27695-8206, USA,* E-mail: healey@ncsu.edu

Other online dashboards also exist: 1Point3Acres (case, test, and vaccination visualization),[6] University of California, Los Angeles (UCLA) Combating COVID-19 (peak date estimates),[7] Los Alamos National Laboratory COVID-19 Dashboard (case visualization),[8] COVID Analytics (effects of policies on cases),[9] and COVID-19 Modeling (case and hospital bed usage and future week predictions).[10] Aggregators are also collecting URLs of COVID-related websites, for example, the Shaman Group's COVID-19 Findings and Simulations list.[11]

Existing dashboards that track the current state of COVID-19 and predict future case totals for a specific region are useful, but they are mainly focused on data reporting. We did not want to duplicate this functionality since numerous sources exist with excellent implementations. Instead, we focused on performing sequence-based pattern matching, time series analysis, regression, and unsupervised machine learning before the application of perceptually optimal visualization strategies to augment existing information with predictive analytics. As noted in the Abstract, our high-level goals are as follows:

(1) To apply predictive analytics to raw COVID-19 data to identify significant events related to region* comparison, case curve "bend" estimates, case trends, and testing, positivity, and vaccine analysis.
(2) To present results using visualizations that are accessible to practitioners, policymakers, and the general public at large.

To achieve these goals, we implemented a dashboard with the following features:

- similarity of a target region's case curves (either fatalities or confirmed cases) to all other regions reporting data, calculated with dynamic time warping;
- estimated dates and maximum case counts for when each region's curve will "bend," and its rate of acceleration will begin to slow, calculated with four-parameter logistic regression;
- week-over-week trend graphs for a target region's case curves, calculated by comparing the linear regression line for adjacent weeks' case totals for direction and statistical significance;
- time series predictions for future estimates of target attributes such as fatalities based on existing data;
- a web-based dashboard written with jQuery and Tableau† to visualize our results (Fig. 1) as well as testing totals, case totals, and case maps to provide context for the current state of the pandemic, without having to switch to a different environment; and
- visual representations selected based on our long-standing knowledge of human visual perception and its appropriate use in visualization.

This offers a window into current and potential future events for the COVID-19 pandemic. Critically, the presentation is specifically designed for interested viewers in terms of complexity: standard charts, graphs, and maps are employed; accessibility: the dashboard is web-based to provide convenient access and dissemination of results; and visual legibility: perceptual representations are used that harness the strengths and avoid the weakness of the low-level human visual system.[12]

Although we focus on analytics and visualization for the COVID pandemic, the underlying approaches are not limited to this domain. The same methodologies could be applied to issues such as vaccine testing and distribution, results from disease mitigation strategies for different segments of the population (e.g., divided by ethnicity, age, gender, socioeconomic status), or how sociopolitical decisions impact disease spread and effect. Given available data, lessons learned from our current analytics and visualization dashboards can be quickly repurposed to new situations.

Visual Analytics for Epidemiology

Analytics and visualization have been applied extensively in the area of disease analysis and epidemiology. COVID-19 is only the most recent disease to undergo this type of investigation and presentation. Many researchers consider John Snow's cholera map of London as the first scientific example of analytics and visualization applied to disease investigation.[13,14] Carroll et al. summarized articles in epidemiological analytics from January 1980 through June 2013, focusing on geographic information systems, molecular epidemiology, and social network analysis.[15]

Visualizing epidemiology data for investigation is a common theme throughout epidemiological dashboards.

*We use the term *region* to refer to geographic regions provided in the JHU data set. This includes countries (e.g., France and Russia) and states or provinces when available (e.g., New York, United States; or Hubei, China).

†https://go.ncsu.edu/iaa-covid-viz

FIG. 1. COVID visualizations: **(a)** trend graphs of fatalities (top) and confirmed cases (bottom) in the United States; **(b)** normalized region–region similarities for U.S. fatalities (top) and confirmed cases (bottom) versus other regions, to compare fatality and confirmed case time sequences between countries; **(c)** estimated fatality (top) and confirmed case (bottom) bend dates to estimate when a country's fatality or confirmed case curve acceleration will slow.

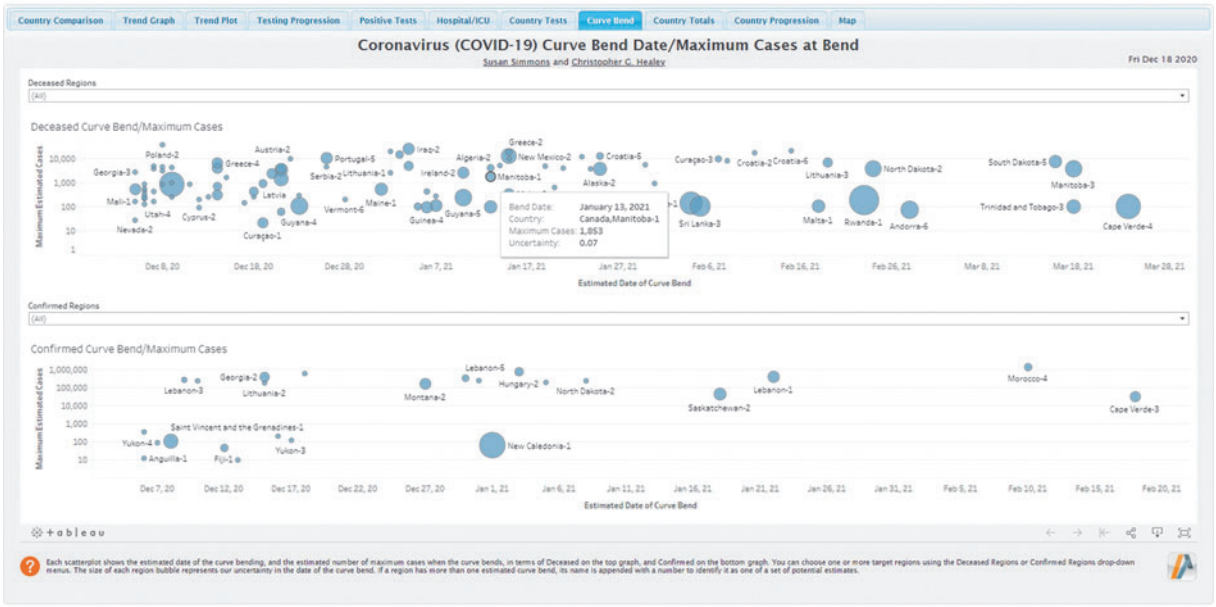

Fig. 1. (Continued).

Epinome presents coordinated views of cases, case histograms, and choropleth maps that allow practitioners to track and evaluate early disease outbreaks.[16] Hamid et al.[17] developed a dashboard that uses the WHOs global FluMART database to visualize both seasonal and novel influenza virus time series plots. Lee et al.[18] built an R+Shiny dashboard to compare risk-standardized mortality rates for sepsis patients in different states in the United States. In recent months, numerous COVID-19 dashboards have been published online, including those by JHU,[3] Tableau,[19] and SAS Institute.[20]

Although the visualization dashboards discussed above offer excellent surveillance capabilities, most of them do not perform analytics to derive new results and insights from their source data sets. Examples of analytics in epidemiology do exist. One study focused on the lack of timely information dissemination related to dengue fever.[21] Researchers applied text analytics to English-language newspaper articles on dengue fever in India to estimate disease cases. This highlighted increases, peaks, and decreases in annual dengue outbreaks in near real time. A dashboard was built using the results to allow practitioners to investigate different aspects of the disease.

To validate their approach, the same technique was applied to newspaper articles on flu in the United States. Results were compared with the Centers for Disease Control and Prevention (CDC) data, showing an 85% accuracy in the estimates. This suggests that text analytics can be used to track diseases through local or social media articles in areas where timely data dissemination is unavailable.

Despite these successes, care must be taken to ensure that analytic strategies perform well as disease characteristics change over time. One well-known example of this was Google Flu Trends.[22] Google mined web search data to predict peak flu levels in the United States. Results pointed to an overwhelming success, producing a mean correlation of 0.90 compared with the CDC data. In subsequent years, however, peak levels were significantly under- or overestimated.[23] Although Google has not publicly commented, researchers believe that pandemics such as H1N1 swine flu in 2009 and the severe flu season in 2013 led to search patterns different from those used to train Google's algorithms.

Rather than considering this a failure, it identified how models must be improved to deal with previously unanticipated information. For example, ARGO (AutoRegression with GOogle search data) extends the work of Google Flu Trends to better predict the pattern of annual flu outbreaks using a self-correcting autoregressive technique.[24]

Region–Region Similarity

One question many people ask is: "How similar is the outbreak in my region to other regions worldwide?" Apart from general interest, an ordered list of regions similar to a target region can offer important benefit. For domain experts, it can highlight regions that may offer clues about how and why the pandemic varies over time in a way that is similar to the target region. Mistakes, mitigation strategies, and policy decisions made by regions with similar pandemic outbreak patterns can offer important clues on how to most effectively fight the pandemic in the target region. For a more general audience, understanding which other regions are similar to a person's home region can highlight where the pandemic is expanding similarly, and where it is expanding in a different manner.

From a high-level view, region–region similarity can be seen as comparing the times series of case totals between pairs of regions. Due to the difference in the onset of the disease around the world, we applied dynamic time warping (DTW),[25] a well-known and robust algorithm for this operation. Although other approaches exist (e.g., pairwise alignment for DNA sequences or progressive alignment construction for multiple sequences), DTW is still considered the most general and computationally tractable approach for sequence pairs.

Consider a target time series $S_1 = \{s_{1,1}, \ldots, s_{1,n}\}$ and a candidate time series $S_2 = \{s_{2,1}, \ldots, s_{2,m}\}$. DTW defines an optimal mapping $s_{2,j} \rightarrow s_{1,i}$, $1 \leq j \leq m$, $1 \leq i \leq n$ between points on the two time series with the minimum "cost," where cost is defined as $\sum_{\text{map}} |s_{2,j} - s_{1,i}|$, the sum of the absolute difference in the values for each point-to-point mapping. DTWs optimal mapping must satisfy four constraints.

(1) Every index in S_1 must match one or more indices in S_2.
(2) $s_{2,1}$ must match $s_{1,1}$.
(3) $s_{2,m}$ must match $s_{1,n}$.
(4) Mappings from S_1 must increase monotonically through S_2.

In terms of implementation, a naive approach to calculating all possible DTW matching costs is expensive but simple, running in $O(n^2)$ quadratic time for two time series with a maximum length of n. Pseudocode for this algorithm is readily available[26] (Algorithm 1). Once all DTW costs are calculated, the cost matrix can be walked from left to right, choosing each column's minimum cost. Rows correspond to points in s_1 and columns to points in s_2. Walking occurs right one step or right and down one step to guarantee a monotonic increase in mappings.

Modern implementations use approximation to significantly reduce computational cost, for example, FastDTW.[27] Here, a DTW solution is recursively refined from an initial coarse estimation. FastDTW showed large improvements in accuracy and runs in $O(n)$ linear time and space complexity for two time series with a maximum length of n, making it the current standard approach for computing DTW solutions on large time series.

We use DTW to match case time series. Figure 2 shows fatality curves for the United States (top), Brazil (middle), and the United Kingdom (bottom). Visually, the U.S. and Brazilian curves appear more similar, compared with the U.S.–U.K. curves. This matches the similarity percentages DTW provides: 99% and 24% for the United States matched to Brazil and the United Kingdom, respectively. DTW costs for each region are normalized to a range of 0 … 1 as follows, with the United States used as an example target region.

1. Compute the DTW cost c_j between the United States and every other region.

Algorithm 1: Calculate DTW for time series s_1 and s_2

Result: The minimum moves needed to align two time series curves
DTW ← *array*[0..*n*, 0..*m*]

```
for i = 1 to n do
    for j = 1 to m do
        DTW [i, j] ← ∞
    end
end
DTW [0, 0] ← 0

for i = 1 to n do
    for j = 1 to m do
        cost ← | s2[i] − s1[j] |
        DTW [i, j] ← cost + min(DTW [i − 1, j], DTW [i, j − 1], DTW [i − 1, j − 1])
    end
end
return DTW [n, m]
```

2. Identify minimum and maximum costs $c_{\min}$ and $c_{\max}$ to define cost range $c_r = c_{\max} - c_{\min}$.
3. Use $c_{\min}$ and c_r to calculate a normalized similarity for every U.S.–region cost c_j as $sim_j = 1 - \left(\left(c_j - c_{\min}\right)/c_r\right)$.

Note that because of normalization, DTW costs are not symmetric. For example, the U.S.–Brazil confirmed case similarity is 99%, but the Brazil–U.S. similarity is only 15%. This is because Brazil's confirmed case curve is more similar to India and Russia versus the United States. An annotated line chart is well suited to visualize these data since most viewers recognize it.

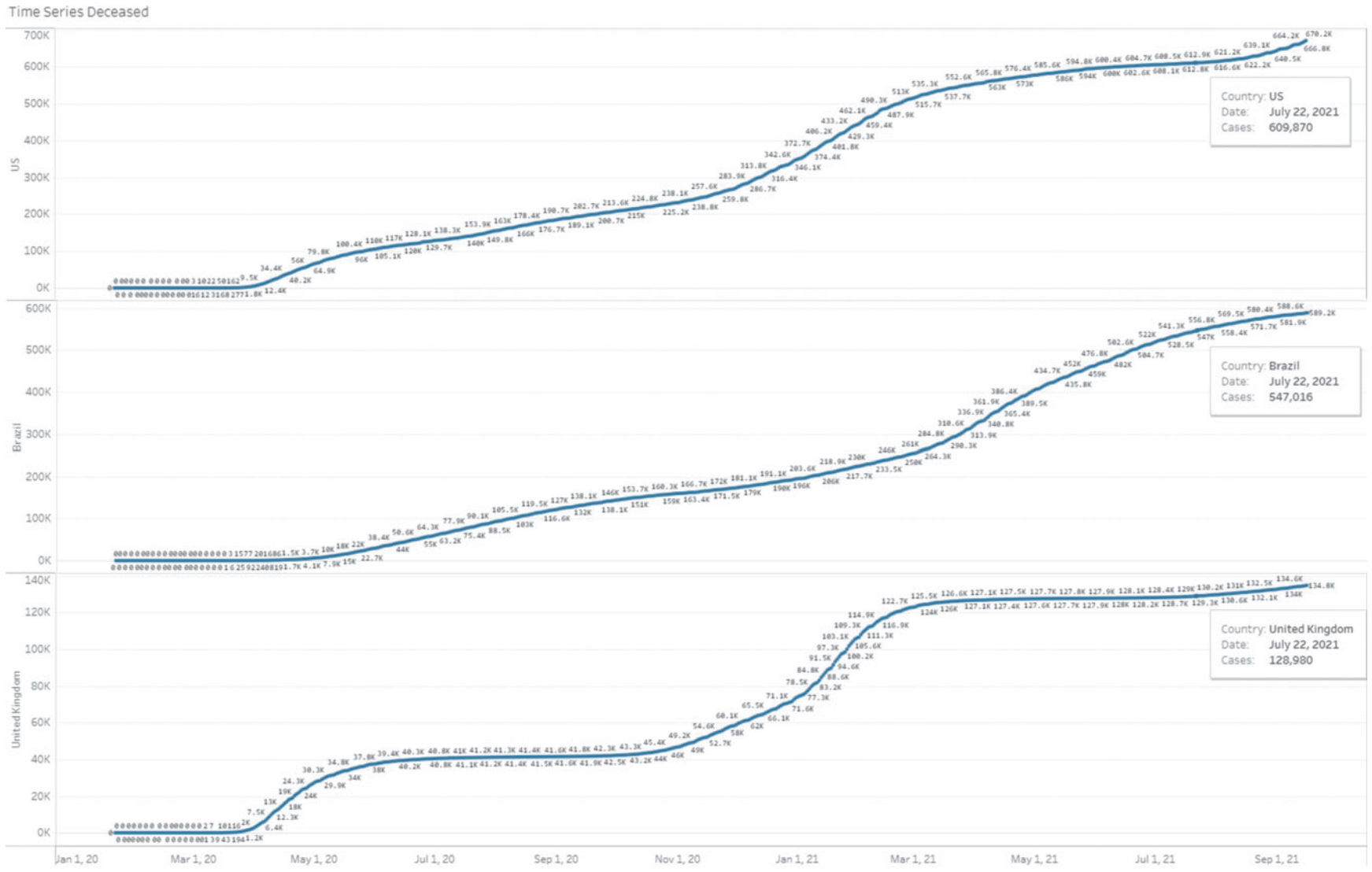

FIG. 2. Three graphs comparing fatality case curves for the United States (top), Brazil (middle), and the United Kingdom (bottom); differences in the United Kingdom versus the United States and Brazil are clearly visible.

A bar chart is also appropriate and can emphasize that the data represent daily snapshots instead of continuous readings. Perceptually, there are few considerations other than ensuring the line–background contrast is sufficient.

Estimated Curve Bend

Public discussion of the COVID-19 pandemic routinely discusses "bending the curve." This is the point where the rate of increase in cases[‡] begins to slow. The initial expectation was that case curves would follow a sigmoid-like shape with an initial rise in the rate of increase in cases (the *exponential* region of the curve), a stabilization in the rate of increase (the *linear* region), and a fall in the rate of increase toward zero as the curve flattens (the *asymptotic* region). Beyond this, the hope is that the rate of increase becomes negligible as the number of cases per day begins to fall.

Four-parameter logistic regression, or 4PL, is used to fit a sigmoid curve to a time series point set. For example, 4PL is commonly used in biology to identify and measure target proteins in enzyme-linked immunosorbent assays. Our interest is in fitting a sigmoid curve to a case time series and then identifying when the curve's rate of acceleration begins to slow. Critically, this can be done on a partial curve, that is, before the bend occurs. This allows us to predict when we expect the rate of increase in cases to fall. 4PL curves are defined by the following four parameters:

(1) a: the minimum value on the curve $g(0)$.
(2) b: the steepness of the curve at its inflection point.
(3) c: the point of inflection $\dot{g}_{max}$.
(4) d: the maximum value on the curve $g(\infty)$.

Our interest is in c, the point of inflection, since this defines the bend's position in the curve. Suppose we plot the independent variable time on the x-axis and the dependent variable cumulative number of cases on the y-axis. Then, the parameters a, b, c, and d define x and y in the 4PL curve as follows:

$$x = c\left(\frac{a-d}{y-b} - 1\right)^{\frac{1}{b}}, \tag{1}$$

$$y = d + \frac{a-d}{1+\left(\frac{x}{c}\right)^{b}}. \tag{2}$$

[‡]For example, if case curves are reported daily, a "bend" in the curve is the point where the day-over-day case count begins to decrease.

A simple way to derive a, b, c, and d from a time series of (x, y) points is to use least squares curve fitting to minimize the error between the point set and the sigmoid curve defined by a, b, c, and d, where error ϵ is measured as follows:

$$\epsilon = y - 4\mathrm{PL}(x,\ a,\ b,\ c,\ d), \tag{3}$$

$$4\mathrm{PL}(x,\ a,\ b,\ c,\ d) = d + \frac{a-d}{1+\left(\frac{x}{c}\right)^{b}}. \tag{4}$$

Once an optimal curve is found, c defines when the case curve bends, and Equations 1 and 2 can be used to calculate the time x and the number of cases y at the bend.

In practice, identifying the inflection point is not as simple as applying the 4PL equations to an entire case time series. 4PL is extremely sensitive to the input point set, especially when used to predict future points. Changing the point set can change the position of c, sometimes dramatically. To address this, we calculate one or more estimates of c_j and its corresponding position (x_i, y_i) as follows for a case time series $S = \{s_0, \ldots, s_n\}$.

(1) Compute 4PL on the entire case time series to date to define c_o and (x_0, y_0).
(2) Remove the first point from the time series and then recompute a new c_1 and corresponding (x_1, y_1).
(3) Continue until y reaches 95% of the maximum y-value of the time series.
(4) Use k-means to cluster all projected (x_i, y_i) points, minimizing the sum of squared errors within each cluster to determine the number of clusters n.
(5) Each cluster forms an estimate of an inflection point. Discard clusters with fewer than four points.
(6) For the m surviving clusters $j = 0, \ldots, m-1$, compute $\mu(a_j)$, $\mu(b_j)$, $\mu(c_j)$, $\mu(d_j)$, and $\sigma(c_j)$ from the points in cluster j.

For each region, the m positions $(x_j, y_j) = 0, \ldots, m-1$ define m estimates of the time and case count of a region's curve bend. $\sigma(c_j)$ defines the confidence in each estimate. The larger the standard deviation, the larger the variation in the cluster's inflection estimates, and the less confidence we have in the cluster's estimated curve bend position.

Figure 1c shows estimated curve bends, together with the estimated number of fatality (top) and confirmed case counts (bottom) at the time of the bend. If a region has more than one estimate, its name is suffixed with $-i$ to denote multiple estimates for the region. Size is used to represent uncertainty. Interestingly, we conducted a short informal experiment to determine whether intuition suggests larger means "more certain" or "more uncertain." Results leaned heavily to "more uncertain," so this is the representation used in the scatterplot. Regions with bend dates more than 10 days in the past are not shown. Ten-day historical estimates are visualized to allow comparison to known data, to see whether the estimate was accurate or not.

Multiple inflection points

Although 4PL worked well initially, it assumes a single inflection point in acceleration, followed by an acceleration plateau. As the pandemic has continued, numerous regions have experienced multiple "spikes" where acceleration decreases and then increases again as external factors such as public holidays, school and university openings, and reductions in safety mandates occur. 4PL is not designed to properly analyze these situations. It can fail to correctly identify the most recent curve bend, whether it has already happened or is estimated to occur in the future.

To address this, we extended 4PL to preprocess a case sequence in a way that determines the most recent case subsequence to analyze. The proper case subsequence is identified through the following steps:

(1) Fit a cubic spline to the case sequence pointset. This smooths the case sequence and also estimates a functional form that underlies the pointset.
(2) Use the spline's functional form to identify second-derivative inflection points, which correspond to positions where the sign of the acceleration changes and the curve "bends."
(3) If fewer than two inflection points exist, the curve is processed in its entirety.
(4) If the last two inflection points represent acceleration changes from $- \rightarrow +$ then $+ \rightarrow -$ (Fig. 3a), the final inflection point is the point where the curve has most recently bent and can be used directly.
(5) If the last two inflection points represent acceleration changes from $+ \rightarrow -$ then $- \rightarrow +$ (Fig. 3b), the curve's acceleration is increasing.

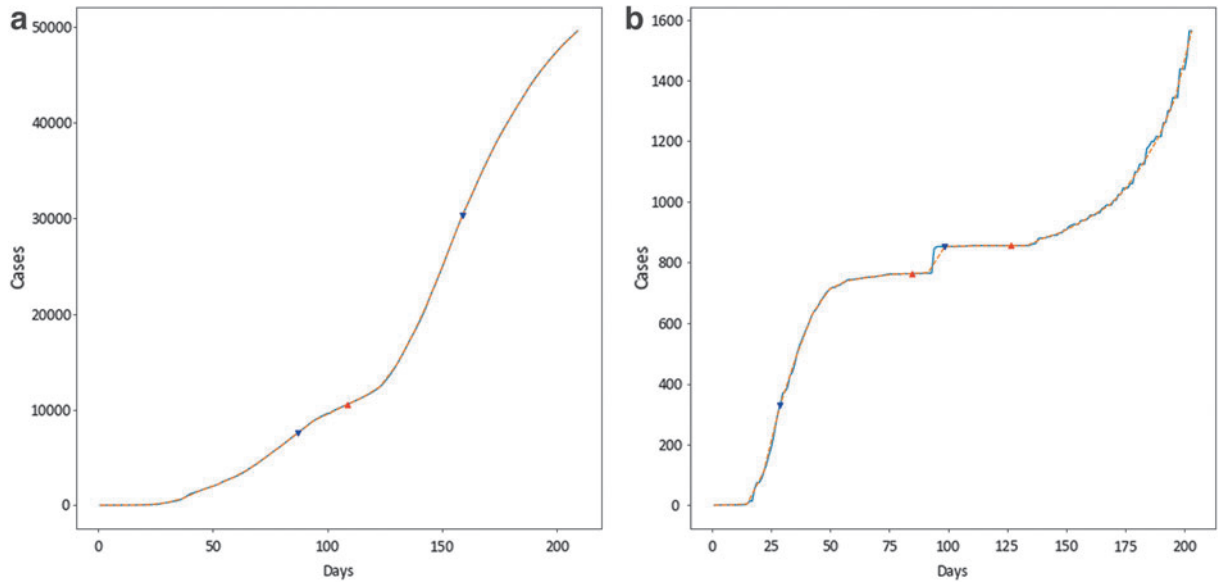

FIG. 3. An extension of 4PL to handle multiple inflection points, $- \rightarrow +$ inflections shown as red upward arrows, $+ \rightarrow -$ inflections shown as blue downward arrows: **(a)** a cubic spline curve (orange) fit to a pointset (blue) with decrease–increase–decrease inflection points; **(b)** a cubic spline curve (orange) fit to a more jagged pointset (blue) with decrease–increase–decrease–increase inflection points.

In this case, 4PL is applied to the curve subsequence starting from the second-last inflection point to estimate the new curve bend position.

Trend Direction

Trend direction provides more localized information on the rate of change in the cumulative number of cases. Perhaps more importantly, a trend direction graph provides a more "intuitive" view of how case rates are changing week-over-week for each region. To the best of our knowledge, this information is not contained in any of the existing visualization dashboards. We use linear regression and regression coefficients to identify both significant and nonsignificant increases and decreases in case rates to calculate this estimate since they are simple and efficient to compute and can be applied with as few as three available data points.

(1) Fit a linear regression line l_1 to the first week's case counts using ordinary least squares (OLS).
(2) Fit a linear regression line l_2 to the second week's case counts using OLS.
(3) Given the slope m_i and its standard error σ_i for each curve, calculate the t-value for the curves' difference.

$$c = \frac{|m_2 - m_1|}{\sqrt{\sigma^2 + \sigma^2}}. \quad (5)$$

(4) Convert the t-value to a p-value to test for significance.
(5) Use the sign of t-value and the p-value to choose between significant upward (t-value >0, $p \leq 0.05$), upward (t-value >0, $p > 0.05$), downward (t-value <0, $p > 0.05$), significant downward (t-value <0, $p \leq 0.05$), or stable (t-value ≈ 0).

Weekly trends for a region are displayed as a graph of line segments, with each segment colored and rotated to differentiate between the five different categories (Fig. 4). The trend graphs for U.S. fatality and confirmed cases are shown in Figure 1a. Upward trends are displayed in shades of red, downward trends in shades of blue, and a stable trend in gray. Perceptually, we selected a red–blue double-ended color scale since these hues are distinguishable to colorblind individuals.[28] Variations in saturation are used to separate significant from nonsignificant.

A secondary data property—the number of cases—is represented with size (width) since experiments have shown that color perceptually dominates size,[29] and we want viewers to see the trend direction first and then the number of cases if this is important to them.

We build out the current week's trend line day-by-day as new case counts are reported. Although this causes potential instability in the trend direction if case counts vary from the current regression line, we decided that individuals would rather see a partial trend, rather than waiting a week to see the next trend line. Three days of data are needed since this is the minimum number of points required to calculate the standard error of the slope σ.

Another important constraint is that the sample sets must be independent. This means if week 1 ends on a Sunday, week 2 starts on the following Monday. It is possible to abut the trend lines. Since this introduces a common point that violates the independence requirement, however, we need to change our significance test method. This can be done using piecewise linear regression. We are currently experimenting with this approach to see if it provides enough value to use as our standard implementation.

To further compare region to region, we provide a second visualization that uses a scatterplot with the most recent week's rate of increase in cases on the x-axis and the number of cases on the y-axis. Again, this visual comparison is unique to our dashboard, allowing viewers to determine the current state of pandemic spread in a given country or region. Both the trend rate and the number of cases are scaled to the range $-1 \ldots 1$. We use this range rather than the traditional $0 \ldots 1$ to imply "low" in the negative area and "high" in the positive area. This divides the scatterplot into four regions with intuitive meanings.

- *managed:* lower-left, low case count and slow case rate increase
- *danger of increase:* lower-right, low case count but accelerating case rate increase
- *accelerating:* upper-right, high case count and rapid case rate increase
- *recovering:* upper-left, high case count but slowing case rate increase

Two common patterns are regions that stay in the lower-left "managed" region (e.g., South Korea) or regions that move counterclockwise from the lower-left to the upper-left region, passing through both the lower-right and upper-right areas. An example is the United States, which spent a significant amount of time in the upper-right region, but has now moved

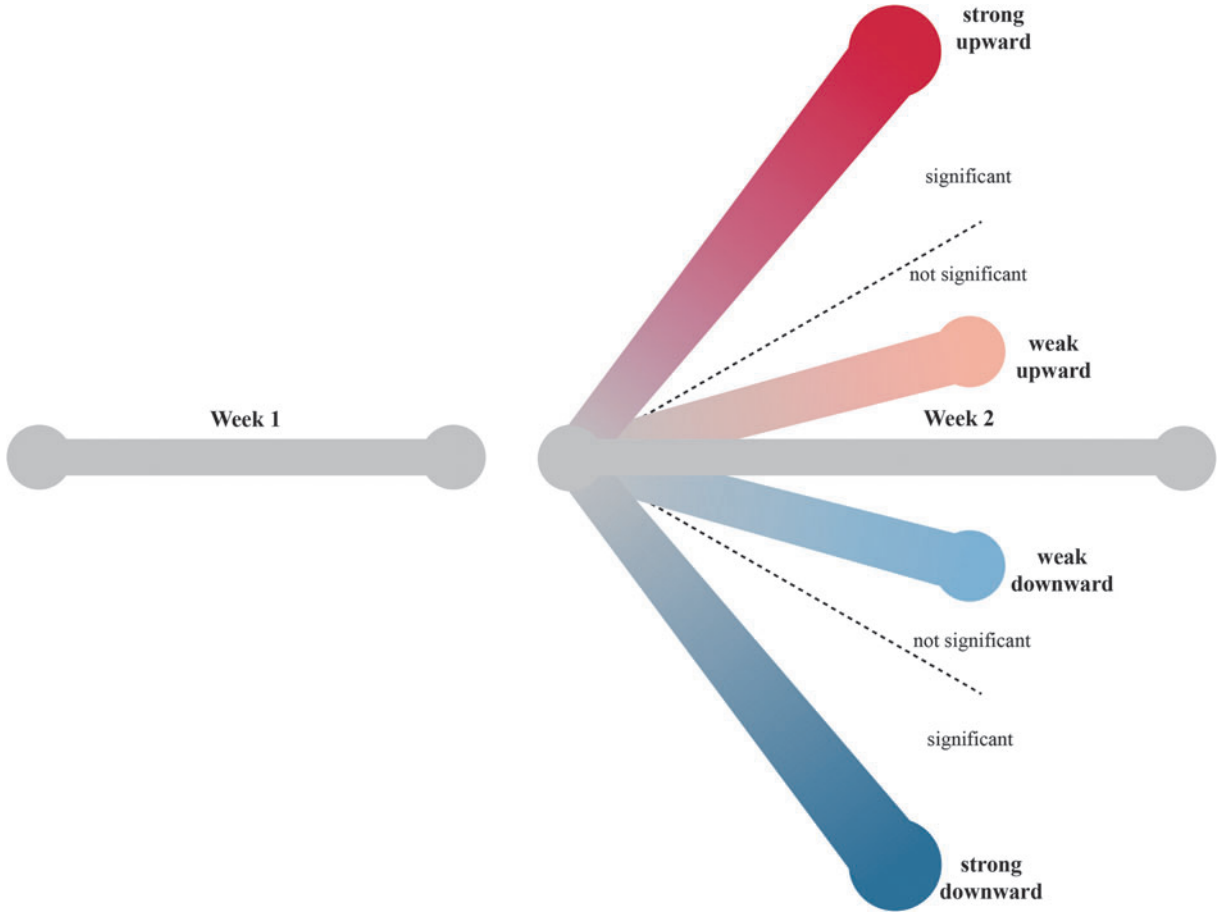

FIG. 4. Weekly trend curve directions: significantly upward (red), upward (pink), downward (light blue), significantly downward (dark blue), or stable (gray); line width represents the number of cases.

into the upper-left "recovering" region for both fatalities and confirmed cases. Figure 5 shows the distribution of regions throughout the four regions for both fatalities (top) and confirmed cases (bottom) in the middle of September 2021.

Time Series Prediction

Predictive analytics provides an understanding of what future values are expected based on existing data. Due to the complexity and volatility of the pandemic, we chose to focus on short-term predictions. Understanding predictive trends can help assess and prepare for potential outcomes, such as increases in intensive care unit admissions and fatalities.

We compared different time series prediction algorithms to see which performed most accurately when predicting fatalities 1 week forward based on a model created using the previous 40 days of confirmed case counts (explanatory variable) and corresponding fatalities (response variable). We investigated three well-known time series models: linear regression (linear), autoregressive integrated moving average (ARIMA), and ARIMA with explanatory variables (ARIMAX).

Another well-known class of time series algorithms is ETS (error, trend, seasonal) exponential smoothing model. We chose to focus on autoregression models since they support multivariate inputs directly. Due to certain regions having little to no variation per week in fatalities at the beginning of the pandemic, we started our analysis $\sim$150 days into the pandemic (specifically, June 18, 2020). The algorithm ran as follows:

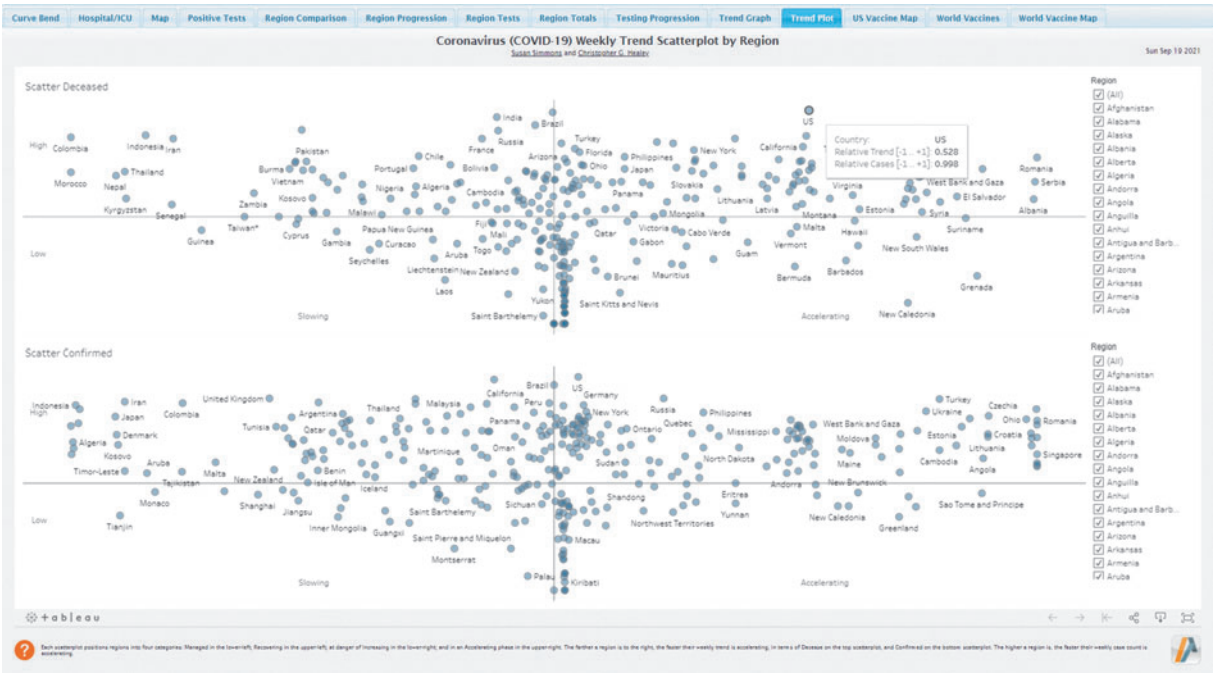

FIG. 5. A distribution of regions across the four trend areas: managed (lower-left), danger of increase (lower-right), accelerating (upper-right), and recovering (upper-left), allowing viewers to identify a region's current pandemic state.

(1) Starting on June 18, for each region, select the next 40 days of (*confirmed case*, *fatality*) pairs for model creation ("training data.")
(2) Fit each algorithm to a region's training set. For linear regression, use a lag of 7 days for confirmed cases since this value is assumed to be unknown when predicting values into the future.
(3) Predict the next 7 days of fatalities and compare them with the known fatality counts ("test data") to calculate an algorithm's mean absolute error (MAE).
(4) Combine the 7 days of test data with the last 33 days of training data to create the next training set and build a new model. Rerun the model to predict new fatality counts and a new MAE score for the next 7 days.
(5) Repeat steps 2–4 until the end of the data set.

Our algorithms created 2458 separate runs of 40 daily sequences with corresponding weekly MAE values. This allowed us to characterize model performance in three separate ways. First, as frequency counts of MAEs between predicted and known fatalities from the test set binned from 0–100, 100–200, …, 900–1000, and >1000 over the 2458 runs. Second, as the algorithms' relative performance on a 1.0 (best) to 0.0 (worst) scale, to see *how much* better a winning algorithm was to the others, as opposed to algorithm order alone. Finally, as order of algorithm performance for each run from lowest to highest based on MAE. This generates a frequency count for how often an algorithm was first, second, or third best for each of the 2458 runs.

To provide context for readers unfamiliar with these algorithms, we provide a brief explanation and references to more detailed descriptions:

Linear: Linear regression (linear) finds the line of best fit using 7-day lagged confirmed case counts as the explanatory variable and fatalities as the response variable for the 40 sample training set. The line is extended to predict the next 7 days of fatalities. This is similar to the simple drift approach, where a line extends from the first training set point to the last. Linear regression minimizes the MSE from the line to each training set sample to optimize its fit.

Autoregressive integrated moving average: Box and Jenkins developed ARIMA in the 1970s.[30] In the ARIMA model, a series must be stationary (a constant mean and variance over time) to begin the modeling process. If a series is not stationary, differencing is applied to the series, which produces an "Integrated" series (the "I" in ARIMA). Once a series is stationary, one can explore fitting autoregressive terms (AR) and/or moving average terms (MA) to the series. We used auto.arima function from Hyndman and Athanasopoulos[31] to search for the best ARIMA model for each training set by iterating through various AR and MA terms, and differencing for the series. The "best" model is defined as the model with lowest corrected Akaike information criterion.

ARIMA with explanatory variables: Incorporating extraneous information in an ARIMA with lagged explanatory variables (ARIMAX) combines the two approaches described above. The explanatory variables in this instance are the 7-day lagged confirmed case counts. Errors from this regression are assumed to follow an ARMA process.[32] We utilize the same auto.arima function from Hyndman and Athanasopoulos[31] with the inclusion of this explanatory variable.

ARIMA and ARIMAX performed well, with MAE scores clustering in the 0–200 range. Very few scores fell outside an MAE of 1000: ARIMA and ARIMAX had 9 and 27 MAEs over 1000, respectively. Due to its simplicity, linear performed worse than the ARIMA-based models, producing MAEs on the 0–2000 range, with 81 MAE values over 1000 and 24 over 2000, respectively.

We plotted relative MAE performance for the algorithms to highlight which algorithms perform best most often and visualize how much better they are than their competitors. Figure 6a–c show the same data, but sorted best-to-worst for each of the three algorithms: linear as the green line in Figure 6a, ARIMA as the purple line in Figure 6b, and ARIMAX as the orange line in Figure 6c.

Linear scored first in only two cases and tied with ARIMAX in two cases. Even when it was first, it was not markedly better than either ARIMA or ARIMAX. This is clear from the almost immediate and sharp drop to zero in the green line in Figure 6a. ARIMA (purple line, Fig. 6b) did much better, maintaining a large relative advantage over both linear and ARIMAX for numerous regions. It placed first for 102 regions. ARIMAX (orange line, Fig. 6c) also performed well, scoring first for 55 regions and showing a large improvement in MAE over the other algorithms in numerous cases.

Figure 7 shows side-by-side bar graphs visualizing the number of times each algorithm placed first, second, or third over the 2458 runs we tested. ARIMA had the most first-place runs (1302), followed by ARIMAX (965), and linear (457). Linear also had the most third-place runs (1595), which is not surprising given the model's simplicity. Alternatively, if we award an algorithm three points for a first-place run, two points for a second place run, and one point for a third place run, then ARIMA and ARIMAX score nearly identically: 2.22 and 2.2, respectively. Linear scores 1.58. Based on these metrics, we chose to use ARIMA since it is fairly simple, well known, and consistently scores best in our experiments.

As an example of our predictions, Figure 8 shows fatality over time in the United States. The most recent 40 days of confirmed case counts and fatalities (the time window shown in light blue) were used to construct an ARIMA model that was then applied to predict the next 1 week of fatalities. This is visualized as the orange line with gray upper and lower boundaries representing the 95% confidence intervals. The same approach can be applied to any region's confirmed case counts and fatalities to predict the number and direction of fatalities over the next 7 days.

Testing, Positivity, and Vaccination

The final analytic component for our work is testing, test positivity rates, and vaccinations. We are currently focused on infection testing, although we will add antibody testing when numbers become available. We are also visualizing data from the U.S.–German Pfizer-BioNTech,[33] U.S. Moderna[34] and Johnson & Johnson,[35,36] U.K. Oxford-AstraZeneca,[37] Russian Sputnik V,[38] and Chinese Sinopharm-CNBG vaccines.[39,40] We will add new candidates (e.g., Novavax[41]) when they are approved for use.

As most epidemiologists have stated, testing is critical to managing the pandemic. First, testing is used in conjunction with contact tracing to identify sick individuals and people they have interacted with, to avoid new disease outbreaks. For example, Apple and Google have proposed a novel method of contact tracing that uses the iOS and Android cell phone operating systems.[42] Low-power Bluetooth will allow iPhones

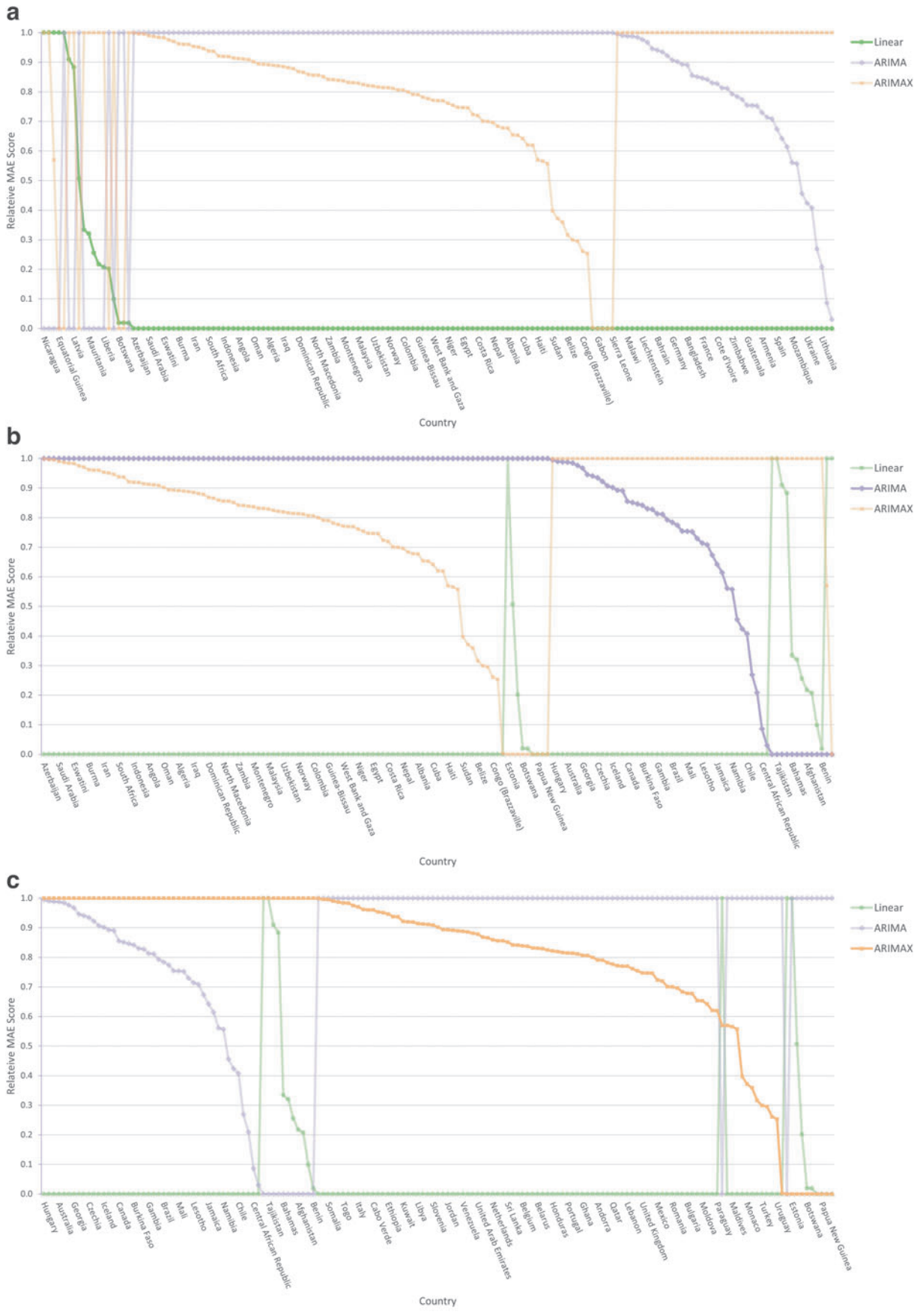

FIG. 6. Line graphs of relative MAE performance by region: **(a)** sorted by linear performance (green line), best to worst; **(b)** sorted by ARIMA (purple); **(c)** sorted by ARIMAX (orange). ARIMA, autoregressive integrated moving average; ARIMAX, ARIMA with explanatory variables; MAE, mean absolute error.

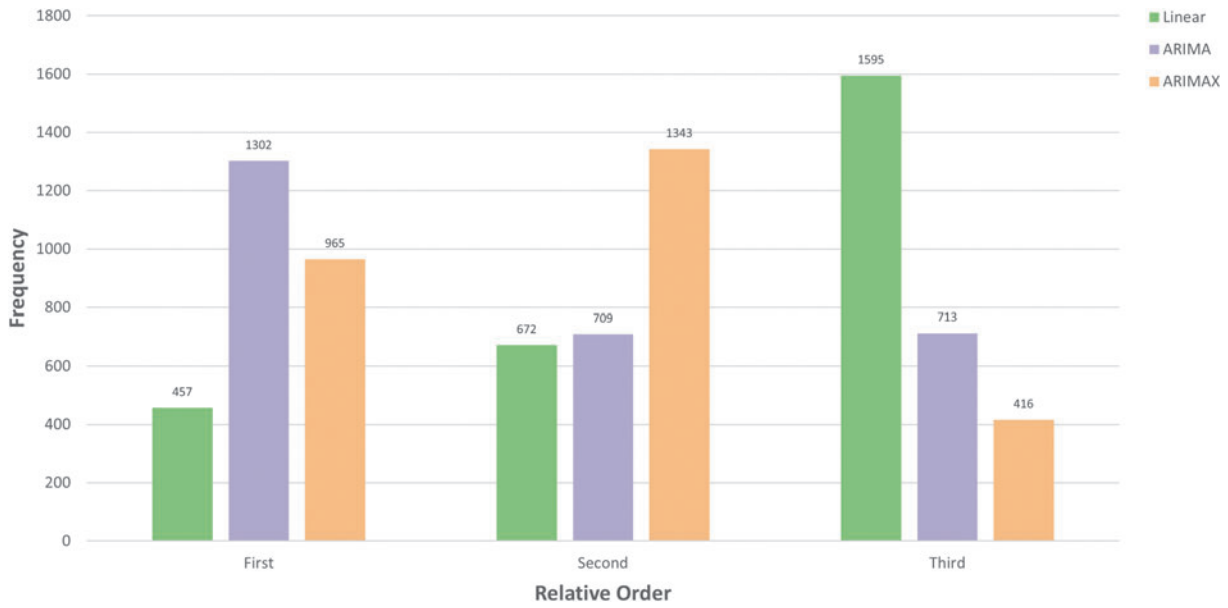

FIG. 7. Side-by-side bar charts shows how often each algorithm placed first, second, and third in our 2548 test runs.

and Android phones to record anonymously who their owners interact with. If a person is diagnosed with COVID-19, their phone can be asked to anonymously send a text message to every person the phone recorded over the last 14 days, informing the recipients that someone they came into contact with has contracted COVID-19.

Code to implement this strategy is included in the recent iOS 13.5, although privacy arguments are still being addressed. Second, testing is meant to inform people if they may have immunity to the disease. This would, for example, allow them to return to work or interact with the public securely. However, it is important to note that there is currently not enough evidence to guarantee immunity if an individual has had COVID-19, nor to know how strong the resistance is, or how long it will last.[43]

Testing data are less readily available, and less complete, than case data. We are currently obtaining daily updates from the Humanitarian Data Exchange (HumData), managed under the United Nations Office for the Coordination of Humanitarian Affairs.[44]

We visualize total tests by region as ordered bar graphs. In addition to the absolute number of total tests, we also include tests per 1000 citizens. This allows us to normalize for a region's size. The United States has performed the most tests of any region (547 million), but Cyprus has delivered the most tests per 1000 citizens (13,926). The United States falls somewhere in the top third of regions reporting testing data per 1000 citizens (1644). Because regions such as the United States and Luxembourg present as outliers, we also overlay a line graph representing $\log_{10}$-corrected test values. This allows a much easier comparison of regions with smaller numbers of tests.

Although testing data are critical, they are influenced by the rate of testing a particular region performs. More tests can result in more confirmed cases, providing a potentially skewed interpretation of the spread of the disease. Experts suggest that positivity rate is a more accurate measure of disease prevalence since it normalizes over number of tests performed, reporting the percentage of tests that report positive COVID-19 infection. Figure 9 shows the total number of tests performed for the United States over time as a blue bar graph, with an orange line graph overlaid to track the 7-day moving average of positivity rates for the same time period. The positivity rate y-axis runs from 0% to 50%, so any rates above this threshold appear as a positivity line leaving the top of the graph.

Figure 9 shows that the United States has performed 551.1 million tests as of September 13, 2021. The positivity rate showed a spike at the beginning of the pandemic as testing procedures were implemented and

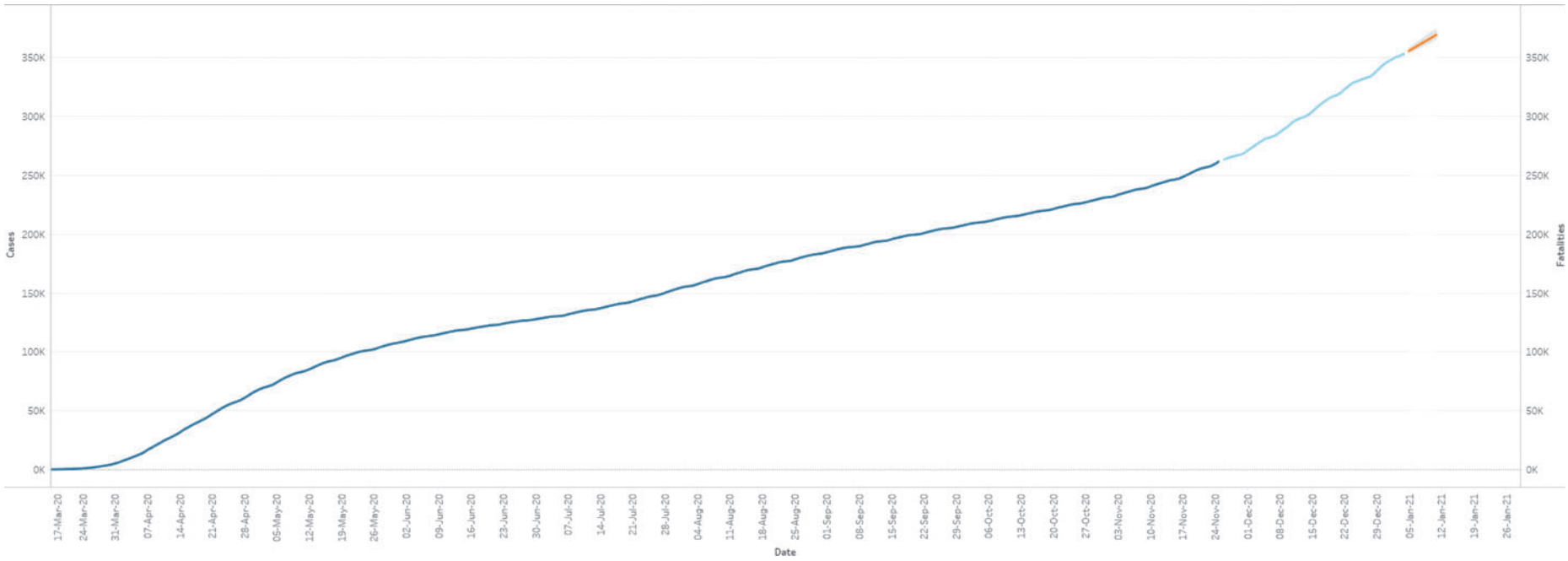

FIG. 8. U.S. fatality graph starting on March 17, 2020, ARIMA fatality predictions for the upcoming 7 days shown in orange, 95% confidence intervals shown in gray, the previous 40 days of (*confirmed case*, *fatality*) pairs (region highlighted in light blue) were used to train the ARIMA model.

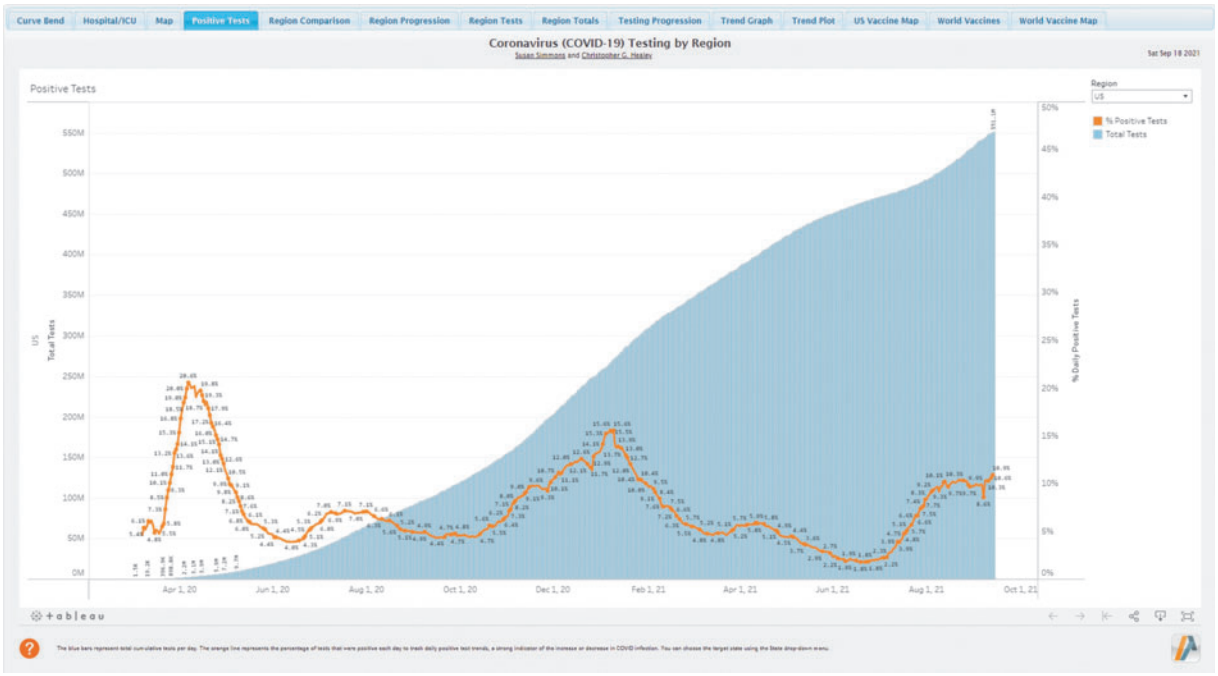

FIG. 9. Line and bar graph of positivity rates and COVID tests administered in the United States, the line graph represents 7-day moving average positivity percentage, the bar graph represents cumulative tests administered, graphs allow viewers to track rate of testing and pandemic spikes.

refined. Positivity rates began to stabilize around the beginning of March, reaching a low of 4.3% in the middle of June, then spiking in August 2020, December 2020, and August 2021. The most recent positivity rate for the United States is 6.4%, with a 7-day moving average of 10.9%. This represents the current rapid spread of the disease, thought to be a combination of vaccine concern coupled with new, more transmissive variants of COVID-19 identified in the United Kingdom, South Africa, and India, and now spreading throughout the world.

Numerous vaccine research projects have been initiated to search for safe and effective vaccines for COVID-19. In the United States and Europe, the Pfizer-BioNTech, Moderna, Oxford-AstraZeneca, and Johnson & Johnson vaccines have been approved for use and were administered starting in mid-December 2020. Russian, Chinese, Latin American, and Middle Eastern countries have started administering the Sputnik V and Sinopharm-CNBG vaccines. Promising late-phase three trials are ongoing for Novavax. Third dose booster shoots have recently been approved or recommended in Israel and for certain groups in the United States. Vaccine distribution is currently focused on non-Western countries, vaccine-hesitant individuals, and children younger than 16 years.

Our interest is in analyzing the distribution of vaccinations by region along three-related dimensions: total number of vaccines administered, vaccines per 1000 citizens, and vaccines per day. The first metric tracks absolute vaccine distribution, the second normalizes for region populations, and the third measures how quickly vaccines are being released to the general public.

Figure 10a shows the three measures represented as individual bar graphs by region: total vaccinations in the top graph, vaccinations per 1000 citizens in the middle graph, and vaccinations per day in the bottom graph. Not surprisingly, the order of countries in each graph varies, often significantly. Although larger countries are more likely to have the capability to release larger numbers of vaccines (e.g., the United States and China), this does not guarantee a corresponding high degree of coverage of the general population.

Pandemic State

In addition to testing and vaccination, we provide a line graph plotting the total daily tests and tests per 1000

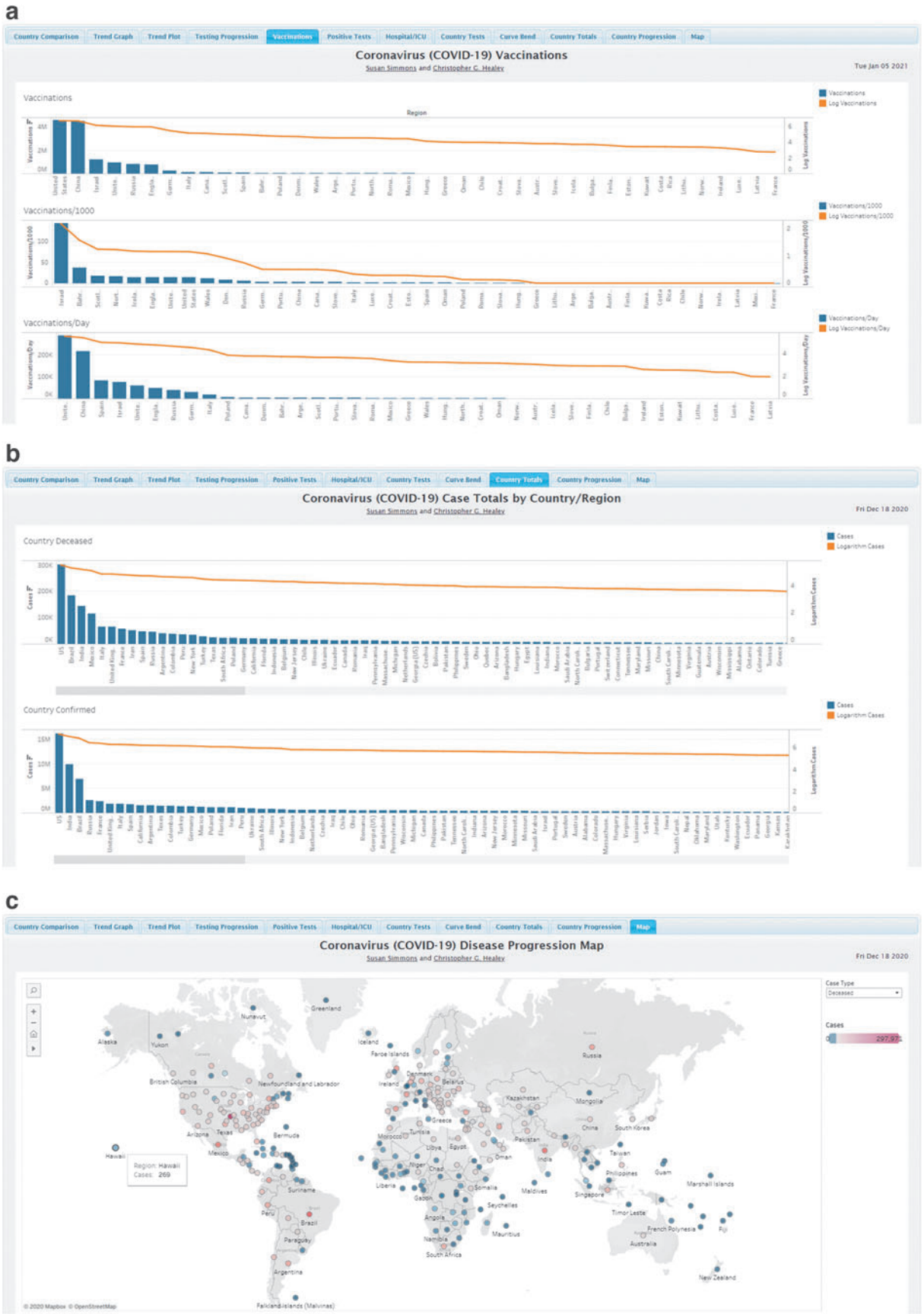

FIG. 10. Pandemic state by region: **(a)** vaccines administered total (top), per 1000 citizens (middle), per day (bottom); **(b)** bar graph of total fatalities (top) and confirmed cases (bottom); **(c)** thematic map of total fatalities, blue for below the median, red for above, all three visualizations are designed to allow viewers to identify and compare pandemic state by region.

citizens for a target region. This is similar to the region total line graph (Fig. 2) and allows investigation of when and how a region's testing has progressed.

Two additional dashboard visualizations are provided: a bar graph of total fatalities and confirmed cases by region, and a world map that allows viewers to switch between fatality and confirmed case totals. Both visualizations represent current state information without any underlying analytics. They provide context, particularly as a basis for evaluating other visualizations such as trend graphs or region comparisons.

The total case bar graph (Fig. 10b) shows bars ordered by fatalities (top) and confirmed cases (bottom). As with the testing bar graph, a $\log_{10}$-corrected line is overlaid in orange to allow comparison of smaller totals due to outliers such as the United States and the United Kingdom.

The geographic map plots circles over each region, using a red–blue double-ended color scale to represent case totals (Fig. 10c). As in the trend graph, red and blue were chosen to support colorblind viewers. Totals below the median are colored blue, with saturation increasing as a region's total moves farther from the median. Totals above the median are colored red, with saturation increasing past the median. Choosing a center point based on the median allows us to avoid bias due to outliers. Perceptually, work in both our laboratory and in cartography has shown that double-ended color scales are best suited to visualizing thematic maps with a semantically meaningful center point.[45,46]

Table 1. A comparison of visualizations included in our proposed dashboard, Johns Hopkins University's coronavirus dashboard, University of California, Los Angeles's combating COVID-19 dashboard, Los Alamos National Laboratory's COVID-19 dashboard, and the CovidNet dashboard

	Proposed	*JHU*	*UCLA*	*LANL*	*CovidNet*
Case trend graph	✓	✓	✓		
Case trend scatterplot	✓				
Estimated curve bend	✓			✓	✓
Hospital/ICU rates	✓	✓			
Hubei, China, graph		✓			
Maps by ethnicity		✓			
Region case totals	✓	✓	✓		
Region progression comparison	✓				
Region progression totals	✓	✓	✓		
Region testing comparison	✓	✓			
Social distancing graph		✓			
Supplies			✓		
Tests/% positive	✓	✓			
Testing progression	✓				
U.S. map current cases	✓	✓	✓	✓	✓
Vaccine map	✓		✓		
Vaccine concern map	✓				
World map current cases	✓	✓	✓		

Although our proposed dashboard often includes a significantly wider geographic coverage (e.g., U.S. states and counties, Canadian and Chinese provinces, and both U.S. and worldwide cases) and data analytic preprocessing, we have credited every dashboard that provides a subset of the given visualization type.

ICU, intensive care unit; JHU, Johns Hopkins University; LANL, Los Alamos National Laboratory; UCLA, University of California, Los Angeles.

Conclusions and Future Work

This article presents an analytics and visualization framework designed to provide future predictions and the current state for different aspects of the coronavirus COVID-19 pandemic. Our overarching goals were to include *both* predictions and state information and present results in ways that are accessible to individuals who are not analytics or visualization experts. We apply dynamic time warping, four-parameter logistic regression, linear regression curve comparison, ARIMA, and knowledge of human visual perception to build a dashboard that includes comparisons between regions, information on case trends, testing, vaccinations, and overall case totals.

To place our work in the context of COVID-19 visualizations, Table 1 enumerates the visualizations in our dashboard, as well as the visualizations in three other well-known dashboards: JHUs Novel Coronavirus Cases Data dashboard,[3] UCLAs Combating COVID-19,[7] Los Alamos's COVID-19 dashboard,[8] and CovidNet.[6]

Numerous avenues of future work exist, and in fact, the dashboard continues to be rapidly updated as pandemic conditions change. For example, during the writing of this article, we added testing and vaccination data and removed some of the curve bend visualizations that were no longer relevant due to the current stage of the pandemic.

One addition we are keenly interested in is incorporating some of the more sophisticated models currently being used to predict the long-term future state of the pandemic, for example, the University of Washington's Institute for Health Metrics and Evaluation (IHME) COVID-19 Projections,[47] the CDCs COVID-19 Forecasts,[48] Imperial College London's forecasts,[49] and Columbia University Epidemiology and Environment Health Sciences' Severe COVID-19 Risk Mapping.[50]

A second area of interest is in applying predictive analytics to testing data. Since these data have only recently become available, we are still considering

possible strategies. One obvious approach that we could quickly include would apply 4PL to test and vaccination curve time series, similar to a case time series. In theory, this would identify the peak of the test and vaccination curves, identifying how many tests and vaccinations total and per 1000 citizens each region was on track to achieve. If information on antibody testing and contact tracing becomes available, these would also be valuable to include in the testing section of our dashboard.

Another potential source of valuable information is the use of mitigation strategies such as masks, social distancing, and limitations on mobility. The challenge here is to identify a source of information for these data that cover a reasonable subset of the countries we are studying. Since different countries have different interpretations of the mitigation strategies, these would need to be considered. Finally, strategies change over time. For example, the CDC in the United States recently reduced the required social distancing for students in K-12 schools from six feet to three feet, with conditions.[51] Care would need to be taken to track these changes and update our models accordingly.

Finally, no visualization can be claimed to be relevant without end user testing. We have received anecdotal feedback on our dashboard's value, for example, during its evaluation as part of the U.S. Health and Human Services COVID-19 design competition.[52]

We plan to collaborate directly with epidemiologist colleagues from our dengue fever research, policymakers, and members of the general public, to obtain feedback on the strengths and limitations of the current visualizations. We expect that, since each group has different skill sets and motivations, their feedback will not be identical nor necessarily even similar. Understanding what each group can use and what they need will answer two important questions: (a) Is the dashboard useful, and how can it be modified to improve its impact? and (b) Can a single set of visualizations serve a diverse group of individuals with different domain expertise and questions of interest? The second question, in particular, is one of significant interest to the visualization community.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

No funding was received for this article.

References

1. World Health Organization. 2020. WHO timeline—COVID-19. Available online at https://www.who.int/news-room/detail/27-04-2020-who-timeline—covid-19 (last accessed December 13, 2021).
2. Holshue ML, DeBold C, Lindquist S, et al. First case of 2019 novel coronavirus in the United States. N Engl J Med. 2020;382:929–936.
3. Johns Hopkins Whiting School of Engineering, Center for Systems Science and Engineering. 2020. Novel Coronavirus (COVID-19) cases data—Humanitarian data exchange. Available online at https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases (last accessed December 13, 2021).
4. data.world. 2020. Global Coronavirus (COVID-19) Data (Johns Hopkins). Available online at https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/file?filename=COVID-19+Cases.csv (last accessed December 13, 2021).
5. Yang T, Shen K, He S, et al. CovidNet: To bring data transparency in the era of COVID-19. *arXiv e-prints*, arXiv:2005.10948, (2020).
6. 1Point3Acres. 2021. Global COVID-19 tracker & interactive charts. Available online at https://coronavirus.1point3acres.com/ (last accessed December 13, 2021).
7. University of California at Los Angeles. 2021. Combating COVID-19. Available online at https://covid19. uclaml.org/ (last accessed December 13, 2021).
8. Los Alamos National Laboratory. 2021. LANL COVID-19 cases and deaths forecast. Available online at https://covid-19.bsvgateway.org/ (last accessed December 13, 2021).
9. COVID Analytics. 2021. Predictions of infections and deaths under a variety of policies. Available online at https://www.covidanalytics.io/policies (last accessed December 13, 2021).
10. GLEAM Project. 2021. Using big data and computational modeling to fight infectious diseases. Available online at https://gleamproject.org (last accessed December 13, 2021).
11. Shaman Group. 2021. COVID-19 findings, simulations. Available online at https://blogs.cuit.columbia.edu/jls106/publications/covid-19-findings-simulations/ (last accessed December 13, 2021).
12. Healey CG, Enns JT. Attention and visual memory in visualization and computer graphics. IEEE Trans Vis Comput Graph. 2012;18:1170–1188.
13. Snow J. On the mode of communication of cholera. London, England: John Churchill, 1855.
14. Frerichs RR. 2016. Mapping the 1854 Broad Street pump outbreak. Available online at https://www.ph.ucla.edu/epi/snow/mapsbroadstreet.html/ (last accessed December 13, 2021).
15. Carroll LN, Au AP, Detwiler LT, et al. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform. 2014;51:287–298.
16. Linvat Y, Rhyne T-M, Samore M. Epinome: A visual-analytics workbench for epidemiological data. Comput Graph Appl. 2012;32:89–95.
17. Hamid S, Bell L, Dueger EL. Digital dashboards as tools for regional influenza monitoring. Western Pac Surveill Response J. 2017;8:1–4.
18. Lee M-T, Lin F-C, Chen S-T, et al. Web-based dashboard for the interactive visualization and analysis of national risk-standardized mortality rates of sepsis in the US. J Med Syst. 2020;44. DOI: 10.1007/s10916-019-1509-9.
19. Tableau. 2020. Coronavirus (COVID-19) global data tracker. Available online at https://www.tableau.com/covid-19-coronavirus-data-resources/global-tracker/ (last accessed December 13, 2021).
20. SAS Institute. 2020. 2019 novel coronavirus. Available online at https://tbub.sas.com/COVID19/ (last accessed December 13, 2021).
21. Villanes A, Griffths E, Rappa M, et al. Dengue fever surveillance in India using text mining in public media. Am J Trop Med Hyg. 2018;98:181–191.
22. Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–1014.
23. Butler D. When Google got flu wrong. Nature. 2013;494:155+.
24. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc Natl Acad Sci USA. 2015;112:14473–14478.
25. Berndt DJ, Clifford J. 1994. Using dynamic time warping to find patterns in time series data. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'94), Seattle, Washington, pp. 359–370.

26. Zhang J. 2020. Dynamic time warping: Explanation and code implementation. Available online at https: //towardsdatascience.com/dynamic-time-warping-3933f25fcdd (last accessed December 13, 2021).
27. Salvador S, Chan P. Toward accurate dynamic time warping in linear time and space. Intell Data Anal. 2007;11:561–580.
28. Shepard RN, Cooper LA. Representation of colors in the blind, color-blind, and normally sighted. Psychol Sci. 1992;3:97–104.
29. Healey CG, Enns JT. Large datasets at a glance: Combining textures and colors in scientific visualization. IEEE Trans Vis Comput Graph. 1999;5:145–167.
30. Box G, Jenkins G. Time series analysis, forecasting, and control, Rev. ed. San Francisco, CA: Holden-Day, 1976.
31. Hyndman R, Athanasopoulos G. Forecasting: Principles and practice, 2nd ed. Melbourne, Australia: OTexts.com/fpp2, 2018.
32. Box GEP, Tiao GC. Intervention analysis with applications to economic and environmental problems. J Am Stat Assoc. 1975;70:70–79.
33. Centers for Disease Control and Prevention. 2020a. Information about the Pfizer-BioNTech COVID-19 vaccine. Available online at https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/Pfizer-BioNTech.html (last accessed December 13, 2021).
34. Centers for Disease Control and Prevention. 2020b. Moderna COVID-19 vaccine. Available online at https://www.cdc.gov/vaccines/covid-19/info-by-product/moderna/index.html (last accessed December 13, 2021).
35. Johnson & Johnson. 2020. Johnson & Johnson announces its first phase 3 COVID-19 vaccine trial ENSEMBLE is fully enrolled. Available online at https://www.jnj.com/our-company/johnson-johnson-announces-its-first-phase-3-covid-19-vaccine-trial-ensemble-is-fully-enrolled (last accessed December 13, 2021).
36. verywell health. 2020. An overview of the Johnson & Johnson COVID-19 vaccine. Available online at https://www.verywellhealth.com/johnson-and-johnson-covid-19-vaccine-5093160 (last accessed December 13, 2021).
37. Knoll MD, Wonodi C. Oxford-AstraZeneca COVID-19 vaccine efficacy. *Lancet* 2020, (in press). DOI: 10.1016/S0140-6736(20)32623-4.
38. Precision Vaccinations. 2020. Sputnik V vaccine. Available online at https://www.precisionvaccinations.Com/vaccines/sputnik-v-vaccine (last accessed December 13, 2021).
39. New York Times. 2020. How the Sinopharm COVID-19 vaccine works. Available online at https://www.nytimes.com/interactive/2020/health/sinopharm-covid-19-vaccine.html (last accessed December 13, 2021).
40. Time. 2020. Chinese COVID-19 vaccine is 86% effective, UAE says. Available online at https://time.com/5919257/china-covid-19-vaccine-cnbg/ (last accessed December 13, 2021).
41. Quartz. 2020. What is the Novavax vaccine, and how does it work? Available online at https://qz.com/1950365/what-is-the-novavax-vaccine-and-how-does-it-work/ (last accessed December 13, 2021).
42. Albanesius C, Muchmore M. 2020. Here's how contact tracing will work on iPhones and Android phones. *PC Magazine*. Available online at https://www.pcmag.com/how-to/heres-how-contact-tracing-will-work-on-iphones-and-android-phones (last accessed December 13, 2021).
43. Shi Y, Wang Y, Shao C, et al. COVID-19 infection: The perspectives on immune responses. Cell Death Differ. 2020;27:1451–1454.
44. UN OCHA Humanitarian Data Exchange. 2020. Total COVID-19 tests performed by country. Available online at https://data.humdata.org/dataset/total-covid-19-tests-performed-by-country/ (last accessed December 13, 2021).
45. MacEachren AM. How maps work. New York, New York: Guilford Publications, Inc. 1995.
46. Slocum TA. Thematic cartography and visualization. Upper Saddle River, New Jersey: Prentice-Hall, Inc. 1998.
47. University of Washington Institute for Health Metrics and Evaluation. 2020. Coronavirus (COVID-19) Projections. Available online at https://covid19.healthdata.org/united-states-of-america/ (last accessed December 13, 2021).
48. Centers for Disease Control and Prevention. 2020. COVID-19 Forecasts. Available online at https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html (last accessed December 13, 2021).
49. Imperial College London MRC Centre for Global Infectious Disease Analysis. 2020. COVID-19 Model. Available online at https://mrc-ide.github.io/covid19-short-term-forecasts/index.html (last accessed December 13, 2021).
50. Columbia University Epidemiology and Environmental Health Sciences. 2020. Severe COVID-19 Risk Mapping. Available online at https://columbia.maps.arcgis.com/apps/webappviewer/index.html?id=ade6ba85450c4325a12a5b9c09ba796c (last accessed December 13, 2021).
51. Centers for Disease Control and Prevention. 2021. CDC updates operational strategy for K-12 schools to reflect new evidence on physical distance in classrooms. Available online at https://www.cdc.gov/media/releases/2021/p0319-new-evidence-classroom-physical-distance.html (last accessed December 13, 2021).
52. United States Department of Health and Human Services. 2020. COVID-19 at-anywhere diagnostics design-a-thon. Available online at https://www.hhs.gov/coronavirus/covid-19-at-anywhere-diagnostics-design-a-thon (last accessed December 13, 2021).

Cite this article as: Healey CG, Simmons SJ, Manivannan C, Ro Y (2022) Visual analytics for the coronavirus COVID-19 pandemic. *Big Data* 10:2, 95–114, DOI: 10.1089/big.2021.0023.

Abbreviations Used

AR = autoregressive terms
ARGO = AutoRegression with GOogle search data
ARIMA = autoregressive integrated moving average
CDC = Centers for Disease Control and Prevention
DTW = dynamic time warping
ETS = error, trend, seasonal
ICU = intensive care unit
IHME = Institute for Health Metrics and Evaluation
JHU = Johns Hopkins University
LANL = Los Alamos National Laboratory
MA = moving average terms
OLS = ordinary least squares
SARS-CoV-2 = severe acute respiratory syndrome coronavirus 2
UCLA = University of California, Los Angeles
WHO = World Health Organization