SOCIETY OF ACTUARIES
TECHNOLOGY SECTION

ACTUARIAL TECHNOLOGY TODAY
OCTOBER 2020

What's in a Number? Considerations for Mapping the COVID-19 Pandemic

By Phil Ellenberg and Kelsie Gosser

Author's note: Using COVID-19 visualizations as a case study, this article highlights the considerations involved with preparing data to tell a story. As actuaries we are often challenged with how to tell a complex story using nuanced data. This article provides a helpful framework.

As COVID-19 remains a mainstay in 2020, it seems like the only certainty we can count on is that we remain uncertain about the global pandemic. In an effort to mitigate this uncertainty, we have collectively turned to data and, by extension, data visualizations. Despite numerous issues with the underlying data, COVID-19 dashboards and other visualizations have become ubiquitous in our daily lives, each with their own recipe for showing the impact of the virus. Using confirmed case count data for the United States from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[1], we intend to explore the interaction between selecting a metric and then visualizing it. We chose the Johns Hopkins' data over other data sources such as The New York Times or The COVID-19 Tracking Project due to its prominence in the media, credibility, and its ease of access.

When we first started writing this article, we went into it with a plan to write strictly about data visualization techniques and how to pick which graph to best illustrate the data it represents—the twist being that we were going to use COVID-19 data. We quickly found ourselves uninterested in the topic, especially after referencing countless other articles on the same topic (although, to be fair, none of them use COVID-19 data). We were, however, intrigued by the variety of metrics used across the endless patchwork of COVID-19 dashboards. If case counts, hospitalizations, fatalities, and tests were the bricks for our house, figuring out how best to report them was like deciding what color to paint the walls. Do we look at cumulative counts going back to March? Do we look at newly reported numbers on a daily basis? Do we adjust the data for population? What about trend? Should we look at day-over-day changes? Week-over-week? Moving averages?

These are questions we failed to thoroughly think through before diving in. This was evident when we first started drafting this article and dropped cumulative case counts by State onto a bubble map (formally known as a proportional symbol map) in Microsoft's Power BI. As shown in Figure 1 (page 2), looking solely at case counts causes New York to eclipse the rest of the country. An argument can be made that this is due primarily to the population density of New York City. States like California, Florida, and Texas also jump out as being hard hit, but these are also big states to begin with. To be sure, these states have all made headlines during the pandemic, but other headline-grabbing states like Louisiana, Illinois, and Arizona are lost in the shuffle.

Figure 1
Cumulative Confirmed Case Counts by State

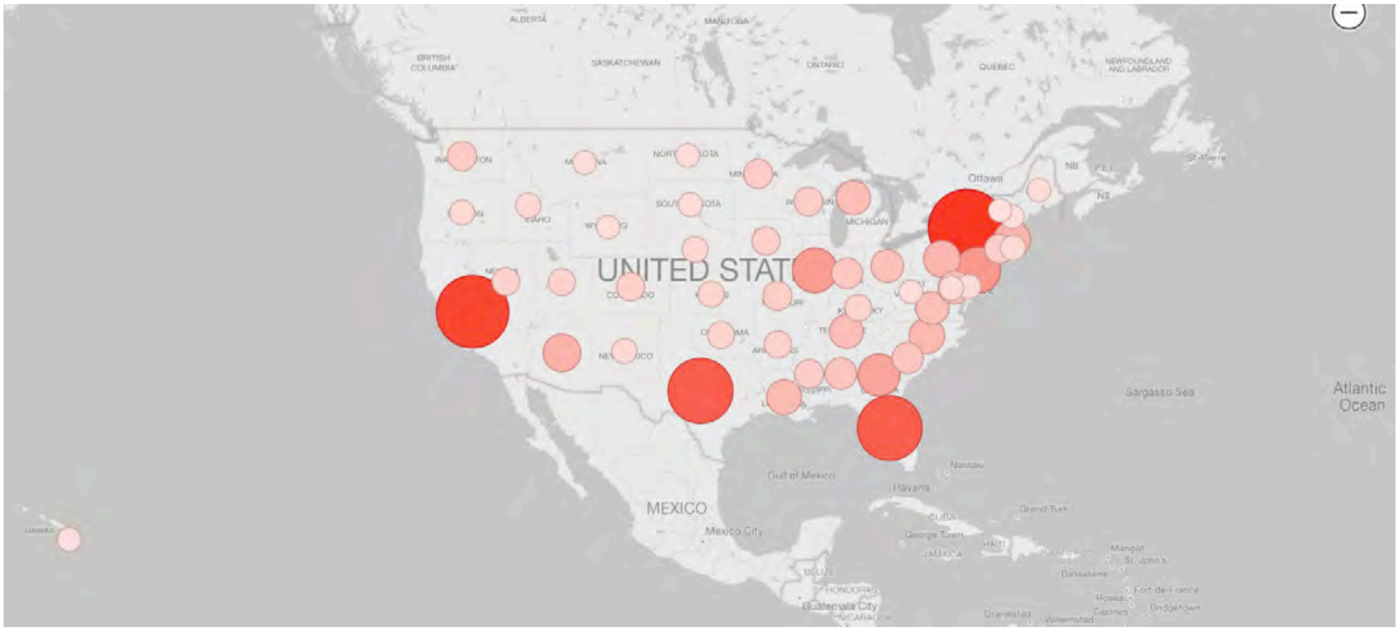
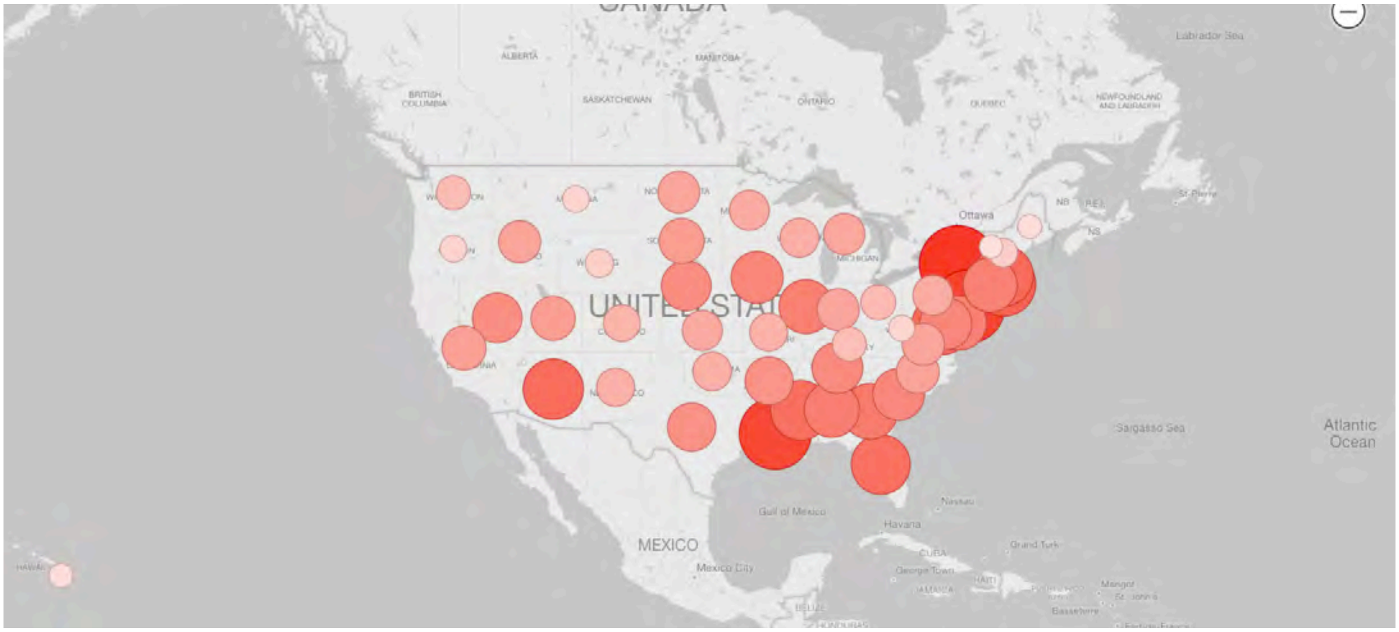


Figure 2
Cumulative Confirmed Case Counts per 100k by State



Simply adjusting the cumulative case count data for population shows a slightly different story. While New York still dominates when showing cumulative case counts per 100k residents in a state, the impacts on other New England and Mid-Atlantic states such as New Jersey, Massachusetts, Rhode Island, Connecticut, and Delaware are immediately more noticeable. Adjusting for population also highlights states like Louisiana, Illinois, and Arizona as being more heavily impacted. (See Figure 2.)

What cumulative case counts miss, however, is how case counts are changing over time. By nearly all accounts, New York's cases peaked in early-April. To somehow capture the trend, we were faced with two decisions: what type of map to use, and what metric to use. Generally speaking, proportional maps should be used to show totals rather than rates. Filled maps, or choropleths, are a standard way of displaying rates or ratios between well-defined geographies, so this was an obvious choice. The less obvious choice was the metric. While some dashboards

Figure 3
Seven-Day Moving Average of New Cases

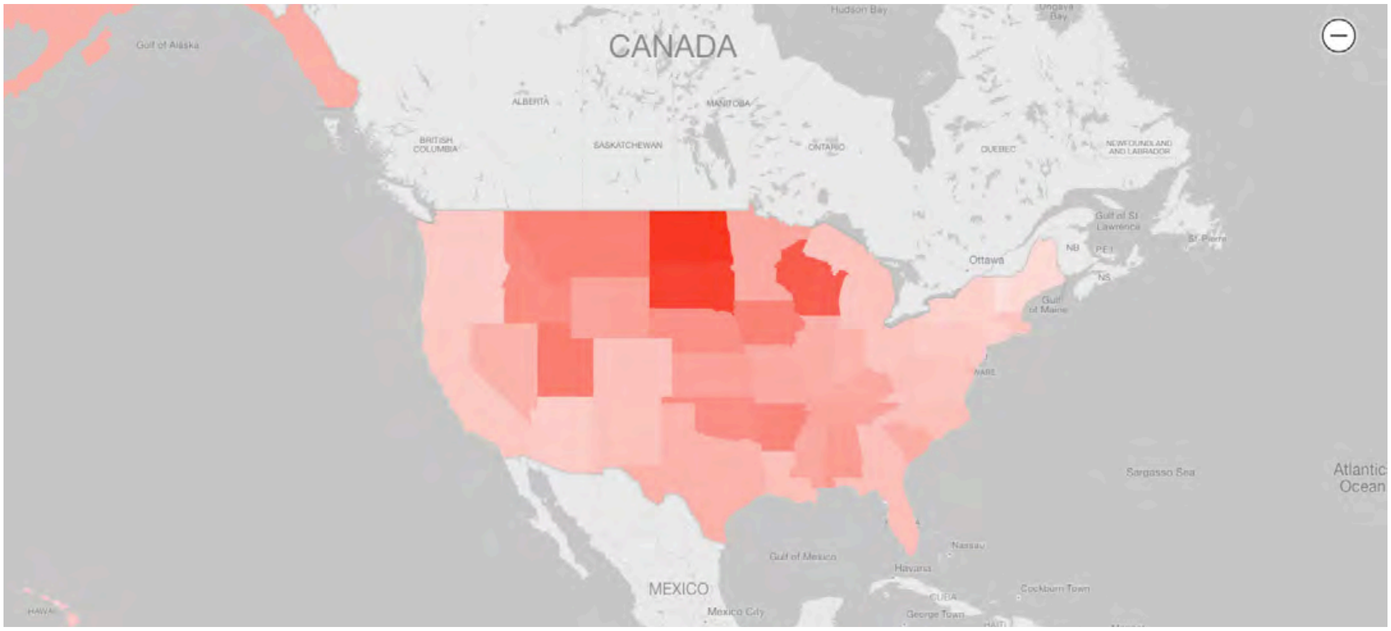
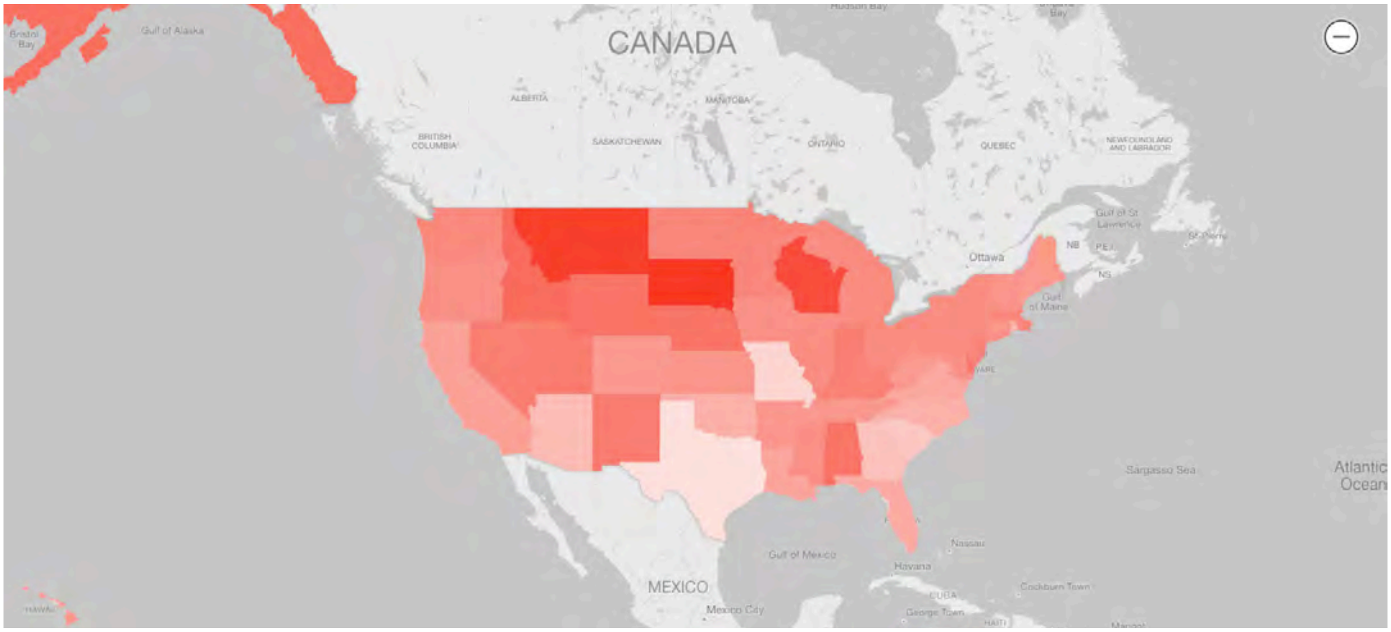


Figure 4
Seven-Day Moving Average of New Cases Week-Over-Week



show the daily change in new cases, we felt that this metric was too sensitive to change and susceptible to large swings following weekends when test volume is typically down. Similarly, weekly change was prone to the same pitfalls in many states where case counts experience weekly seasonality with lower volume during weekends. We eventually found that using a seven-day moving average of new cases per 100k managed to smooth over these types of swings without distorting the direction or strength of new case counts. (See Figure 3.)

By definition, a simple moving average takes the mean of a given set of data over a specific number of time periods in the past. In effect, this minimizes the effect of random fluctuations in the short-term. To truly understand which states should be concerned about becoming a hot spot, we want to be aware of short-term fluctuations. Comparing the most recent seven-day moving average of new cases to the previous seven-day period captures the momentum of new case counts, which we concluded was a more meaningful way of looking at these data. (See Figure 4.)

Figure 5a
7-Day Moving Average of New Cases

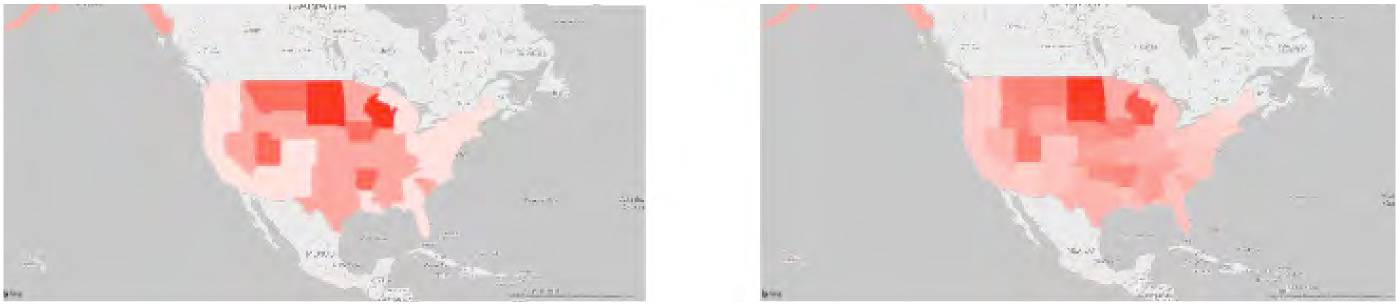
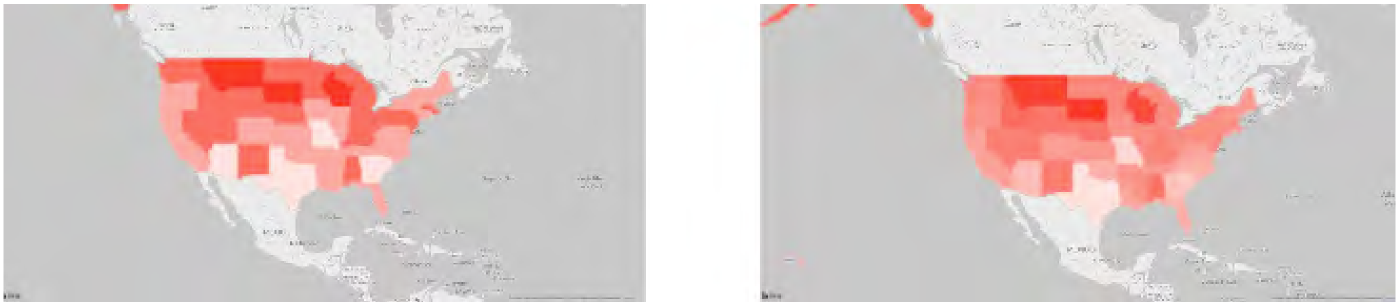


Figure 5b
7-Day Moving Average of New Cases Week-Over-Week



Once we settled on a more useful metric, we wanted to ensure that we were using the best color scheme for the graphic. We started by looking for common heat map color palettes, which led us to "magma," a popular palette in the even more popular Matplotlib Python library. Originally, we simply used Power BI's out-of-the-box color gradient to represent the lowest and highest values with darker colors representing a higher week-over-week change. This, unfortunately, resulted in a muddy graph that failed to highlight the hot spots. After some testing, we found that using a simple gradient with a soft, light pink color as the low value and an intense bright red as the high value clearly highlighted the states that were being impacted the most based on our metric. Even though magma is an awesome color palette, and one that finds its way into many Jupyter notebooks, it felt too intense for the state-level data we are displaying here.

Next, we had to decide if we wanted to continue with the continuous gradient effect, or if we wanted to create color steps by binning the data. Determining the number of steps requires a balancing act between showing too many steps, which can cause the change to seem more drastic than it really is, and showing too few steps, which can have the opposite effect. We landed on using four steps based on both the distribution of the week-over-week change in the seven-day moving average as well as the readability of the map itself. To be fair, both the gradient and the steps for these maps make sense. If we were to show the data at the county level, bins of color would likely be more appropriate as you can more easily discern each small piece of the larger puzzle and where it stands with cases. When it comes to 50 states, the gradient works just fine. (See Figure 5a and 5b.)

It goes without saying that there are many ways to display data. Similar to data analysis, data visualization requires practitioners to journey through a garden of forking paths. From determining what metric to display to deciding on how best to display it, there is no perfect way to design a data visualization. Exploring metrics, graphics, and colors will get you closer to your intended goal, but oftentimes the goal itself can be a moving target. Moreover, the availability and quality of your underlying data may lead you down a rabbit hole of developing metrics or drill downs. For example, the visualizations in this article could benefit from the ability to filter by date, or the ability to drill down to county-level detail. That said, it is imperative to recognize when your visualization has met its intended goal.

We started out with the goal of simply developing a visualization related to COVID-19. Instead, we found ourselves thinking more critically about the entire process. We ultimately presented our journey of using different metrics, their effect on developing an appropriate map, and how to think about making the map as readable as possible without losing critical information in the data. We hope to not only inspire readers to thoroughly consider both their data and purpose when developing any type of visualization, but also to recognize the need to be flexible throughout the process. Being conscious of your underlying data, the potential for errata, and how your data is prepared are important steps in telling the data's story. An even more important step is to tell the story.

LIMITATIONS

This article is written based on data available as of the time of writing. The underlying data and circumstances related to COVID-19 are subject to uncertainty, given the emerging experience of the pandemic. These data are dependent on many factors at the state and local levels such as shelter in place orders, availability of testing, and reporting agency capabilities. The visualizations presented herein are based on past data and are not projections of the future. ■

This article is intended only for the purpose of approaching a data visualization task and should not be used for other purposes.

Phil Ellenberg is a healthcare consultant with Milliman. He can be reached at *phil.ellenberg@milliman.com.*

Kelsie Gosser is a user experience (UX) engineer with Milliman. She can be reached at *kelsie.gosser@milliman.com.*

ENDNOTES

[1] *https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv*