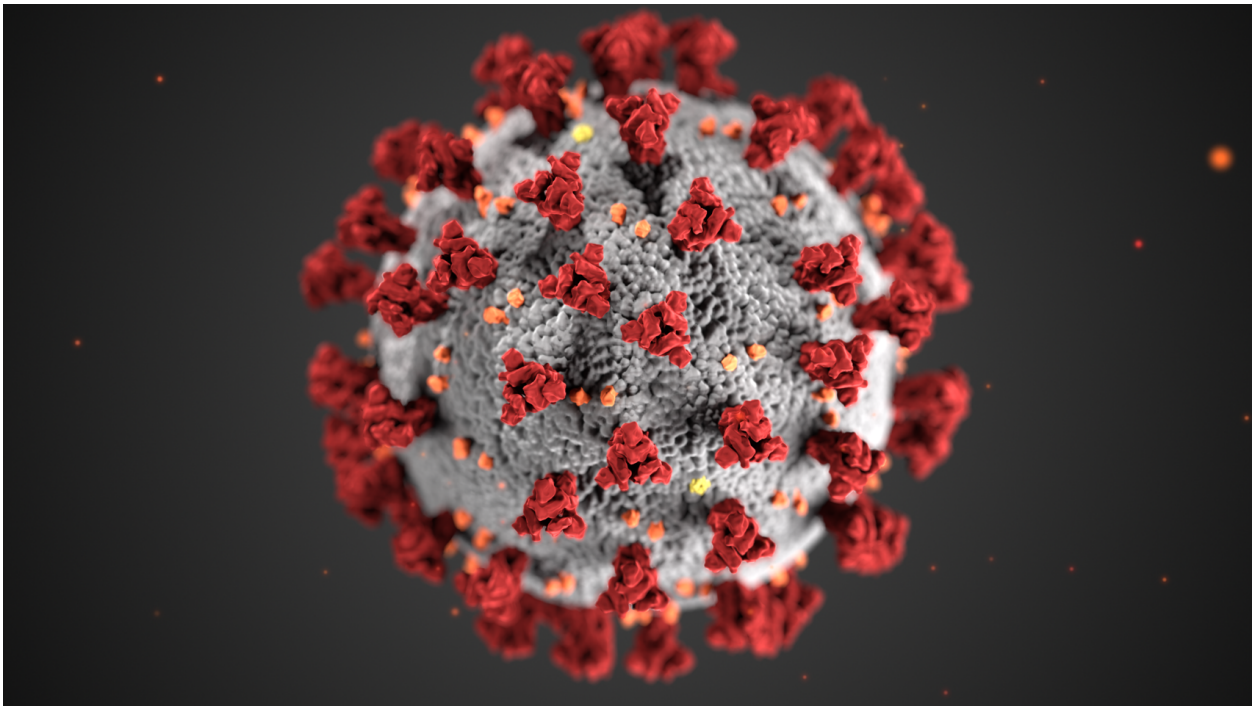
COVID-19 Risk Modeling

June 02, 2020



Background

The U.S. country, state, and county health data used for this analysis is pulled directly from the New York Times COVID-19 GitHub repository (https://github.com/nytimes/covid-19-data) using `git pull` requests. The predictive analytics model is constructed in the R programming language using RStudio and the Tidyverse family of packages.

> The New York Times is releasing a series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time. We are compiling this time series data from state and local governments and health departments in an attempt to provide a complete record of the ongoing outbreak.
>
> Since late January, The Times has tracked cases of coronavirus in real time as they were identified after testing. Because of the widespread shortage of testing, however, the data is necessarily limited in the picture it presents of the outbreak.
>
> We have used this data to power our maps and reporting tracking the outbreak, and it is now being made available to the public in response to requests from researchers, scientists and government officials who would like access to the data to better understand the outbreak.
>
> The data begins with the first reported coronavirus case in Washington State on Jan. 21, 2020. We will publish regular updates to the data in this repository.

Data Analysis

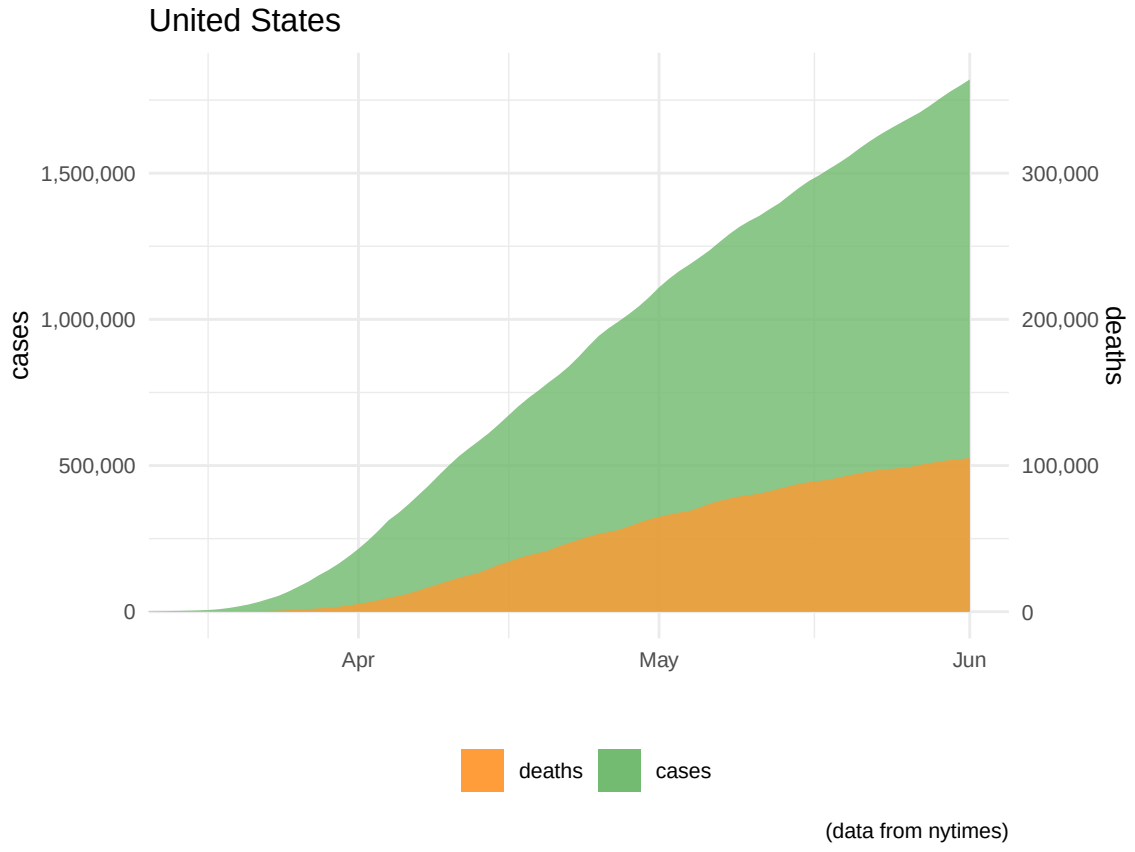
The COVID-19 data in the New York Times GitHub repository is structured as three main comma-separated value data files—one top-level country summary file, one state-level summary file, and one data file containing reported case and death data for each individual U.S. county. Each of these is used for this analysis. The data from each of these files is used to calculate the rate of reported new cases and deaths for each state and county, and these rates are used to build a predictive model by linear regression using least-squares methods for each entity. A risk estimate is generated from these models, and the states and counties with the highest estimated risk are compared in the charts shown in this document. In the charts showing new reported cases and deaths, a generalized additive model (GAM) smoothing function was fit to each data set.

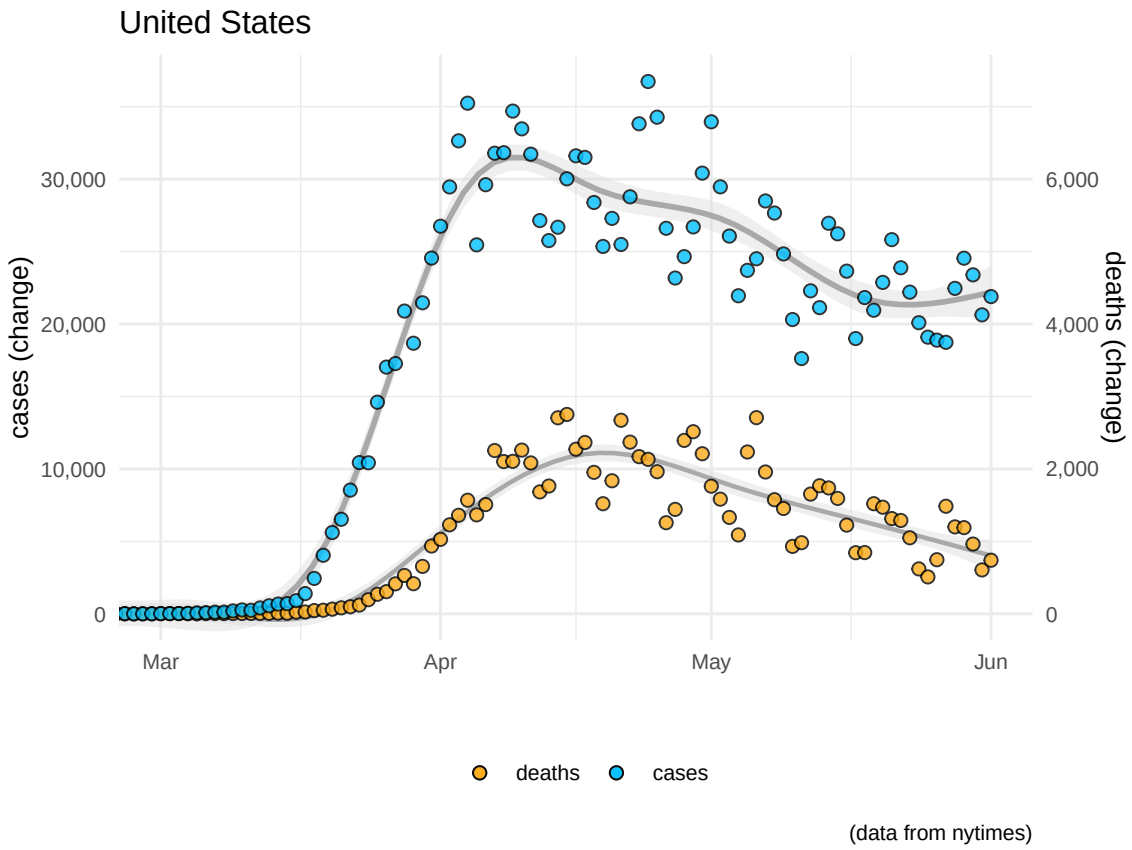
The risk assessment methodology used in this analysis has not been validated and is subject to noise in the data. There is a phenomenon that has been reported in the White House press briefings about the COVID-19 response whereby some counties report updates to the county data on Mondays for the incremental changes over the weekend. This will negatively affect the accuracy of the model to some degree. To enable more robustness in the risk estimation algorithm, data over a 10-day period us used as a compromise between speed of detection of a change in risk and errors due to high sensitivity to noise in the data.

Summary Results

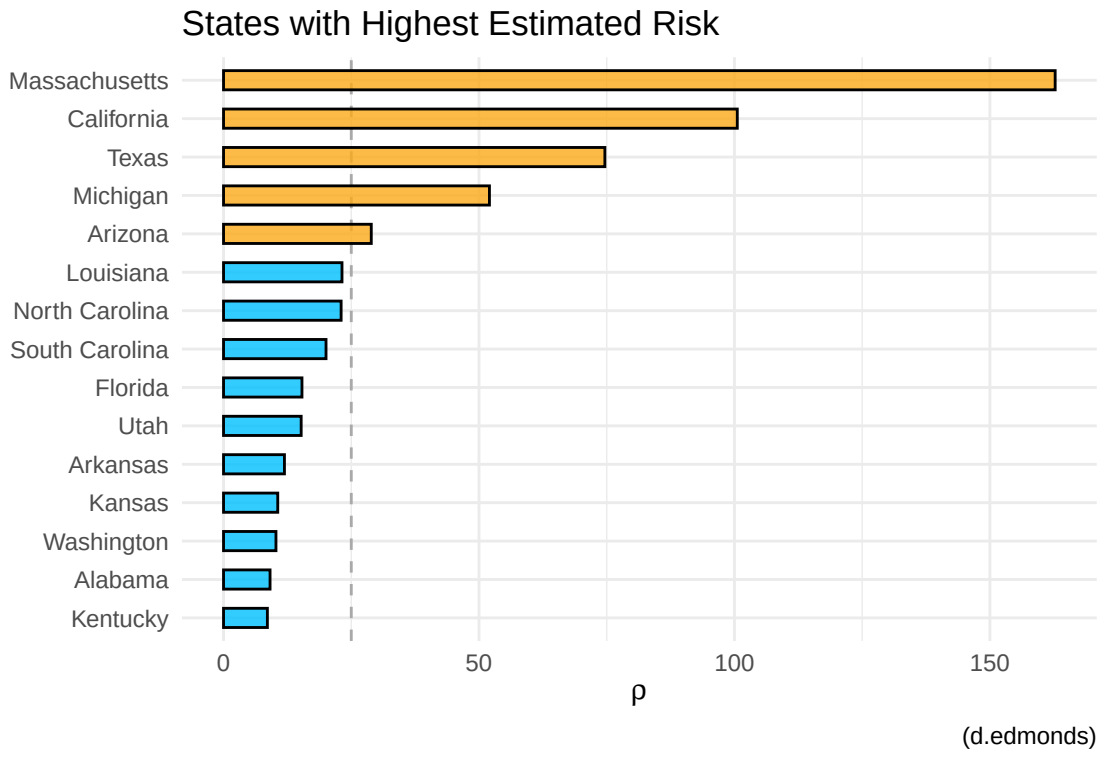
United States

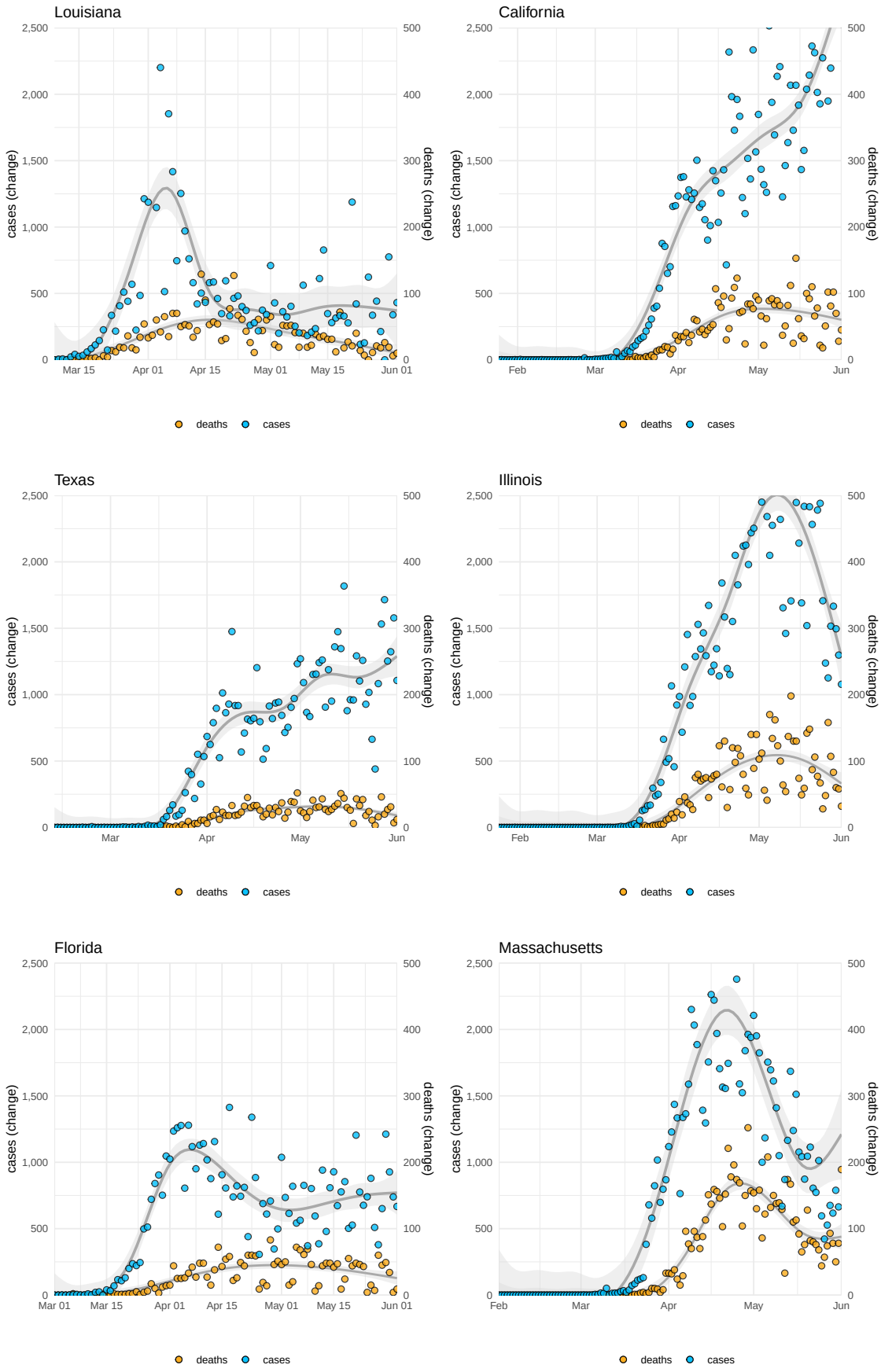
There have been **1,820,808** COVID-19 cases (21,894 new cases per day) and **105,124** deaths (740 new deaths per day) in the United States.





Individual States




Louisiana
California
Texas
Illinois
Florida
Massachusetts
cases (change)
deaths (change)
deaths
cases
Mar 01
Mar 15
Apr 01
Apr 15
May 01
May 15
Jun 01
Feb
Mar
Apr
May
Jun

Individual Counties