Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-07-04

Contents

Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by a new type of coronavirus: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak first started in Wuhan, China in December 2019. The first kown case of COVID-19 in the U.S. was confirmed on January 20, 2020, in a 35-year-old man who teturned to Washington State on January 15 after traveling to Wuhan. Starting around the end of Feburary, evidence emerge for community spread in the US.

We, as all of us, are indebted to the heros who fight COVID-19 across the whole world in different ways. For this data exploration, I am grateful to many data science groups who have collected detailed COVID-19 outbreak data, including the number of tests, confirmed cases, and deaths, across countries/regions, states/provnices (administrative division level 1, or admin1), and counties (admin2). Specifically, I used the data from these three resources:

- JHU (https://coronavirus.jhu.edu/)
  - The Center for Systems Science and Engineering (CSSE) at John Hopkins University.
  - World-wide counts of coronavirus cases, deaths, and recovered ones.
  - https://github.com/CSSEGISandData/COVID-19
- NY Times (https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html)
  - The New York Times
  - "cumulative counts of coronavirus cases in the United States, at the state and county level, over time"
  - https://github.com/nytimes/covid-19-data

- COVID Trackng (https://covidtracking.com/)
  - COVID Tracking Project
  - "collects information from 50 US states, the District of Columbia, and 5 other US territories to provide the most comprehensive testing data"
  - https://github.com/COVID19Tracking/covid-tracking-data

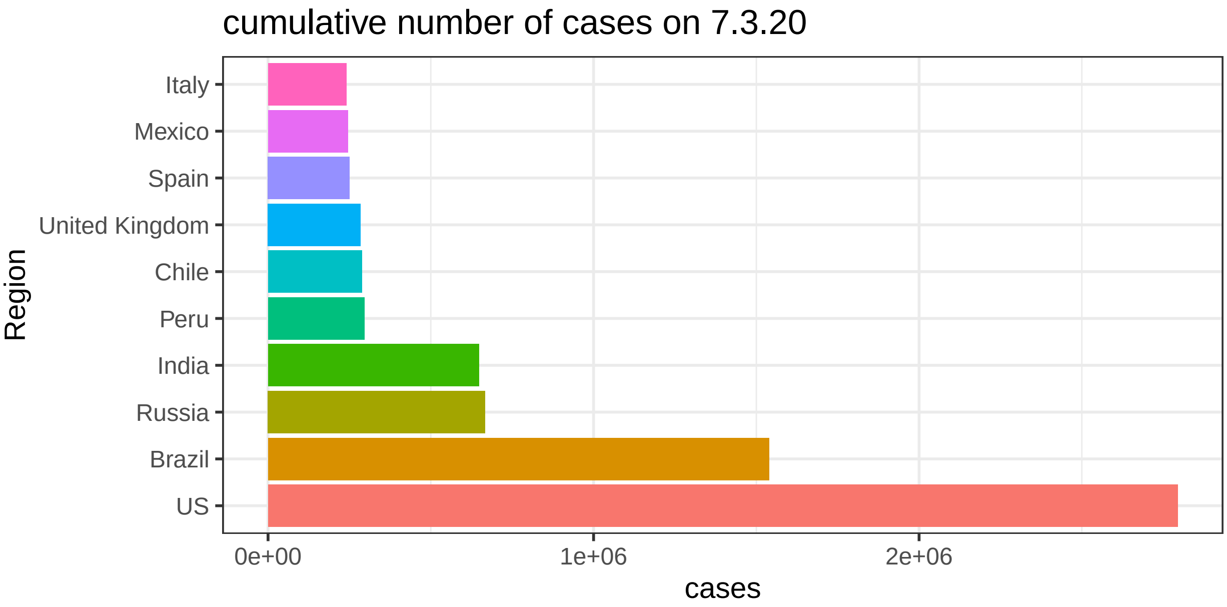
JHU

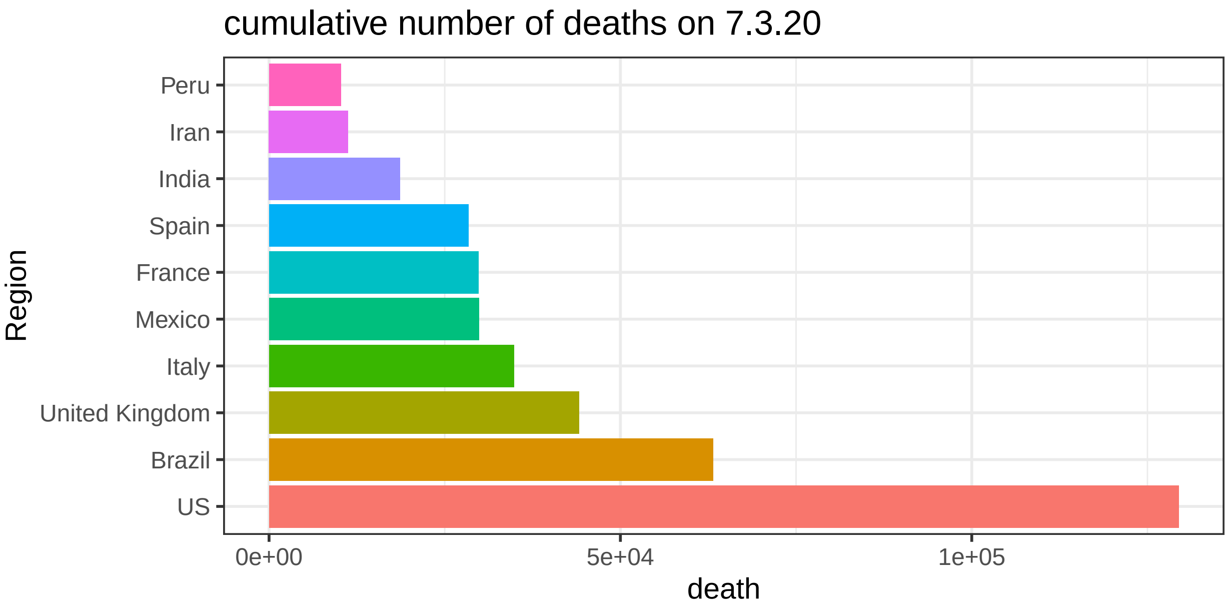
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

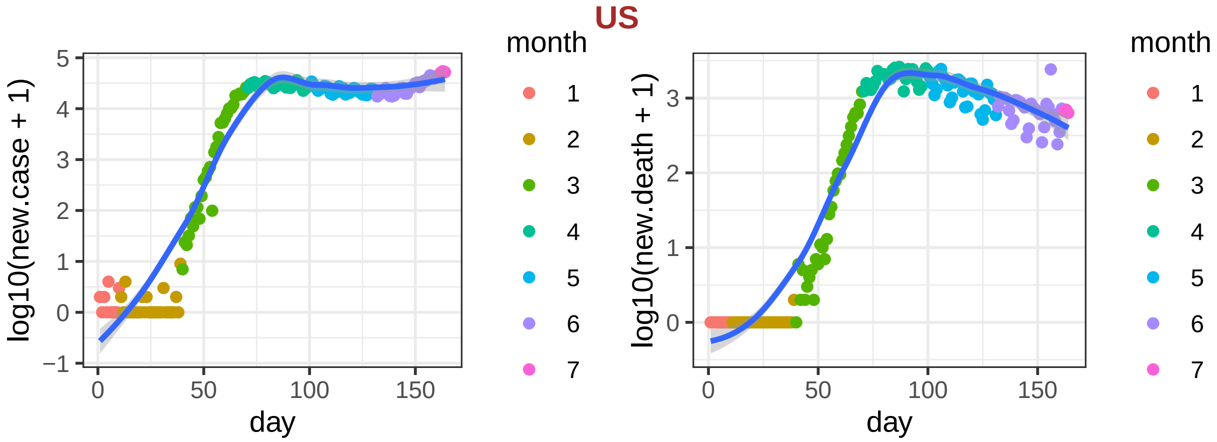
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

Here is the list of 10 records with the largest number of cases or deaths on the most recent date.

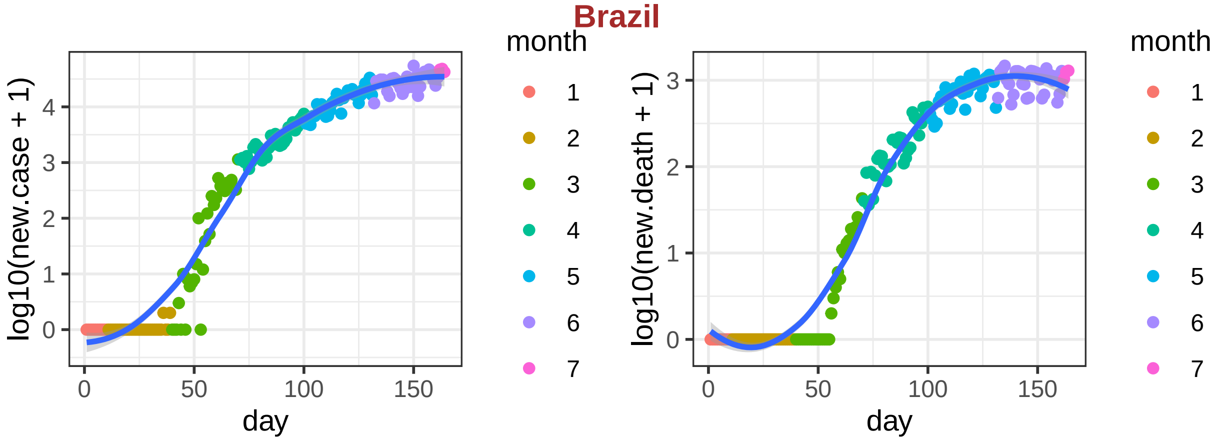



cumulative number of deaths on 7.3.20

Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

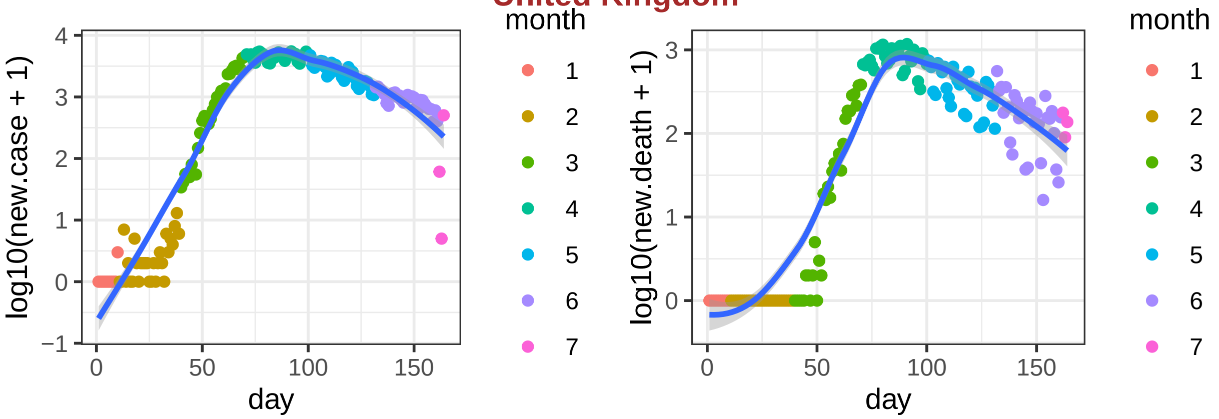

US

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020


Brazil

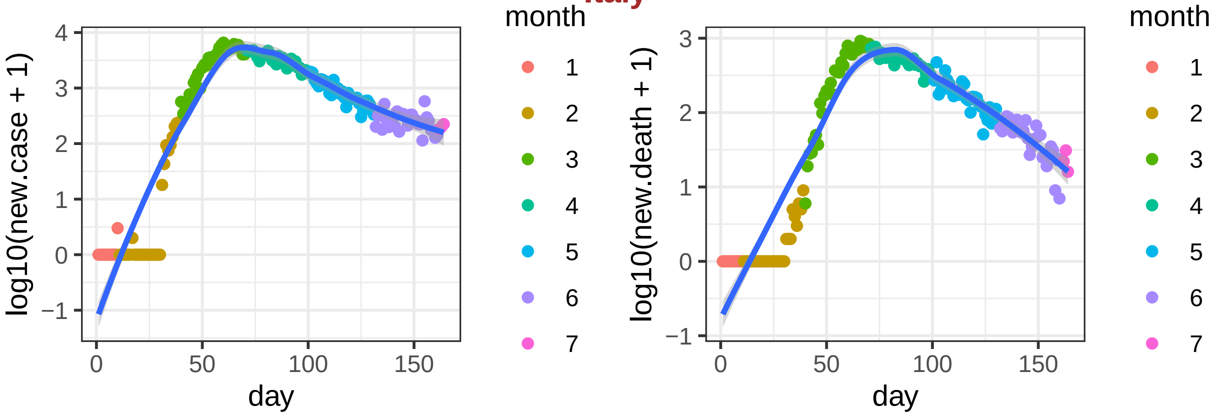
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



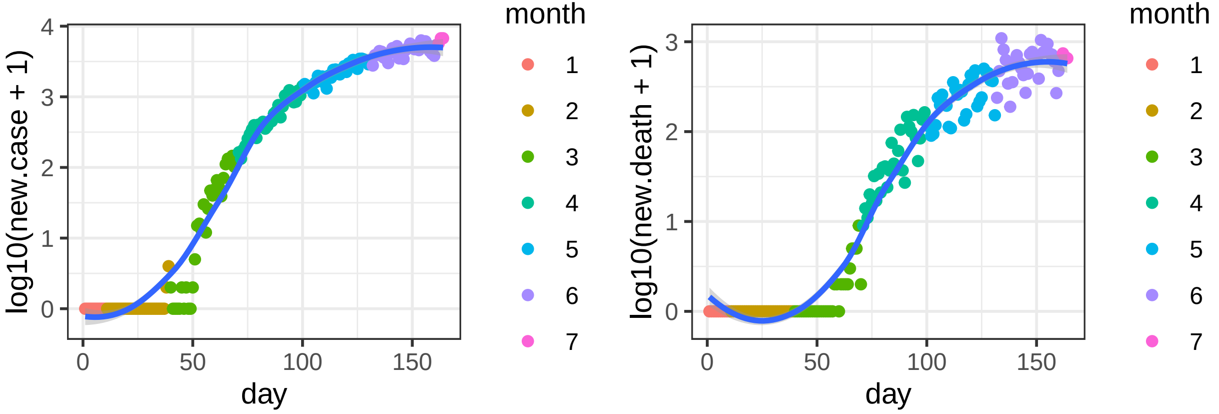
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



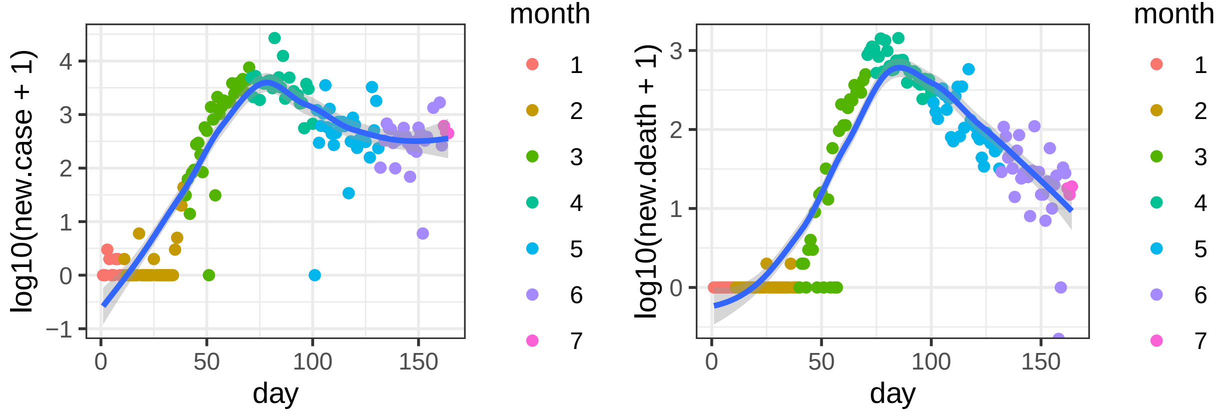
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico



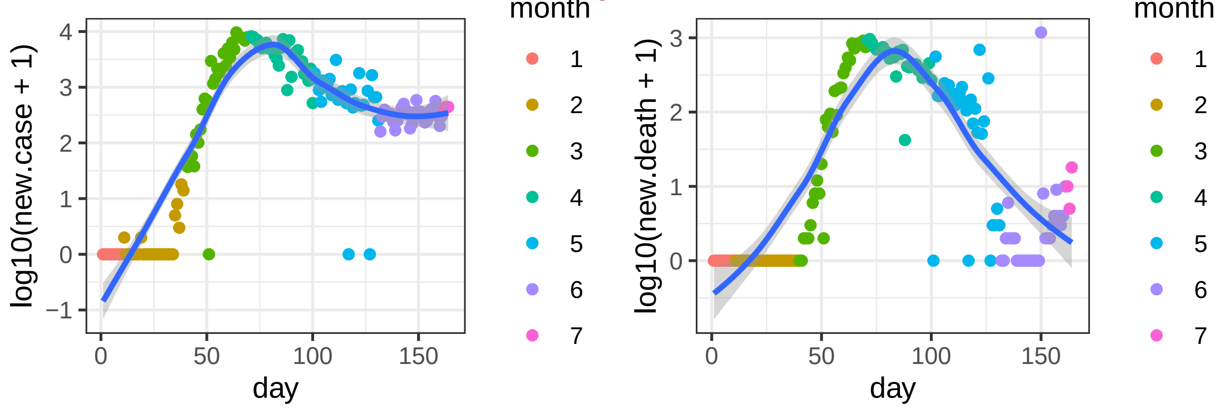
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



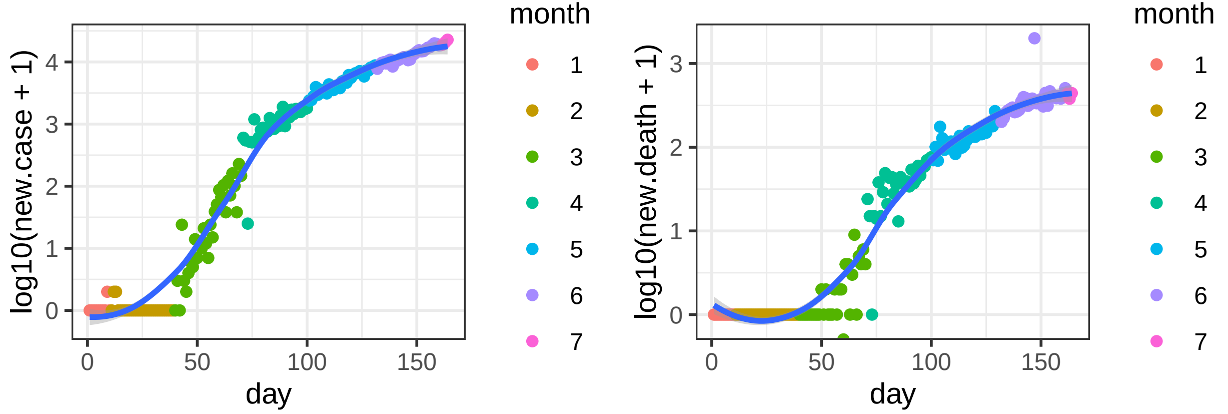
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain

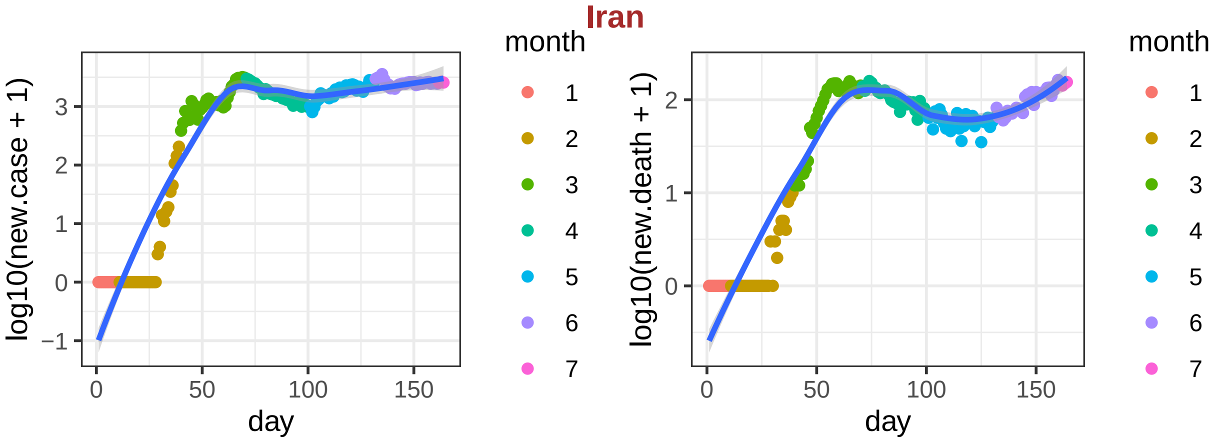


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

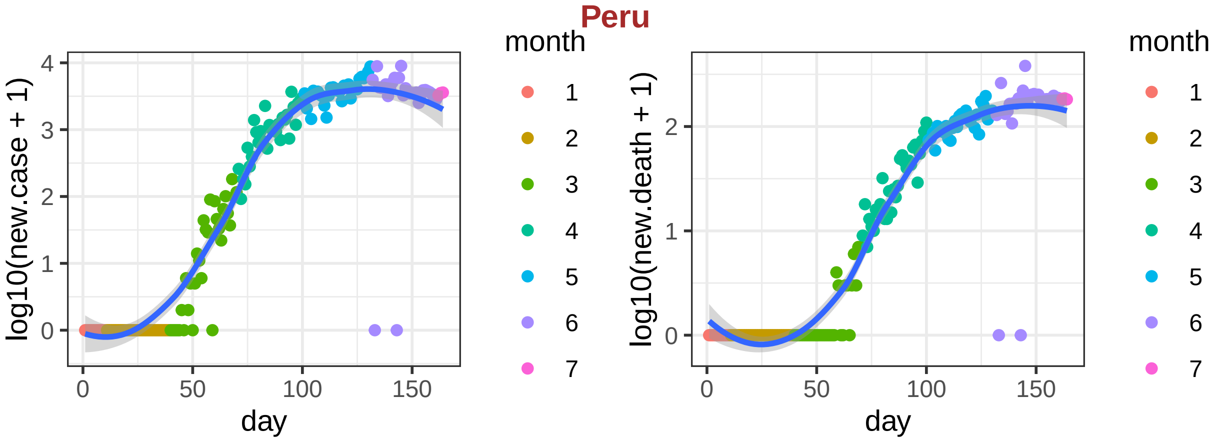
India



data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



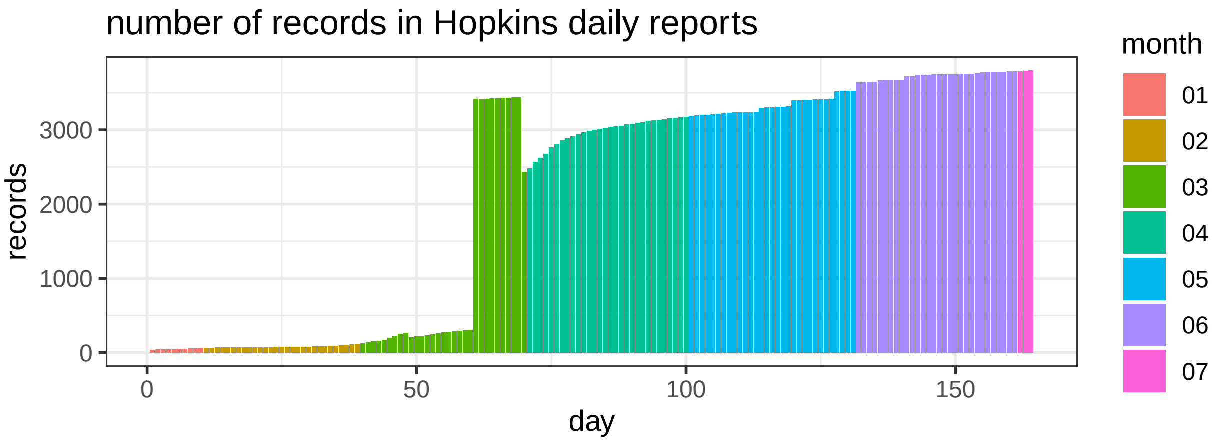
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

daily reports data

The raw data from Hopkins are in the format of daily reports with one file per day. More recent files (since March 22nd) inlcude information from individual states of US or individual counties, as shown in the following figure. So I turn to NY Times data for informatoin of individual states or counties.



data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

NY Times

The data from NY Times are saved in two text files, one for state level information and the other one for county level information.

The currente date is

```
## [1] "2020-07-03"
```

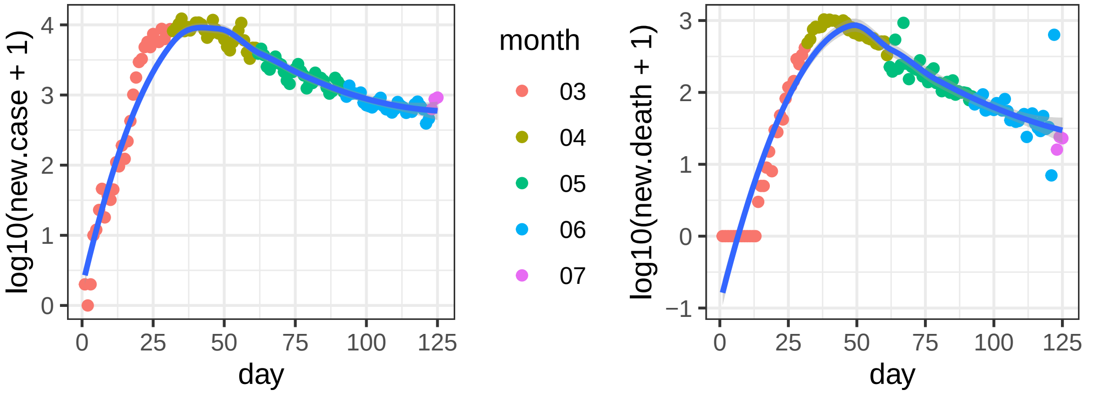
state level data

First check the 30 states with the largest number of deaths.

```
##            date          state fips  cases deaths
## 6758 2020-07-03       New York   36 400561  31836
## 6756 2020-07-03     New Jersey   34 174598  15164
## 6747 2020-07-03  Massachusetts   25 109628   8149
## 6739 2020-07-03       Illinois   17 147293   7222
## 6765 2020-07-03   Pennsylvania   42  93418   6790
## 6729 2020-07-03     California    6 253655   6315
## 6748 2020-07-03       Michigan   26  72306   6219
## 6731 2020-07-03    Connecticut    9  46717   4335
## 6734 2020-07-03        Florida   12 178586   3683
## 6744 2020-07-03      Louisiana   22  63397   3278
## 6746 2020-07-03       Maryland   24  69422   3223
## 6762 2020-07-03           Ohio   39  55257   2903
## 6735 2020-07-03        Georgia   13  85079   2808
## 6740 2020-07-03        Indiana   18  48099   2682
## 6771 2020-07-03          Texas   48 188834   2601
## 6775 2020-07-03       Virginia   51  64393   1845
## 6727 2020-07-03        Arizona    4  91894   1801
## 6730 2020-07-03       Colorado    8  33607   1701
## 6749 2020-07-03      Minnesota   27  37661   1503
## 6759 2020-07-03 North Carolina   37  70447   1413
## 6776 2020-07-03     Washington   53  36417   1353
## 6750 2020-07-03    Mississippi   28  29684   1103
## 6751 2020-07-03       Missouri   29  23620   1060
## 6725 2020-07-03        Alabama    1  41865   1006
## 6767 2020-07-03   Rhode Island   44  16991    960
## 6778 2020-07-03      Wisconsin   55  33565    804
## 6768 2020-07-03 South Carolina   45  41532    793
## 6741 2020-07-03           Iowa   19  30631    721
## 6770 2020-07-03      Tennessee   47  47882    626
## 6743 2020-07-03       Kentucky   21  16655    608
```

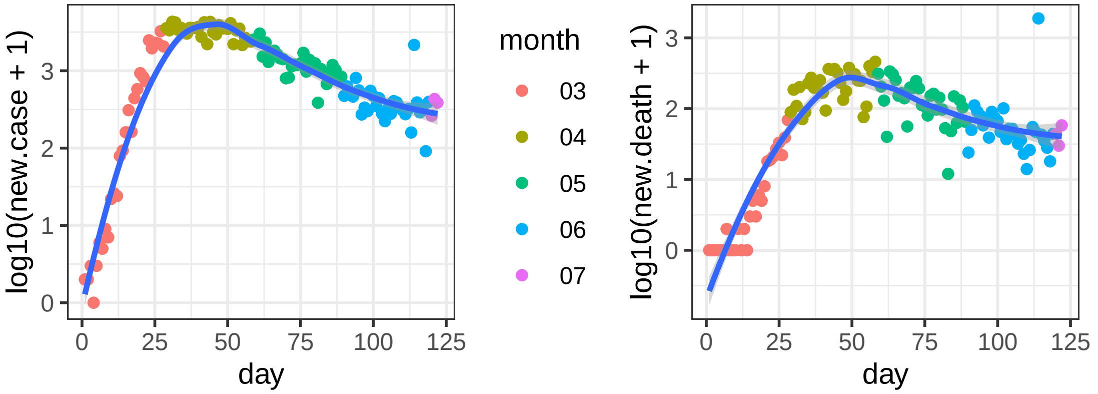
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.

New York



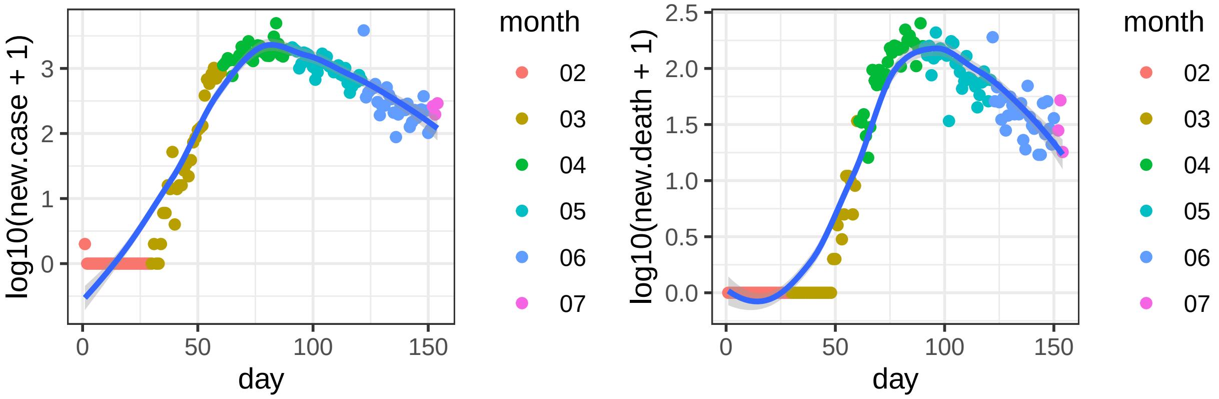
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



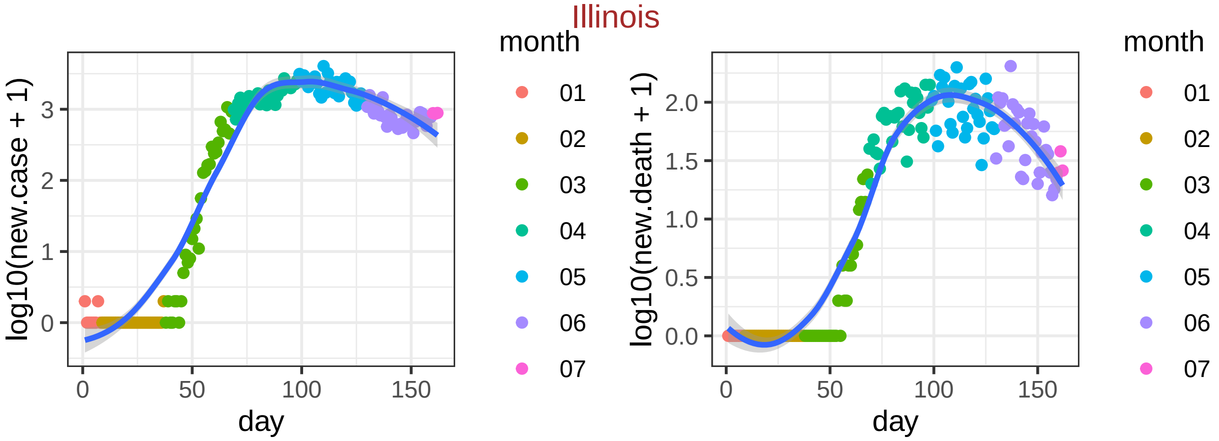
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



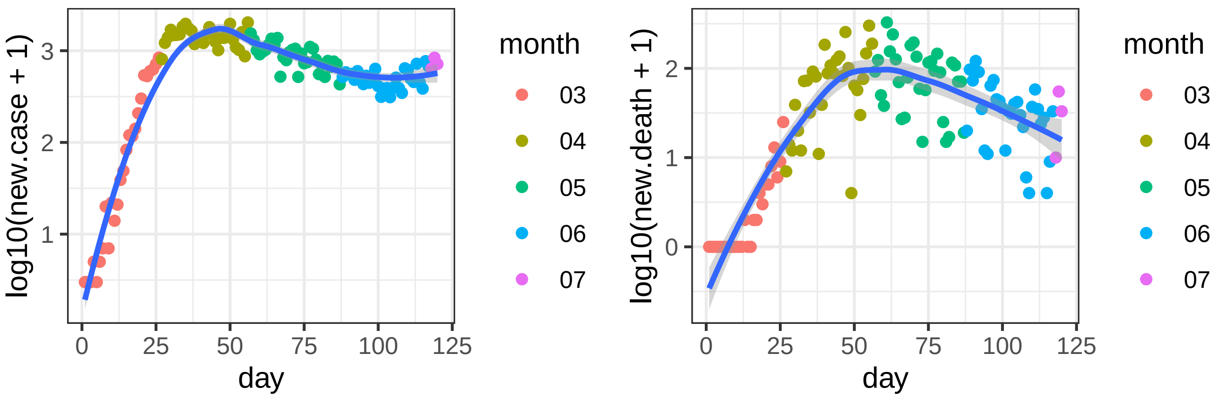
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



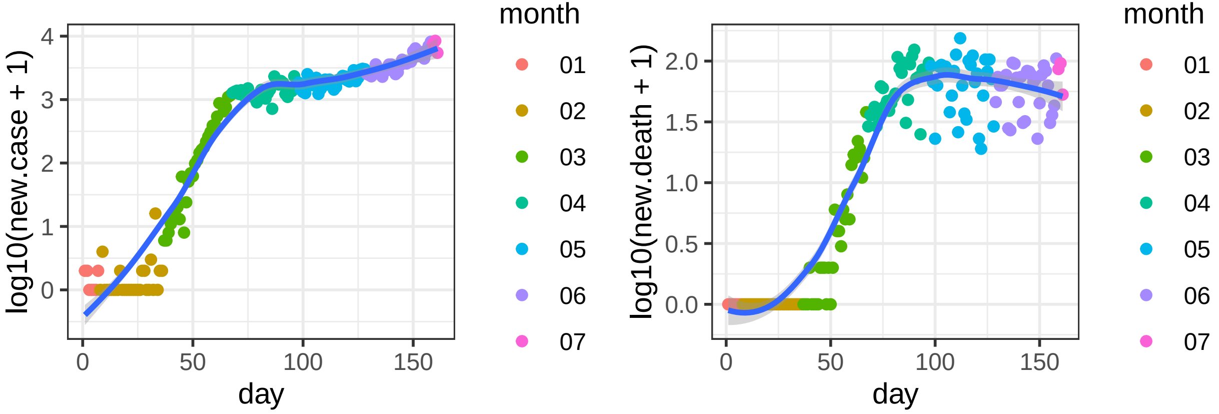
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Pennsylvania



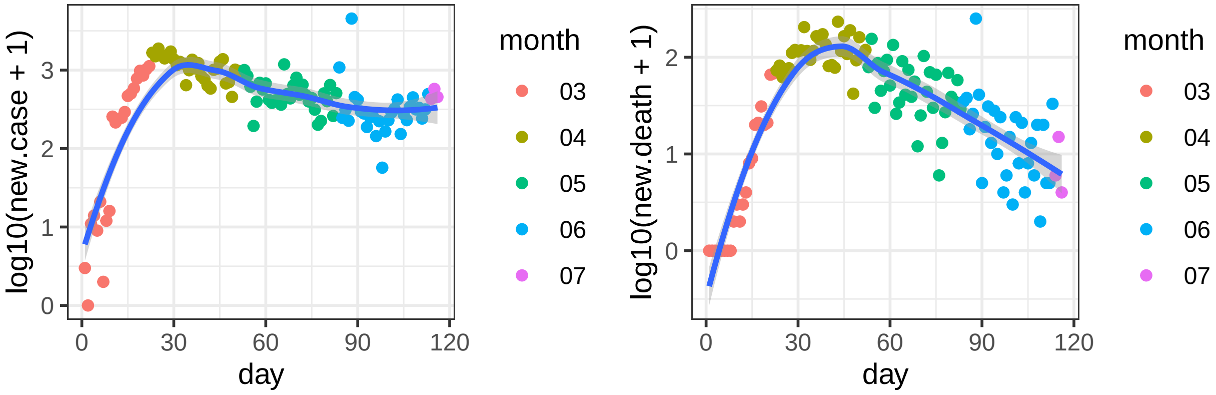
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

California



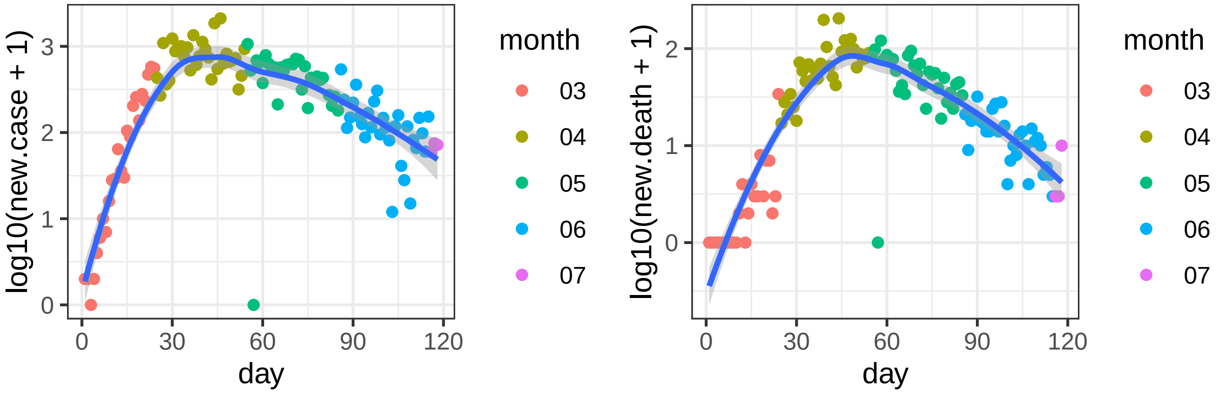
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Michigan



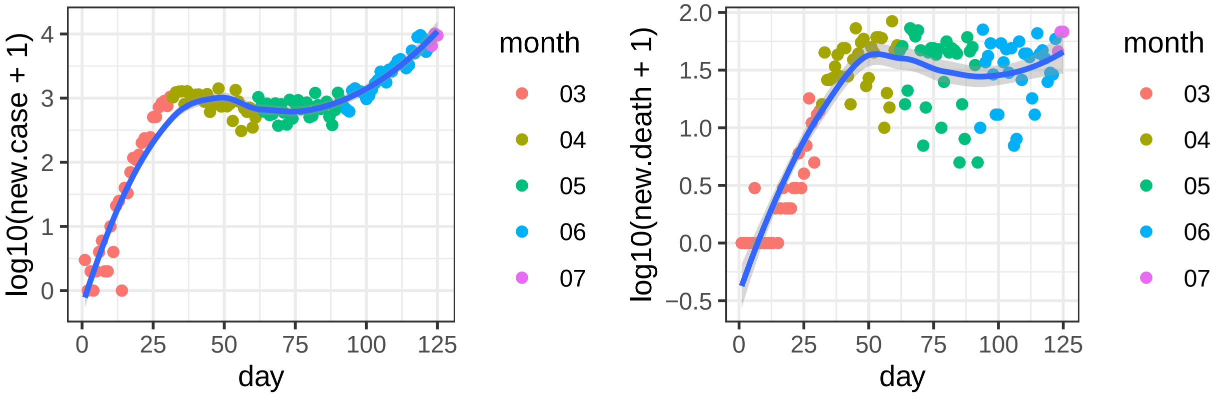
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Connecticut



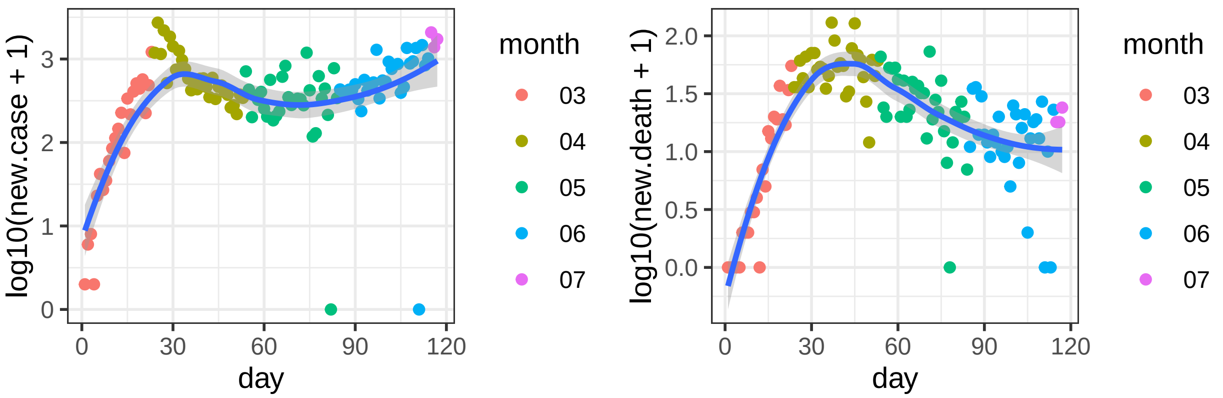
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Florida



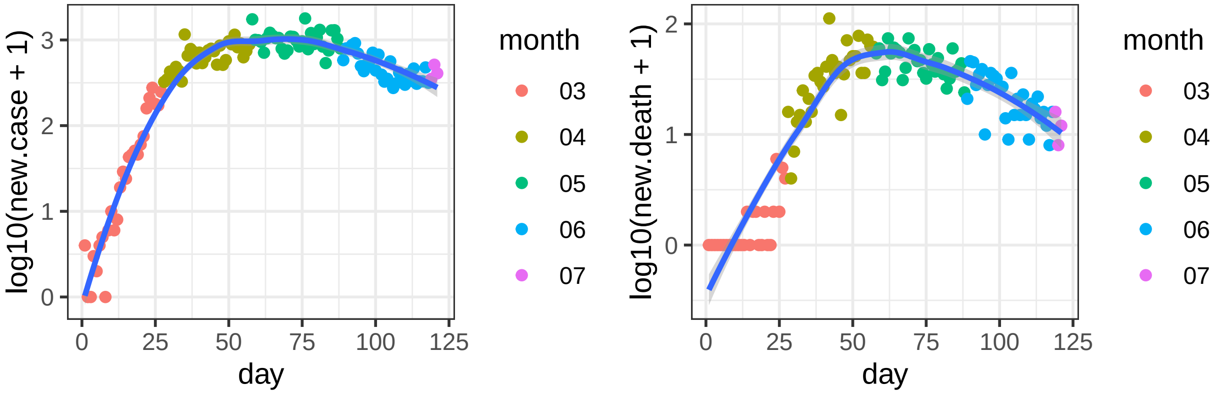
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Louisiana



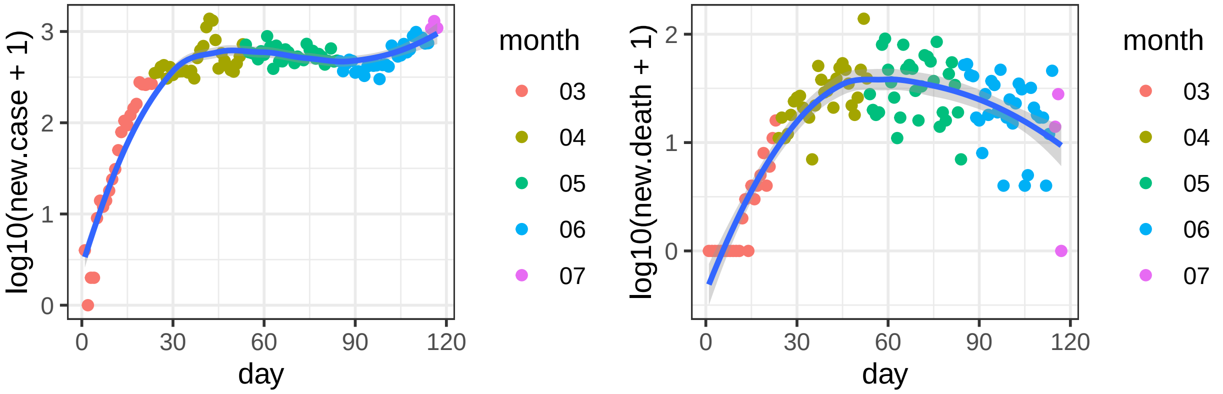
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



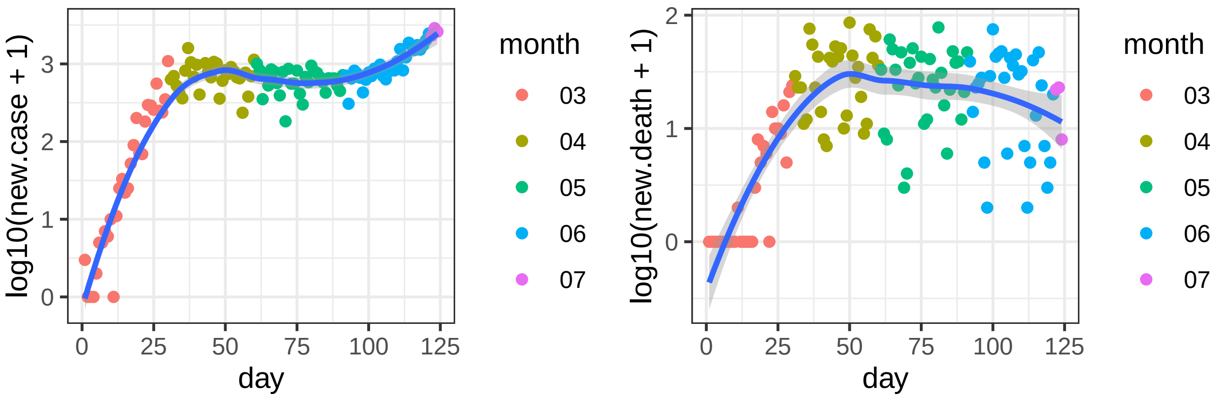
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Ohio



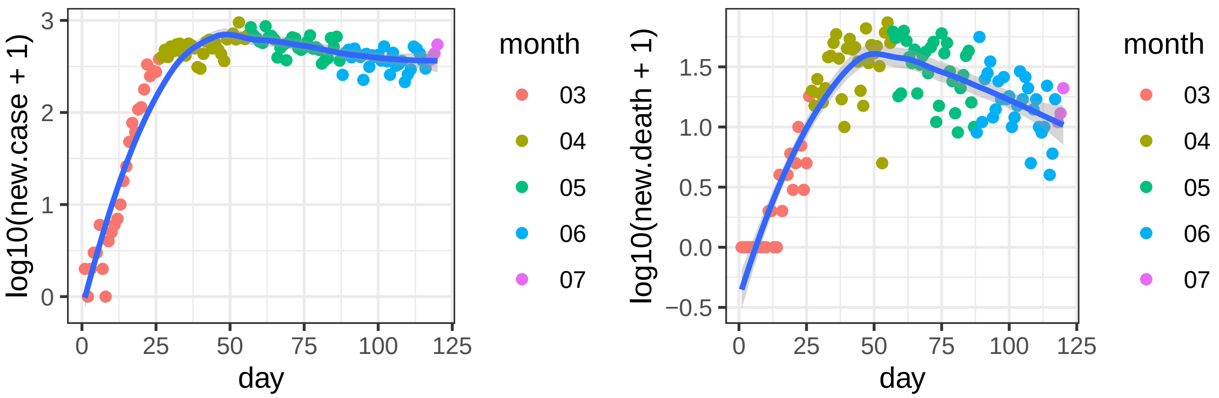
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Georgia



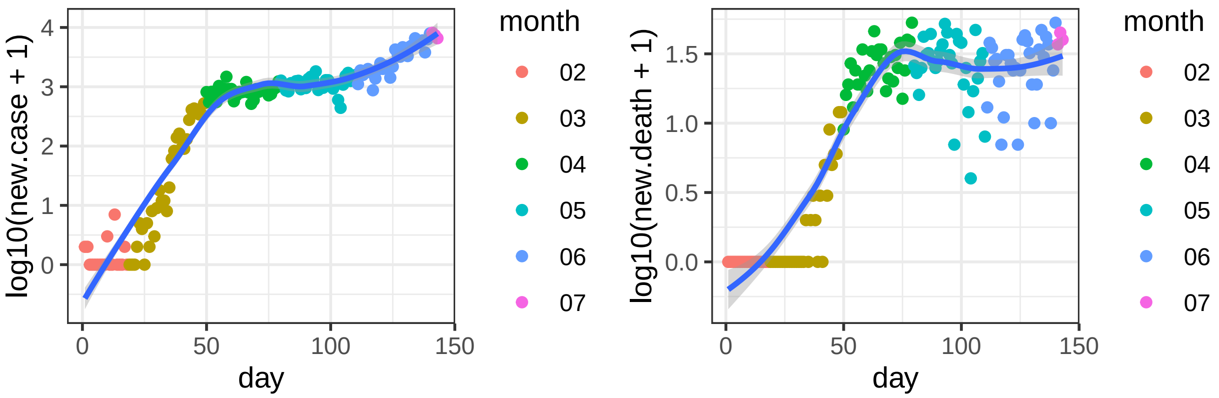
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Indiana

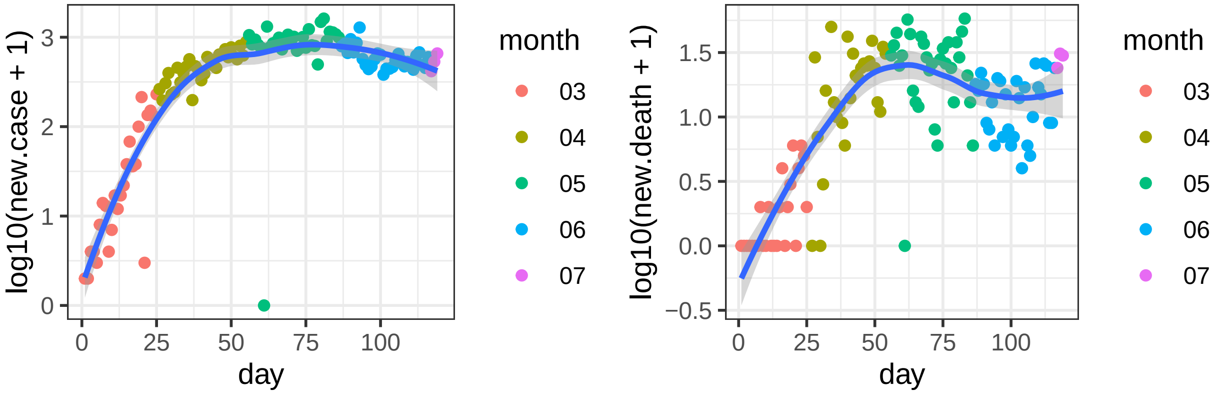


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Texas

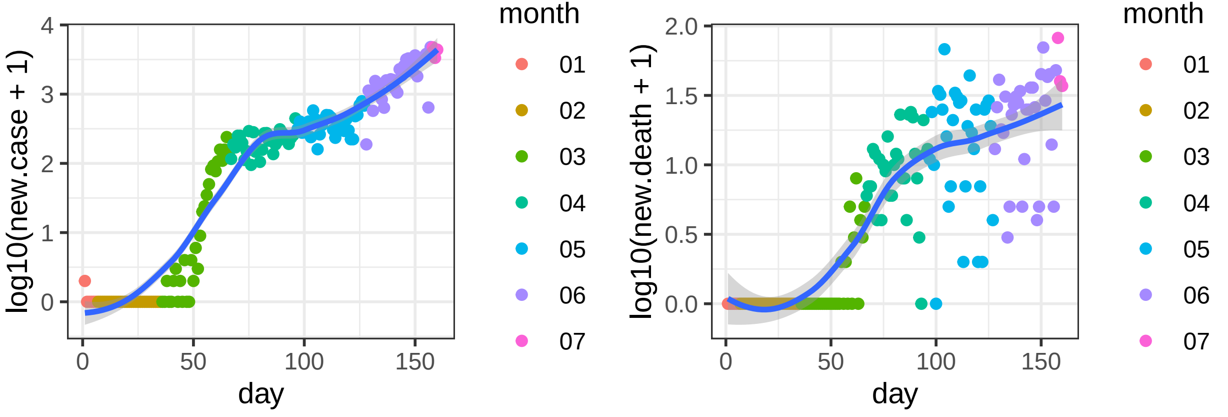


data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12



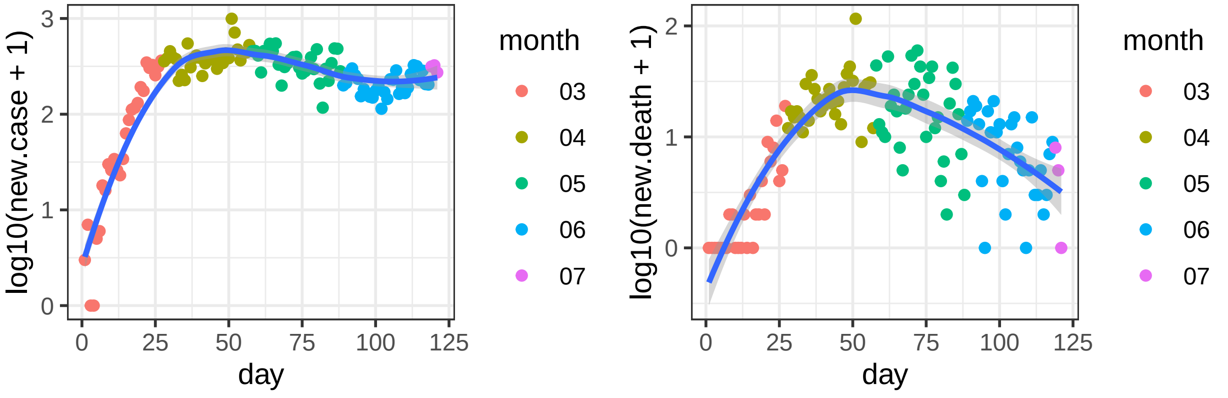
Virginia

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



Arizona

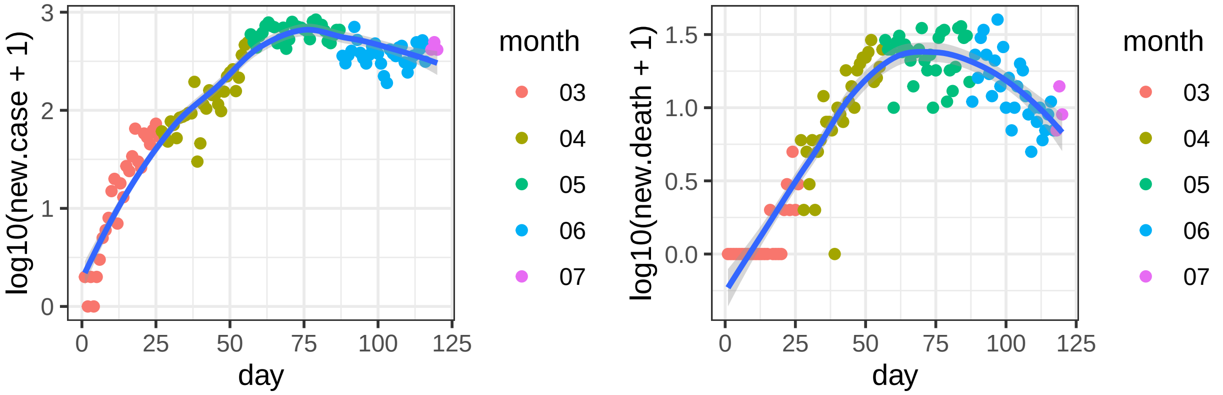
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26



Colorado

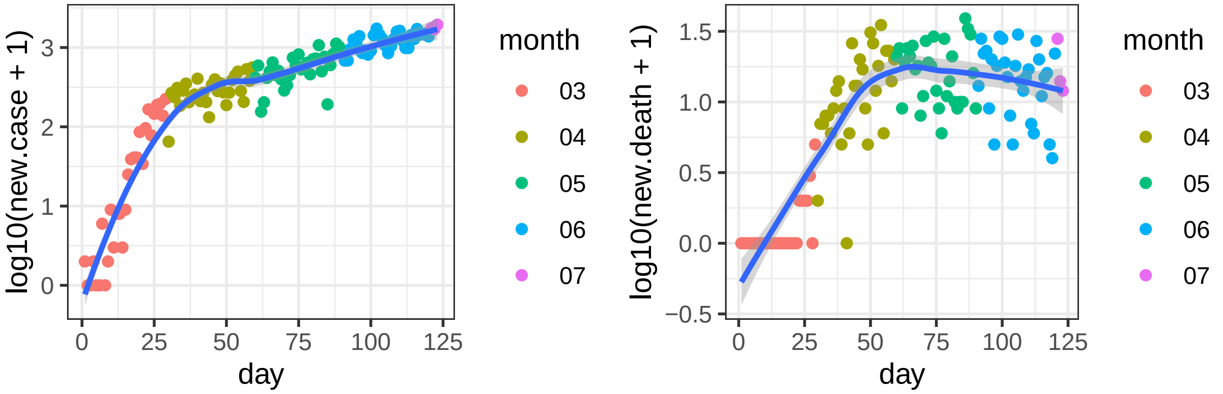
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Minnesota



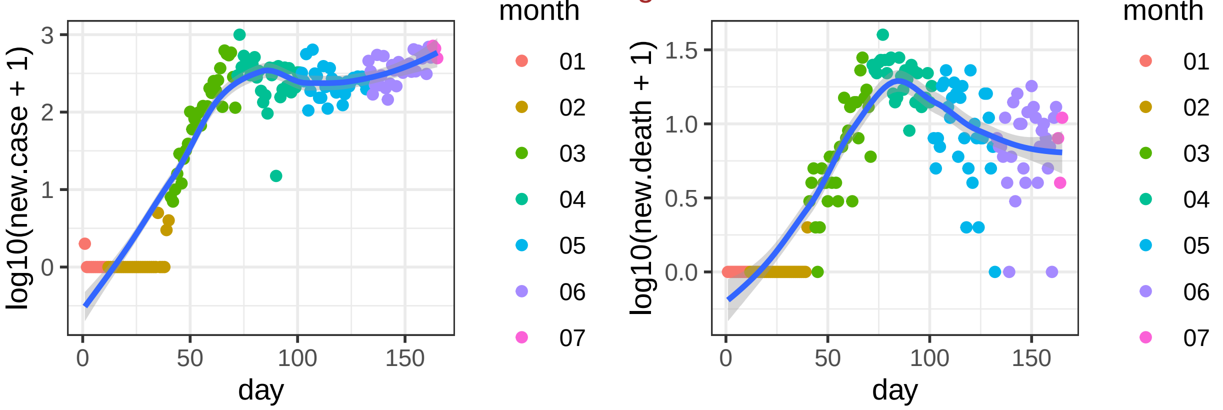
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

North Carolina



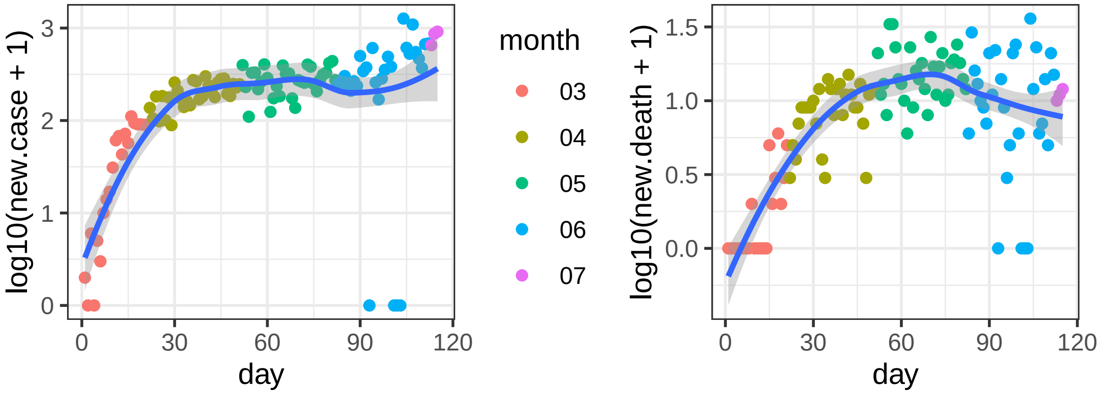
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Washington



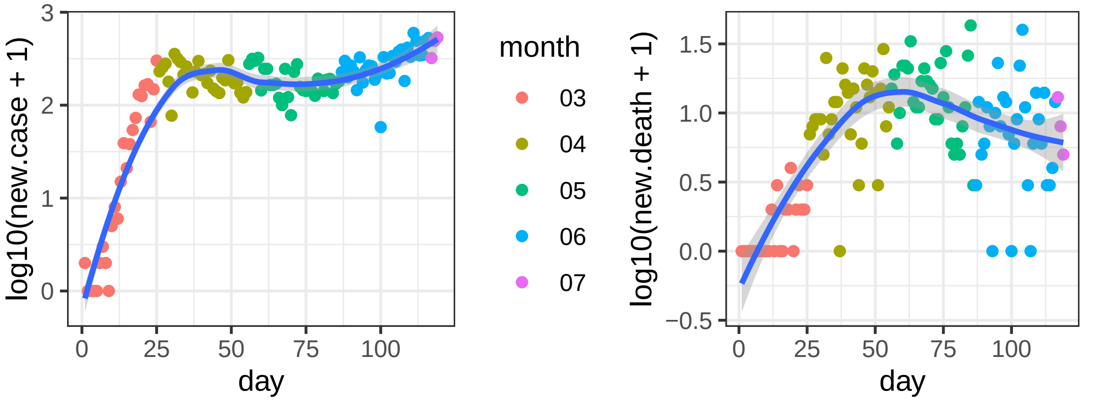
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Mississippi



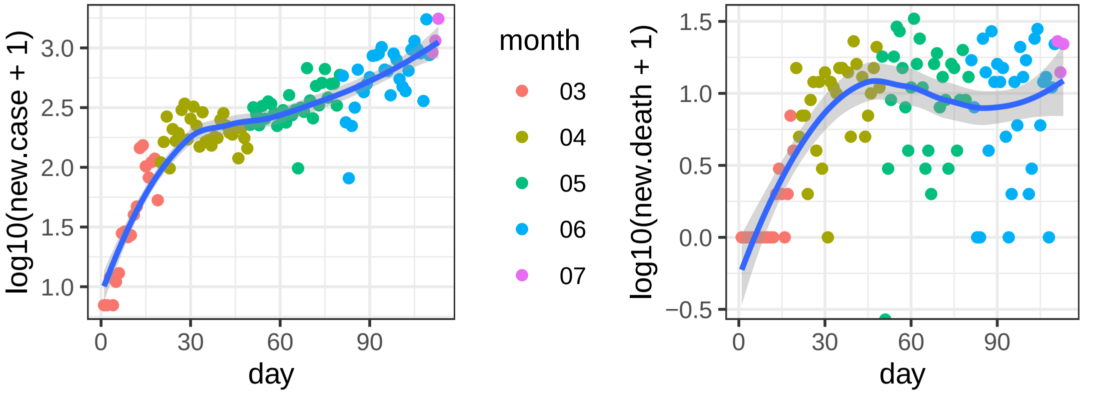
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Missouri



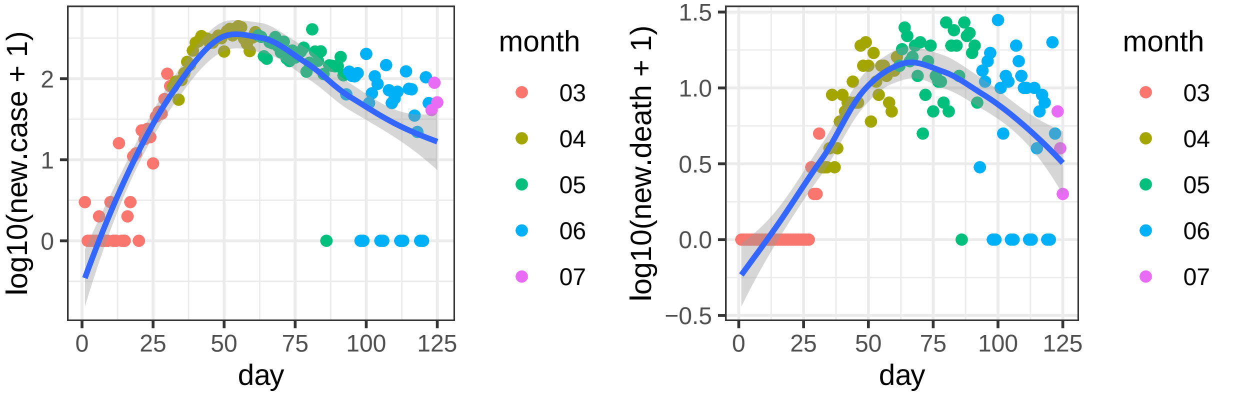
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Alabama



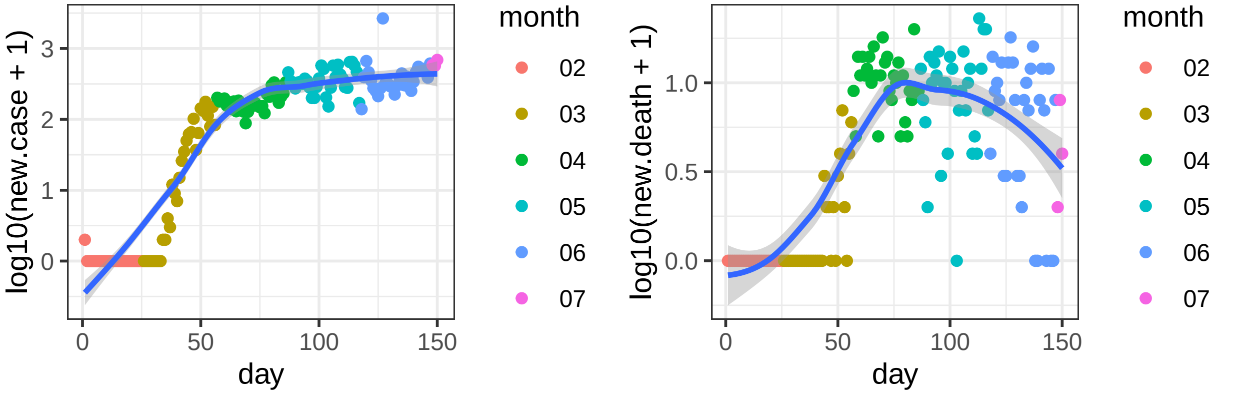
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Rhode Island



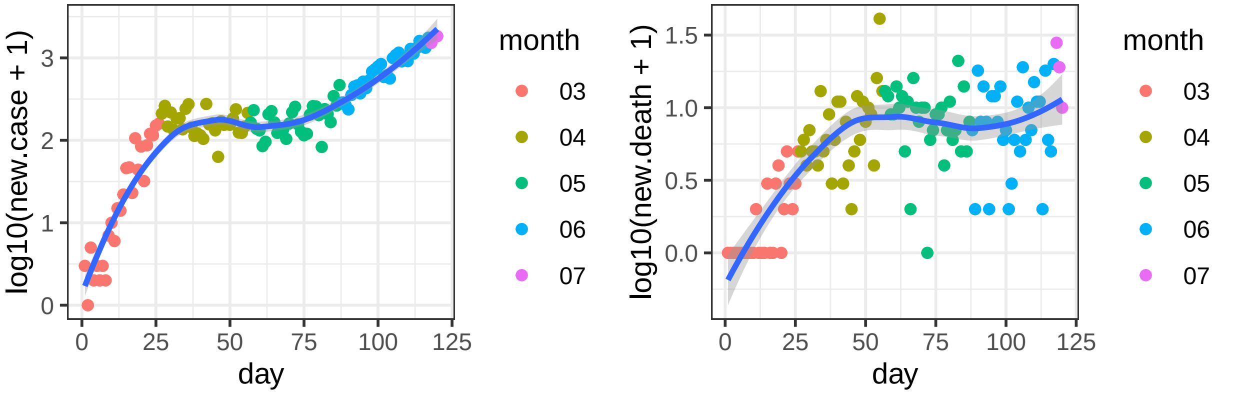
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wisconsin

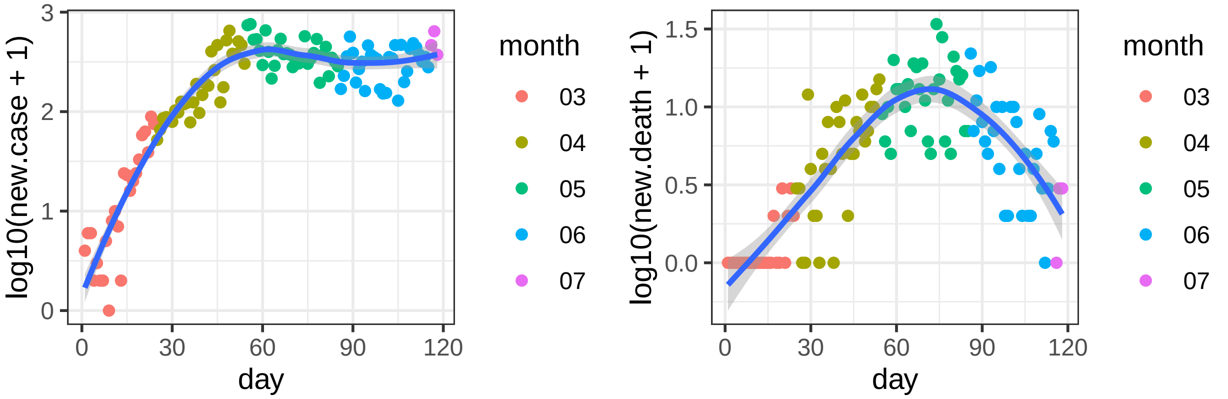


data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

South Carolina

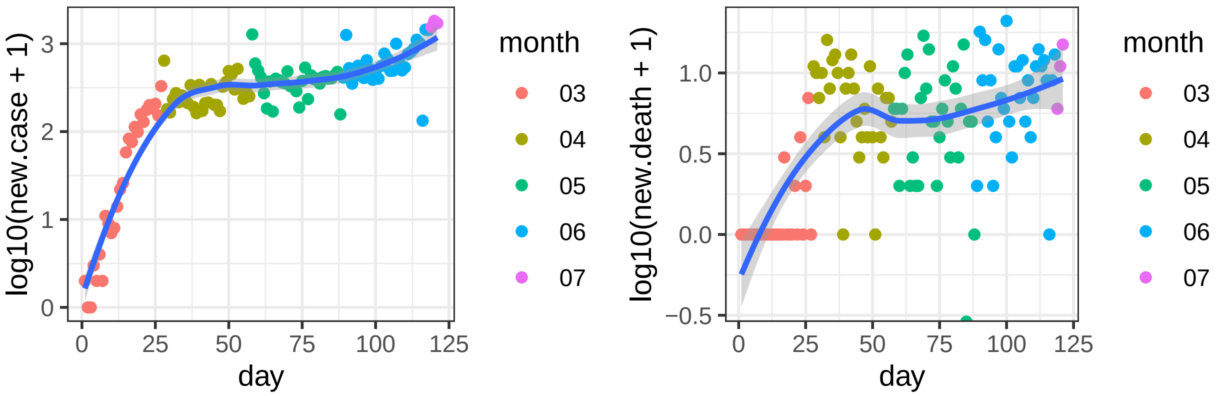


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



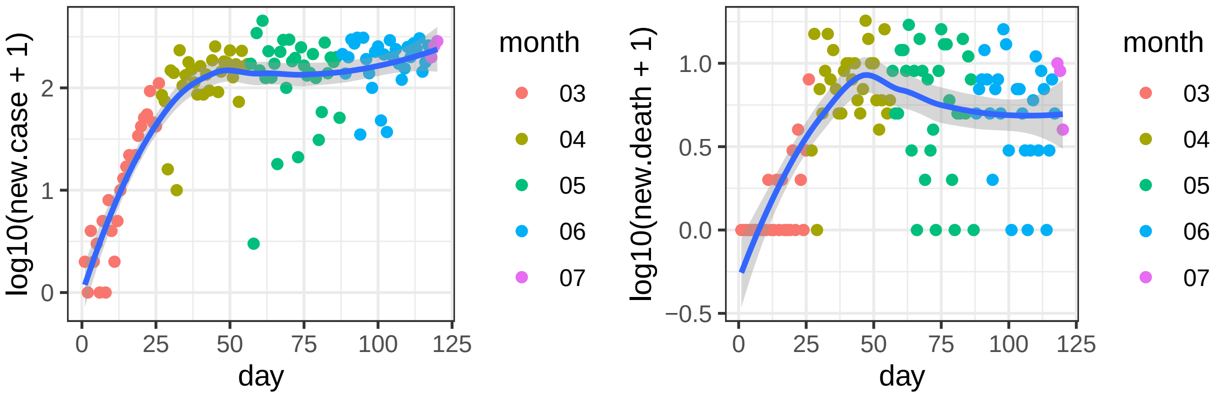
Iowa

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



Tennessee

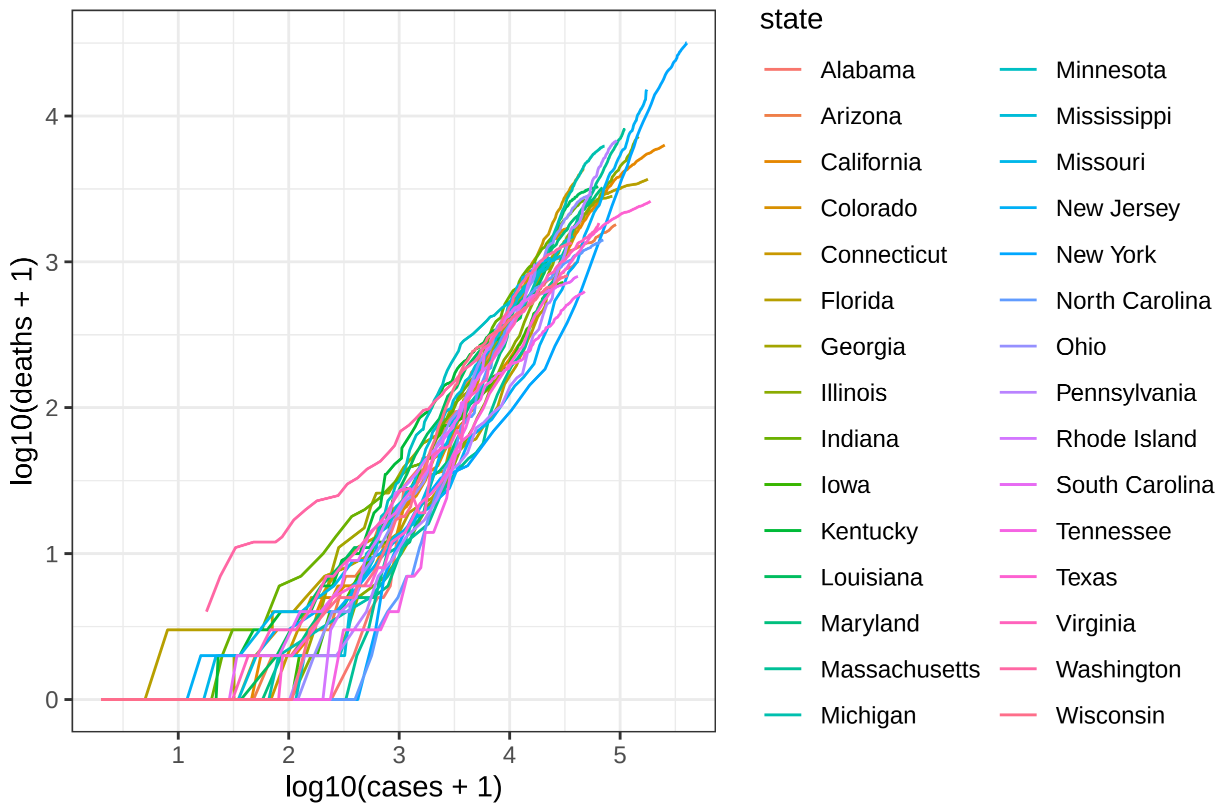
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



Kentucky

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Next I check the relation between the **cumulative** number of cases and deaths for these 10 states, starting on March



data source: https://github.com/nytimes/covid−19−data

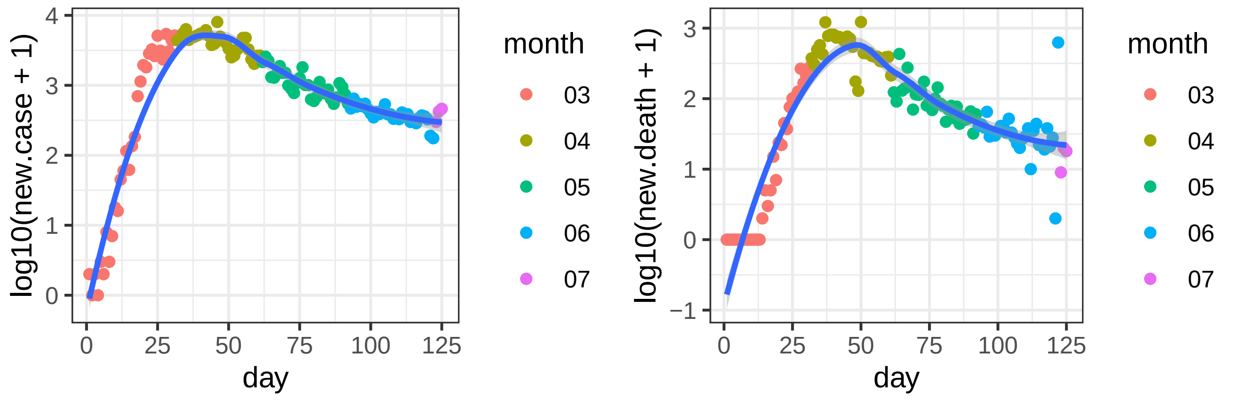
county level data

First check the 50 counties with the largest number of deaths.

```
##              date               county                state  fips  cases deaths
## 293579 2020-07-03        New York City             New York    NA 221028  22610
## 292366 2020-07-03                 Cook             Illinois 17031  91774   4618
## 291963 2020-07-03          Los Angeles           California  6037 107667   3454
## 293065 2020-07-03                Wayne             Michigan 26163  23078   2730
## 293578 2020-07-03               Nassau             New York 36059  41947   2697
## 293504 2020-07-03                Essex           New Jersey 34013  19083   2034
## 293598 2020-07-03              Suffolk             New York 36103  41538   2029
## 293499 2020-07-03               Bergen           New Jersey 34003  19793   2000
## 292977 2020-07-03            Middlesex        Massachusetts 25017  24083   1870
## 294008 2020-07-03         Philadelphia         Pennsylvania 42101  26536   1619
## 293606 2020-07-03          Westchester             New York 36119  34979   1558
## 293506 2020-07-03               Hudson           New Jersey 34017  19080   1454
## 292064 2020-07-03             Hartford          Connecticut  9003  11728   1378
## 292063 2020-07-03            Fairfield          Connecticut  9001  16757   1377
## 293509 2020-07-03            Middlesex           New Jersey 34023  17080   1328
## 293517 2020-07-03                Union           New Jersey 34039  16591   1326
## 293513 2020-07-03              Passaic           New Jersey 34031  17074   1192
## 292973 2020-07-03                Essex        Massachusetts 25009  16210   1116
## 293046 2020-07-03              Oakland             Michigan 26125  12160   1091
## 292067 2020-07-03            New Haven          Connecticut  9009  12409   1077
## 292119 2020-07-03           Miami-Dade              Florida 12086  42310   1034
## 292981 2020-07-03              Suffolk        Massachusetts 25025  19936   1009
## 293512 2020-07-03                Ocean           New Jersey 34029   9734    967
## 292983 2020-07-03            Worcester        Massachusetts 25027  12443    935
## 292979 2020-07-03              Norfolk        Massachusetts 25021   9242    934
## 293033 2020-07-03               Macomb             Michigan 26099   7742    924
## 291861 2020-07-03             Maricopa              Arizona  4013  57929    865
## 293510 2020-07-03             Monmouth           New Jersey 34025   9344    817
## 294003 2020-07-03           Montgomery         Pennsylvania 42091   8562    810
## 293511 2020-07-03               Morris           New Jersey 34027   6937    805
## 293093 2020-07-03             Hennepin            Minnesota 27053  12150    785
## 292959 2020-07-03           Montgomery             Maryland 24031  15059    747
## 294029 2020-07-03           Providence         Rhode Island 44007  13144    747
## 292502 2020-07-03               Marion              Indiana 18097  11630    730
## 293980 2020-07-03             Delaware         Pennsylvania 42045   7299    702
## 292960 2020-07-03      Prince George's             Maryland 24033  19221    687
## 292980 2020-07-03             Plymouth        Massachusetts 25023   8722    665
## 292975 2020-07-03              Hampden        Massachusetts 25013   6834    662
## 294670 2020-07-03                 King           Washington 53033  10721    620
## 293564 2020-07-03                 Erie             New York 36029   7427    598
## 292971 2020-07-03              Bristol        Massachusetts 25005   8295    587
## 293508 2020-07-03               Mercer           New Jersey 34021   7754    585
## 293337 2020-07-03            St. Louis             Missouri 29189   6755    583
## 293966 2020-07-03                Bucks         Pennsylvania 42017   5829    567
## 292076 2020-07-03 District of Columbia District of Columbia 11001  10435    555
## 292126 2020-07-03           Palm Beach              Florida 12099  15322    536
## 292897 2020-07-03              Orleans            Louisiana 22071   8031    534
## 293515 2020-07-03             Somerset           New Jersey 34035   4999    529
## 293501 2020-07-03               Camden           New Jersey 34007   7470    513
## 294559 2020-07-03              Fairfax             Virginia 51059  13965    494
```

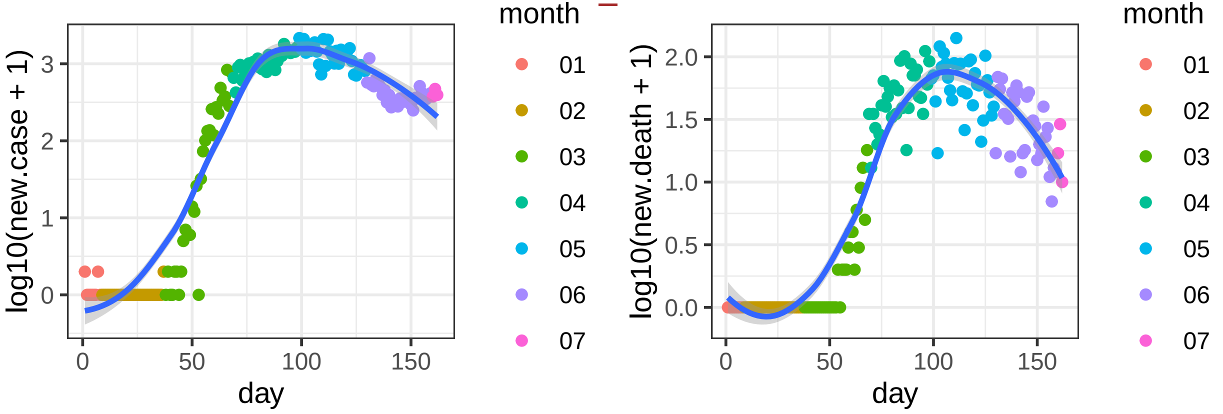
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



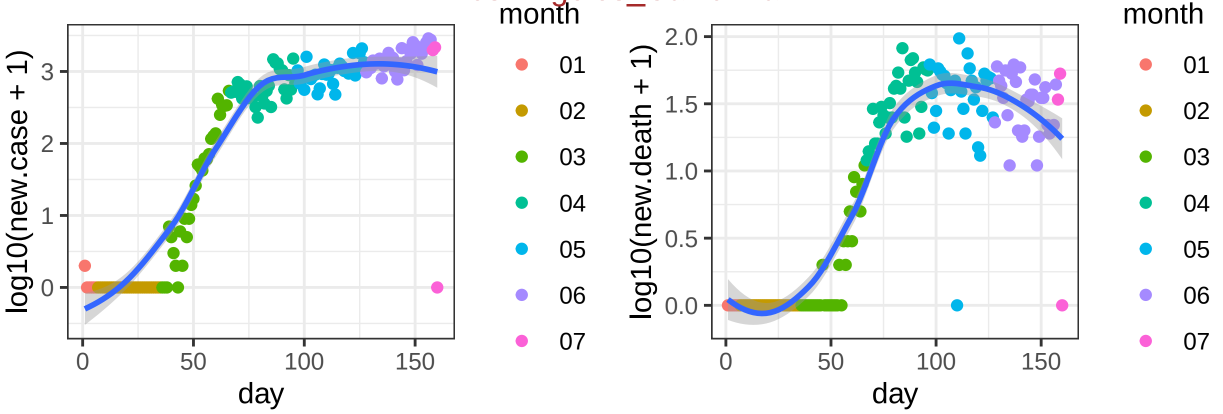
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



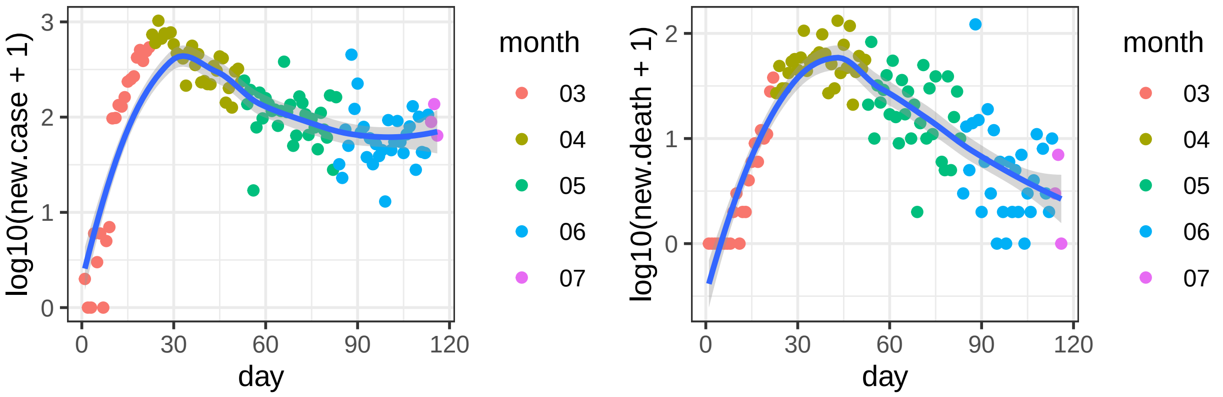
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



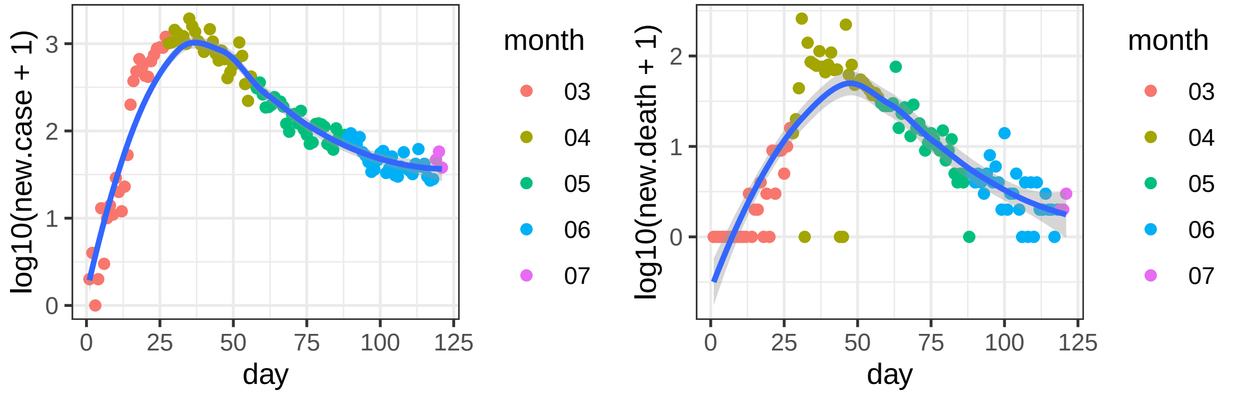
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



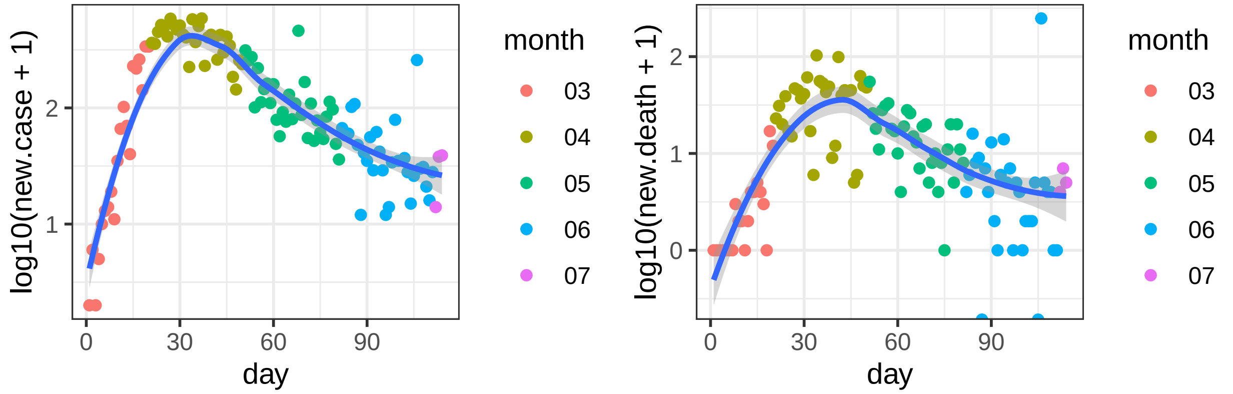
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



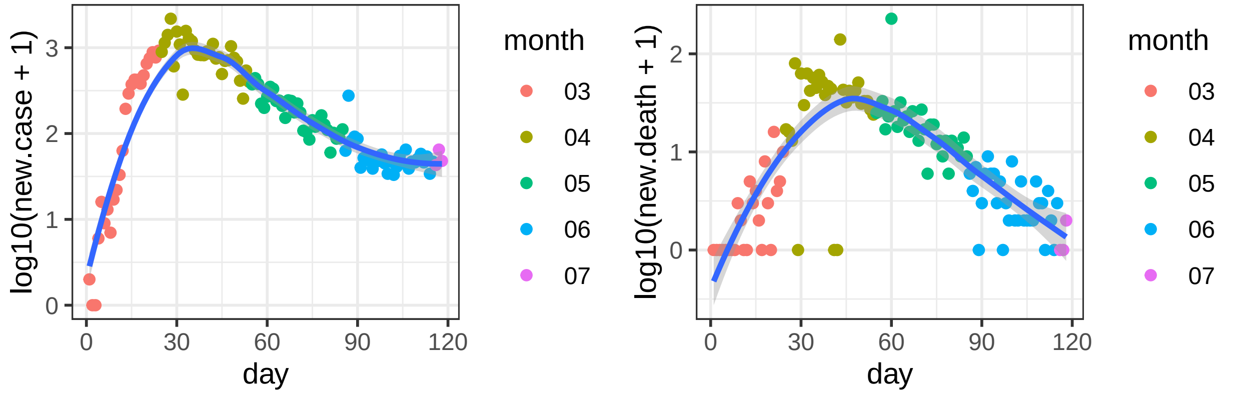
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



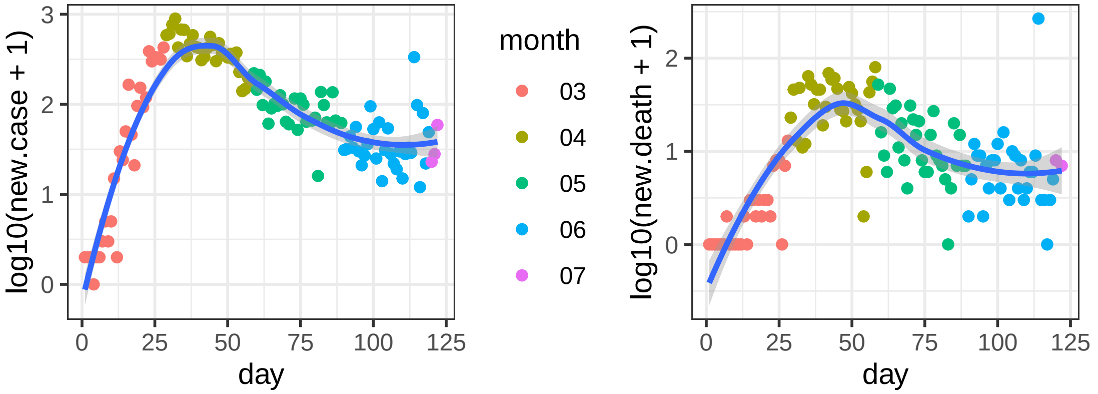
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Suffolk_New York



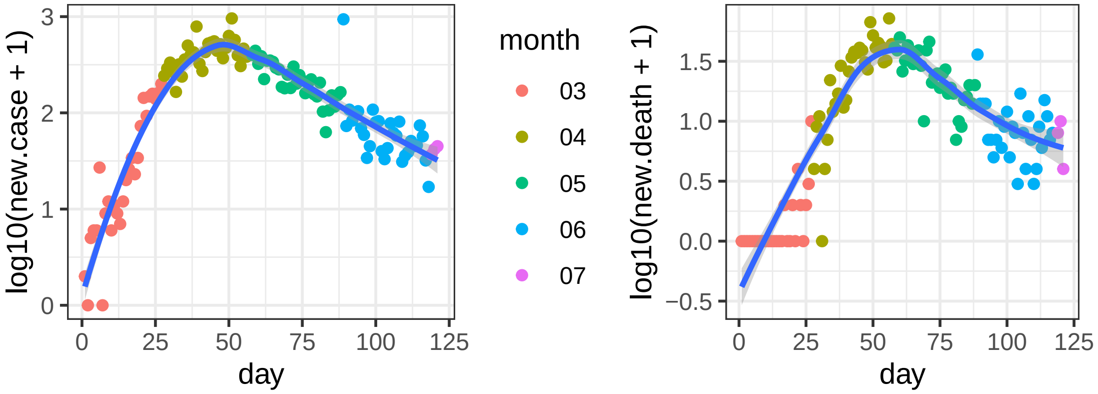
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Bergen_New Jersey



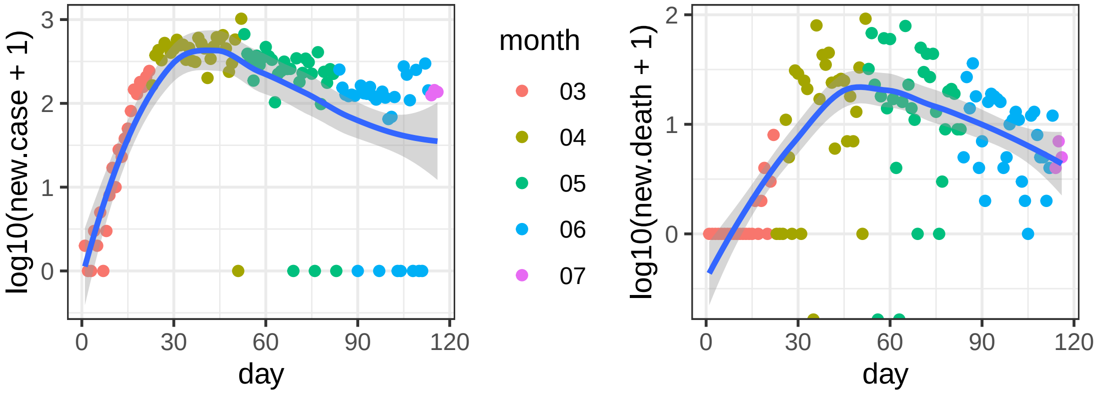
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



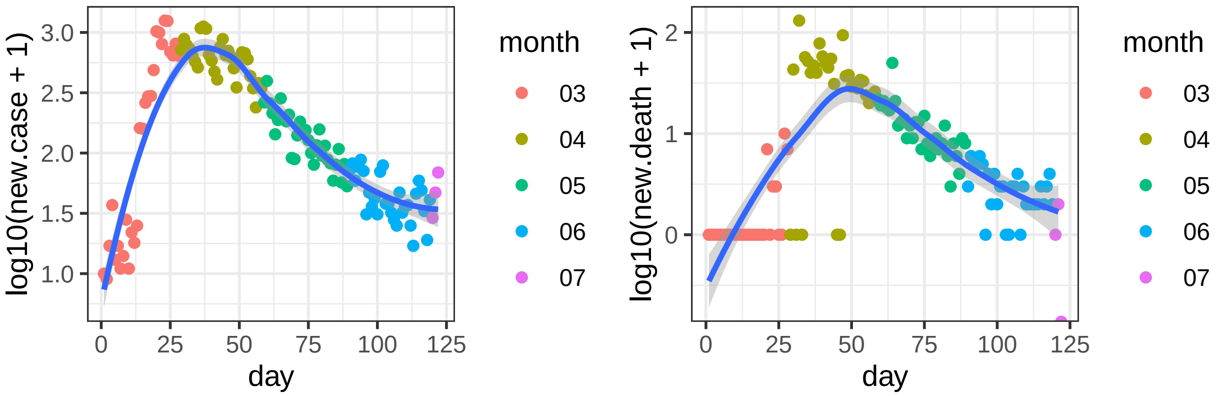
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Philadelphia_Pennsylvania



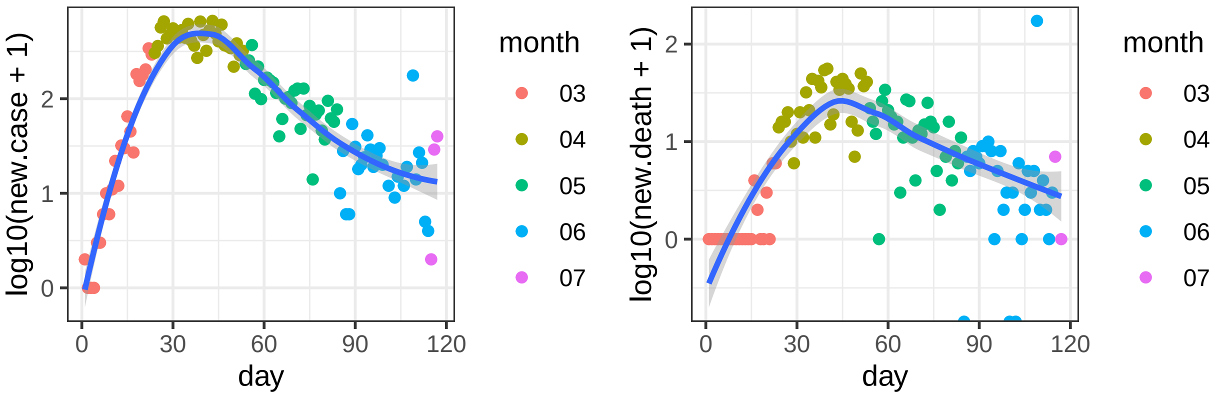
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Westchester_New York



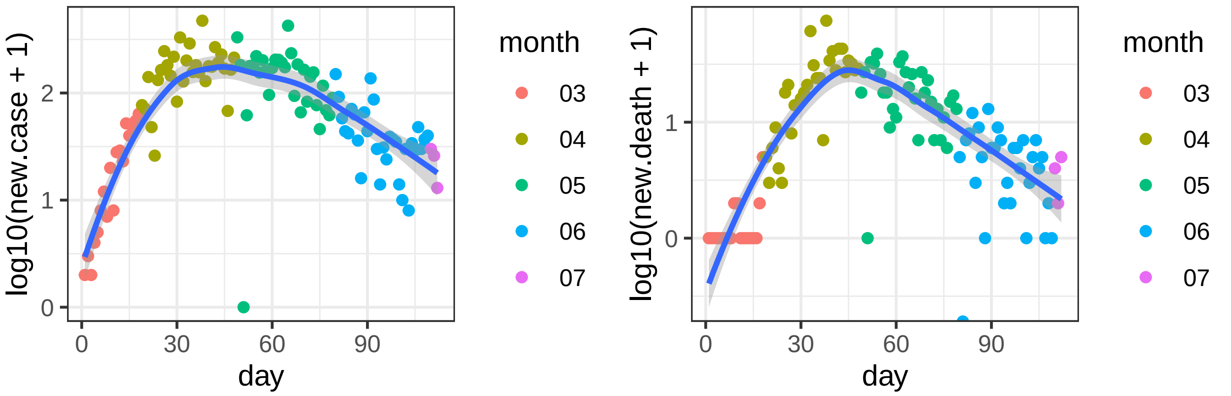
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hudson_New Jersey



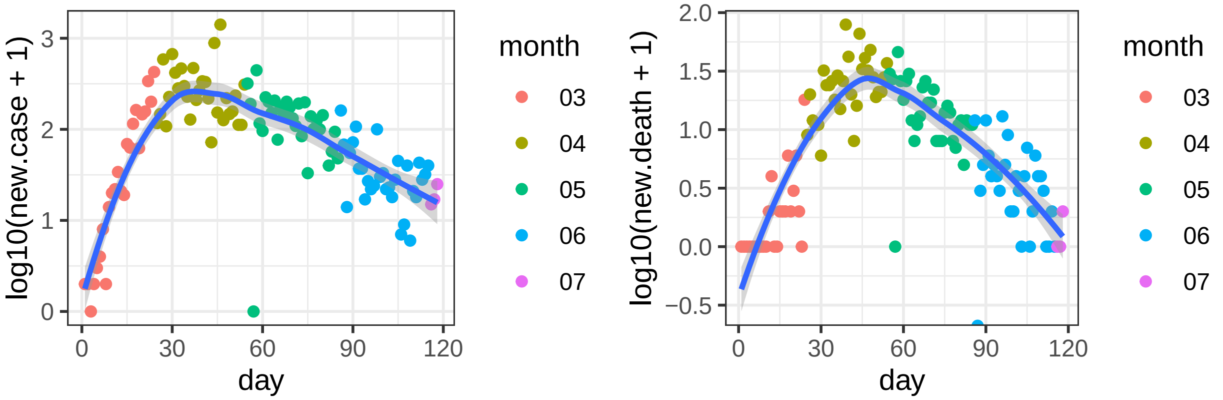
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hartford_Connecticut



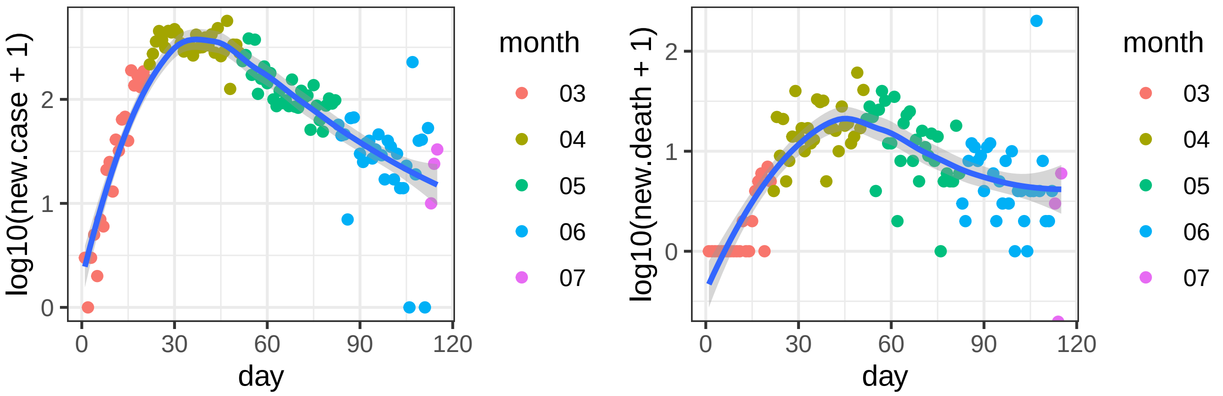
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Fairfield_Connecticut



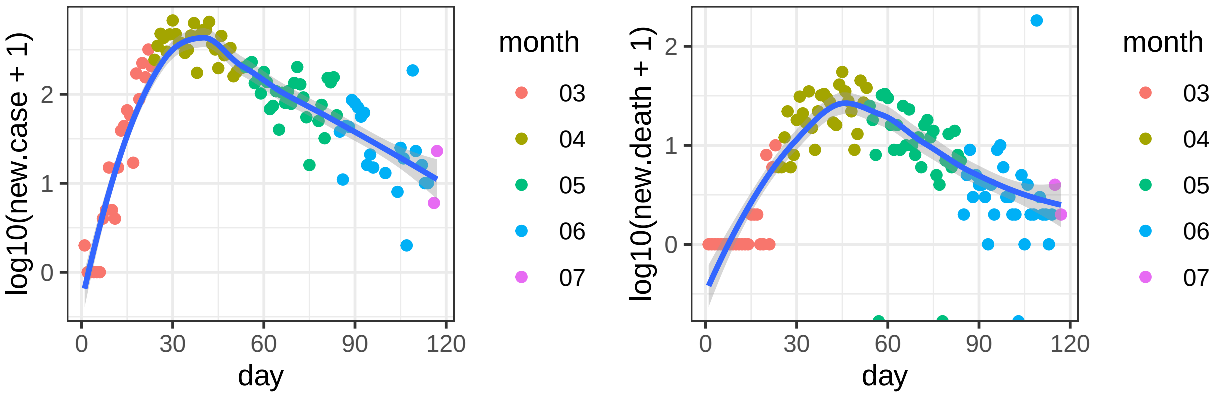
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_New Jersey



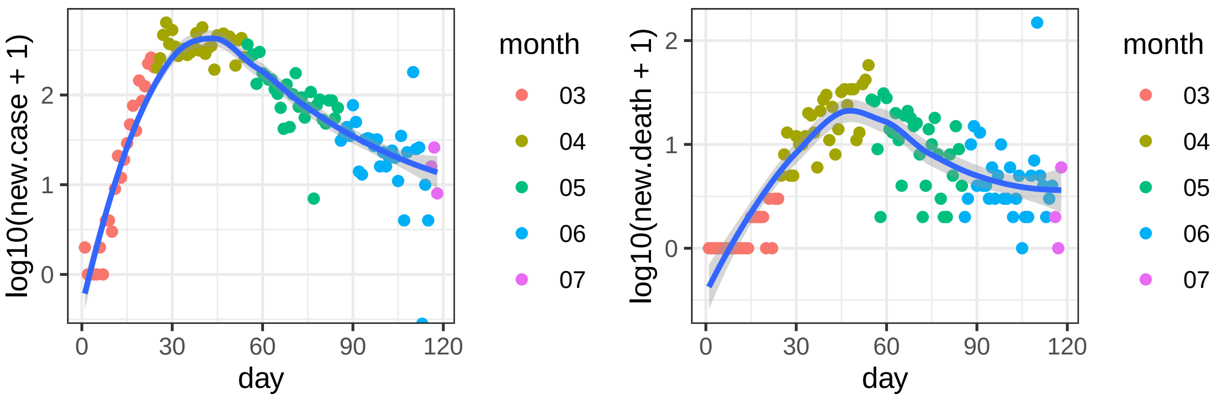
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Union_New Jersey



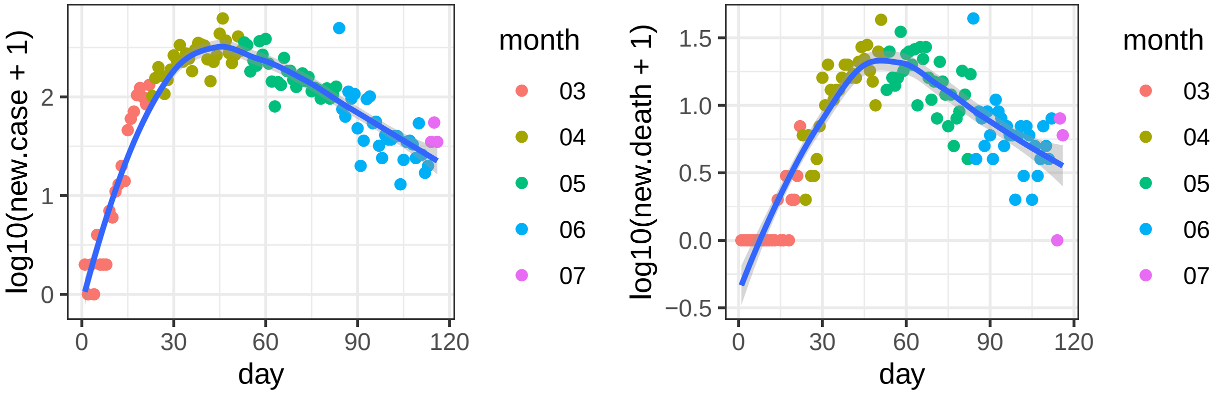
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Passaic_New Jersey



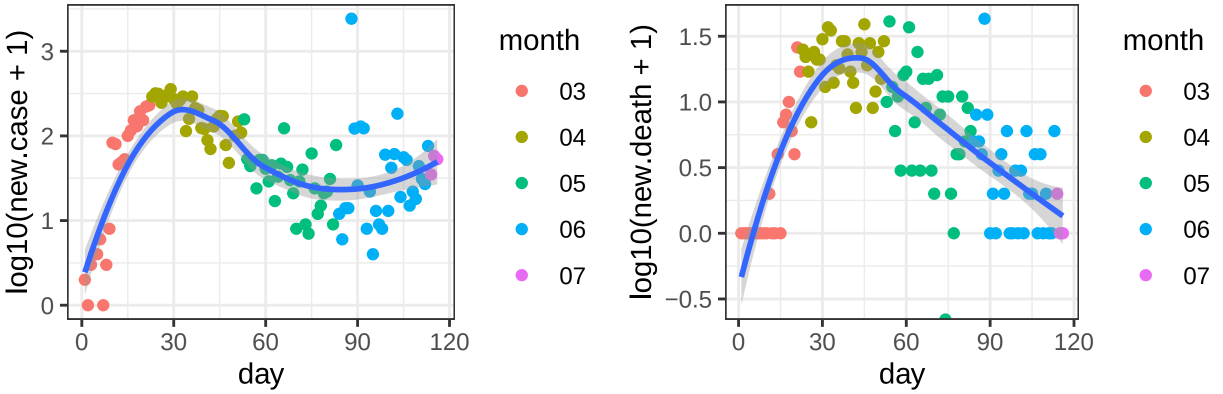
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



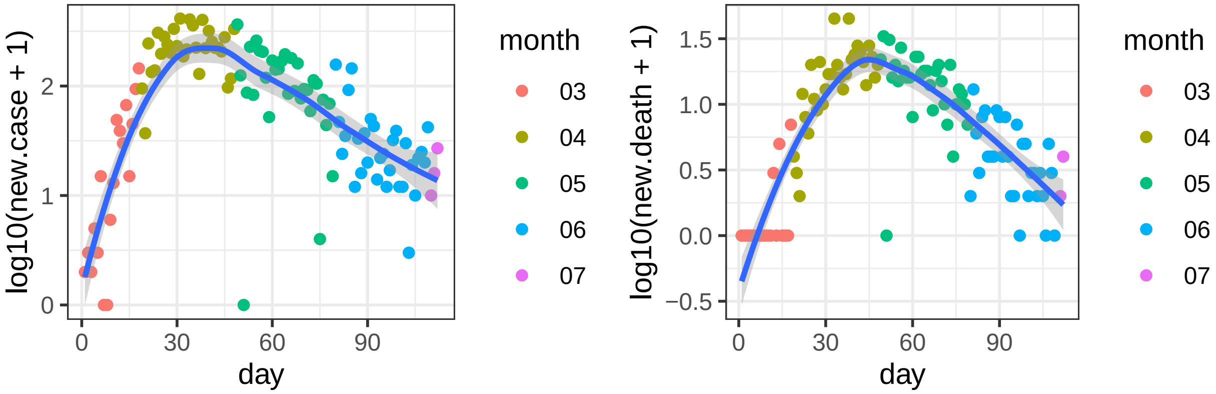
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



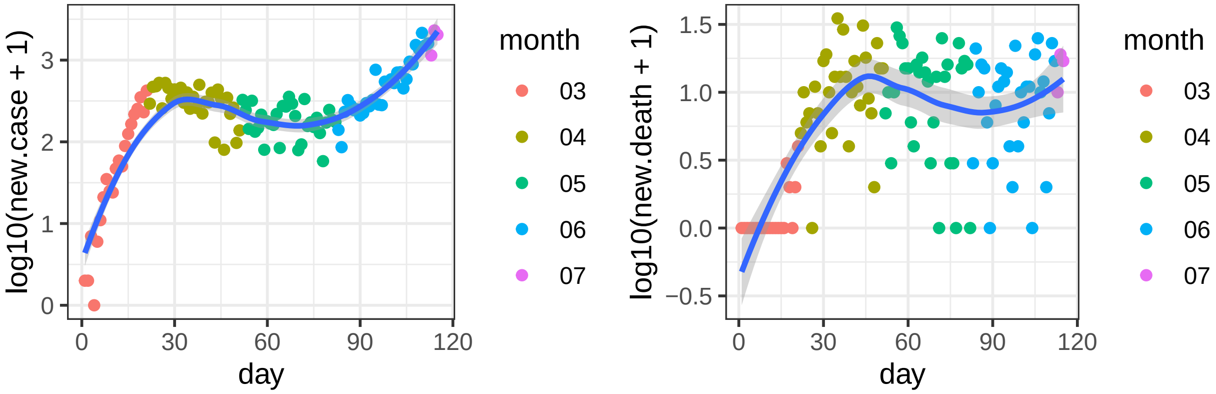
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



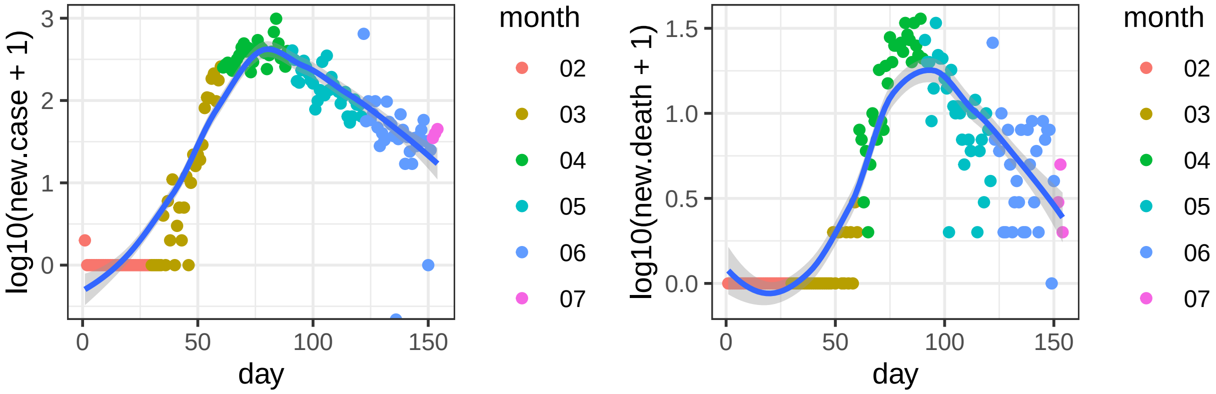
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Miami−Dade_Florida



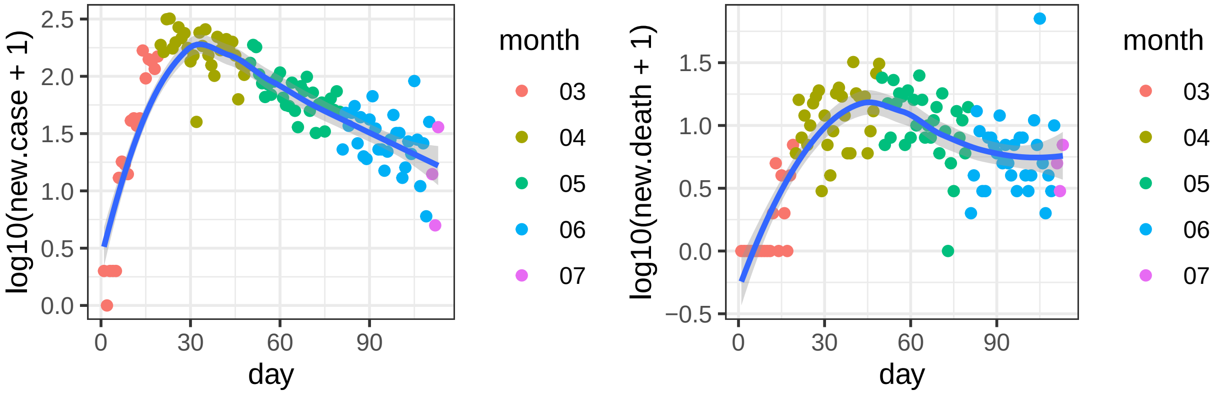
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Suffolk_Massachusetts



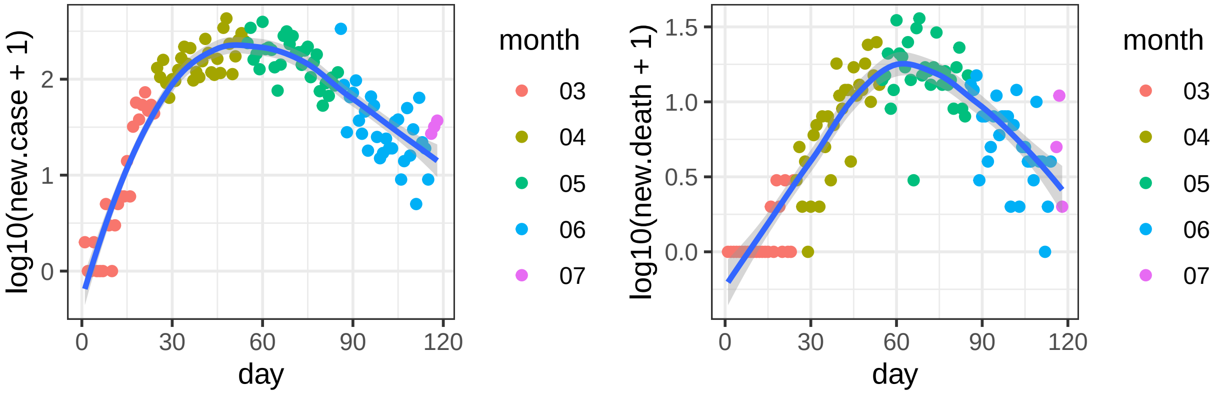
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Ocean_New Jersey



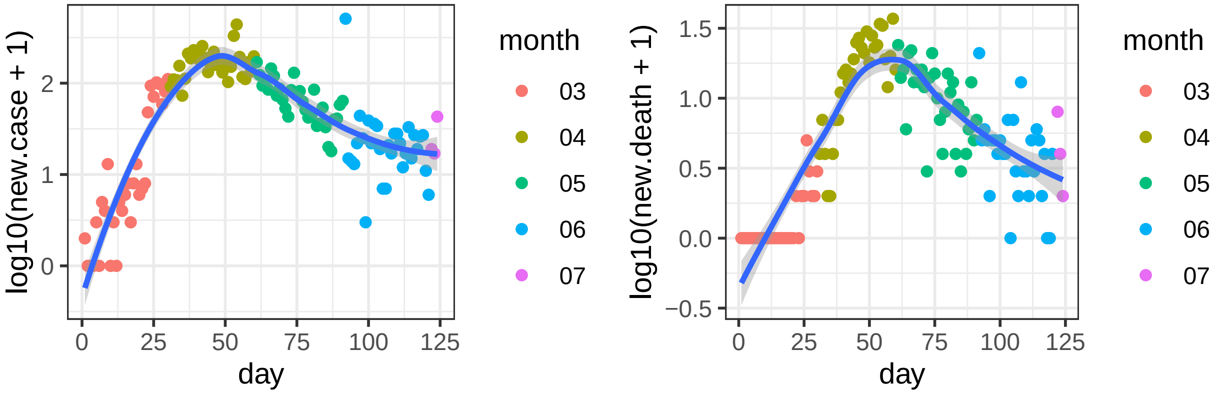
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



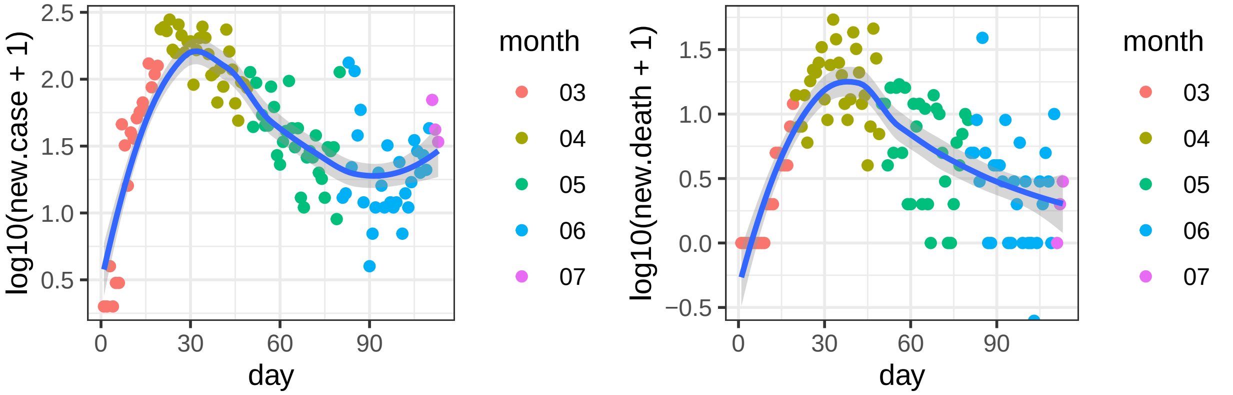
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



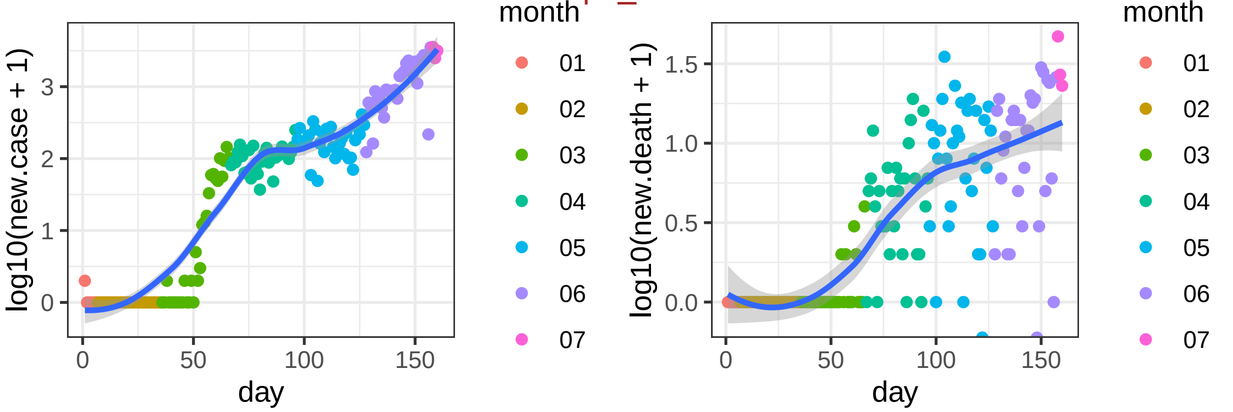
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



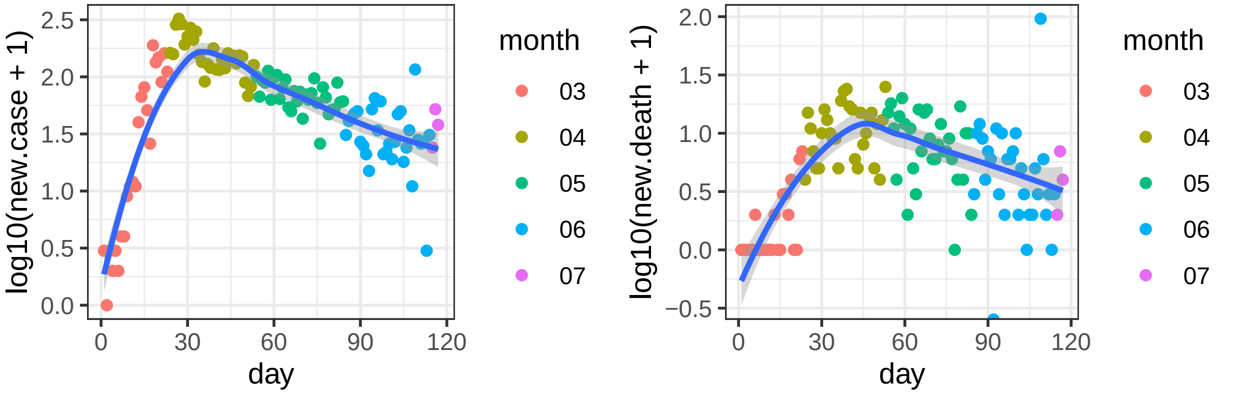
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Maricopa_Arizona



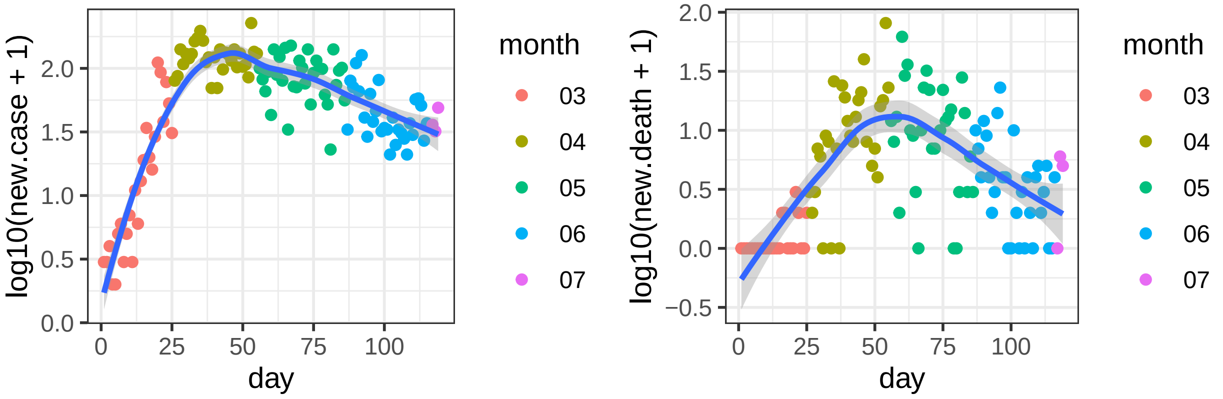
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Monmouth_New Jersey



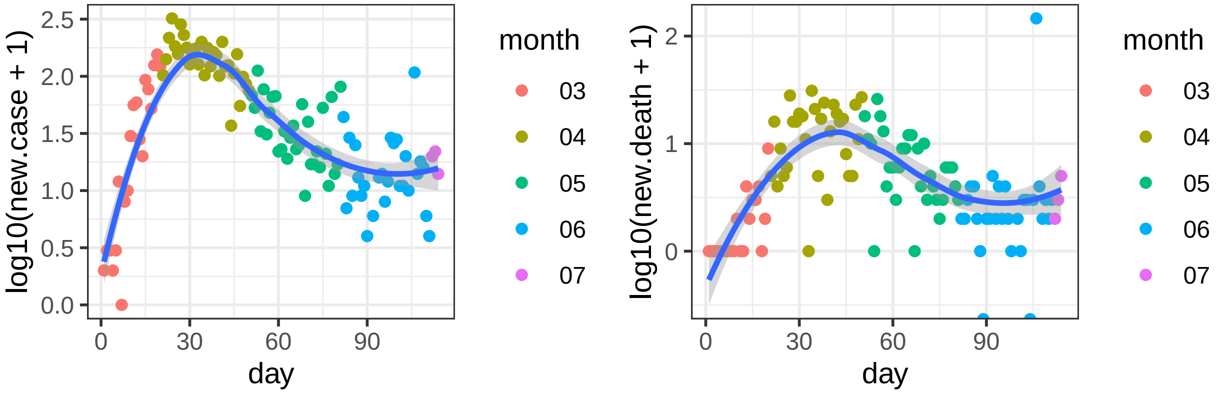
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Montgomery_Pennsylvania



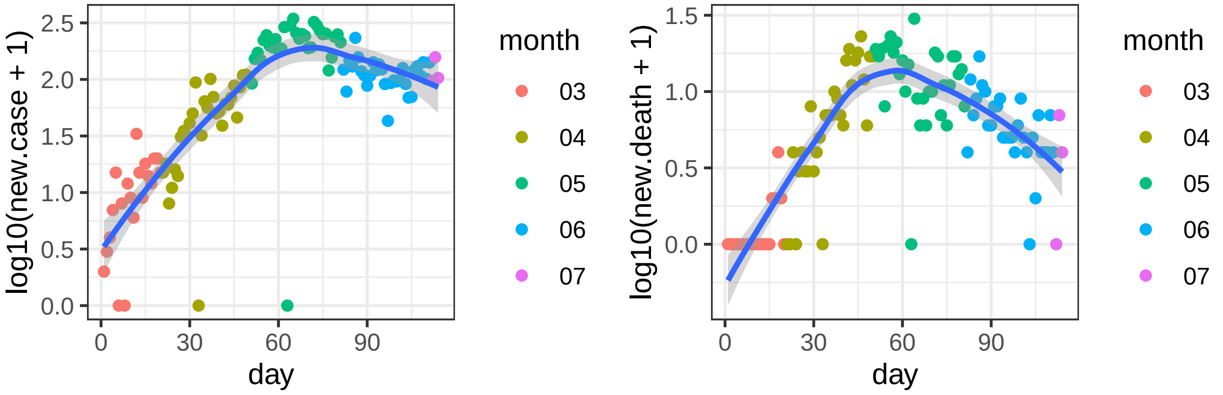
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



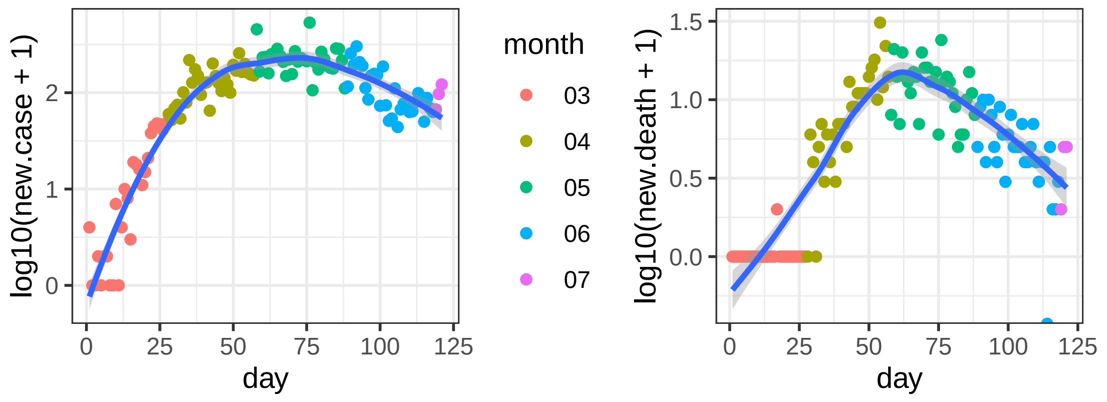
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hennepin_Minnesota



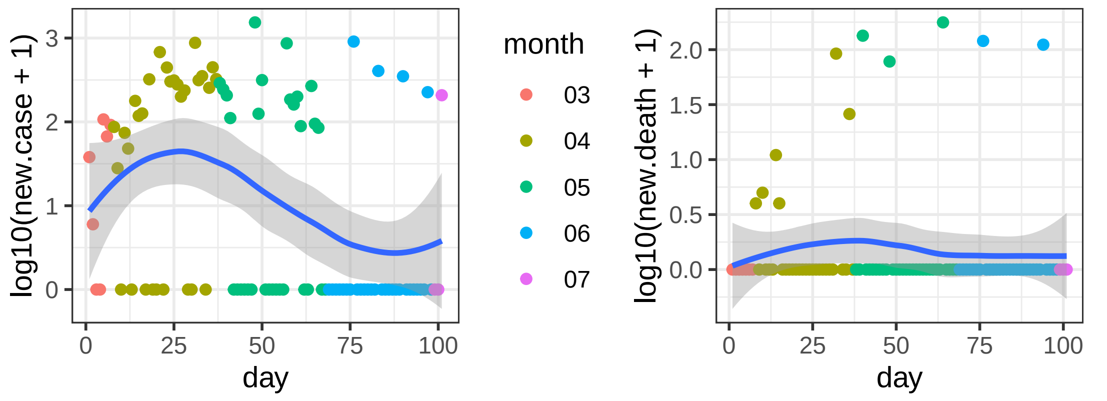
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Montgomery_Maryland



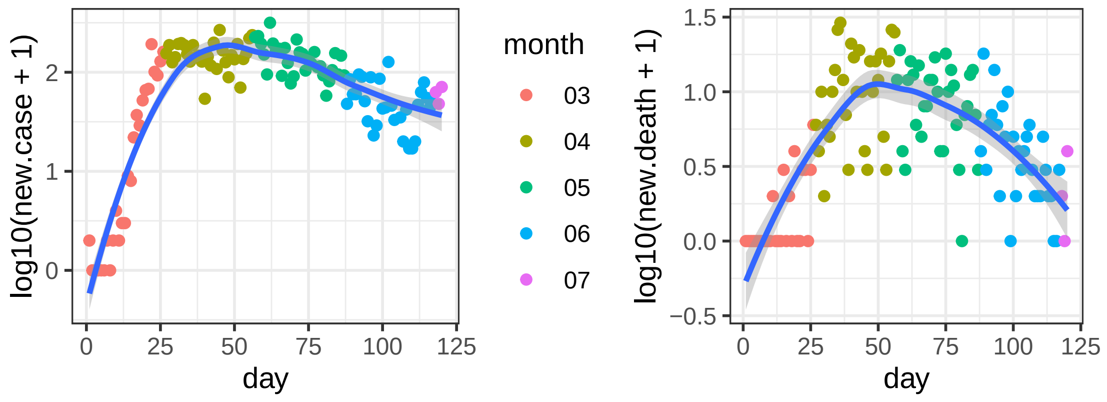
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Providence_Rhode Island



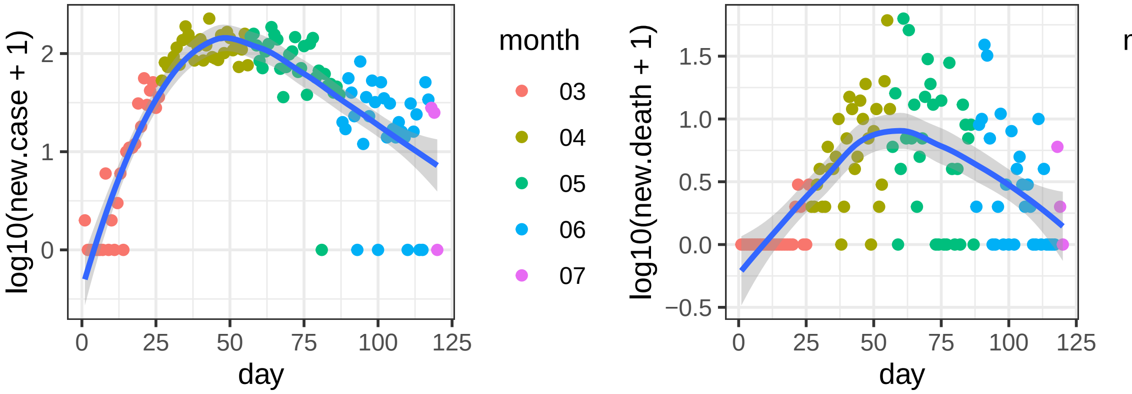
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Marion_Indiana



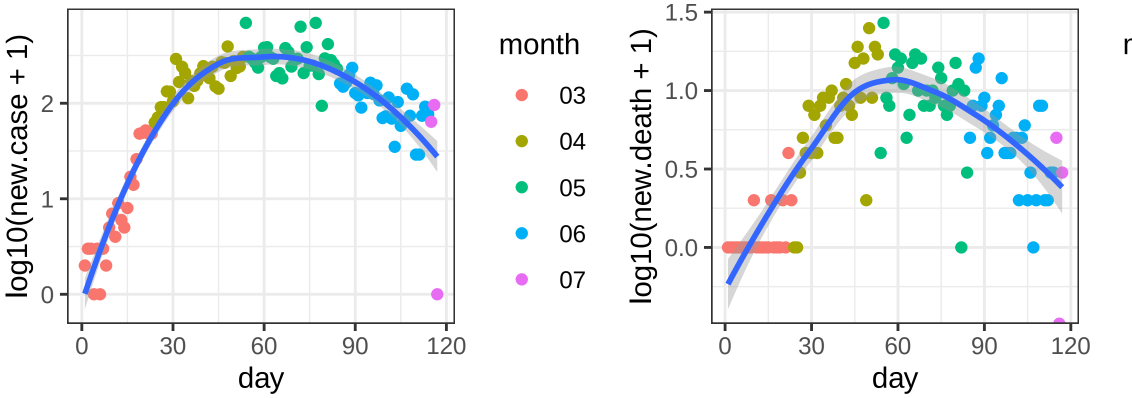
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Delaware_Pennsylvania



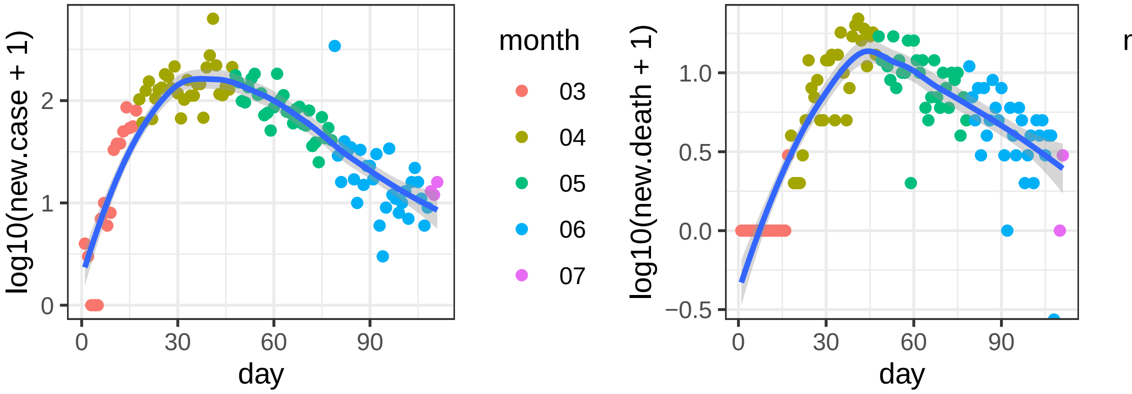
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



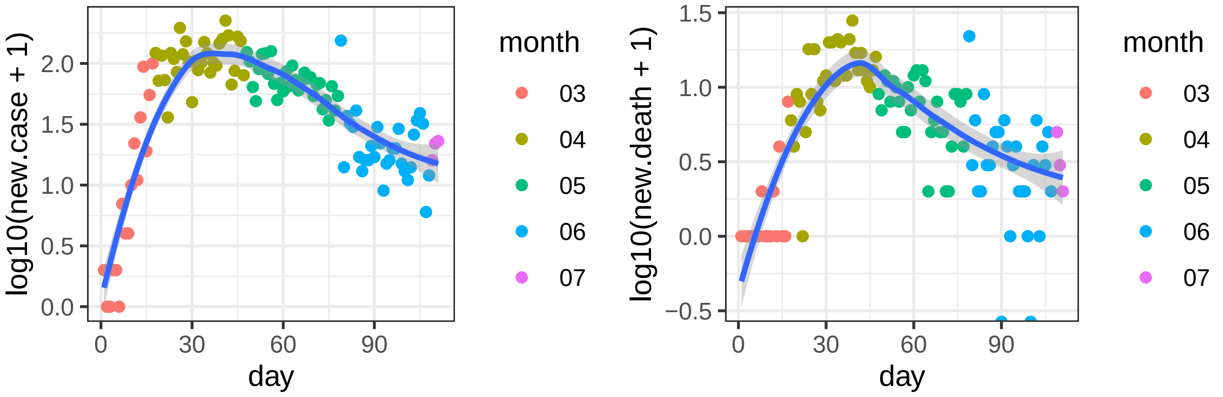
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Plymouth_Massachusetts



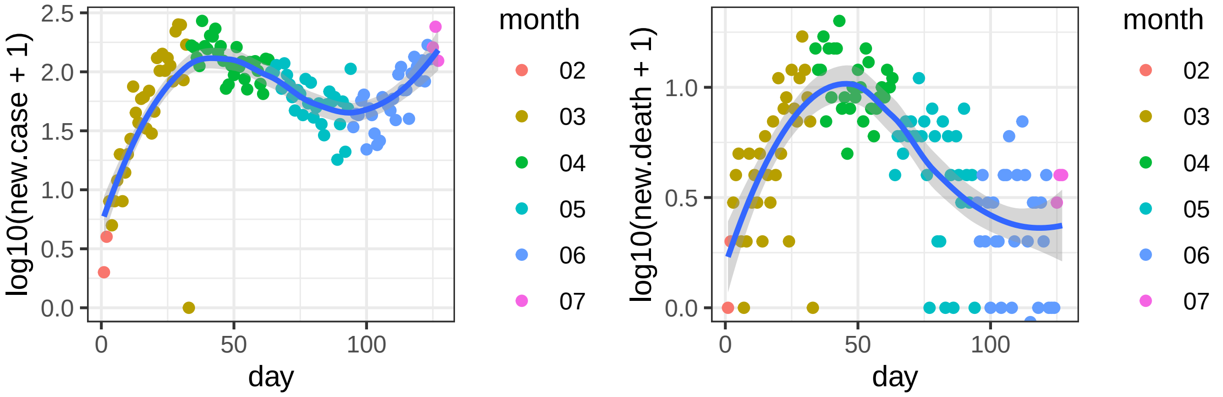
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Hampden_Massachusetts



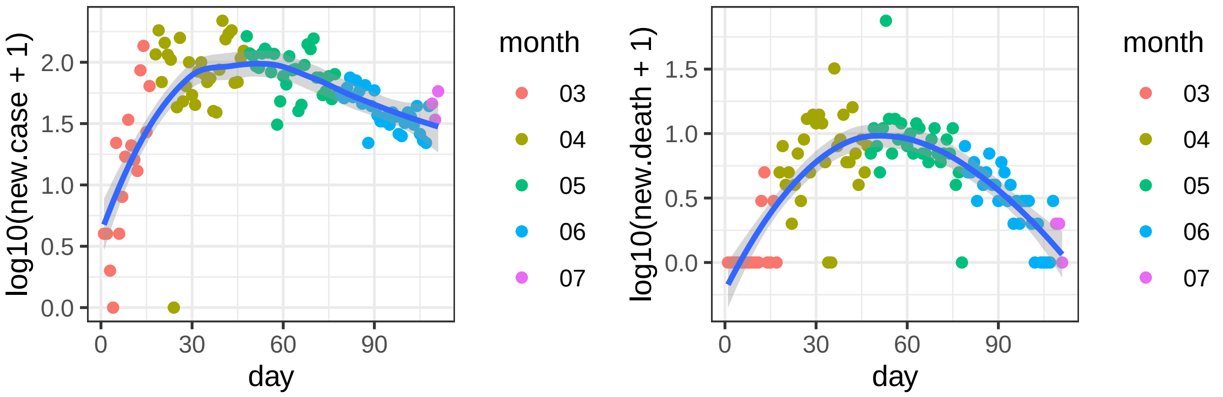
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington



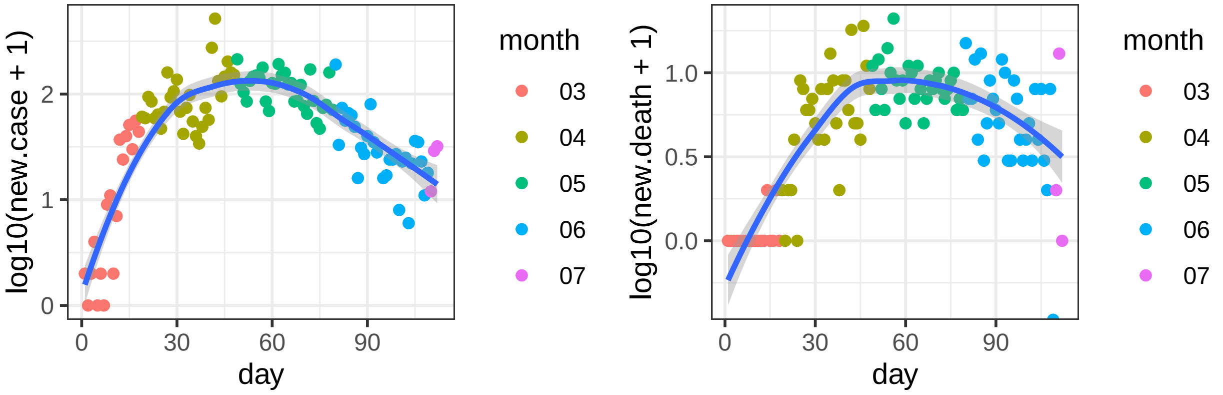
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Erie_New York



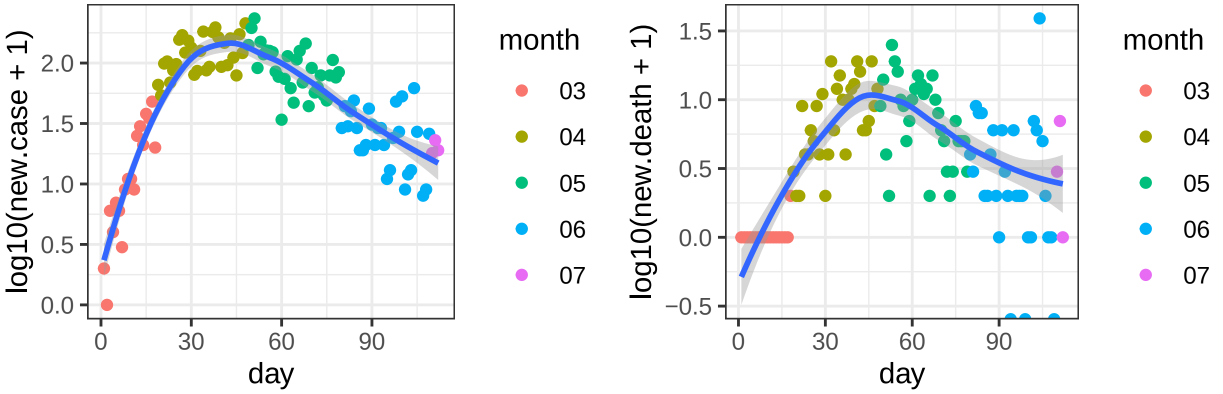
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Bristol_Massachusetts



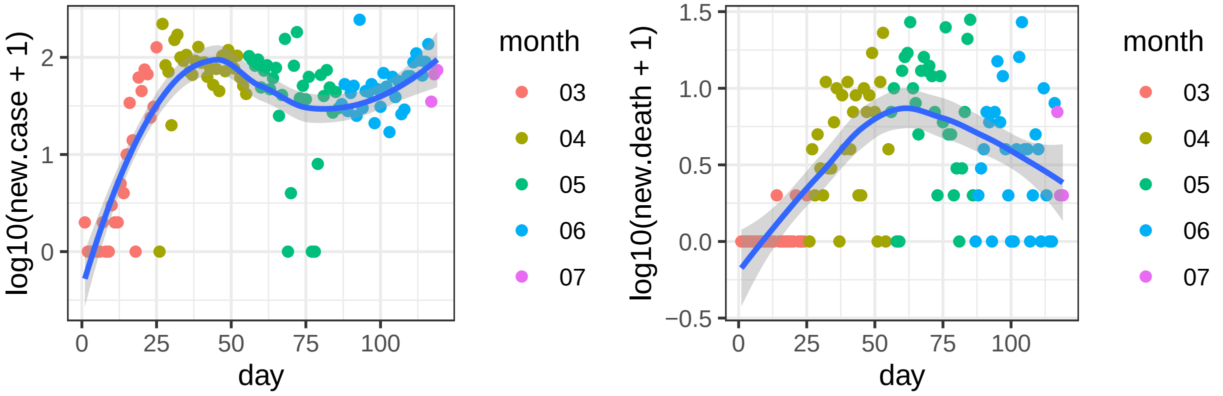
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Mercer_New Jersey



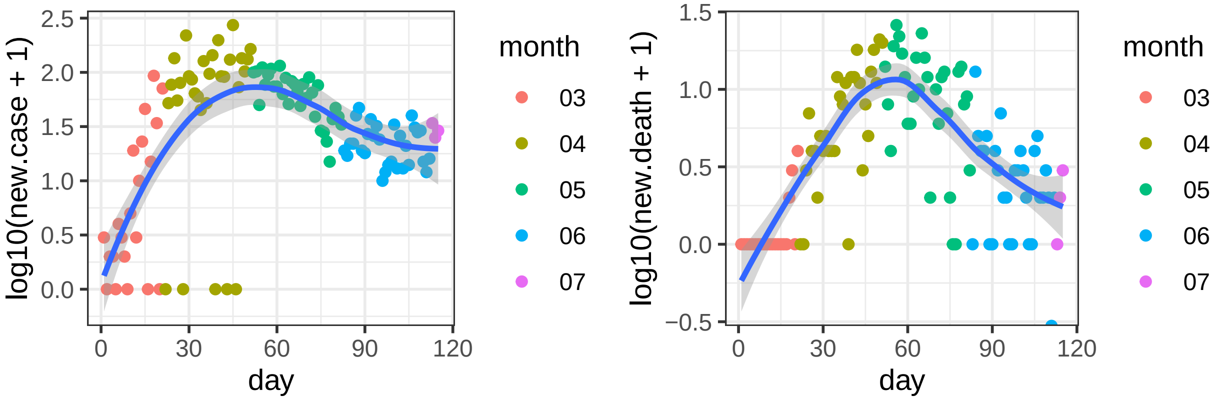
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

St. Louis_Missouri



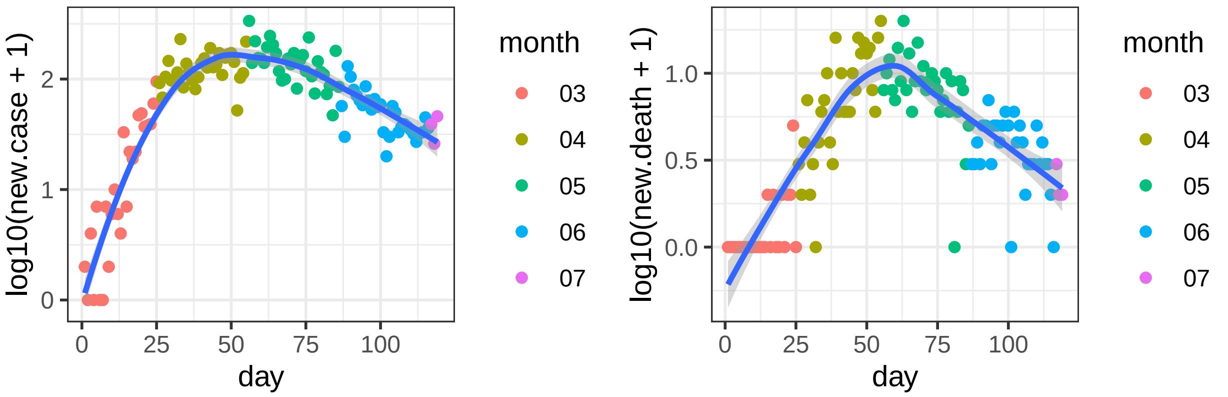
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Bucks_Pennsylvania



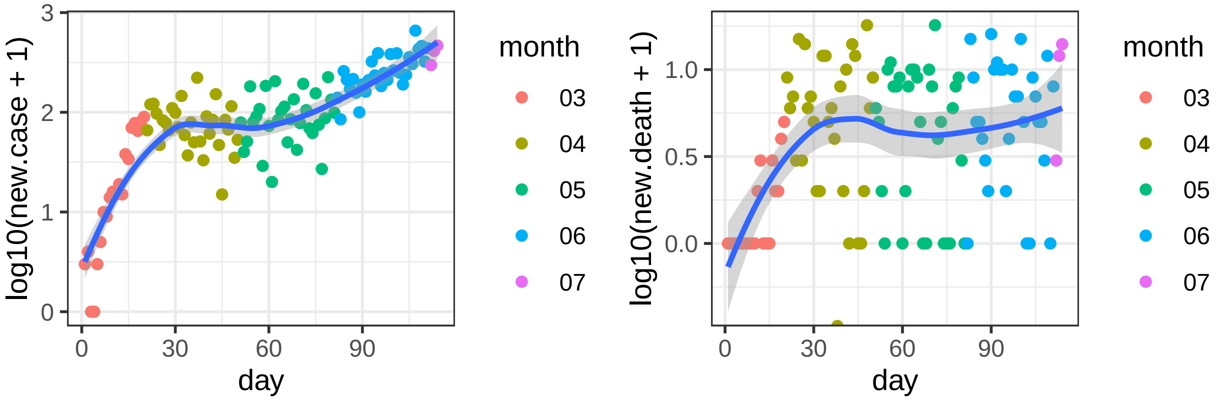
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

District of Columbia_District of Columbia



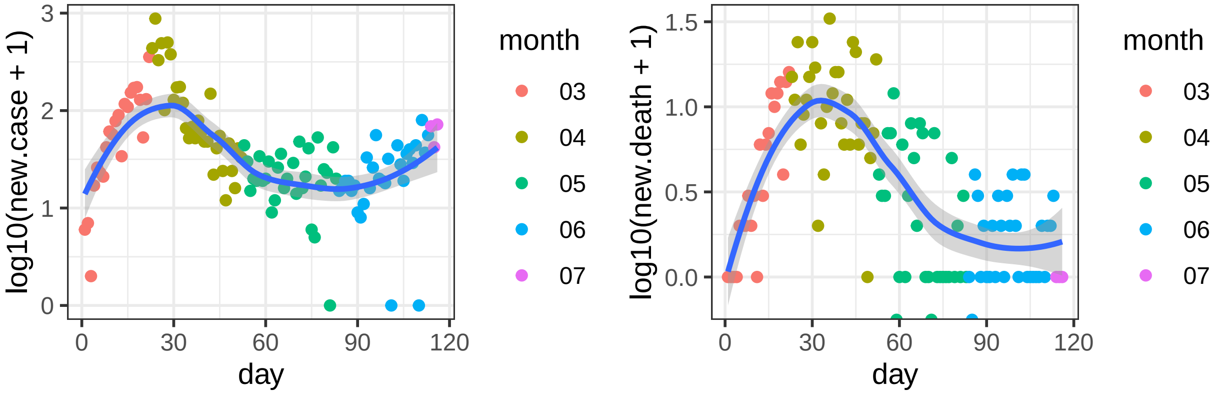
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Palm Beach_Florida



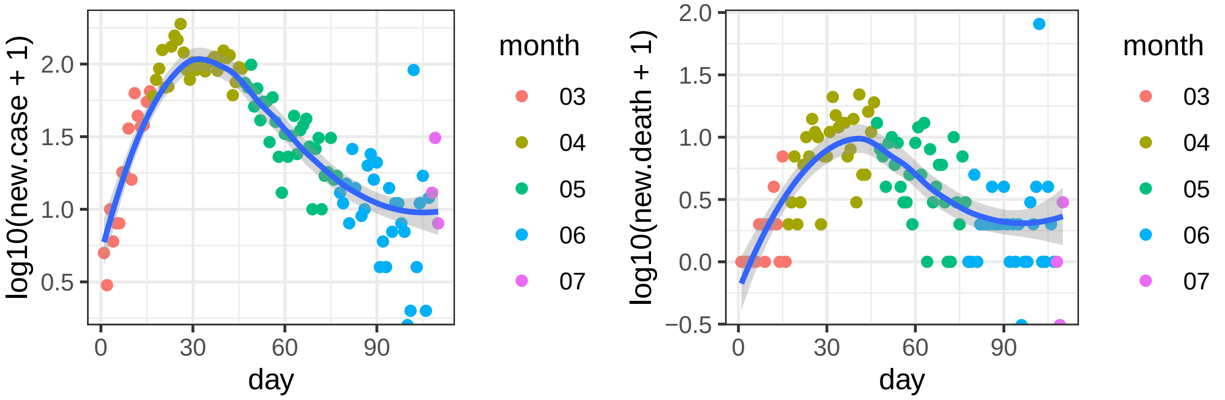
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Orleans_Louisiana



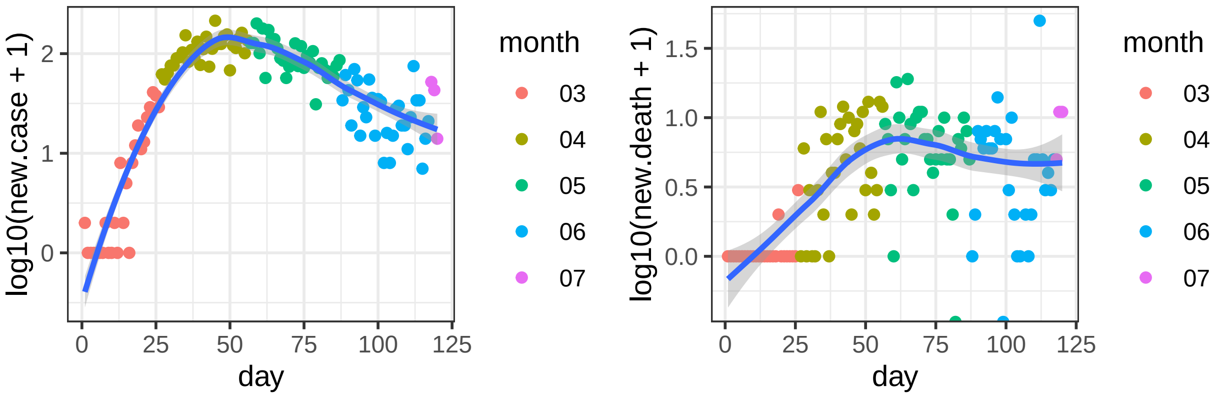
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Somerset_New Jersey

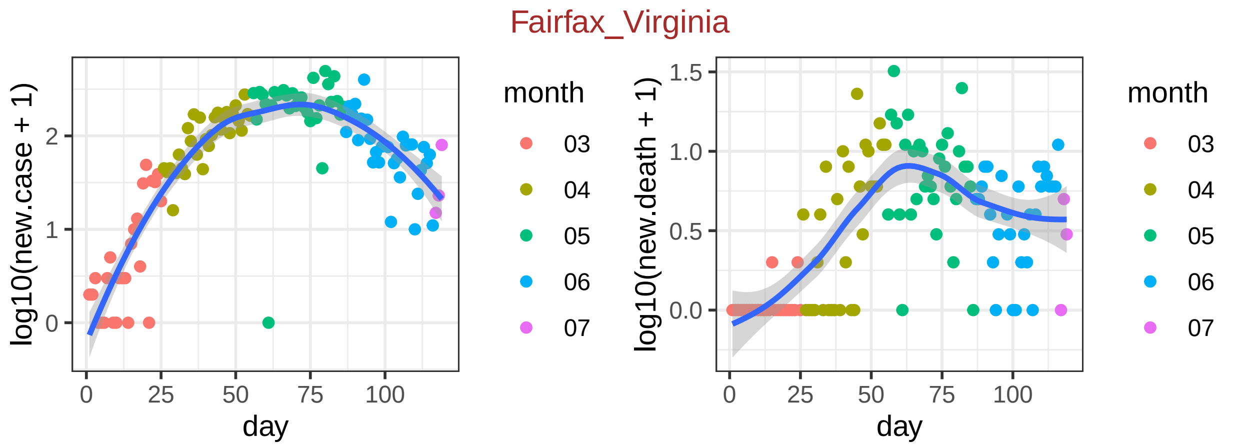


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−16

Camden_New Jersey



data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

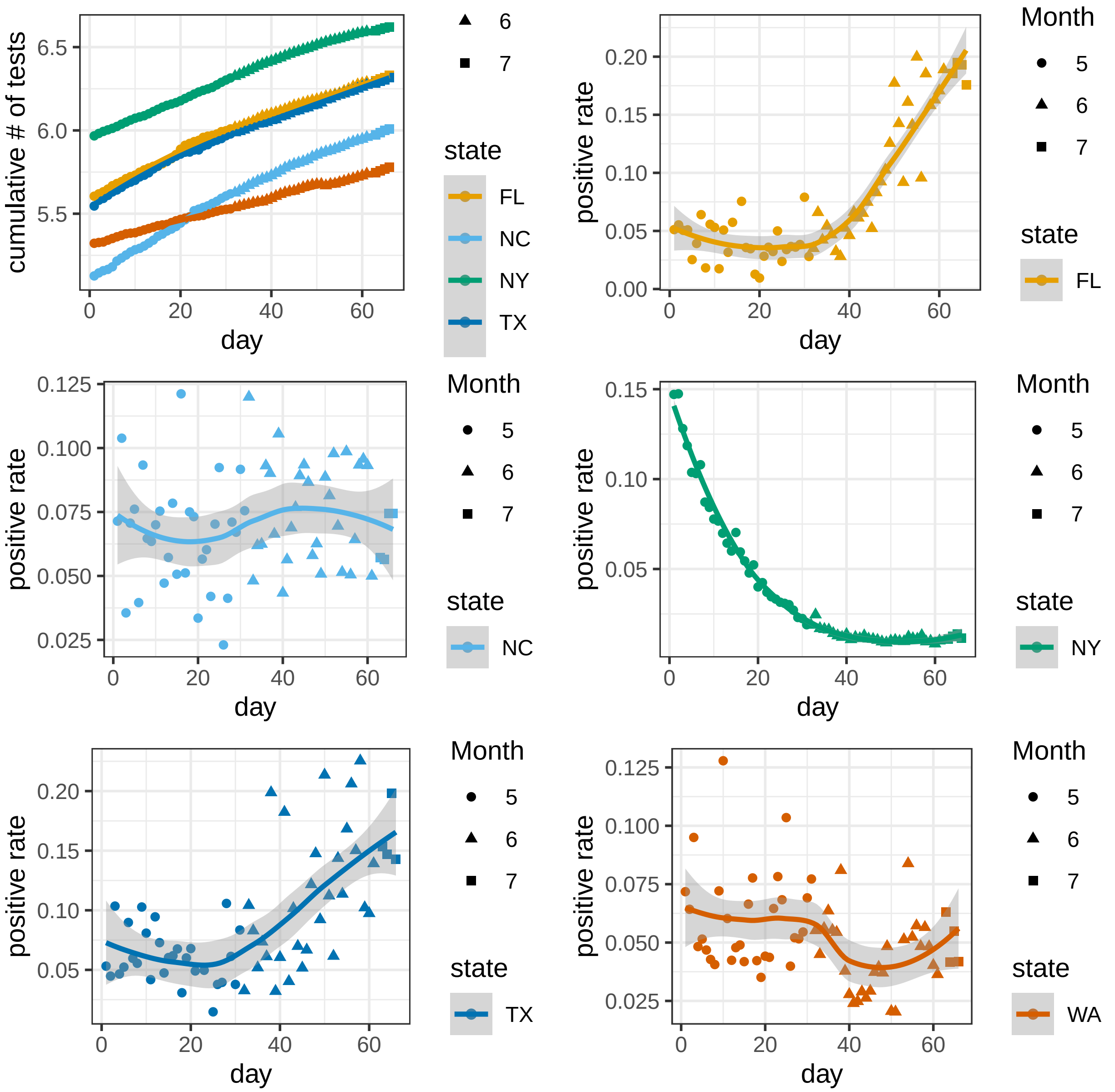


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

COVID Trackng

The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they can be much more noisy data since the negative testing resutls are often not reported and the tests are almost surely taken on a non-representative random sample of the population. The COVID traking project proides a grade per state: “If you are calculating positive rates, it should only be with states that have an A grade. And be careful going back in time because almost all the states have changed their level of reporting at different times.” (https://covidtracking.com/about-tracker/). The data are also availalbe for both counties and states, here I only look at state level data.

The grades of the states may change over timea and I strongly recommend checking their webiste before puting serious interpretation on the following plot.



github.com/COVID19Tracking/, positive rate on 0704: 0.18(FL) 0.07(NC) 0.01(NY) 0.14(TX) 0.04(WA)

Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
```

```
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] httr_1.4.1    ggpubr_0.2.5  magrittr_1.5  ggplot2_3.3.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3       pillar_1.4.3     compiler_3.6.2   tools_3.6.2
##  [5] digest_0.6.23    lattice_0.20-38  nlme_3.1-144     evaluate_0.14
##  [9] lifecycle_0.2.0  tibble_3.0.1     gtable_0.3.0     mgcv_1.8-31
## [13] pkgconfig_2.0.3  rlang_0.4.6      Matrix_1.2-18    yaml_2.2.1
## [17] xfun_0.12        gridExtra_2.3    withr_2.1.2      stringr_1.4.0
## [21] dplyr_0.8.4      knitr_1.28       vctrs_0.3.0      cowplot_1.0.0
## [25] grid_3.6.2       tidyselect_1.0.0 glue_1.3.1       R6_2.4.1
## [29] rmarkdown_2.1    purrr_0.3.3      farver_2.0.3     splines_3.6.2
## [33] scales_1.1.0     ellipsis_0.3.0   htmltools_0.4.0  assertthat_0.2.1
## [37] colorspace_1.4-1 ggsignif_0.6.0   labeling_0.3     stringi_1.4.5
## [41] munsell_0.5.0    crayon_1.3.4
```