Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-08-23

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

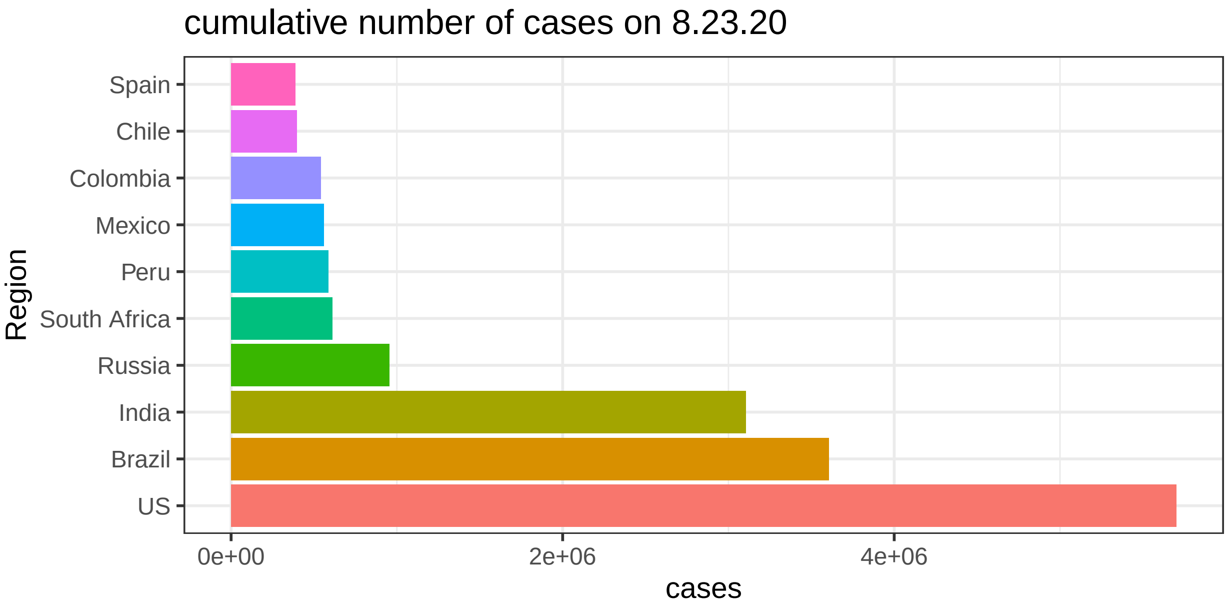
JHU

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

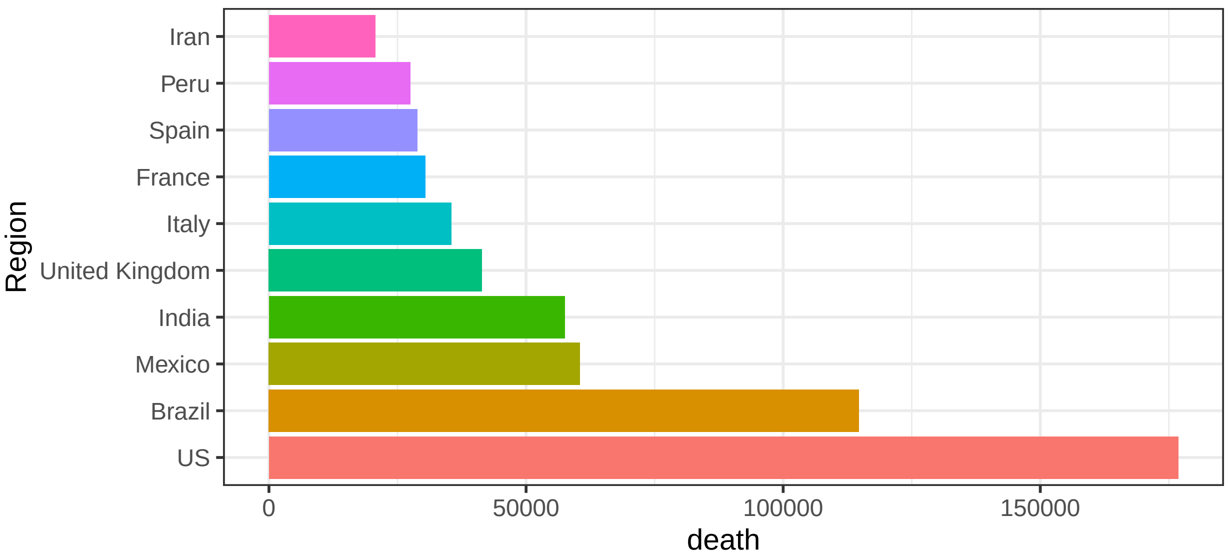
time series data

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.

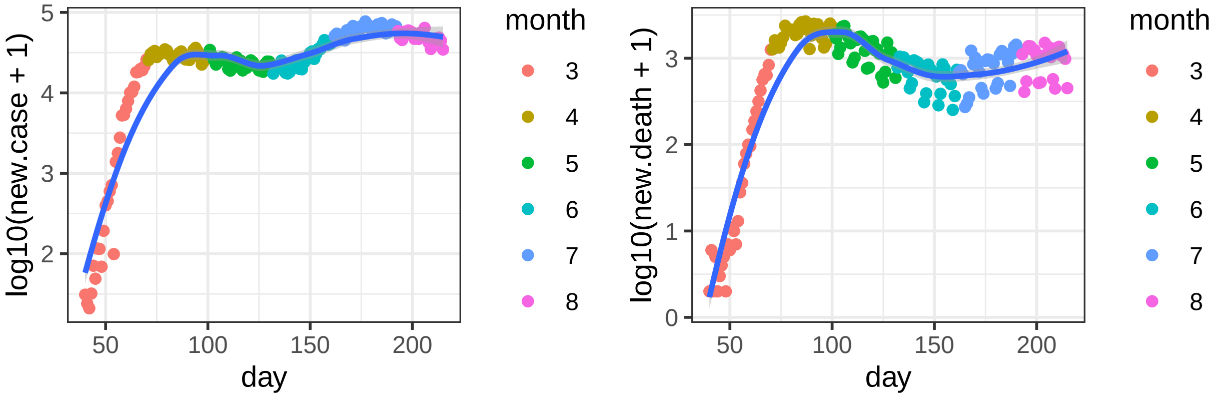


cumulative number of deaths on 8.23.20



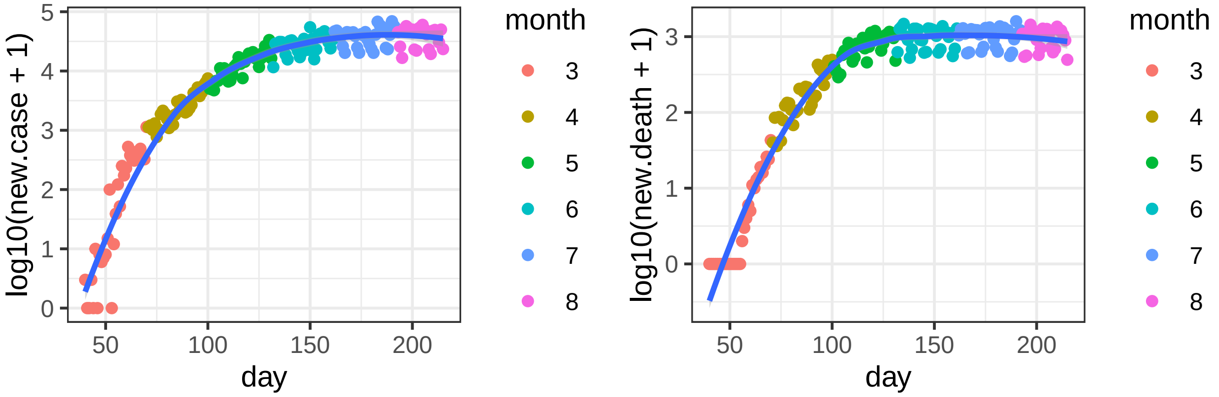
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



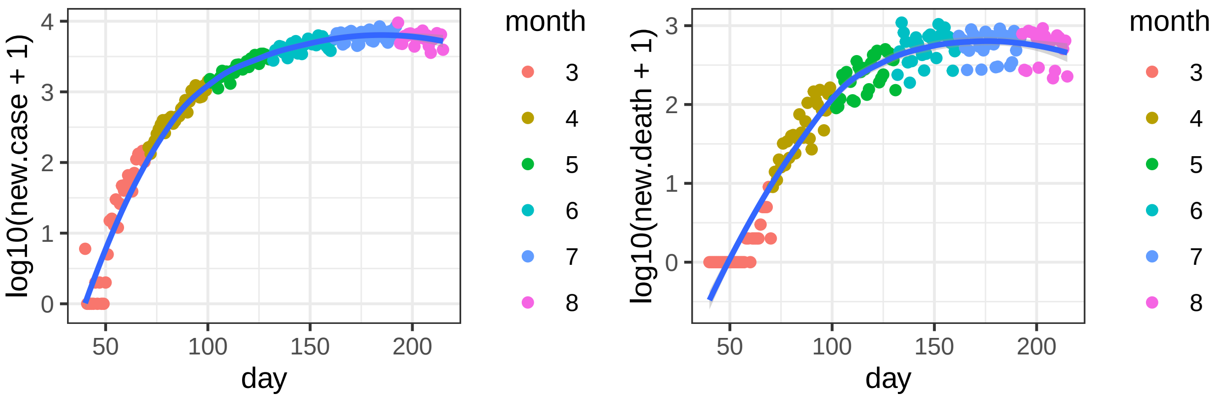
data source: https://github.com/CSSEGISandData/COVID−19

Brazil



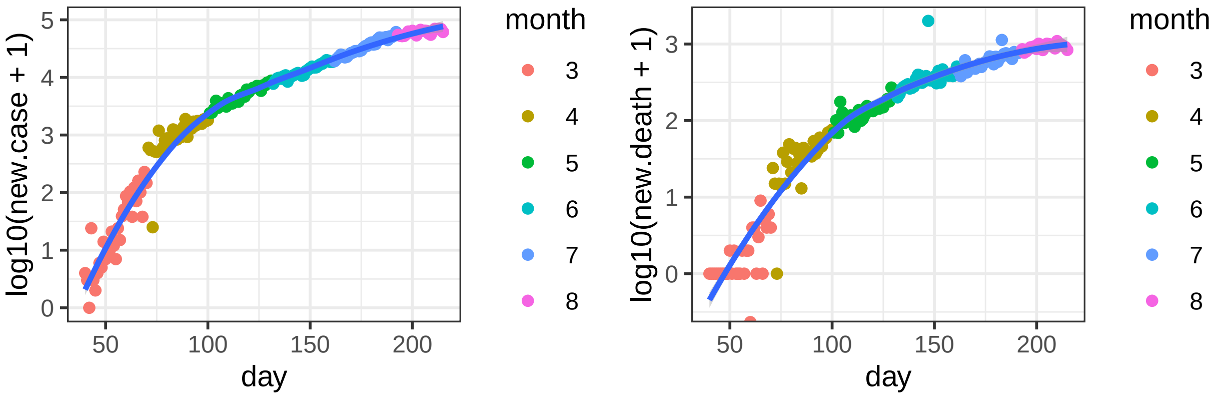
data source: https://github.com/CSSEGISandData/COVID−19

Mexico



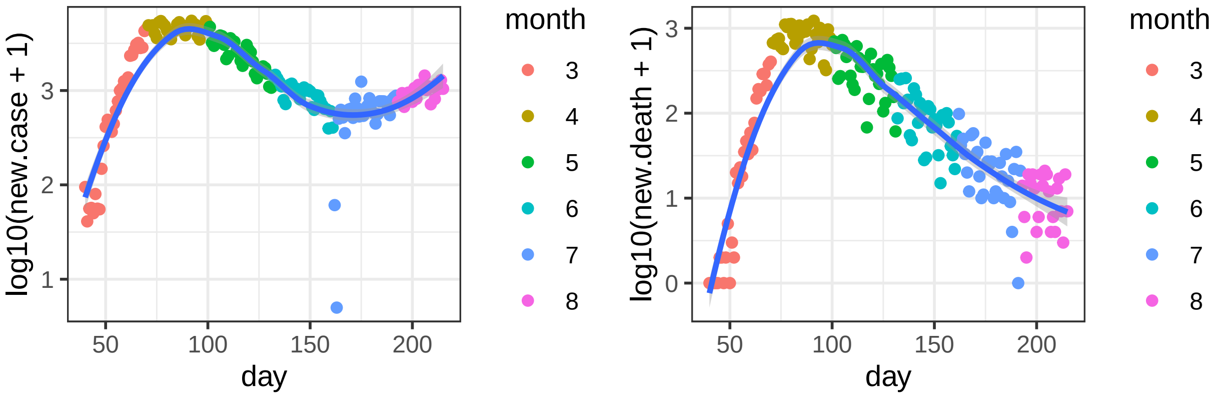
data source: https://github.com/CSSEGISandData/COVID−19

India

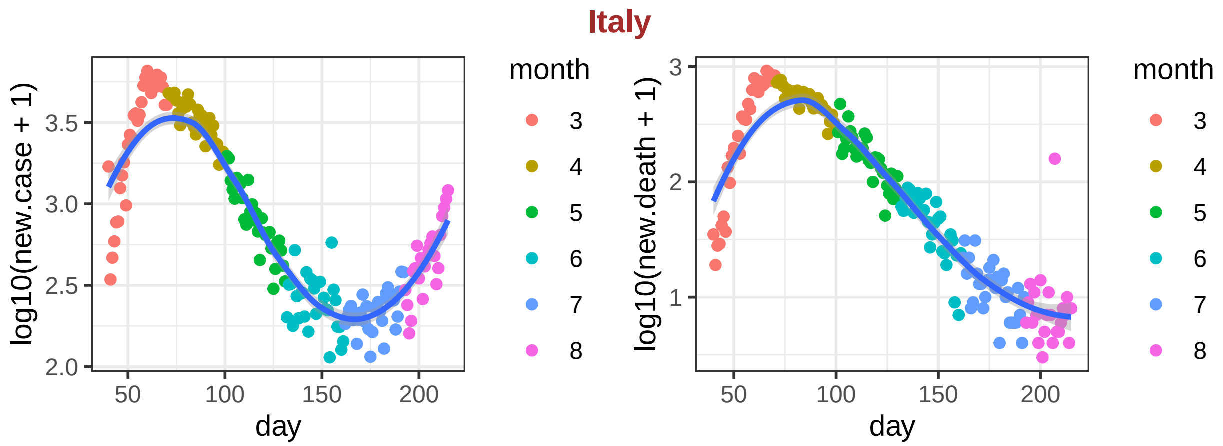


data source: https://github.com/CSSEGISandData/COVID−19

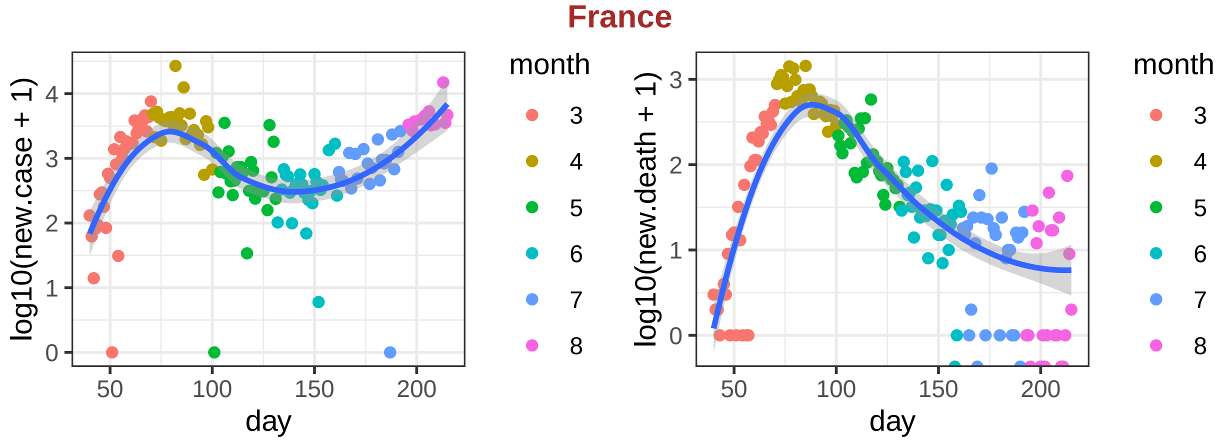
United Kingdom



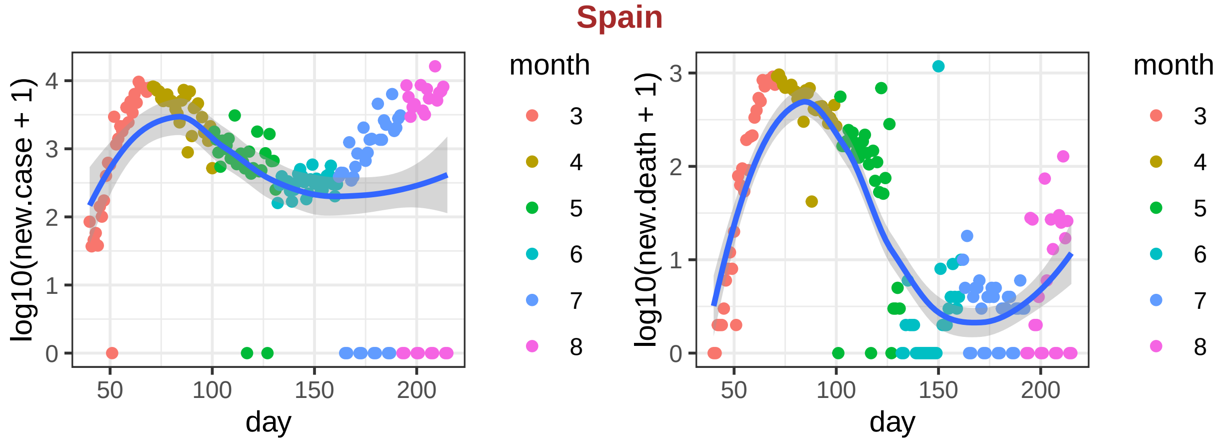
data source: https://github.com/CSSEGISandData/COVID−19



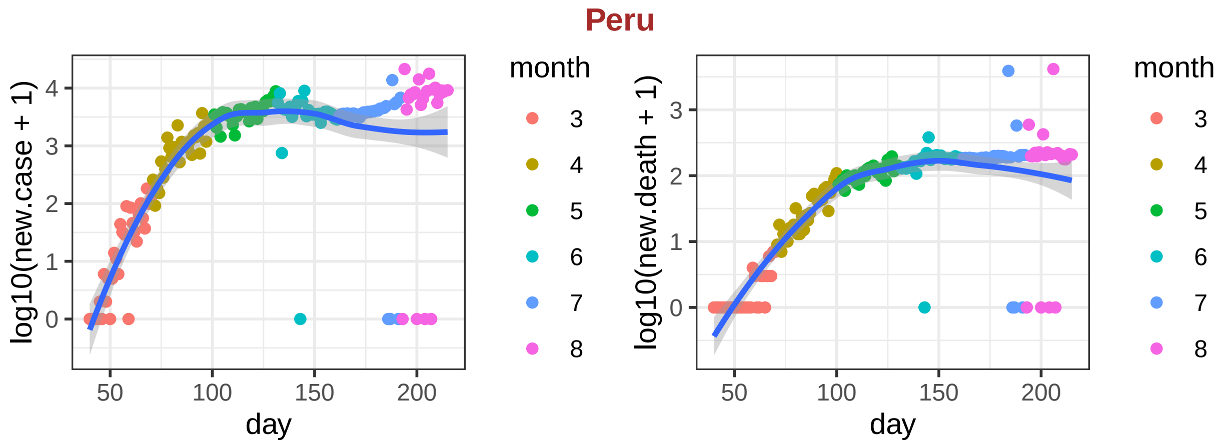
data source: https://github.com/CSSEGISandData/COVID−19



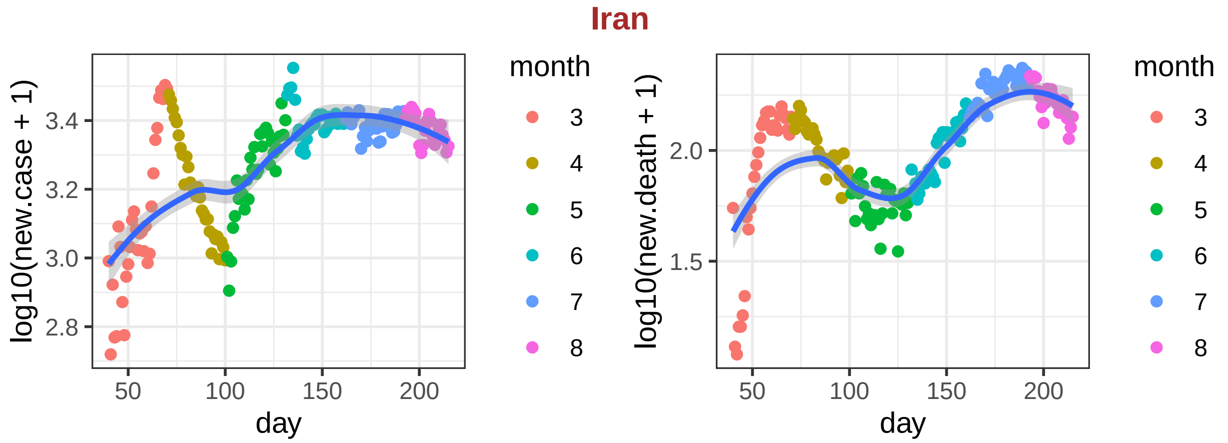
data source: https://github.com/CSSEGISandData/COVID−19



data source: https://github.com/CSSEGISandData/COVID−19



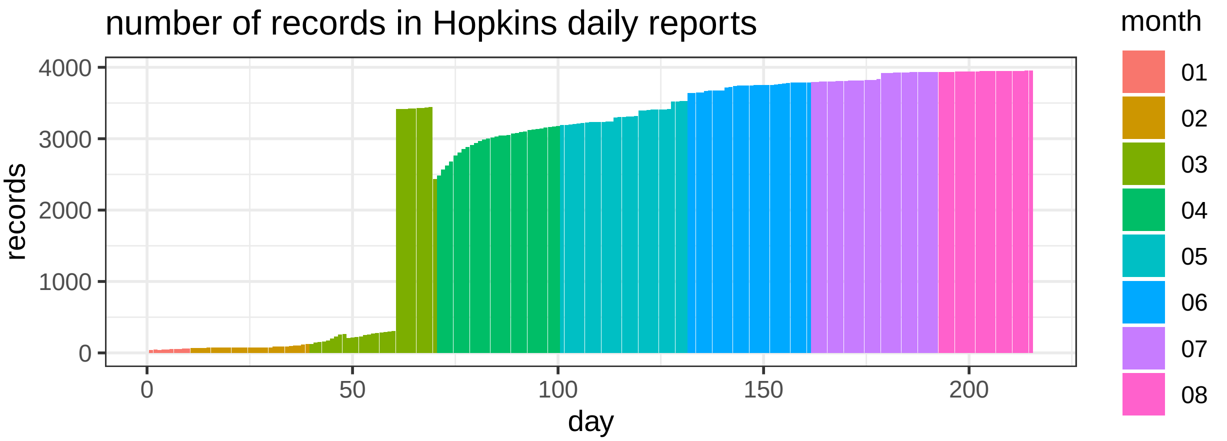
data source: https://github.com/CSSEGISandData/COVID−19



data source: https://github.com/CSSEGISandData/COVID−19

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] "2020-08-22"
```

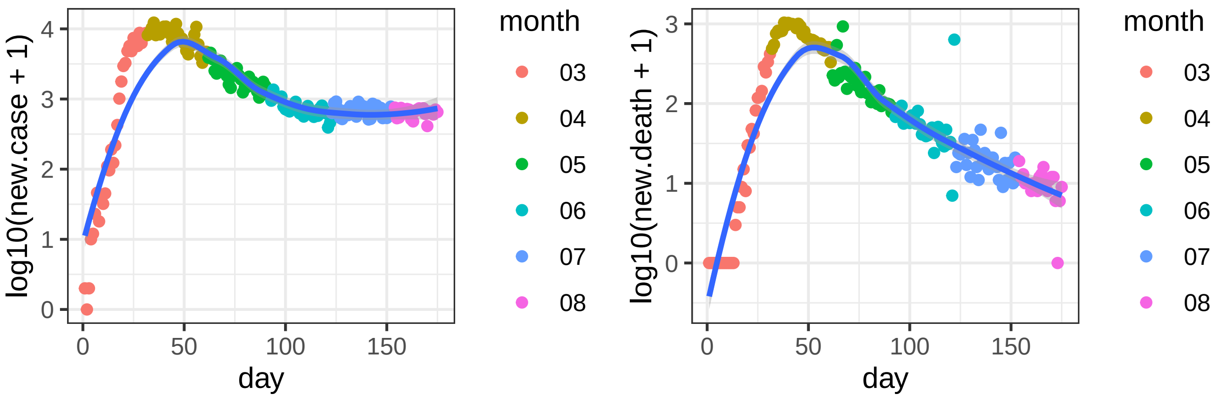
state level data

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

```
##            date          state fips  cases deaths
## 9508 2020-08-22       New York   36 433881  32464
## 9506 2020-08-22     New Jersey   34 191175  15943
## 9479 2020-08-22     California    6 665325  12137
## 9521 2020-08-22          Texas   48 599156  11650
## 9484 2020-08-22        Florida   12 597589  10273
## 9497 2020-08-22  Massachusetts   25 125360   8921
## 9489 2020-08-22       Illinois   17 220459   8107
## 9515 2020-08-22   Pennsylvania   42 133160   7643
## 9498 2020-08-22       Michigan   26 106141   6657
## 9485 2020-08-22        Georgia   13 235783   4982
## 9477 2020-08-22        Arizona    4 197909   4760
## 9494 2020-08-22      Louisiana   22 141861   4687
## 9481 2020-08-22    Connecticut    9  51519   4460
## 9512 2020-08-22           Ohio   39 114165   3975
## 9496 2020-08-22       Maryland   24 104040   3685
## 9490 2020-08-22        Indiana   18  87325   3218
## 9509 2020-08-22 North Carolina   37 153966   2546
## 9518 2020-08-22 South Carolina   45 111295   2493
## 9525 2020-08-22       Virginia   51 112072   2443
## 9500 2020-08-22    Mississippi   28  77268   2237
## 9475 2020-08-22        Alabama    1 114532   2011
## 9526 2020-08-22     Washington   53  73354   1945
## 9480 2020-08-22       Colorado    8  54939   1923
## 9499 2020-08-22      Minnesota   27  68913   1807
## 9520 2020-08-22      Tennessee   47 139366   1542
## 9501 2020-08-22       Missouri   29  75409   1519
## 9504 2020-08-22         Nevada   32  65150   1197
## 9528 2020-08-22      Wisconsin   55  74740   1092
## 9491 2020-08-22           Iowa   19  55996   1033
## 9517 2020-08-22   Rhode Island   44  21022   1030
```

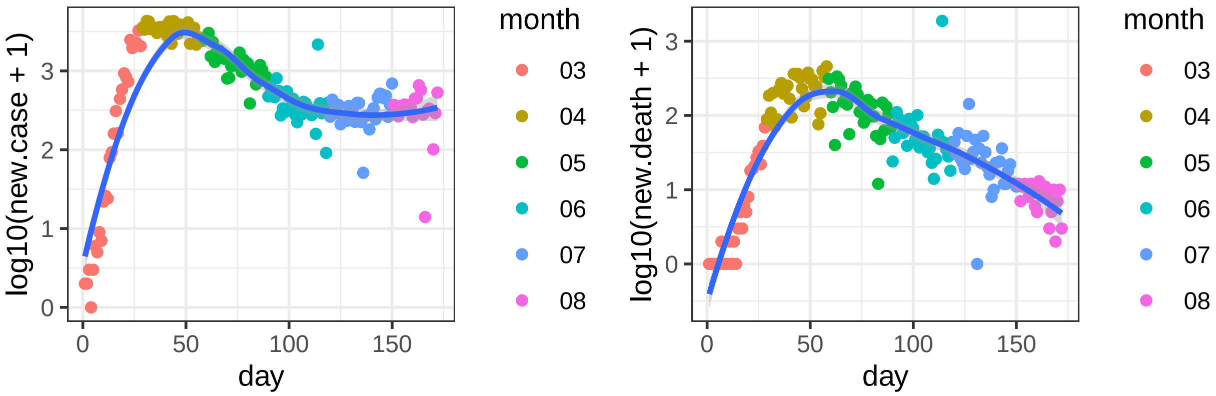
For these 30 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.

New York



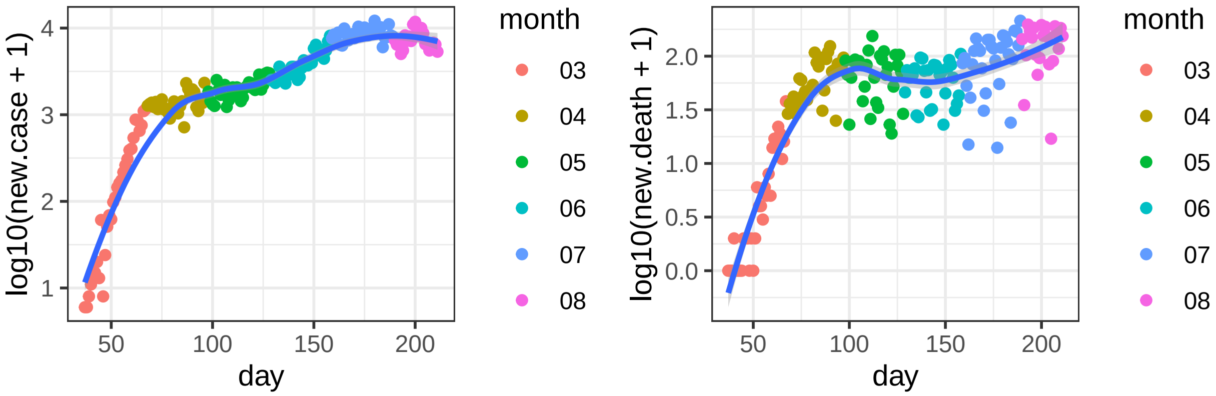
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



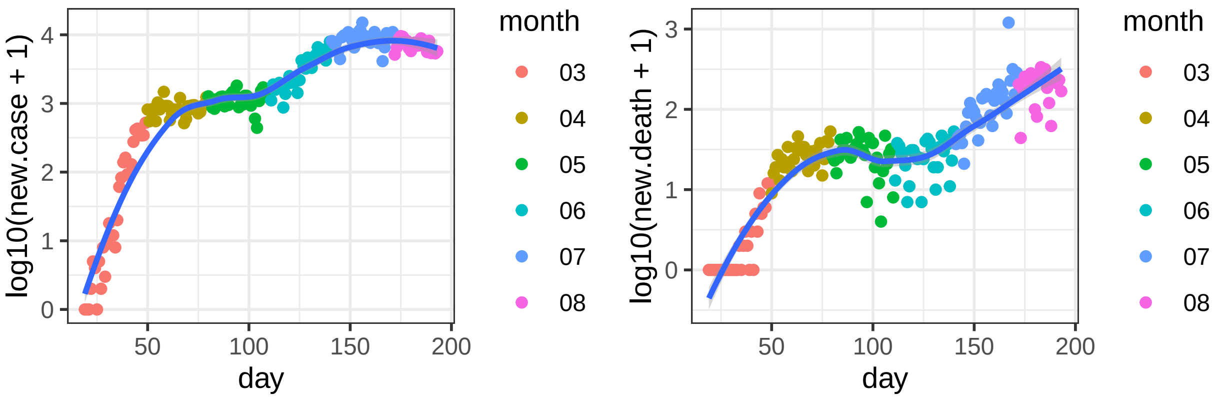
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

California



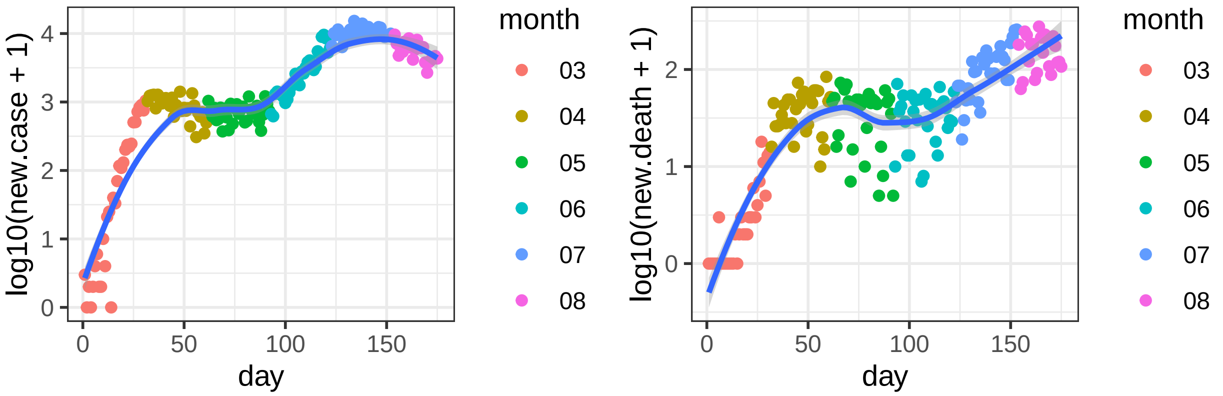
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Texas



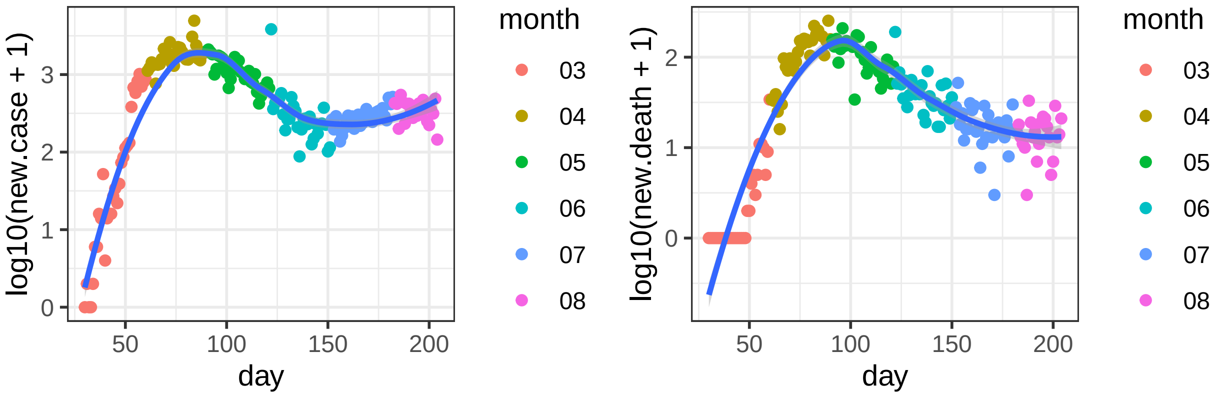
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Florida

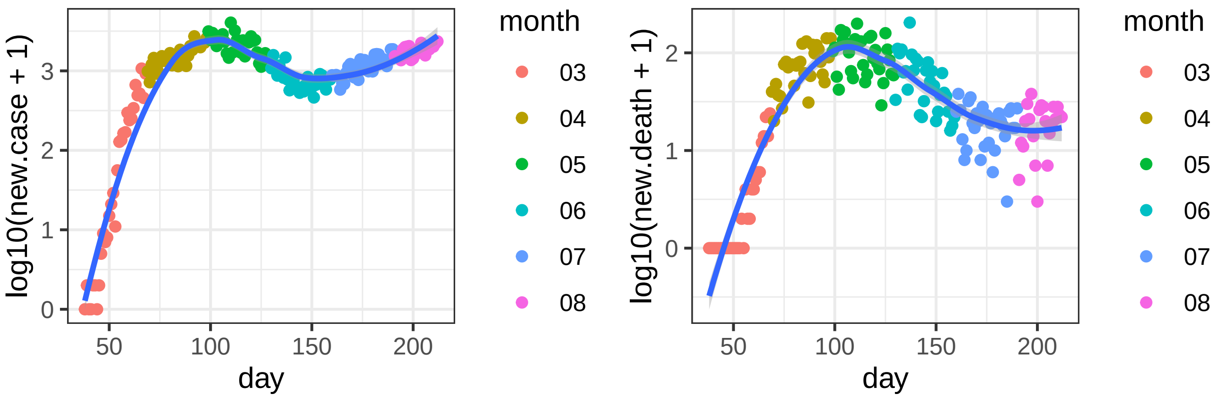


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Massachusetts

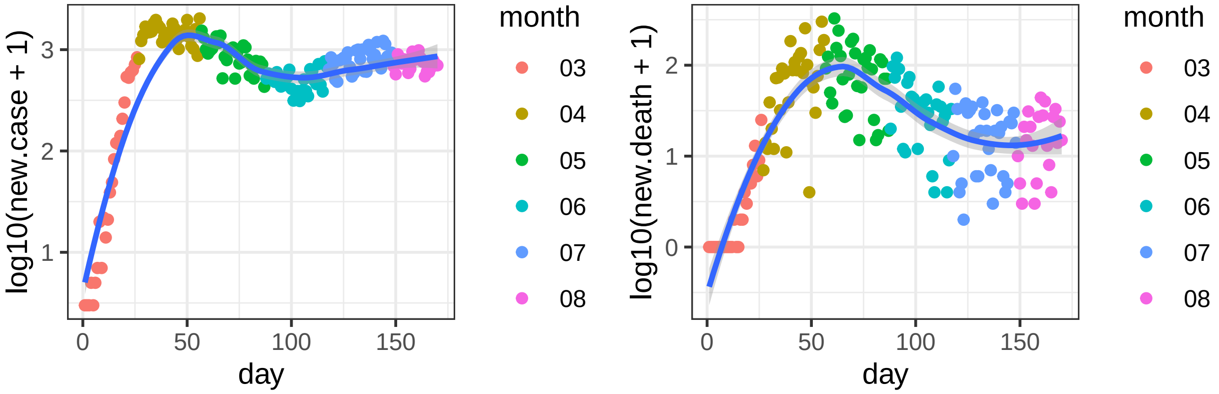


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



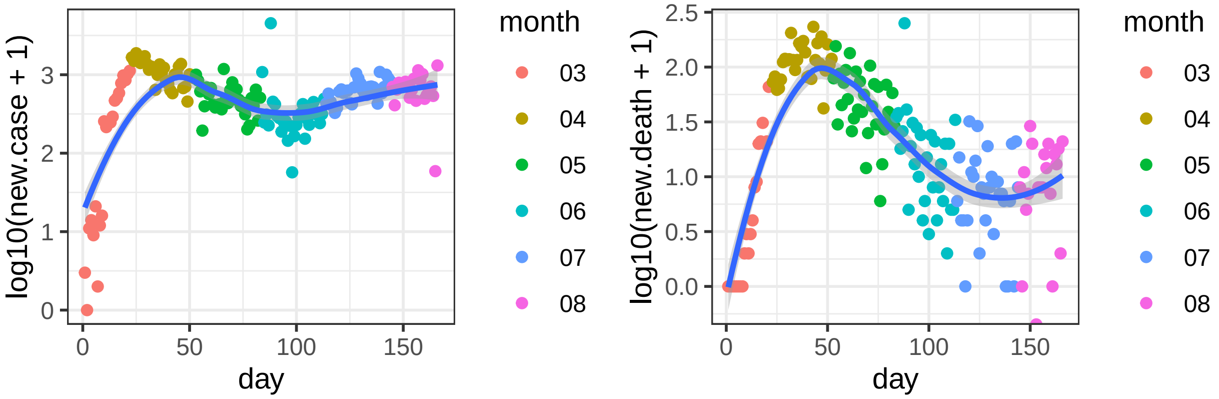
Illinois

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



Pennsylvania

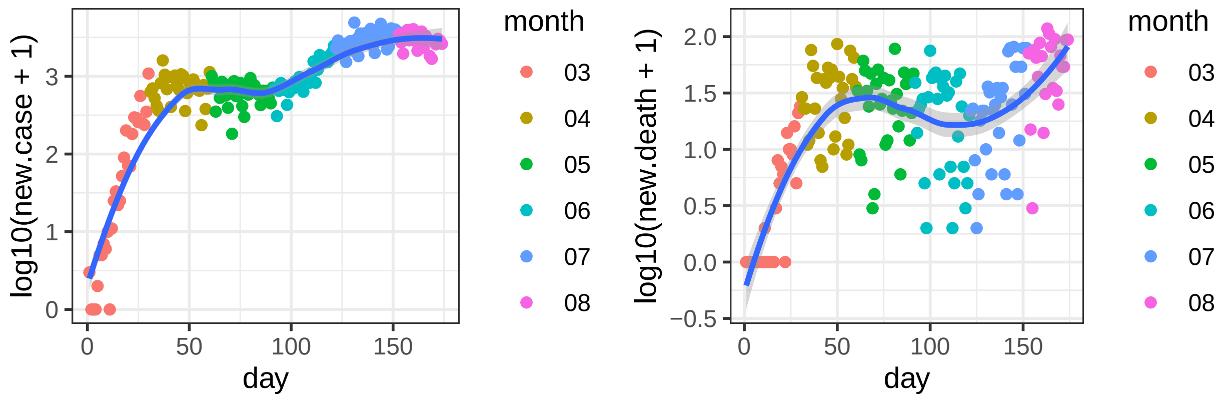
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



Michigan

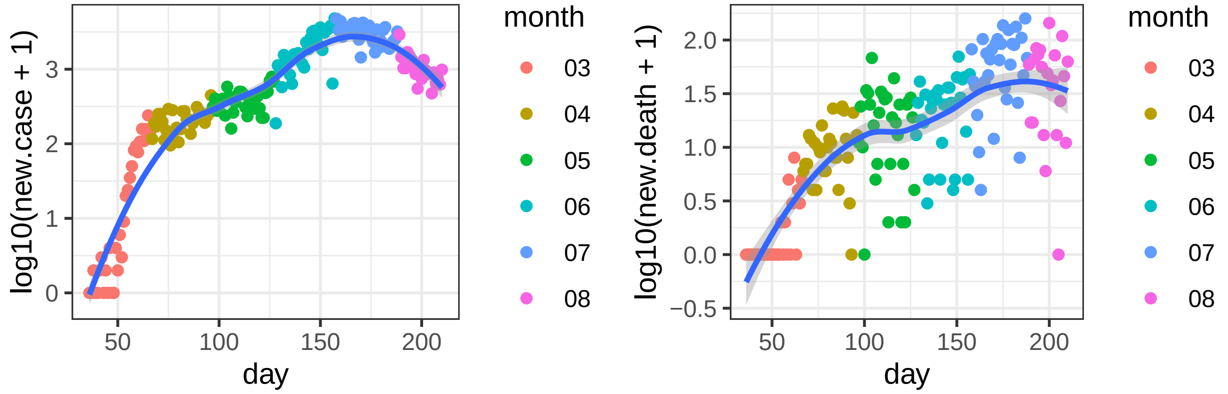
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Georgia



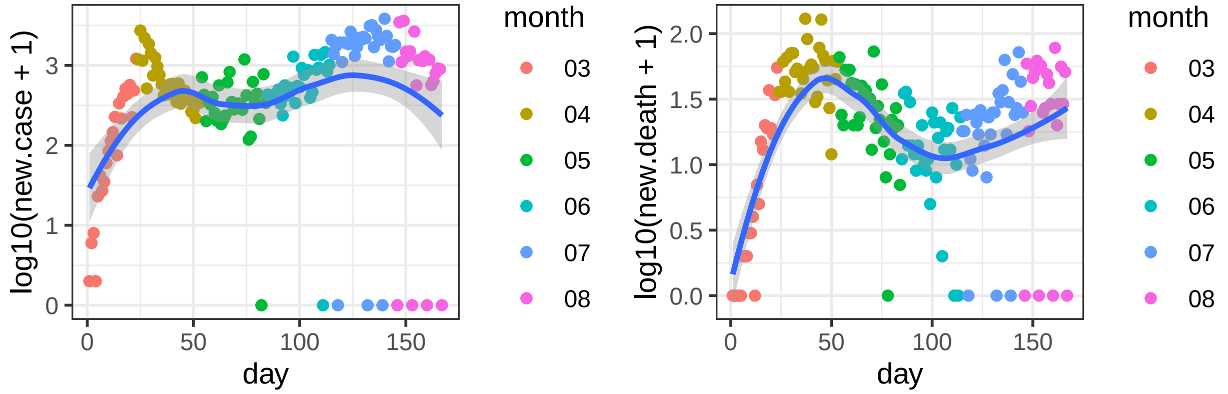
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Arizona



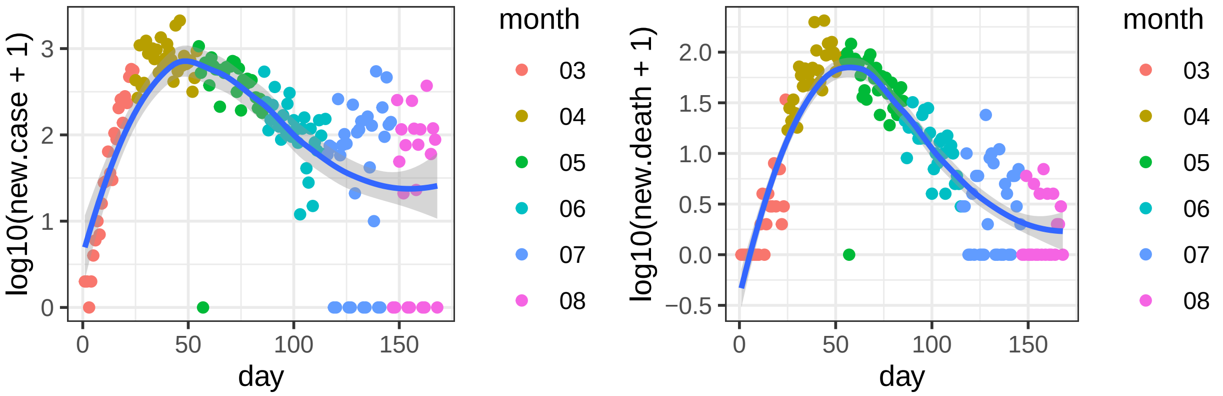
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Louisiana



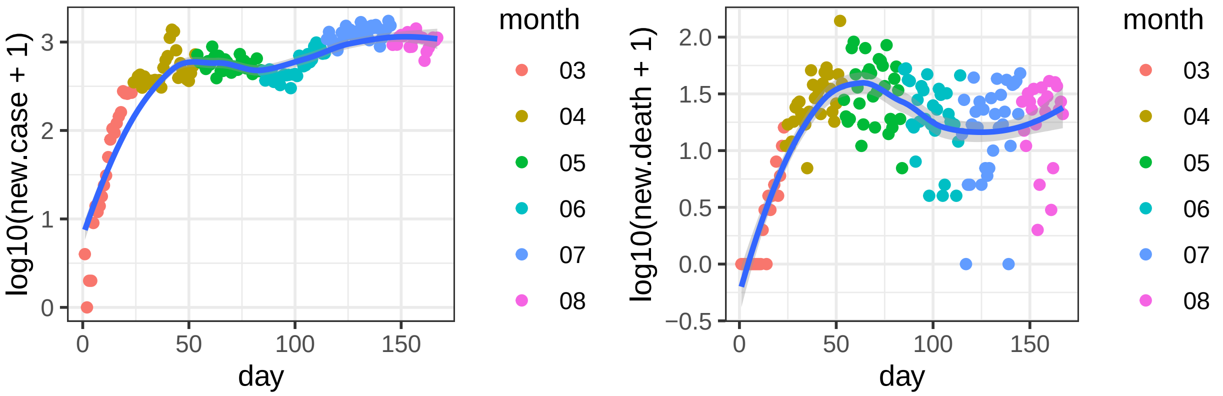
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Connecticut



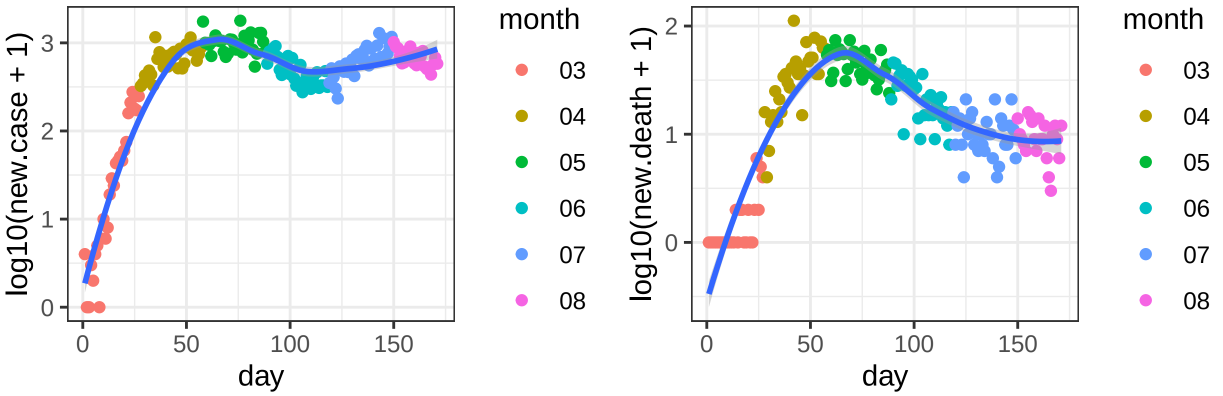
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Ohio

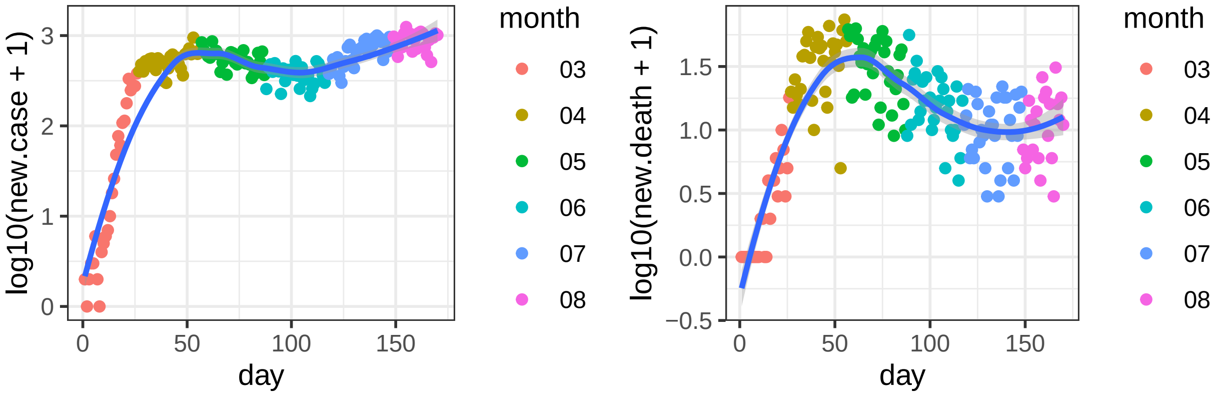


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland

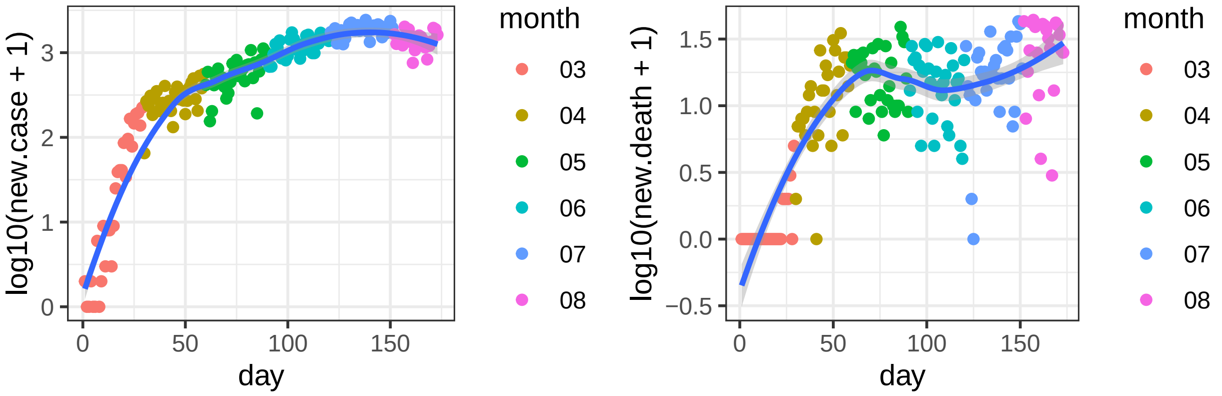


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



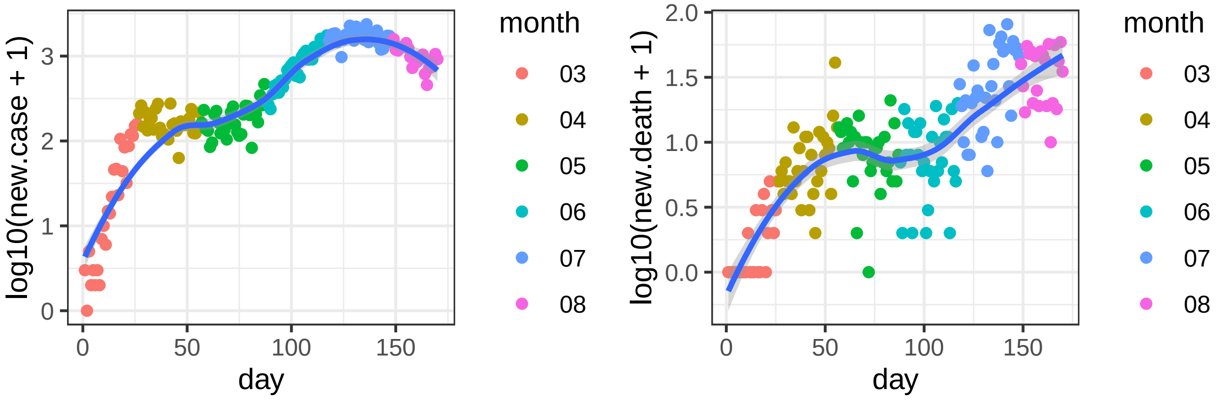
Indiana

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



North Carolina

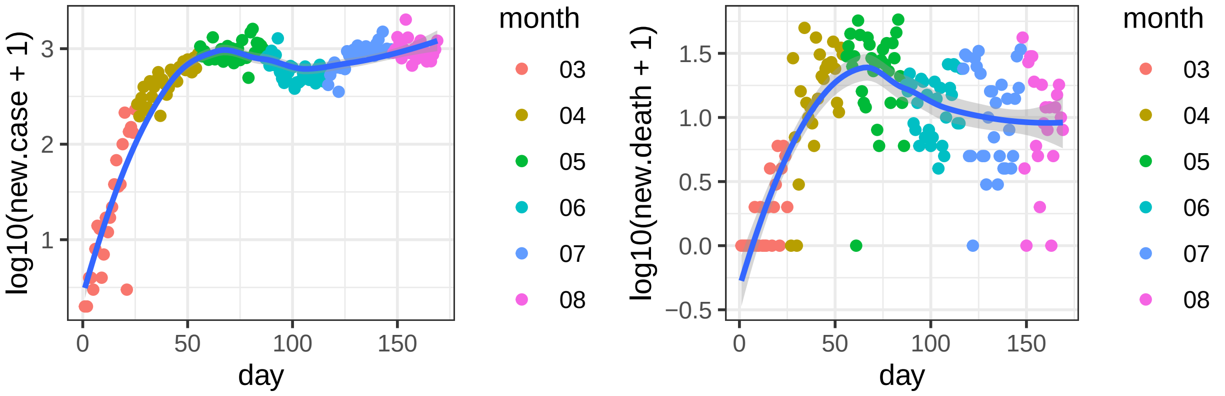
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03



South Carolina

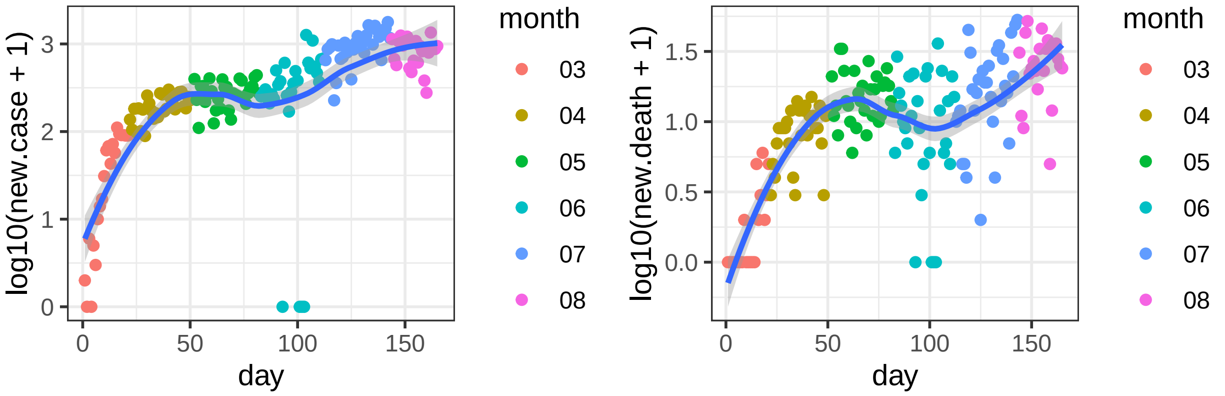
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Virginia



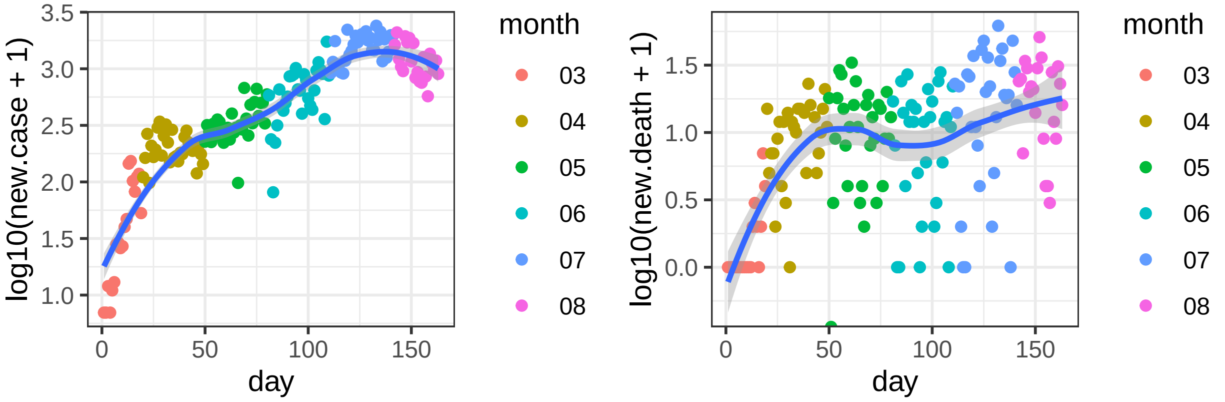
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi

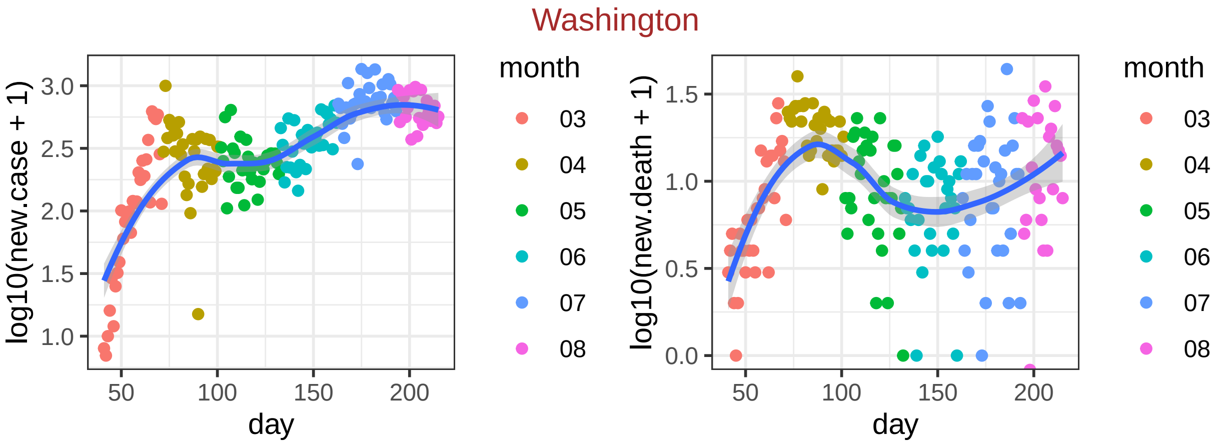


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

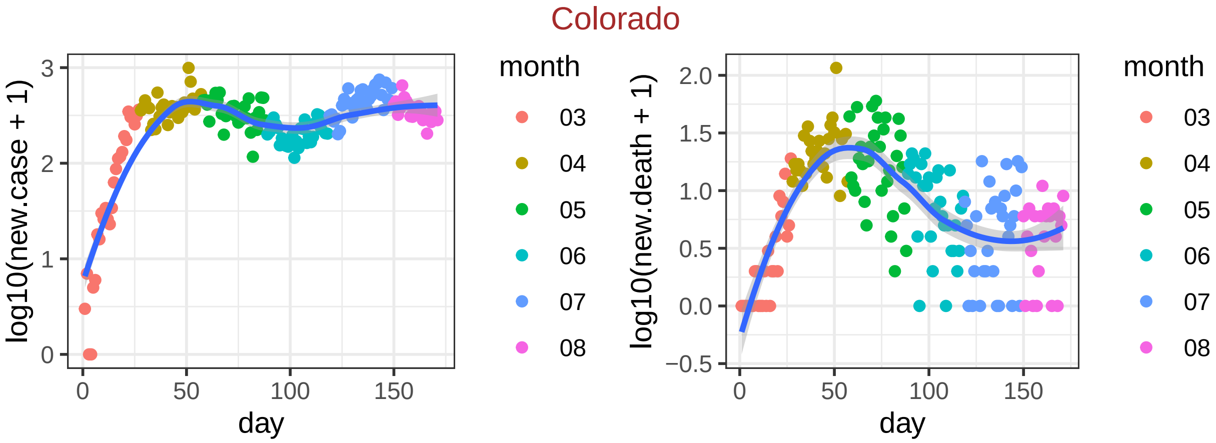
Alabama



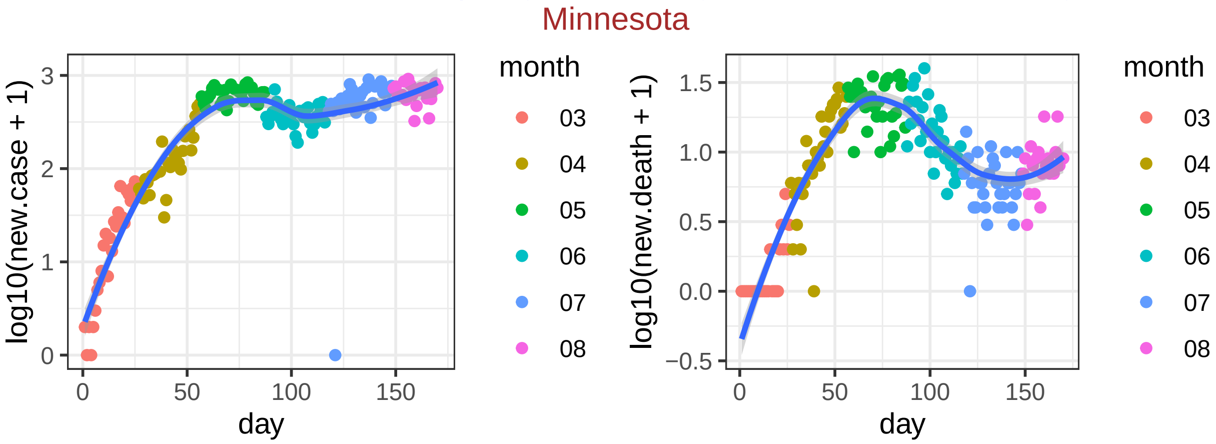
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13



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

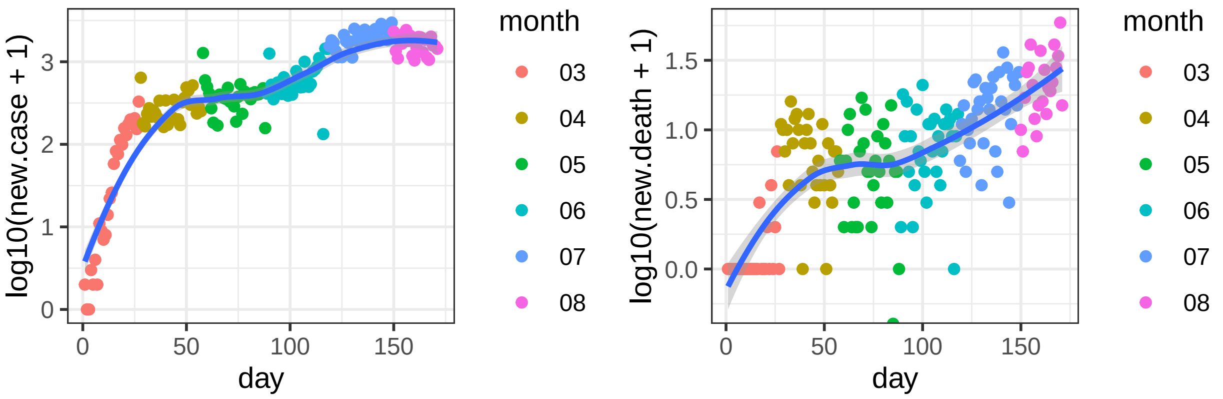


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



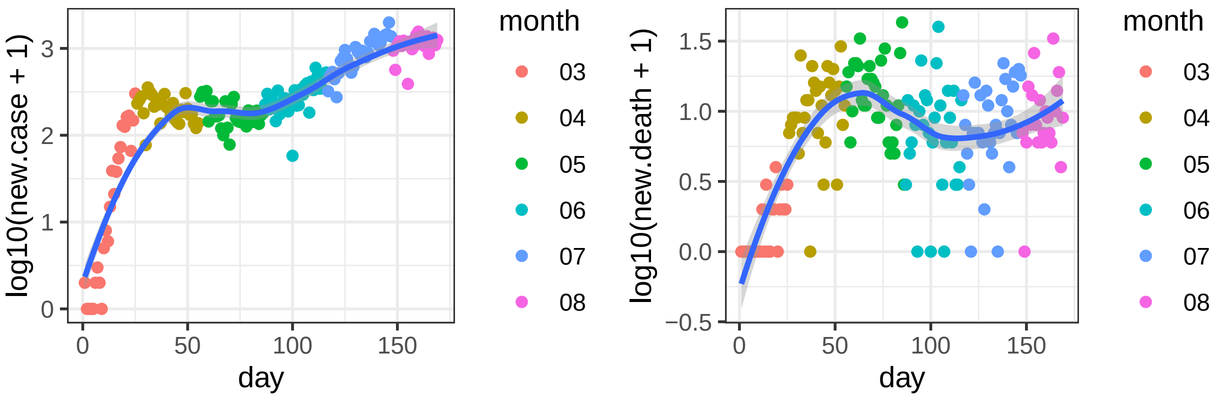
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Tennessee



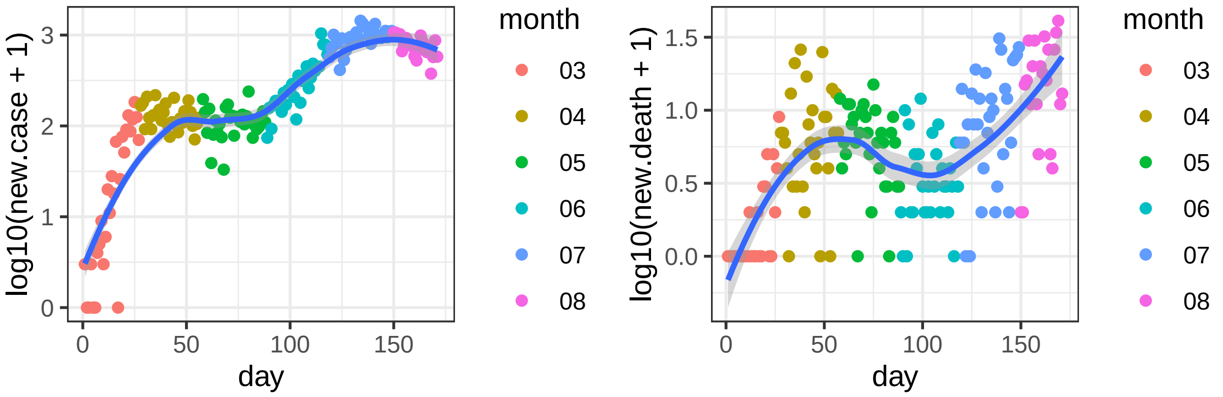
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Missouri



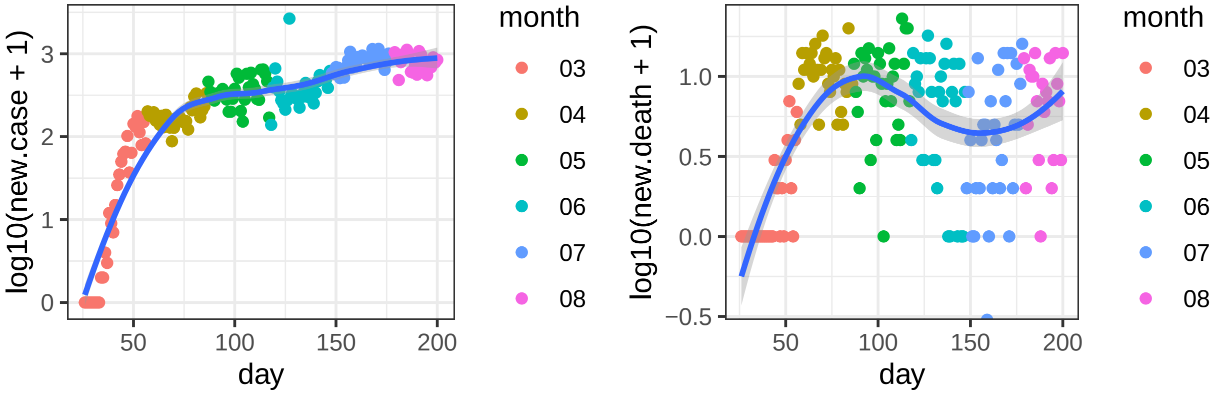
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Nevada



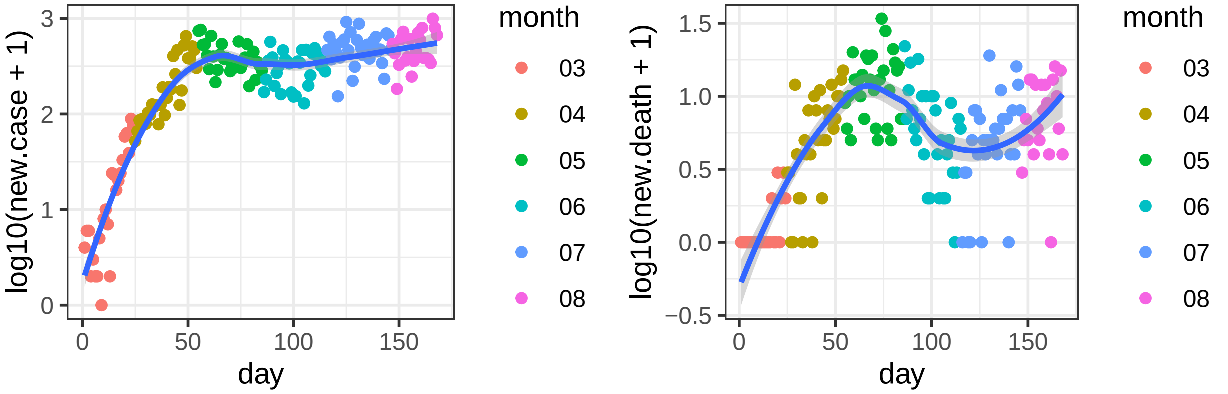
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wisconsin



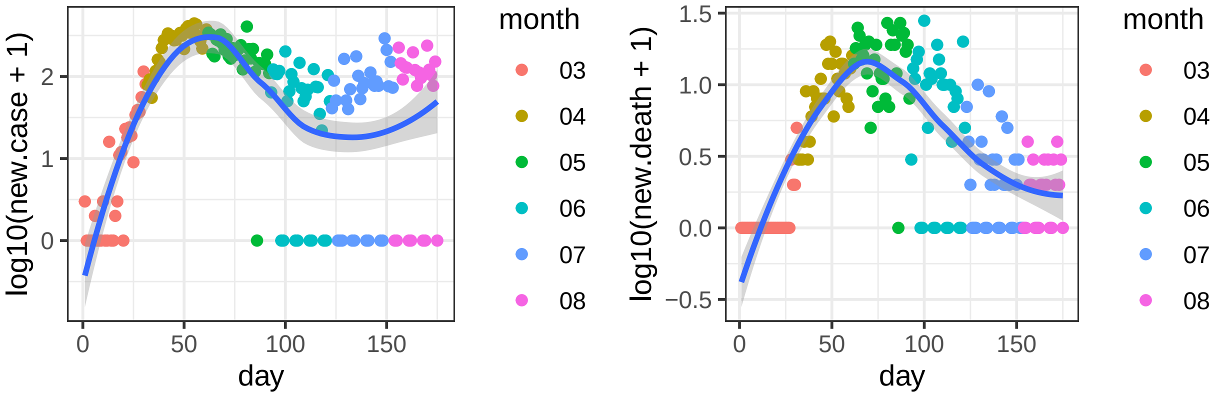
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Iowa



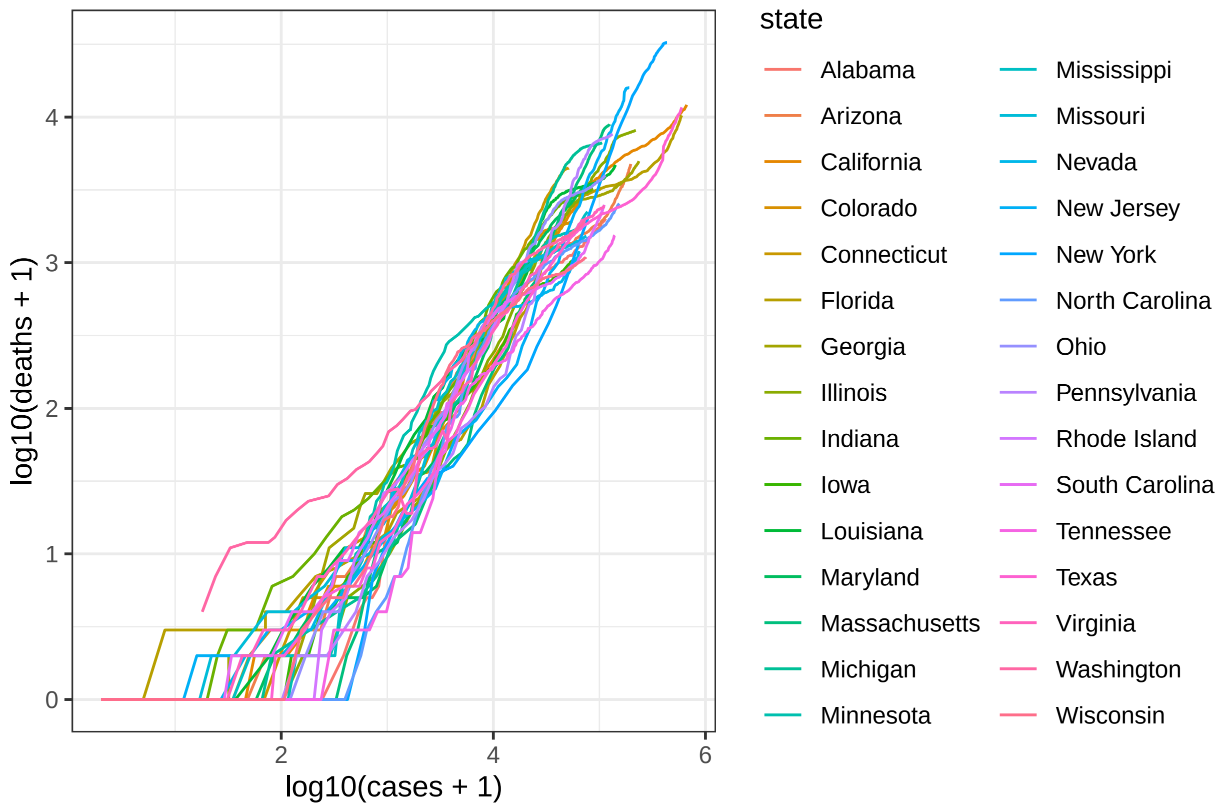
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Rhode Island



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

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

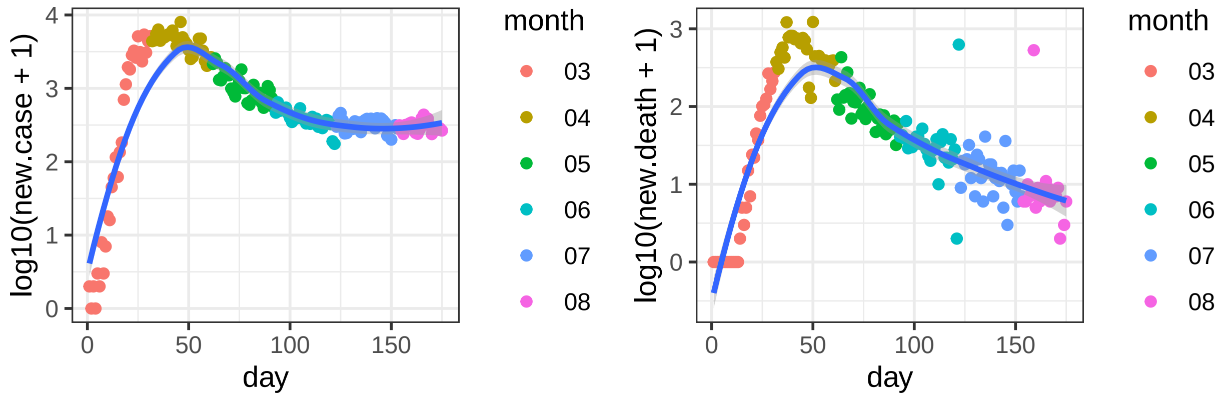
county level data

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

```
##              date          county         state  fips  cases deaths
## 458742 2020-08-22   New York City      New York    NA 236534  23646
## 457086 2020-08-22     Los Angeles    California  6037 230662   5537
## 457495 2020-08-22            Cook      Illinois 17031 120567   5008
## 458201 2020-08-22           Wayne      Michigan 26163  29879   2859
## 456984 2020-08-22        Maricopa       Arizona  4013 131685   2776
## 457245 2020-08-22      Miami-Dade       Florida 12086 151213   2238
## 458741 2020-08-22          Nassau      New York 36059  44205   2196
## 458665 2020-08-22           Essex    New Jersey 34013  20322   2112
## 458112 2020-08-22       Middlesex Massachusetts 25017  27056   2043
## 458660 2020-08-22          Bergen    New Jersey 34003  21560   2031
## 459587 2020-08-22          Harris         Texas 48201  97745   2011
## 458761 2020-08-22         Suffolk      New York 36103  44456   2001
## 459179 2020-08-22    Philadelphia  Pennsylvania 42101  32936   1758
## 458667 2020-08-22          Hudson    New Jersey 34017  20184   1510
## 458769 2020-08-22     Westchester      New York 36119  36650   1449
## 457190 2020-08-22        Hartford   Connecticut  9003  13002   1422
## 458670 2020-08-22       Middlesex    New Jersey 34023  18454   1420
## 457189 2020-08-22       Fairfield   Connecticut  9001  18434   1411
## 458678 2020-08-22           Union    New Jersey 34039  17133   1352
## 458674 2020-08-22         Passaic    New Jersey 34031  18291   1246
## 458108 2020-08-22           Essex Massachusetts 25009  18430   1215
## 458181 2020-08-22         Oakland      Michigan 26125  17106   1151
## 457193 2020-08-22       New Haven   Connecticut  9009  13442    1111
## 458116 2020-08-22         Suffolk Massachusetts 25025  22706    1093
## 457208 2020-08-22         Broward       Florida 12011  68891    1088
## 459594 2020-08-22         Hidalgo         Texas 48215  23993    1071
## 457252 2020-08-22      Palm Beach       Florida 12099  40385    1059
## 458118 2020-08-22       Worcester Massachusetts 25027  13955    1036
## 458635 2020-08-22           Clark        Nevada 32003  56010    1027
## 458673 2020-08-22           Ocean    New Jersey 34029  10975    1025
## 458114 2020-08-22         Norfolk Massachusetts 25021  10845    1004
## 459502 2020-08-22           Bexar         Texas 48029  45168     992
## 458168 2020-08-22          Macomb      Michigan 26099  12121     963
## 457100 2020-08-22       Riverside    California  6065  49482     927
## 457097 2020-08-22          Orange    California  6059  45801     896
## 459543 2020-08-22          Dallas         Texas 48113  71148     878
## 458229 2020-08-22        Hennepin     Minnesota 27053  21466     865
## 459174 2020-08-22      Montgomery  Pennsylvania 42091  10667     861
## 458671 2020-08-22        Monmouth    New Jersey 34025  10669     860
## 458672 2020-08-22          Morris    New Jersey 34027   7546     830
## 459278 2020-08-22      Providence  Rhode Island 44007  16142     826
## 458094 2020-08-22      Montgomery      Maryland 24031  19424     814
## 457631 2020-08-22          Marion       Indiana 18097  17502     794
## 458095 2020-08-22 Prince George's      Maryland 24033  25746     777
## 459151 2020-08-22        Delaware  Pennsylvania 42045  10011     770
## 458115 2020-08-22        Plymouth Massachusetts 25023   9443     736
## 458110 2020-08-22          Hampden Massachusetts 25013   7791     734
## 459935 2020-08-22            King    Washington 53033  18589     731
## 458475 2020-08-22       St. Louis      Missouri 29189  17721     713
## 457103 2020-08-22  San Bernardino    California  6071  44603     691
```

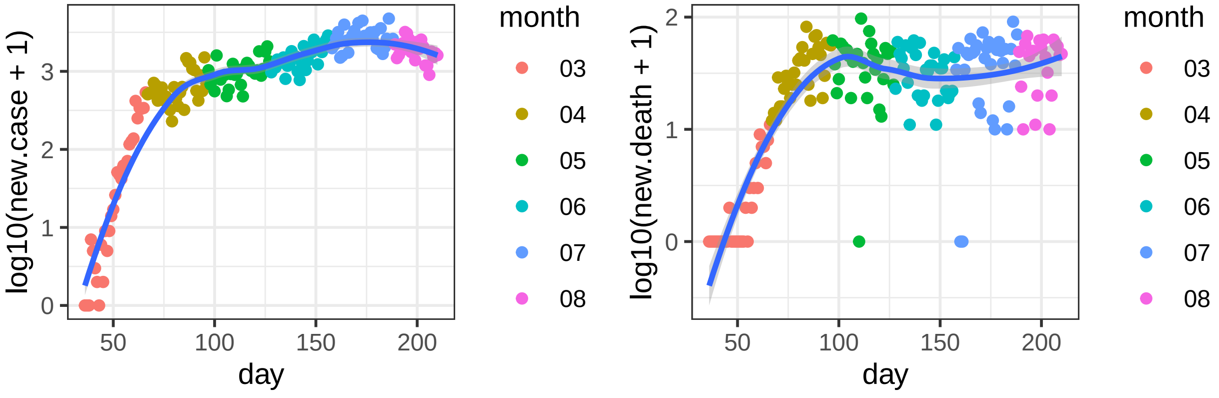
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



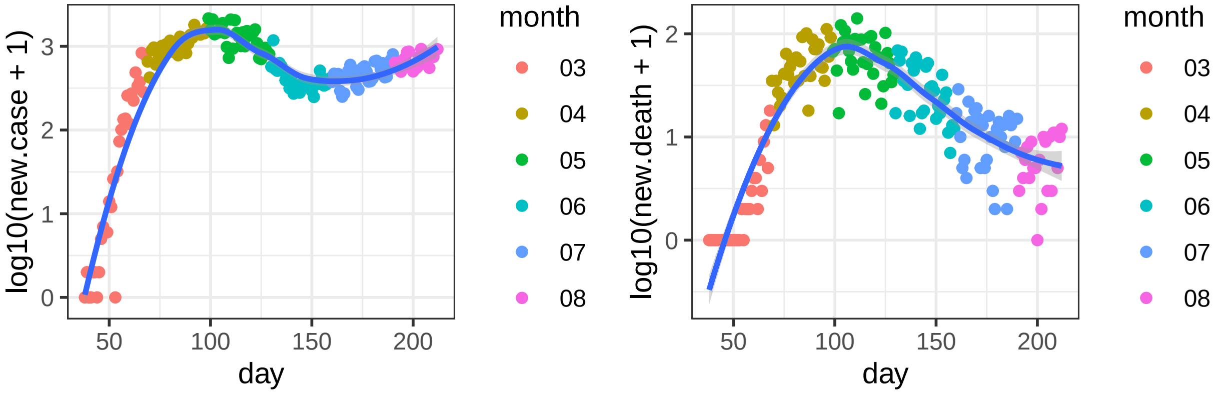
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Los Angeles_California



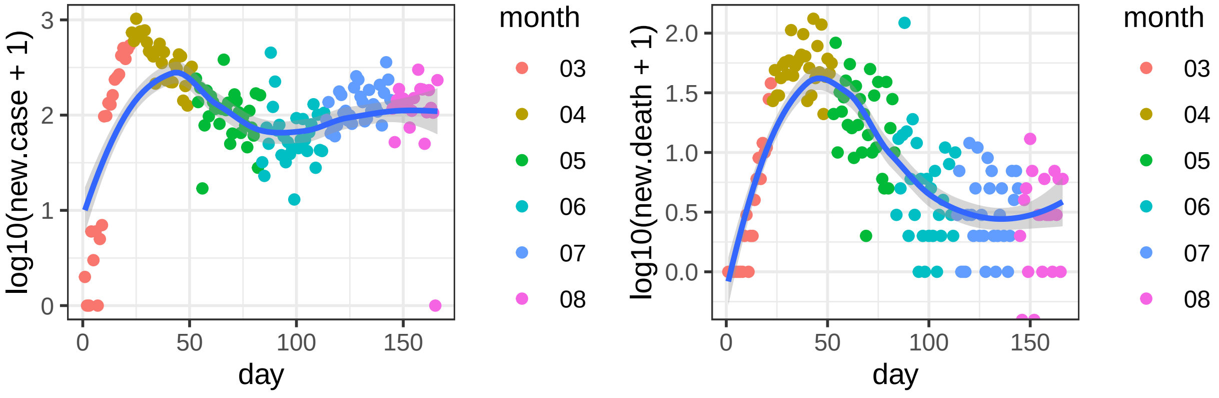
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



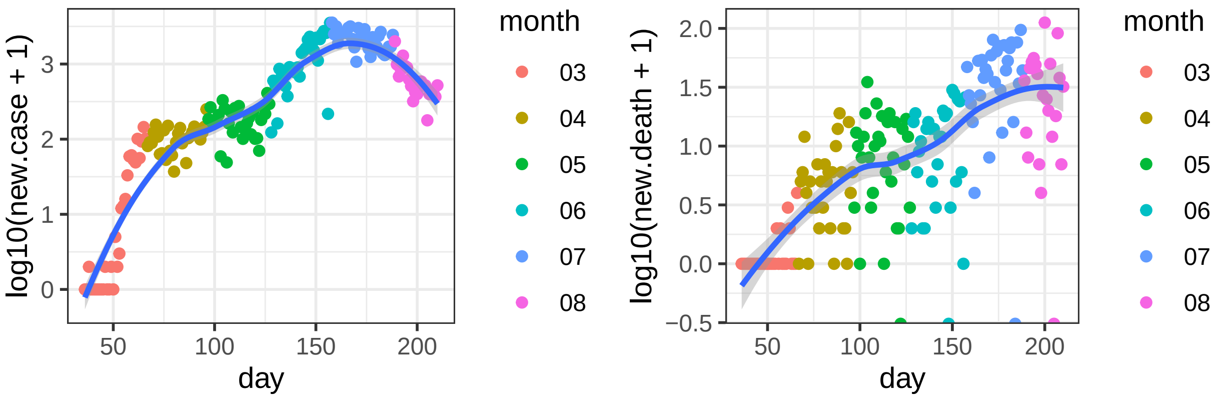
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wayne_Michigan



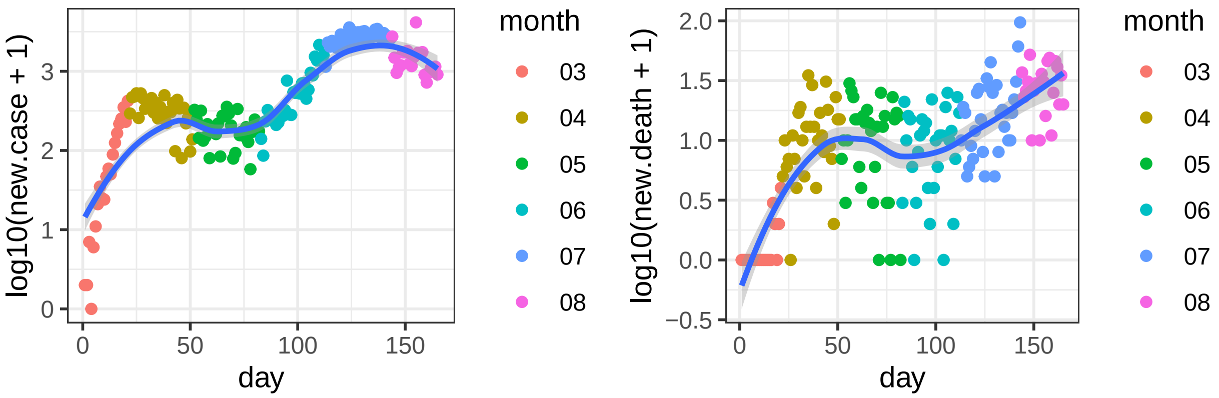
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Maricopa_Arizona



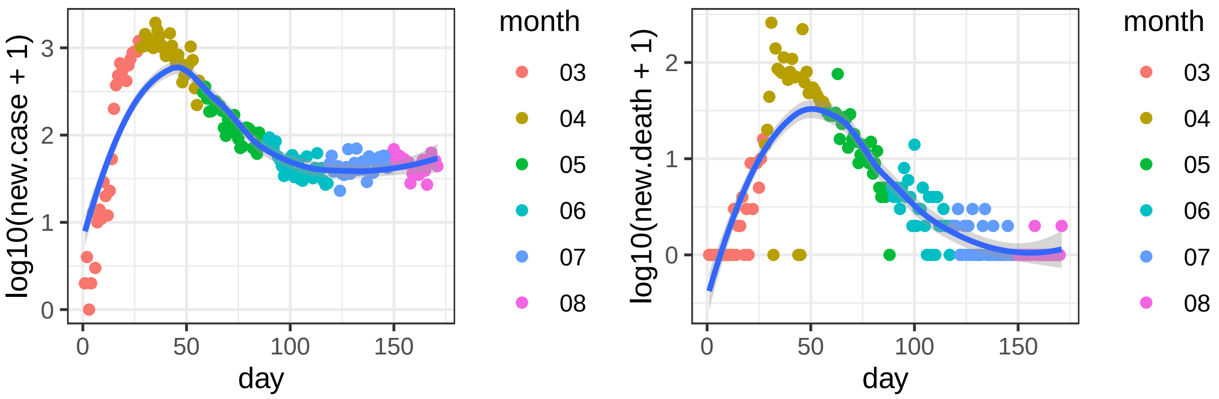
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Miami−Dade_Florida

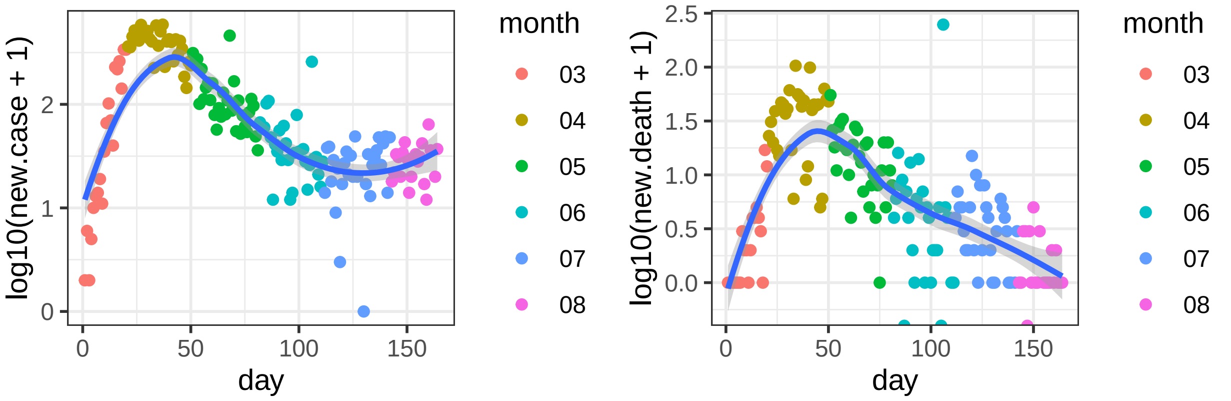


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Nassau_New York

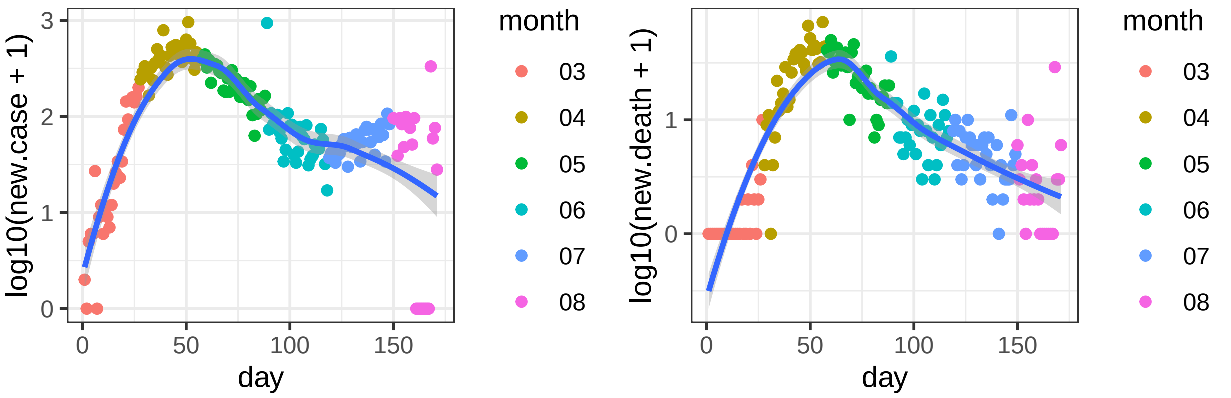


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



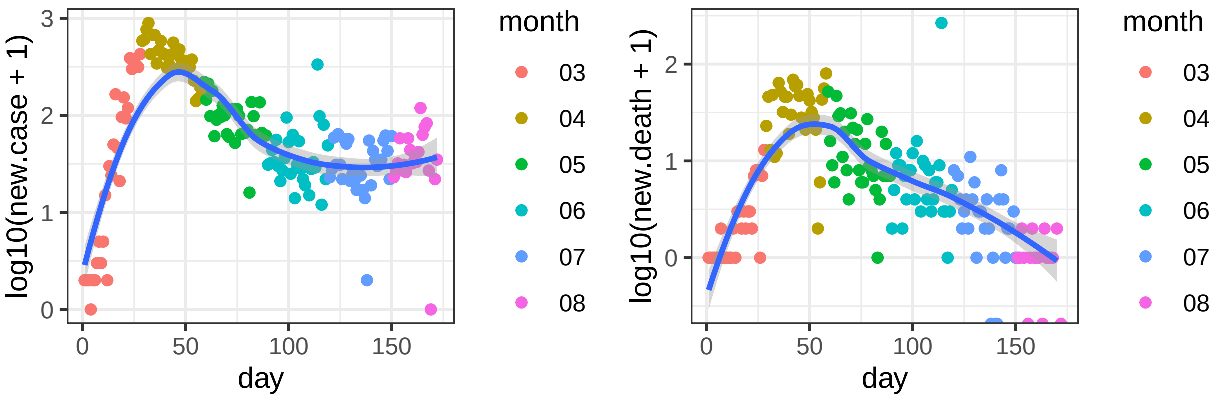
Essex_New Jersey

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



Middlesex_Massachusetts

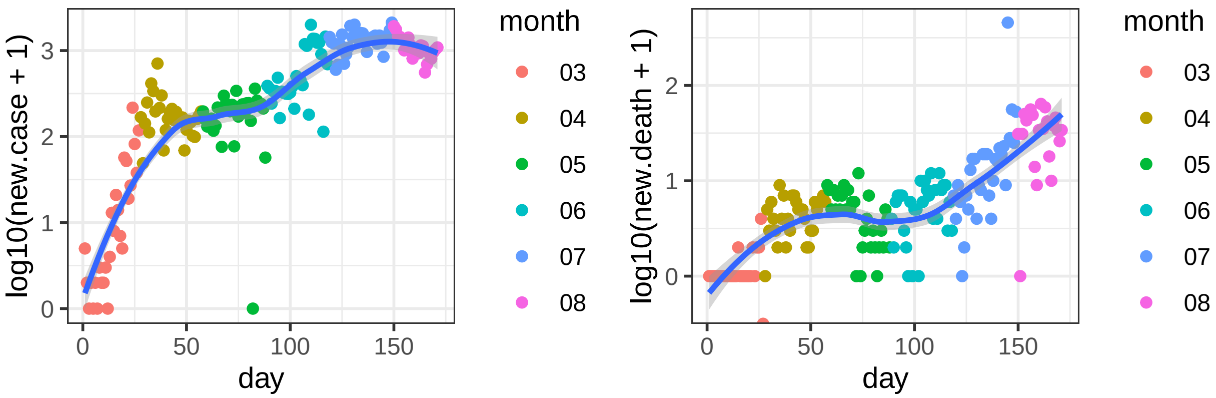
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



Bergen_New Jersey

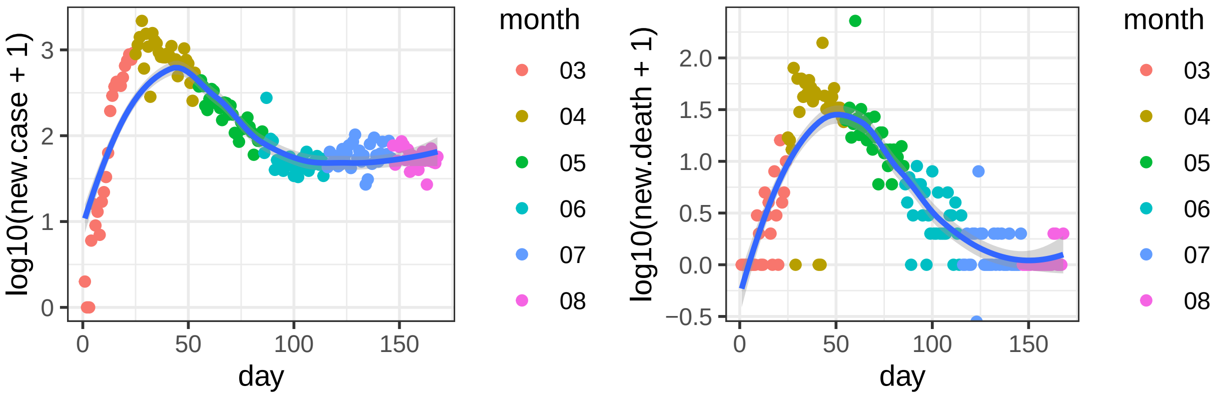
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Harris_Texas



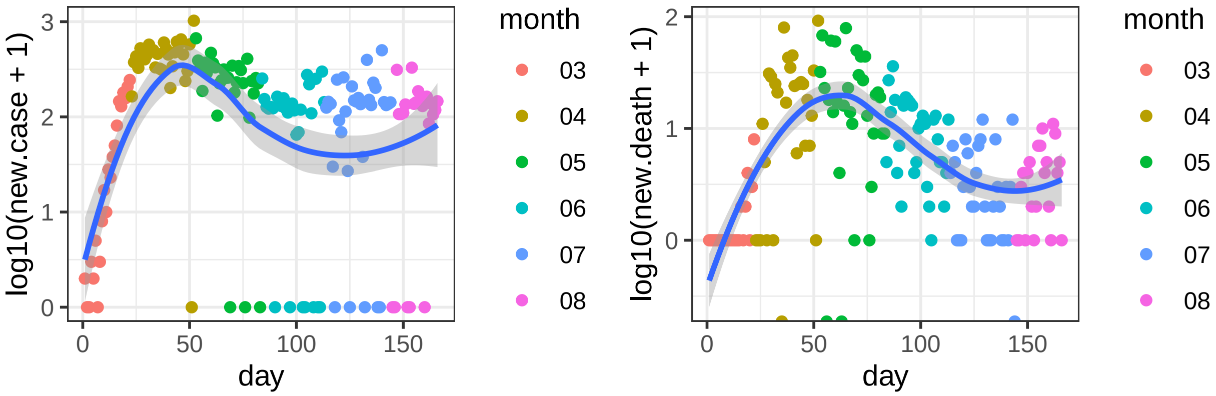
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Suffolk_New York



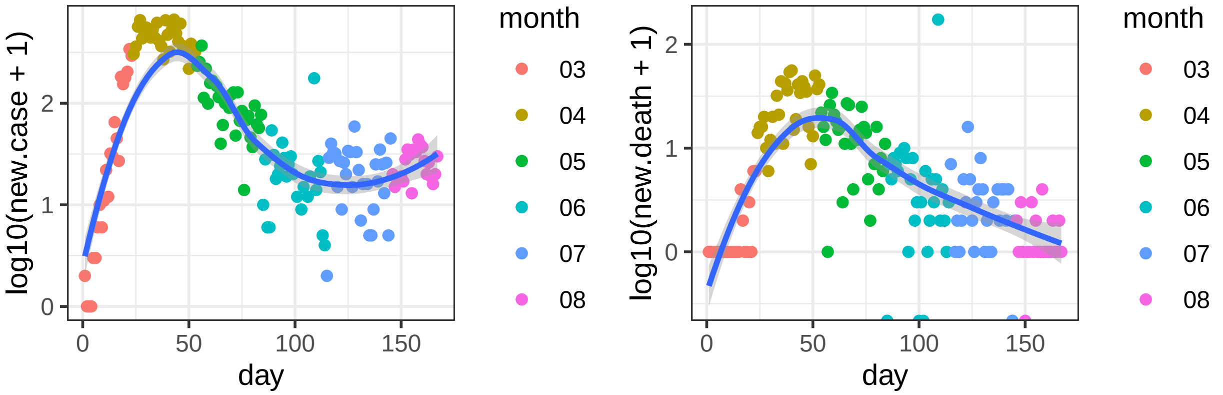
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Philadelphia_Pennsylvania



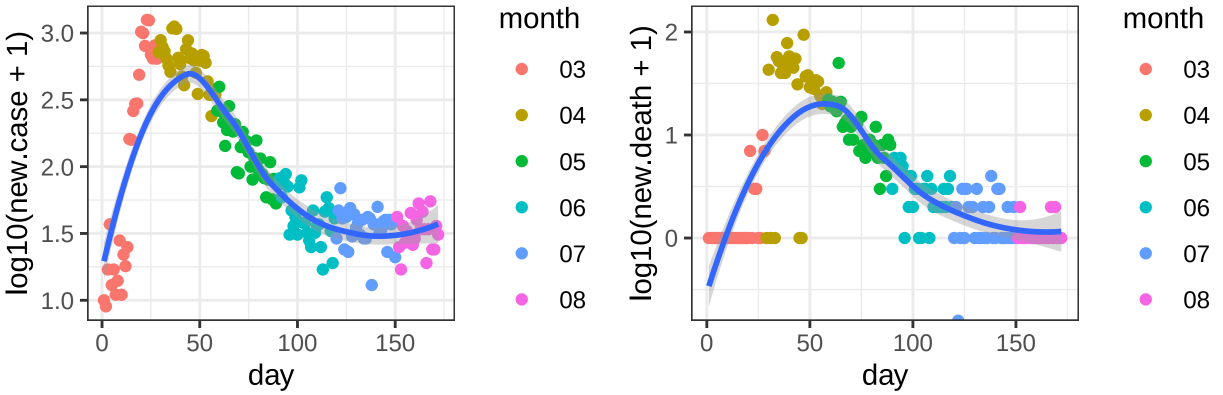
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Hudson_New Jersey



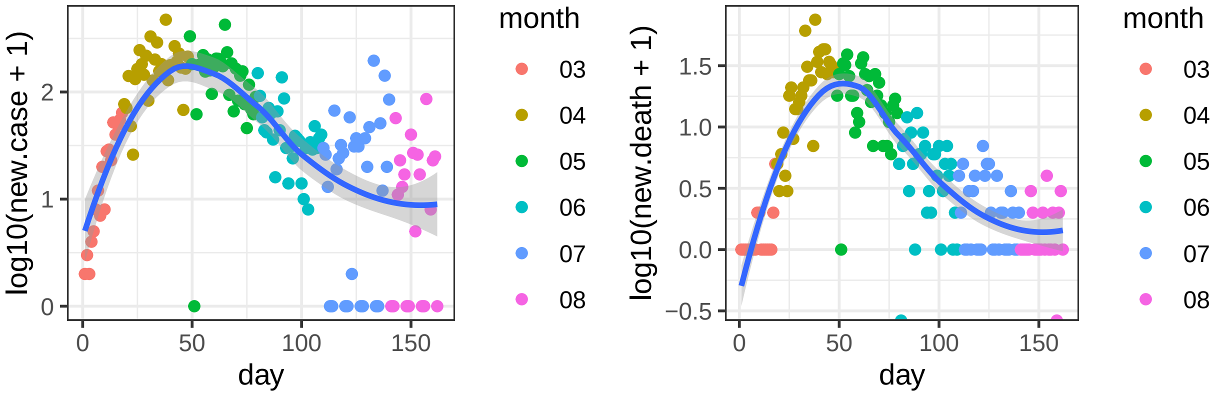
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Westchester_New York



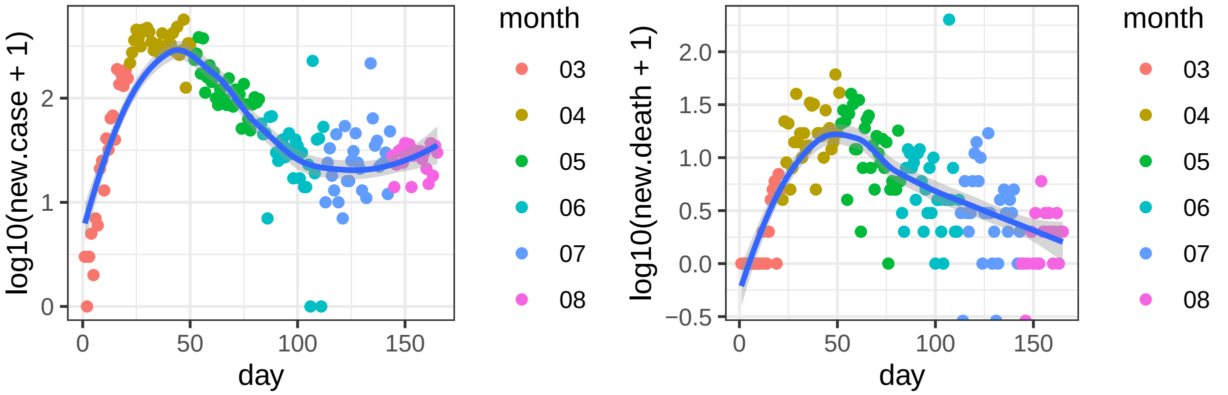
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hartford_Connecticut



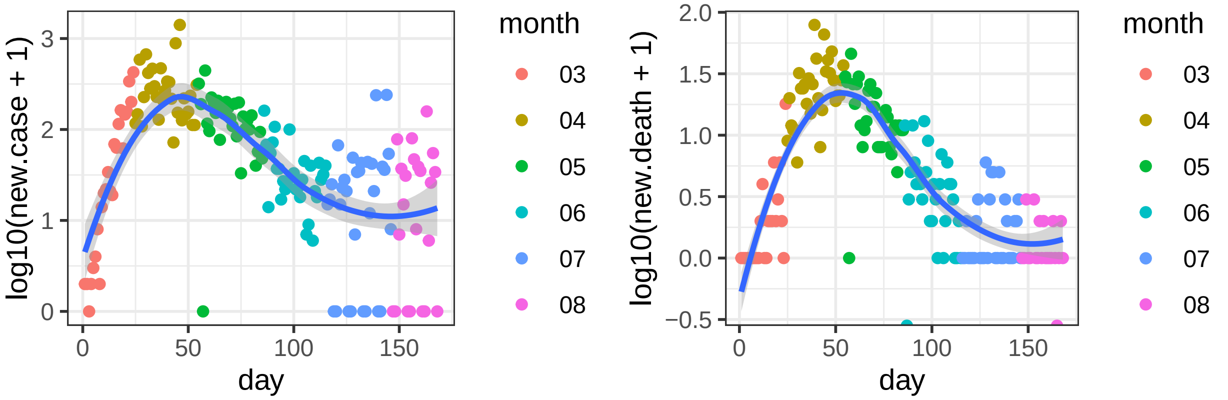
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Middlesex_New Jersey



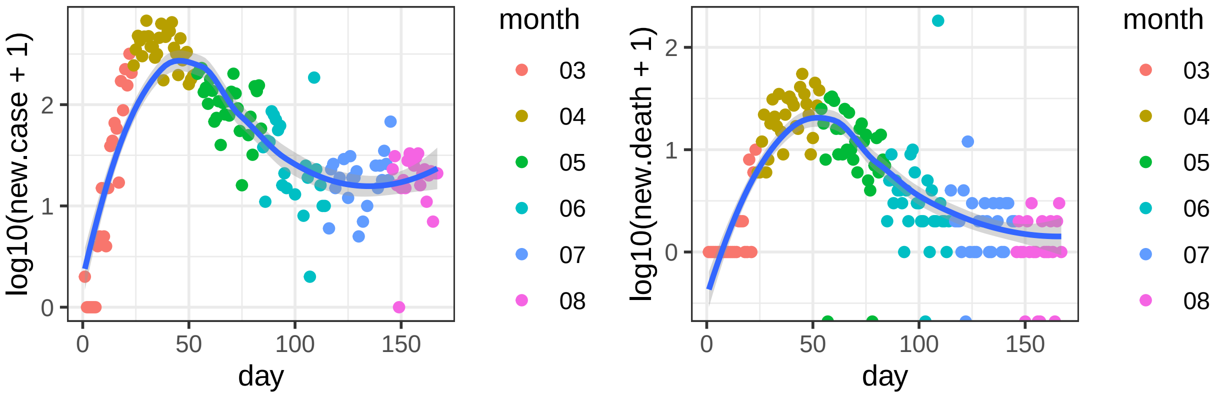
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Fairfield_Connecticut



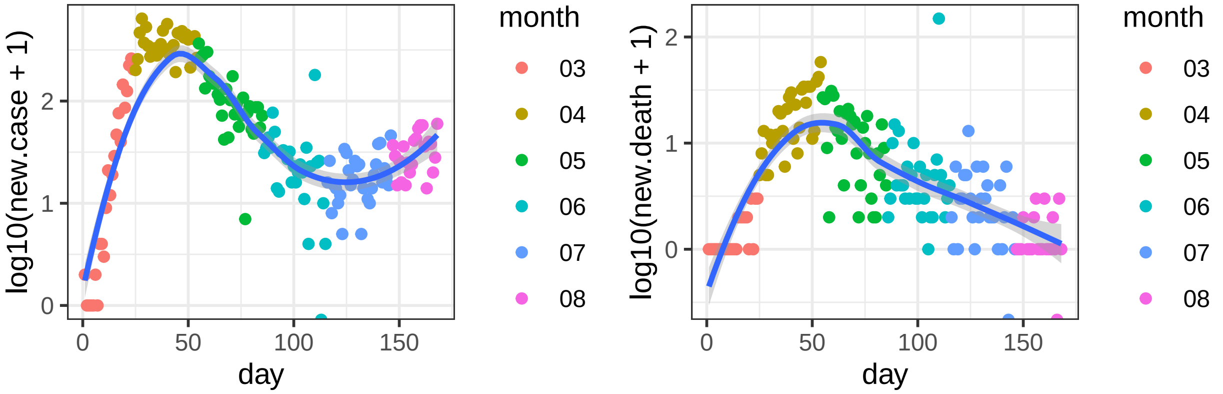
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Union_New Jersey



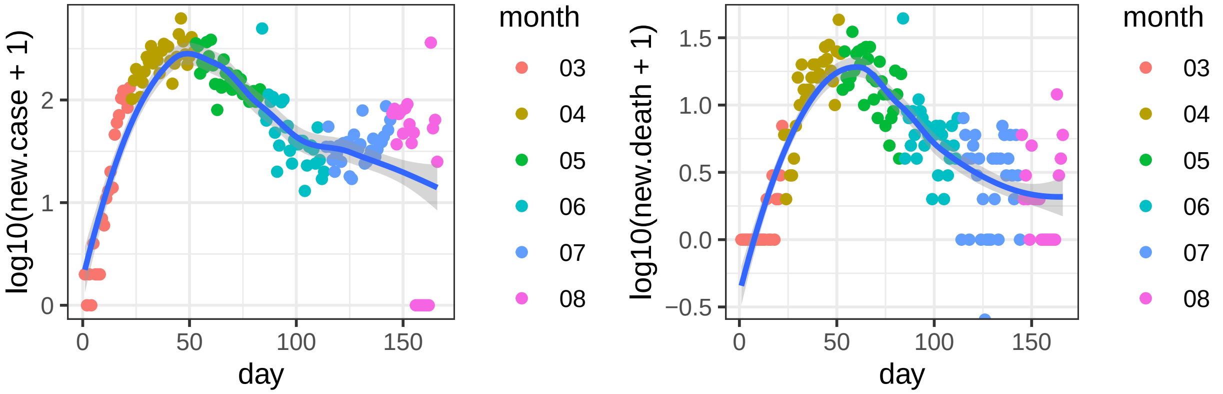
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Passaic_New Jersey



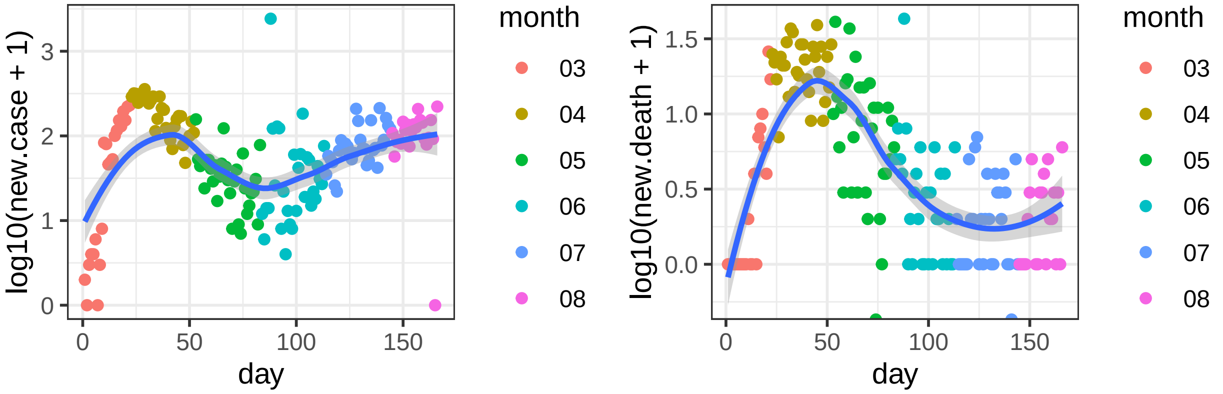
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



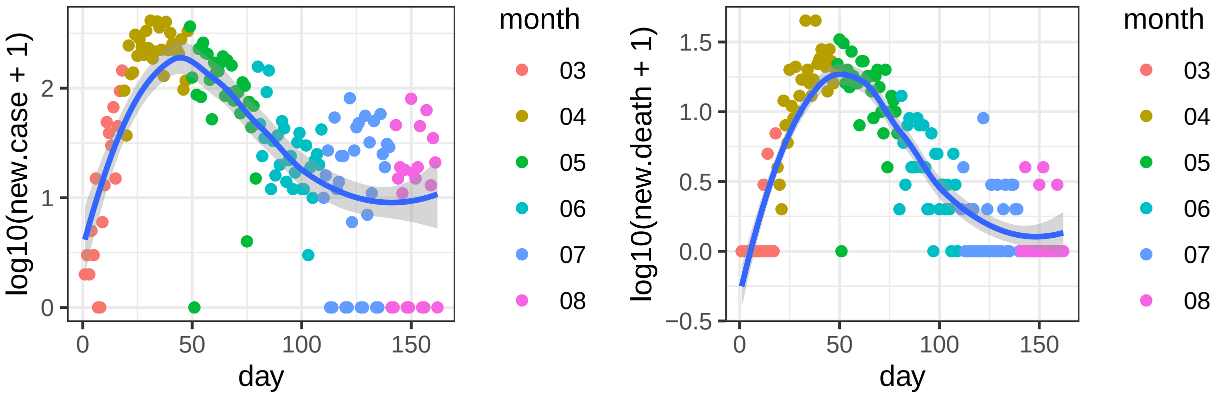
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



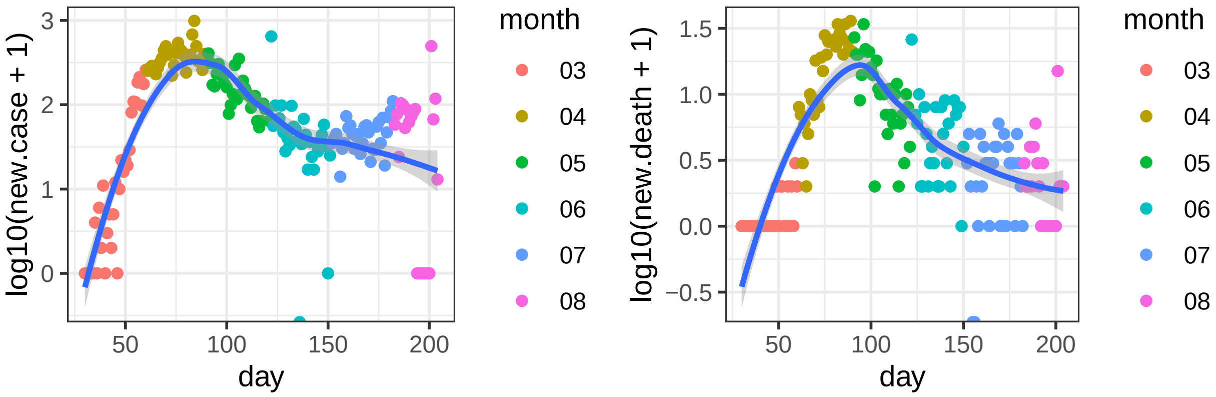
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



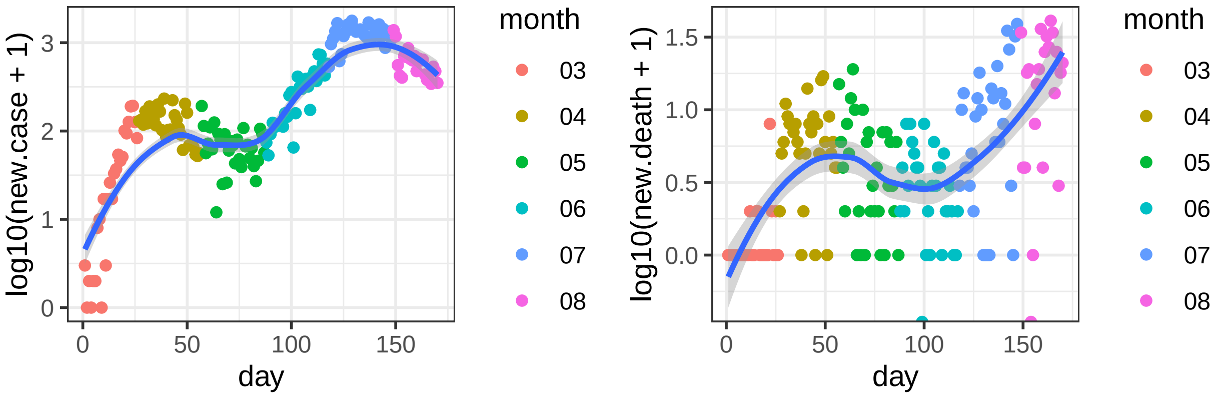
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



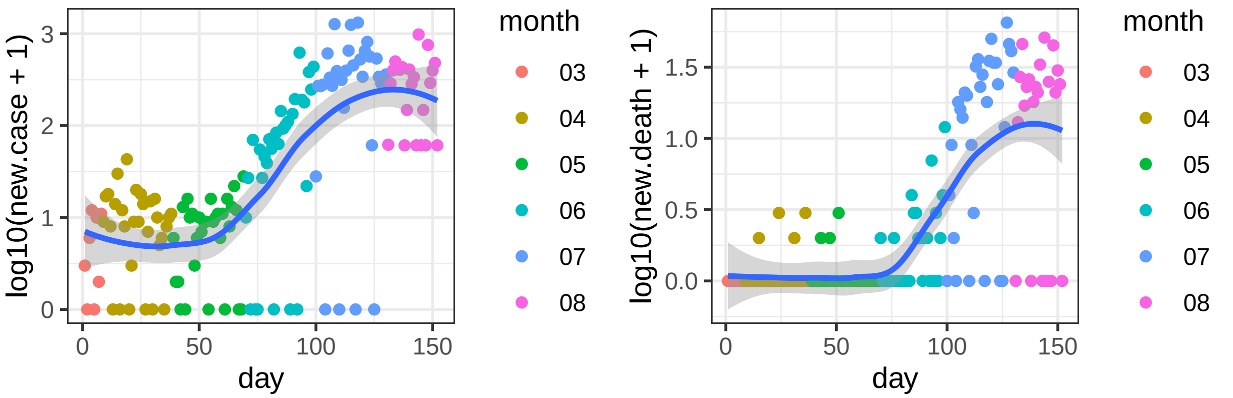
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Broward_Florida



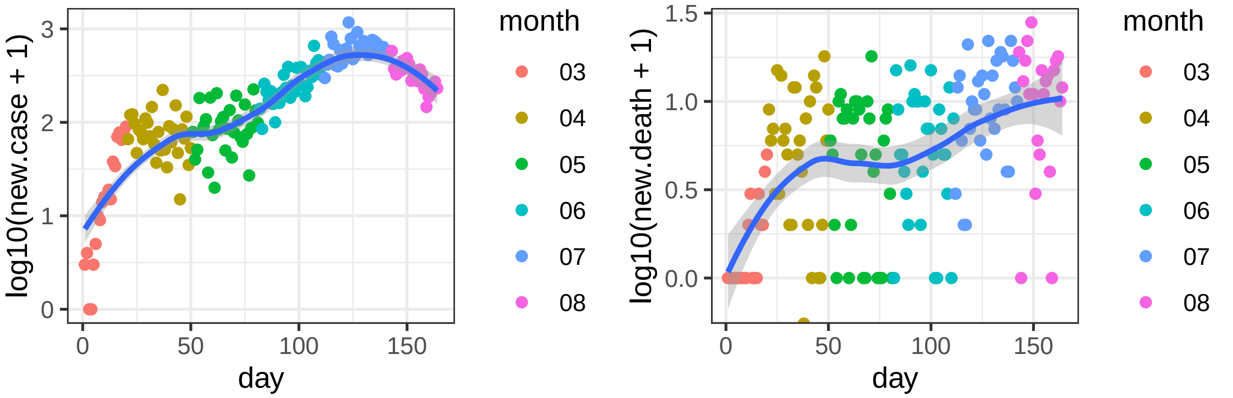
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Hidalgo_Texas



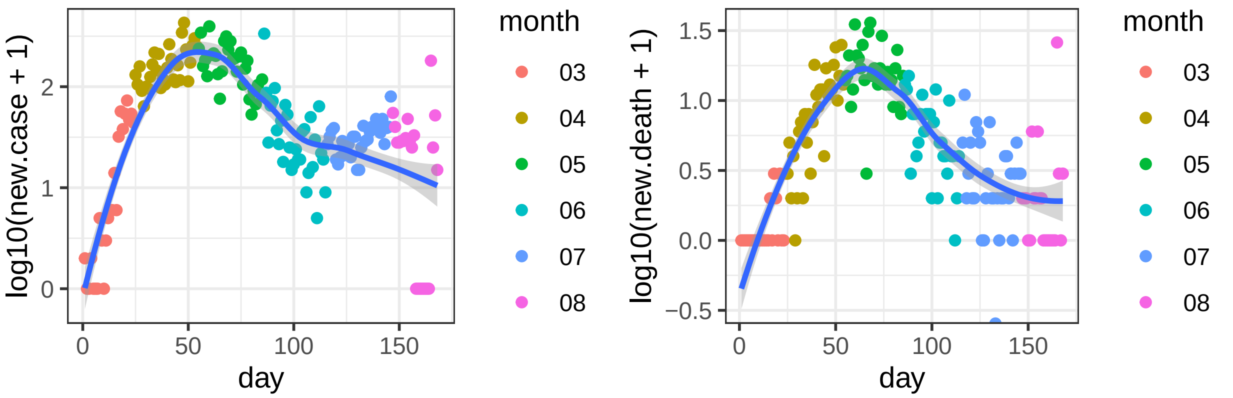
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24

Palm Beach_Florida



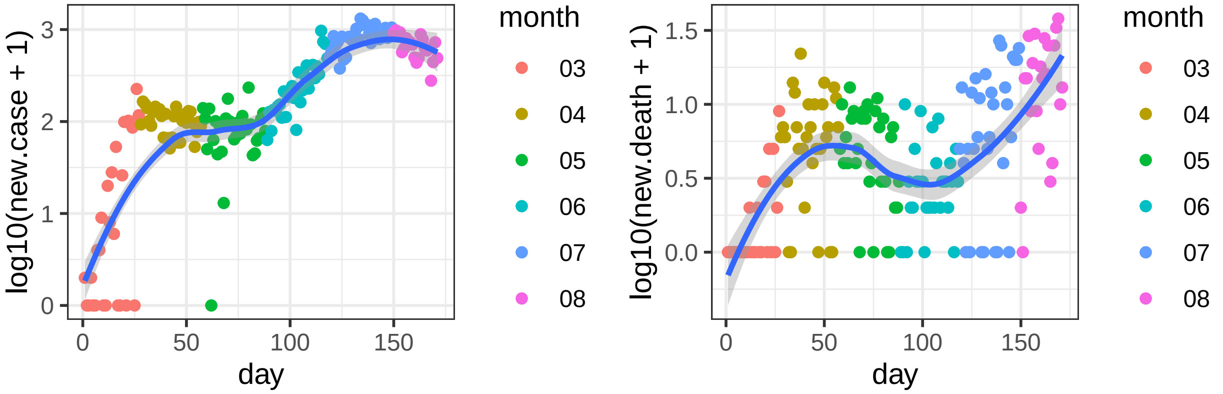
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Worcester_Massachusetts



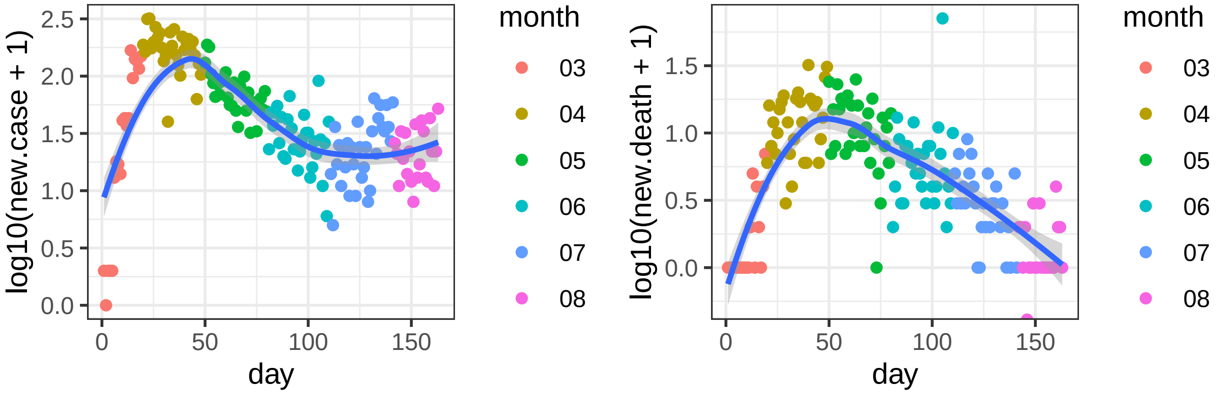
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Clark_Nevada



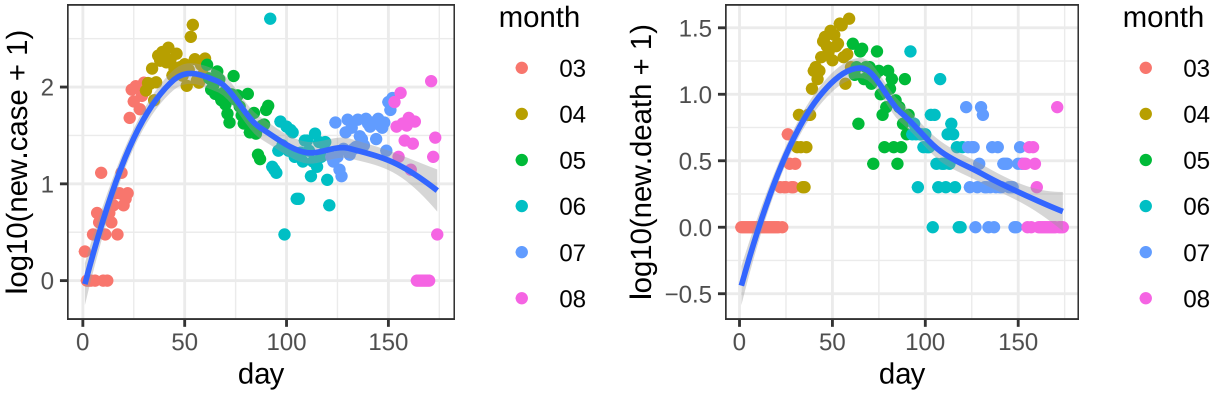
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Ocean_New Jersey



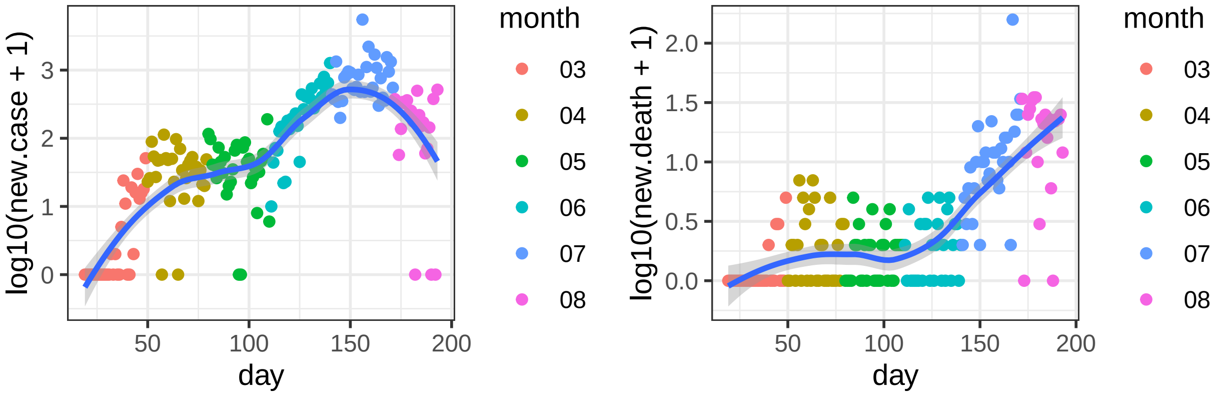
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Norfolk_Massachusetts



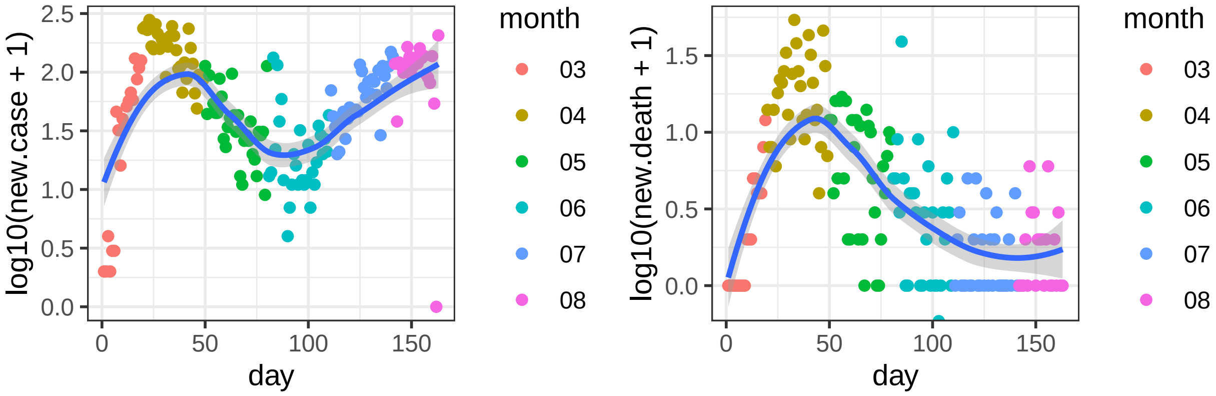
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Bexar_Texas



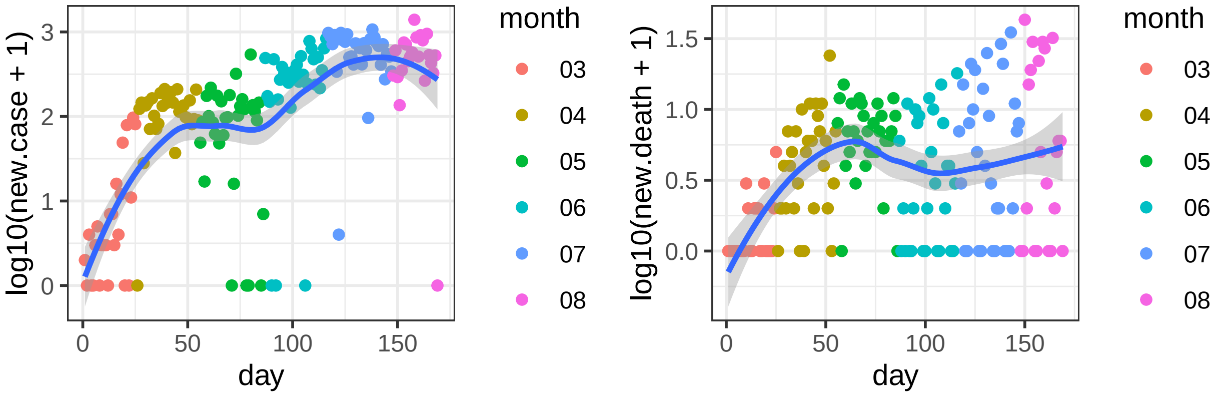
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Macomb_Michigan



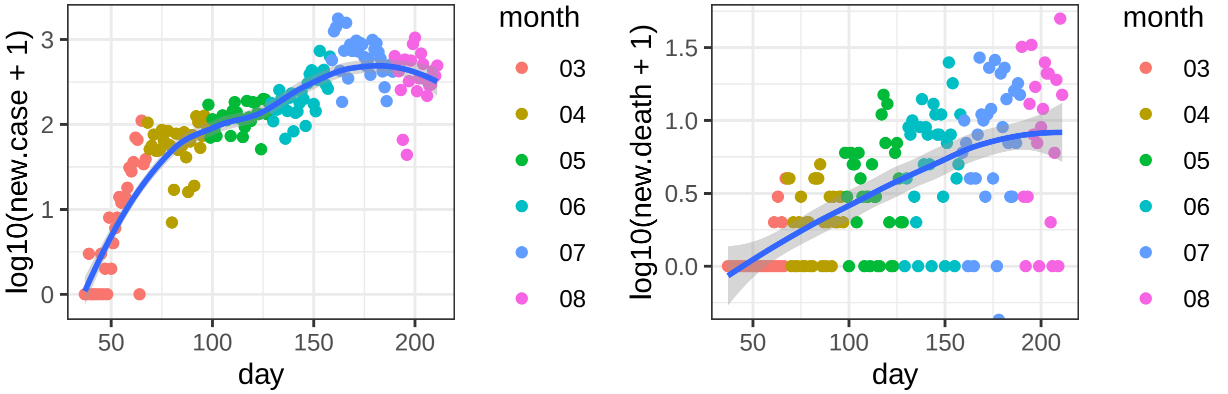
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Riverside_California



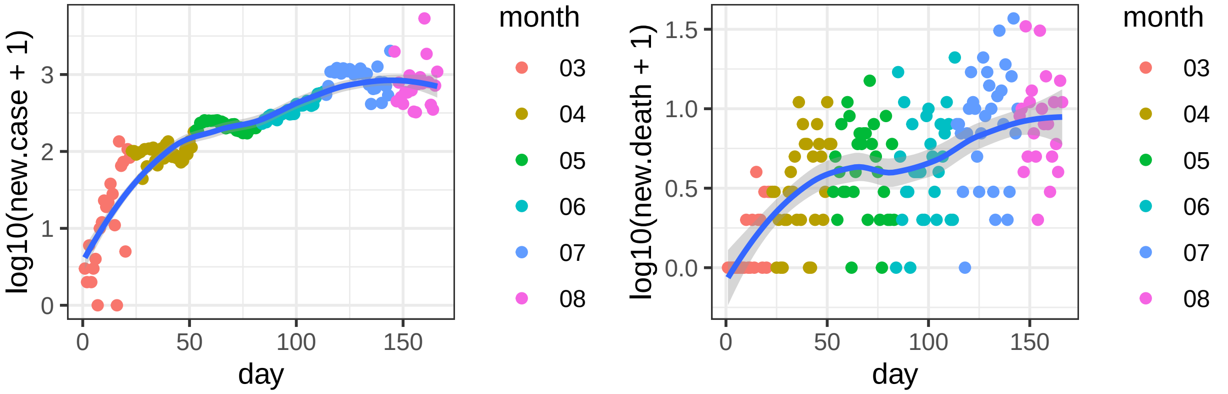
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Orange_California



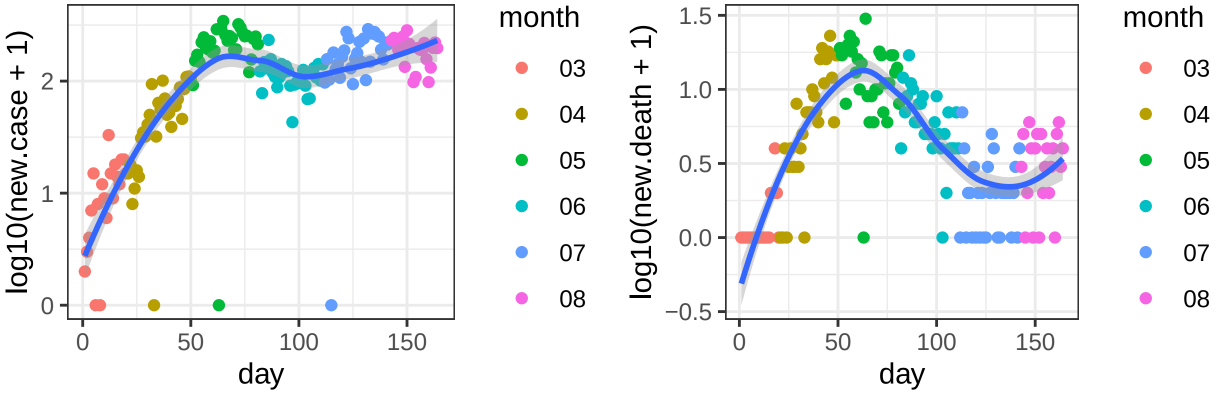
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Dallas_Texas



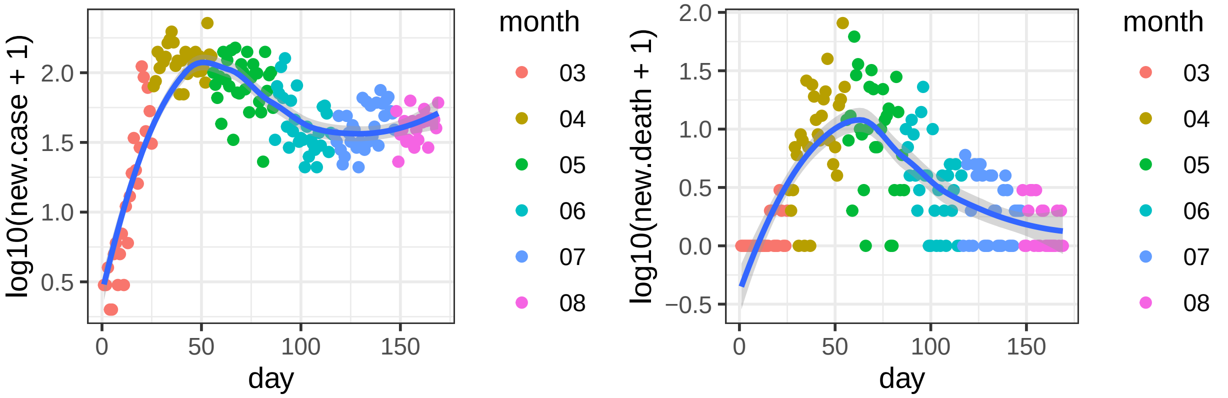
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Hennepin_Minnesota



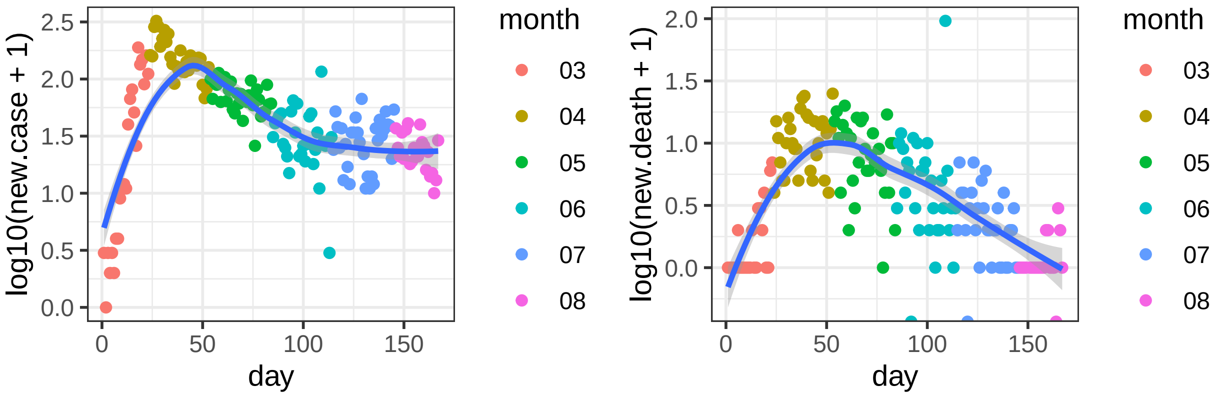
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Montgomery_Pennsylvania



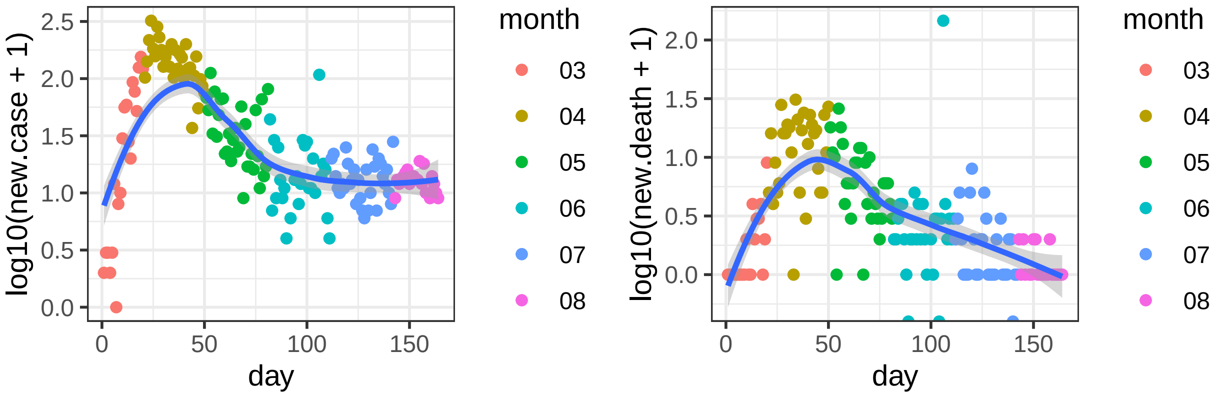
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Monmouth_New Jersey



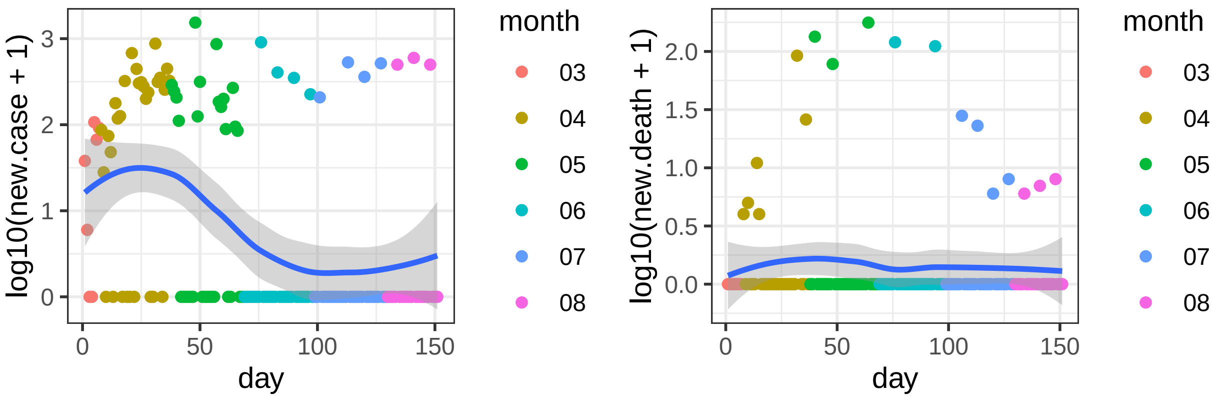
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Morris_New Jersey



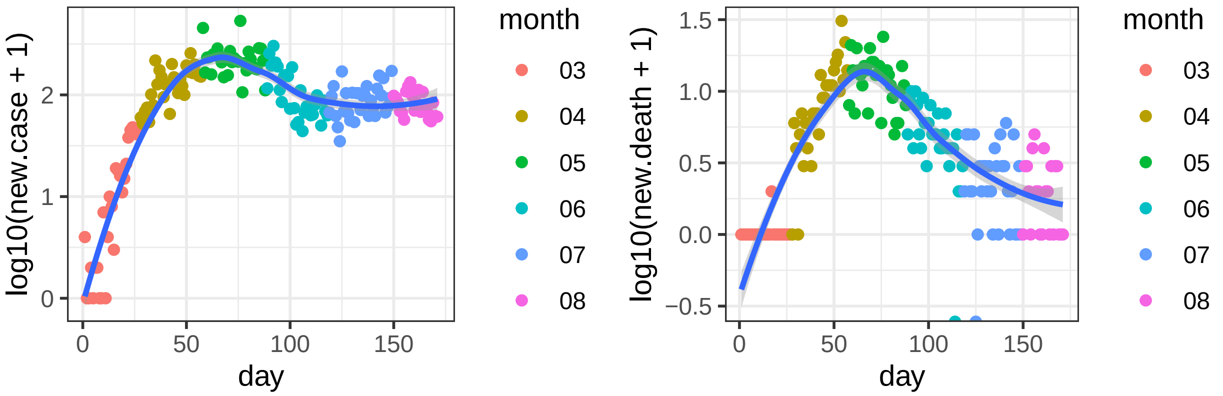
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



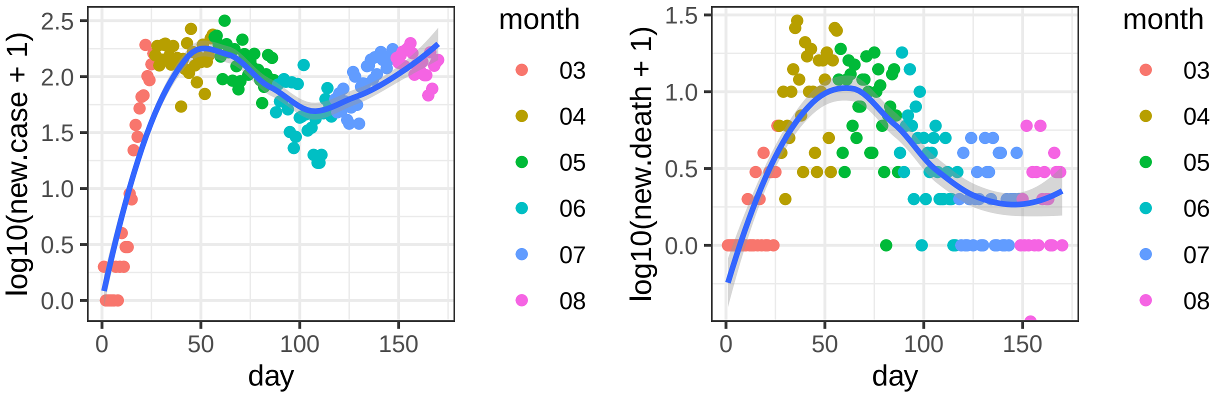
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Maryland



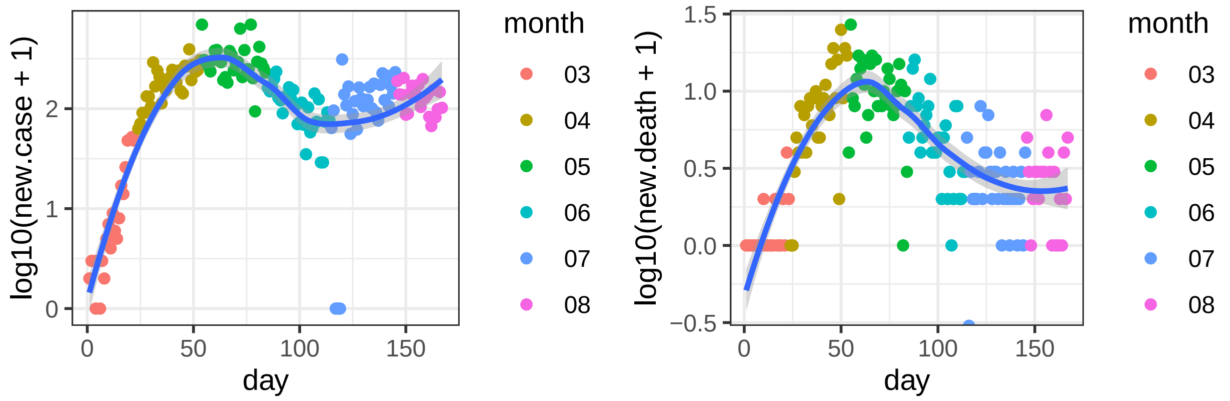
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Marion_Indiana



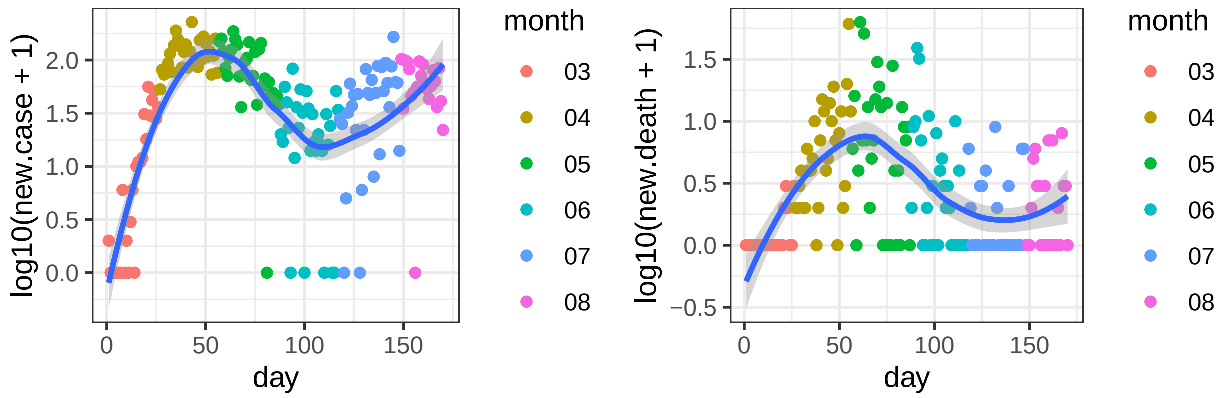
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



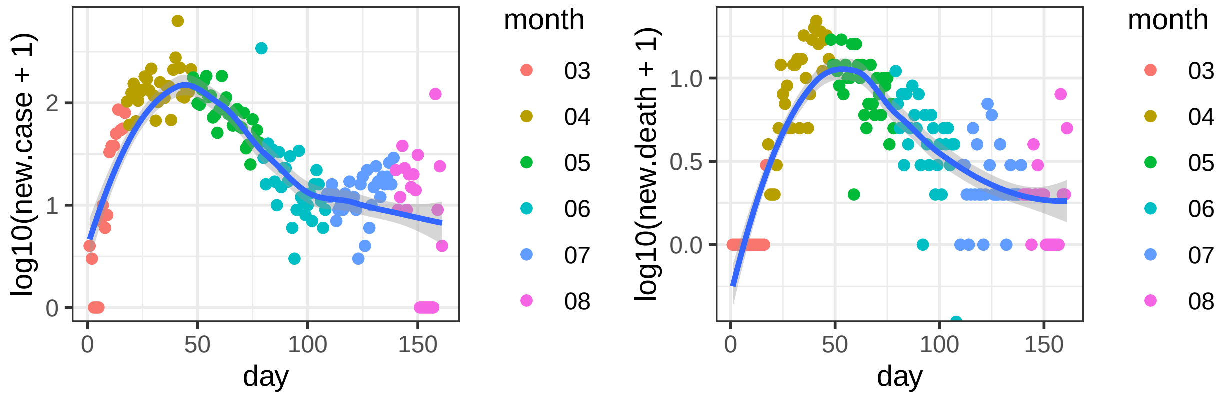
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



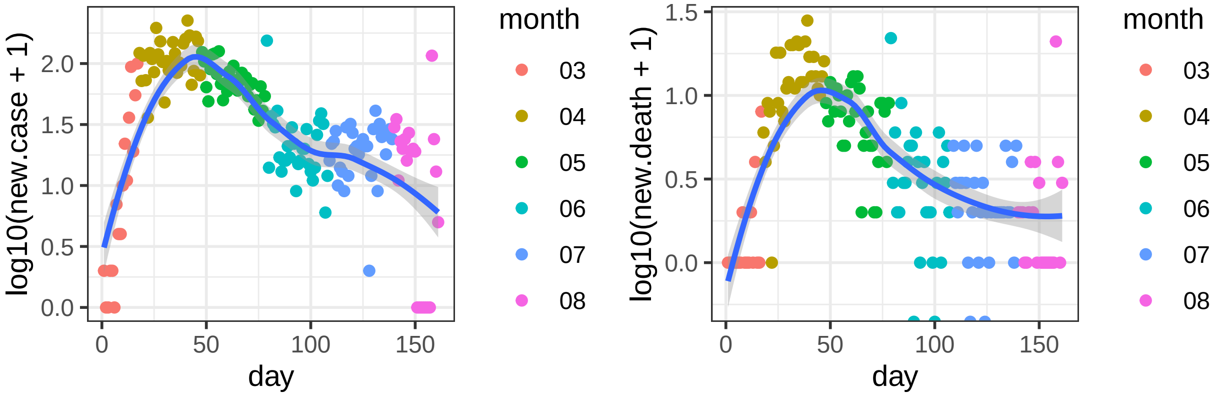
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Plymouth_Massachusetts



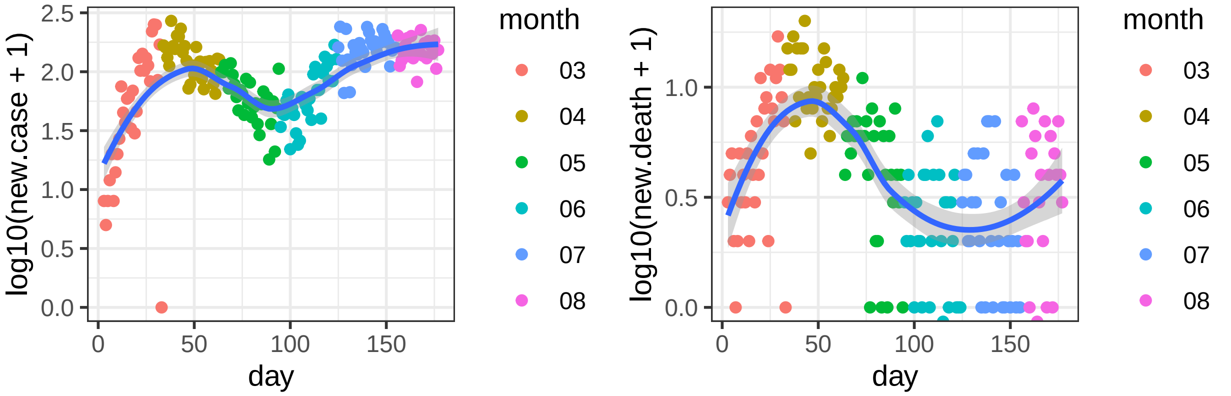
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Hampden_Massachusetts



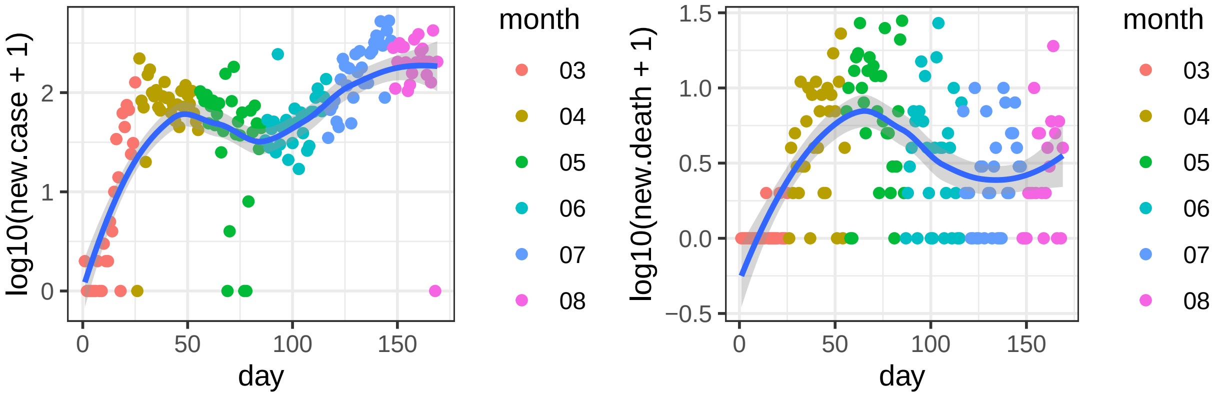
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington

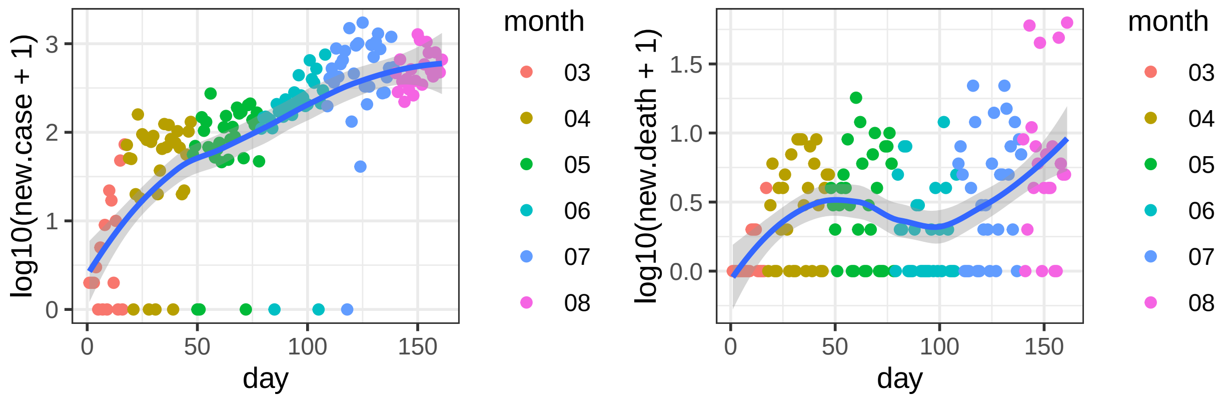


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

St. Louis_Missouri



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



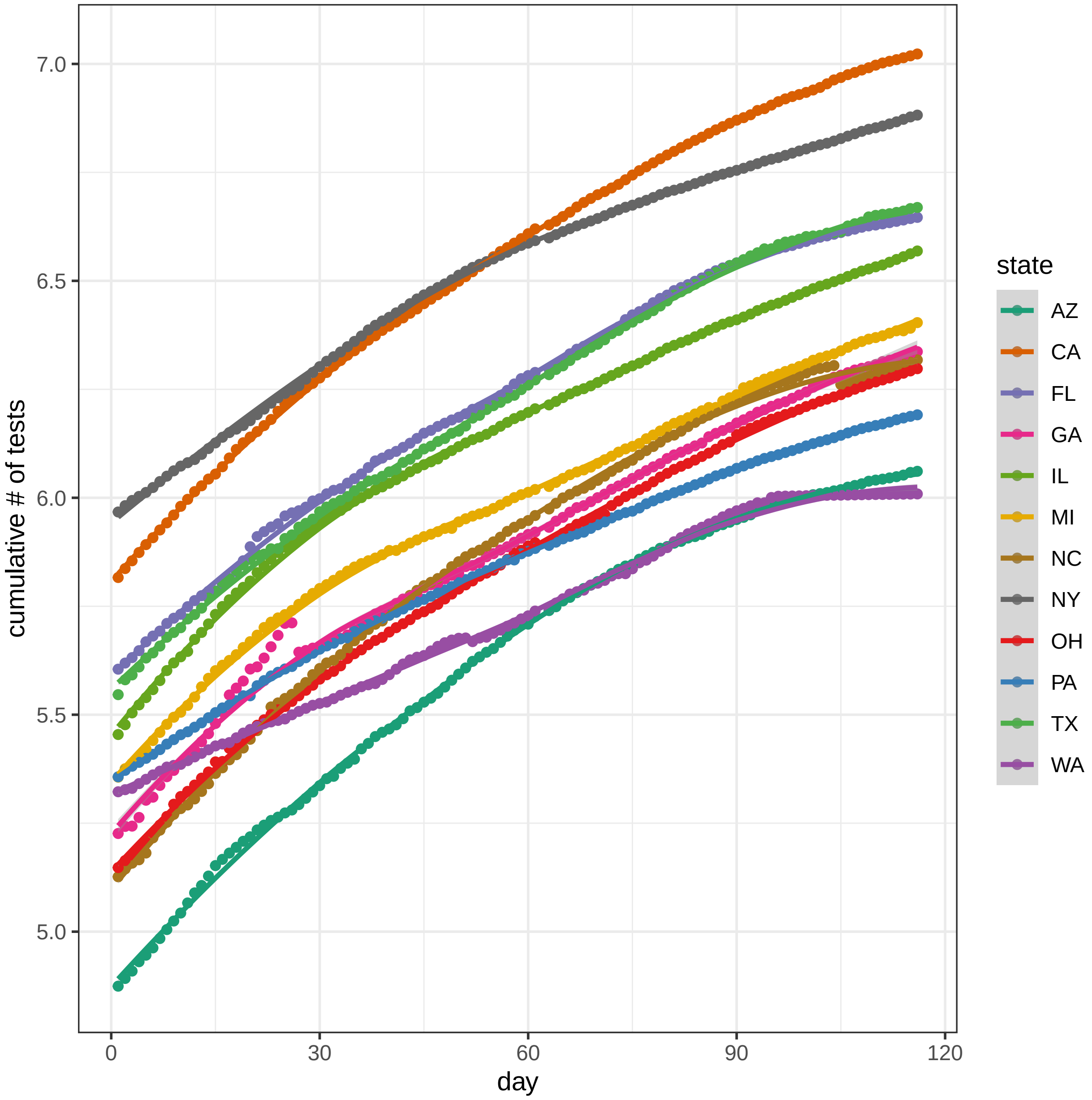
San Bernardino_California

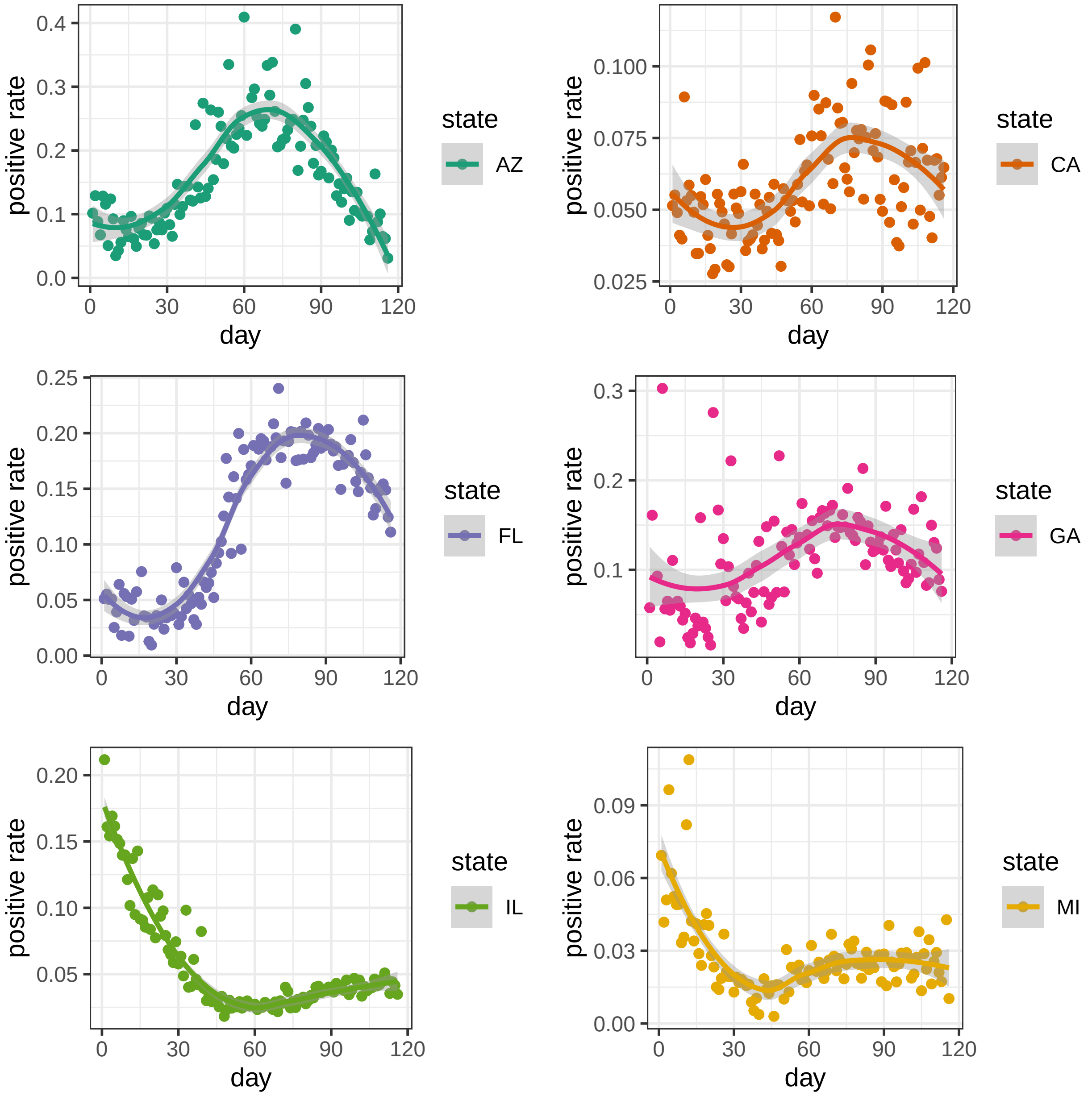
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

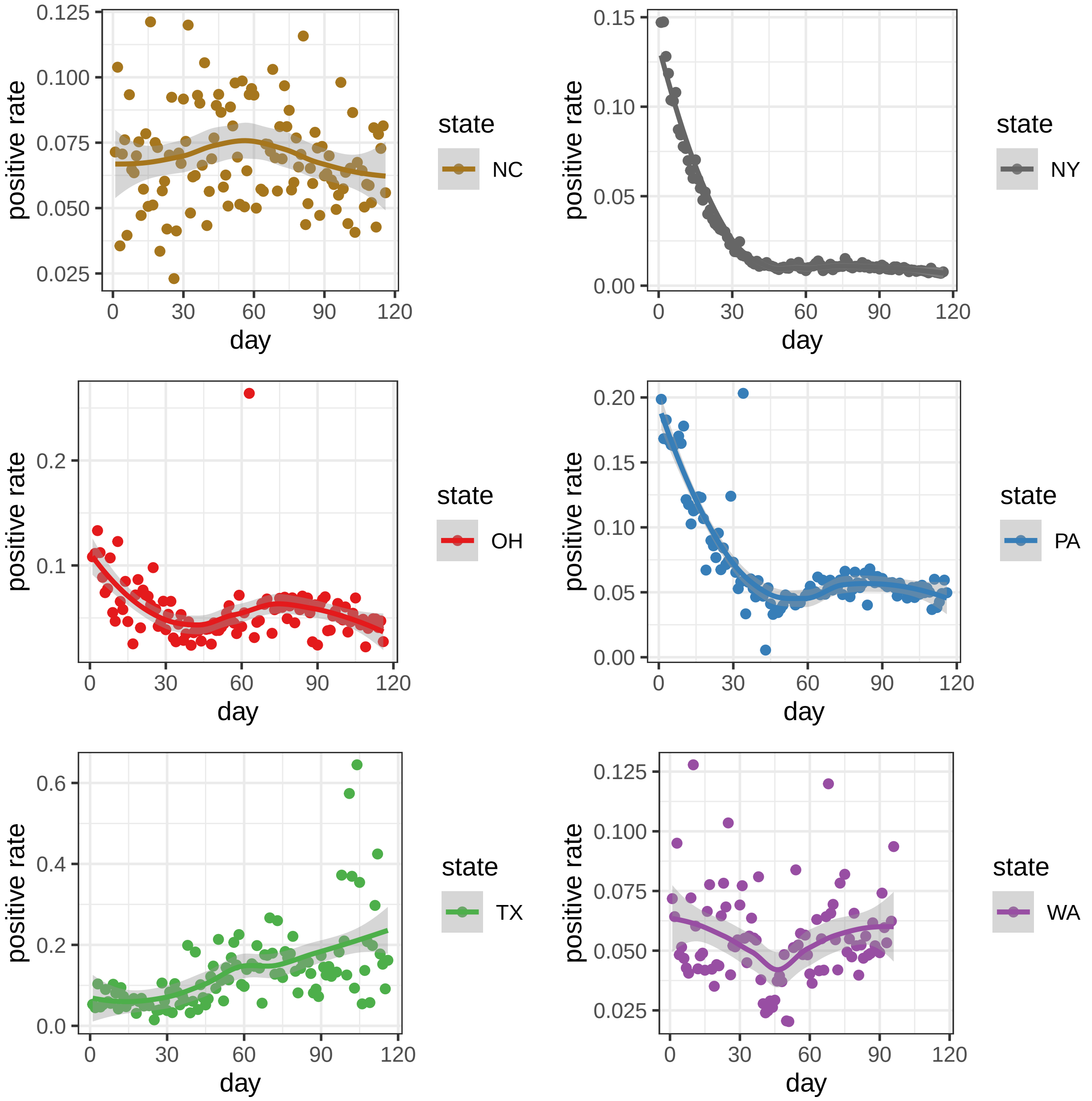
COVID Trackng

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.


7.0
6.5
6.0
5.5
5.0
cumulative # of tests
0
30
60
90
120
day
state
AZ
CA
FL
GA
IL
MI
NC
NY
OH
PA
TX
WA


positive rate
0.0
0.1
0.2
0.3
0.4
0
30
60
90
120
day
state
AZ
positive rate
0.025
0.050
0.075
0.100
0
30
60
90
120
day
state
CA
positive rate
0.00
0.05
0.10
0.15
0.20
0.25
0
30
60
90
120
day
state
FL
positive rate
0.1
0.2
0.3
0
30
60
90
120
day
state
GA
positive rate
0.05
0.10
0.15
0.20
0
30
60
90
120
day
state
IL
positive rate
0.00
0.03
0.06
0.09
0
30
60
90
120
day
state
MI



Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.6
##
## 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] RColorBrewer_1.1-2 httr_1.4.1         ggpubr_0.2.5       magrittr_1.5
## [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    farver_2.0.3     purrr_0.3.3      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
```