Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-08-13

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

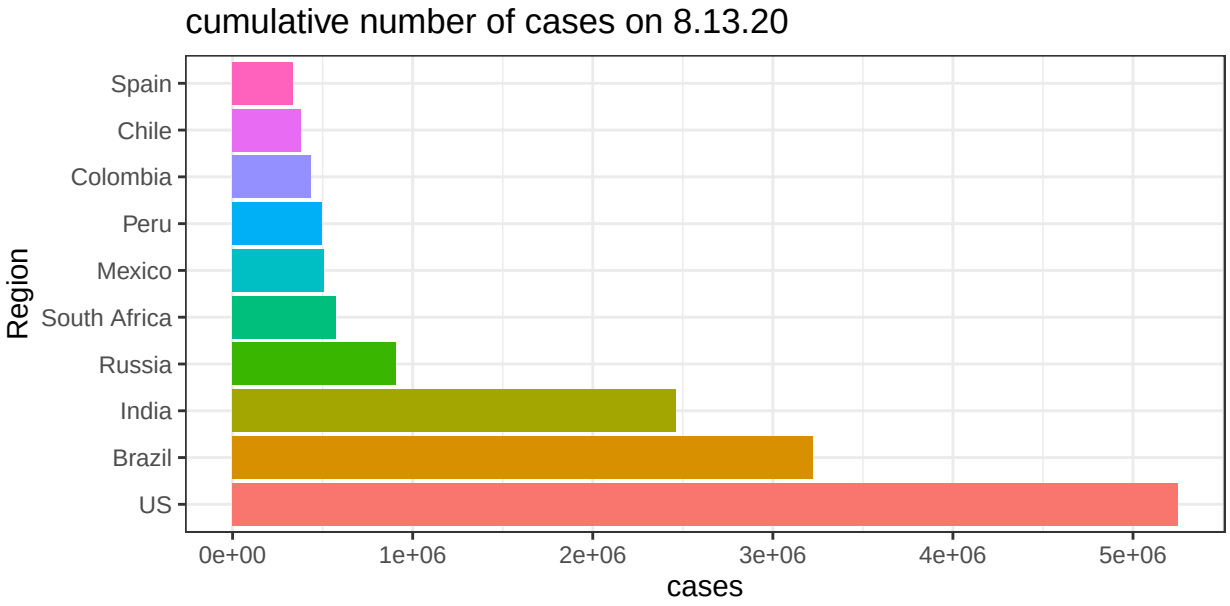
JHU

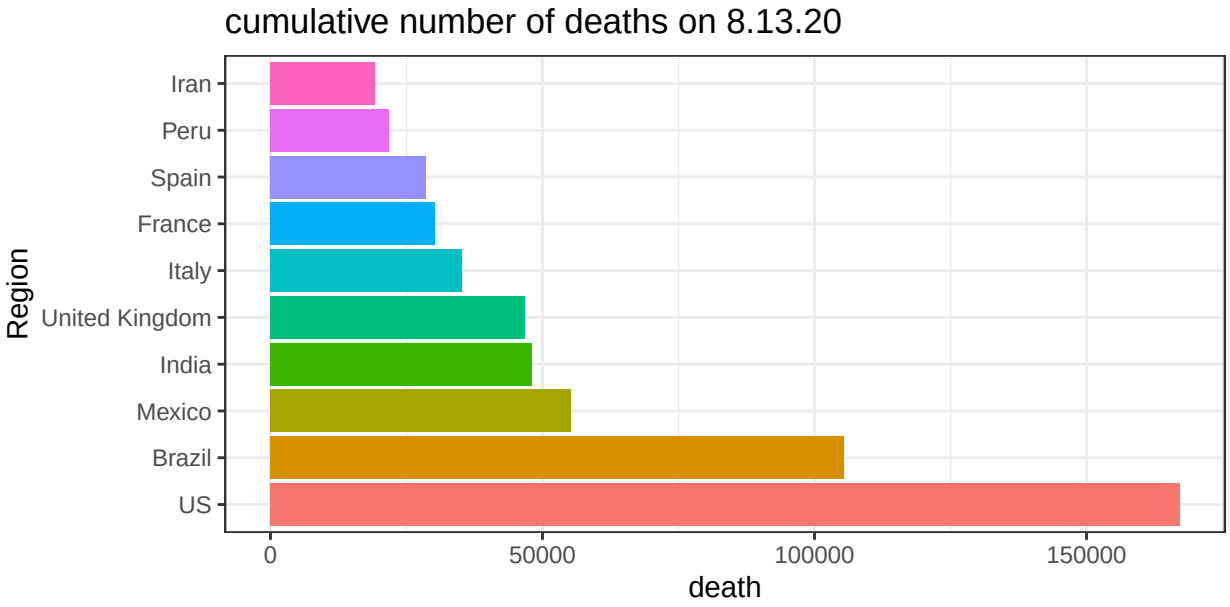
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

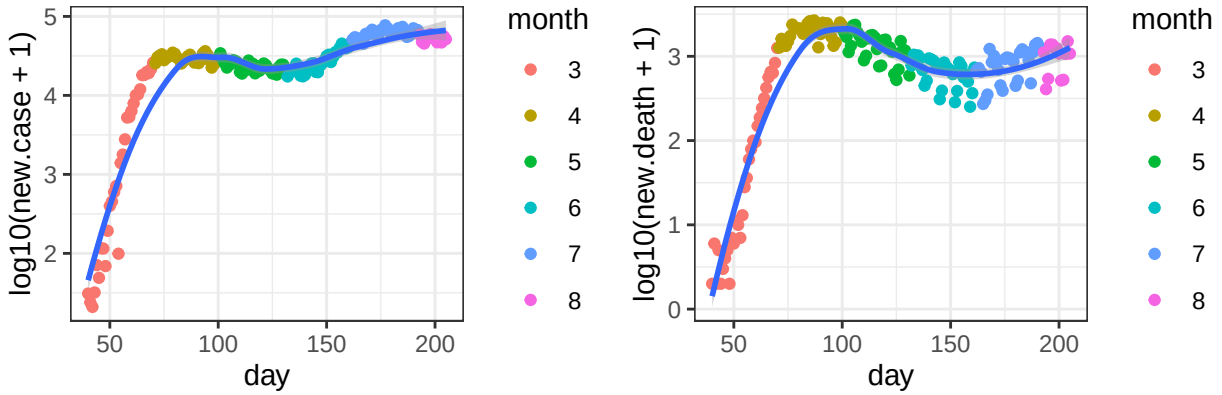
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





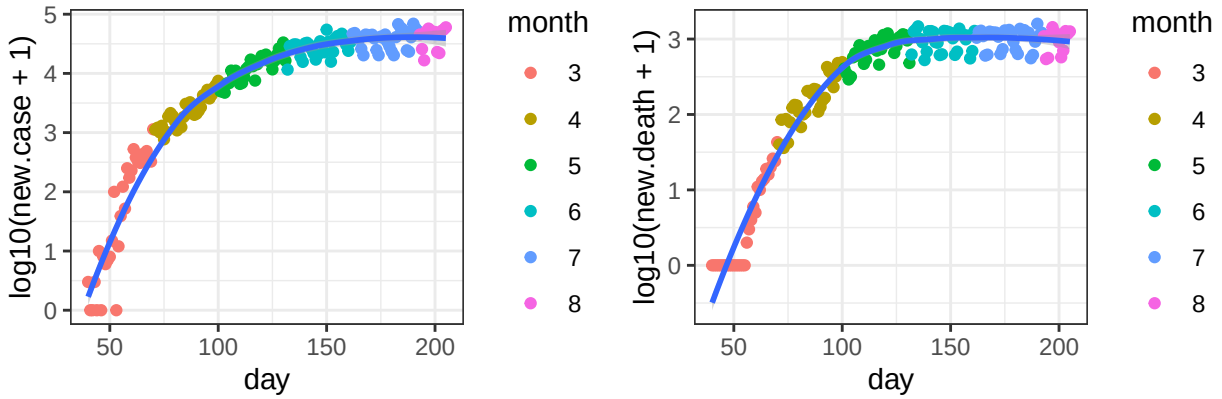
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



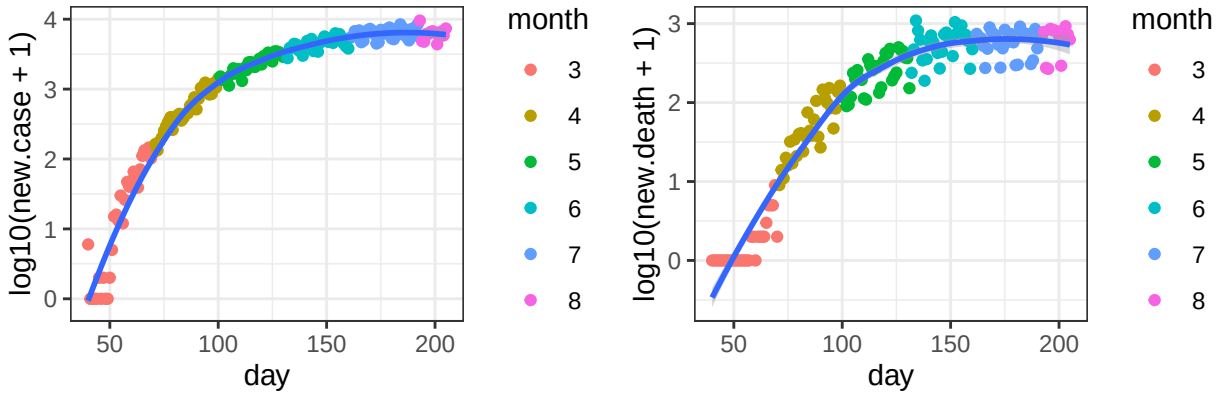
data source: https://github.com/CSSEGISandData/COVID−19

Brazil



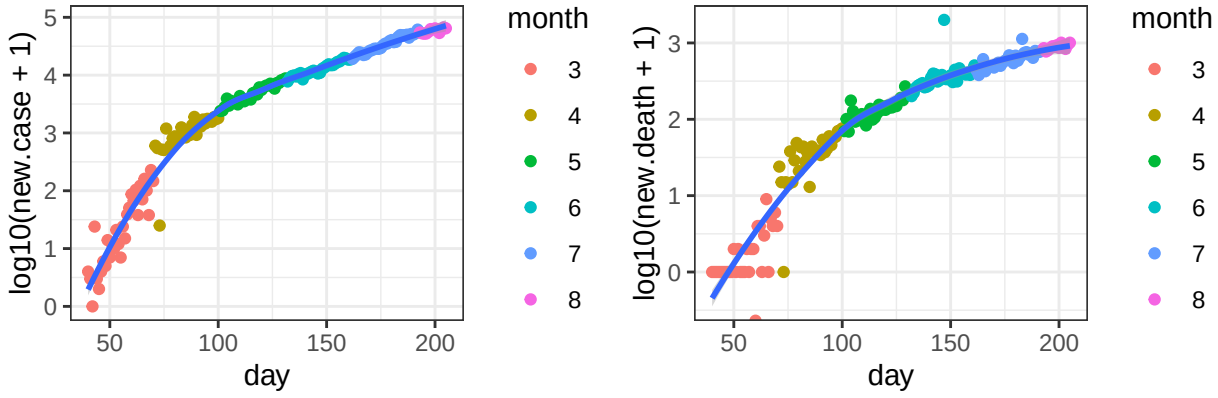
data source: https://github.com/CSSEGISandData/COVID−19

Mexico



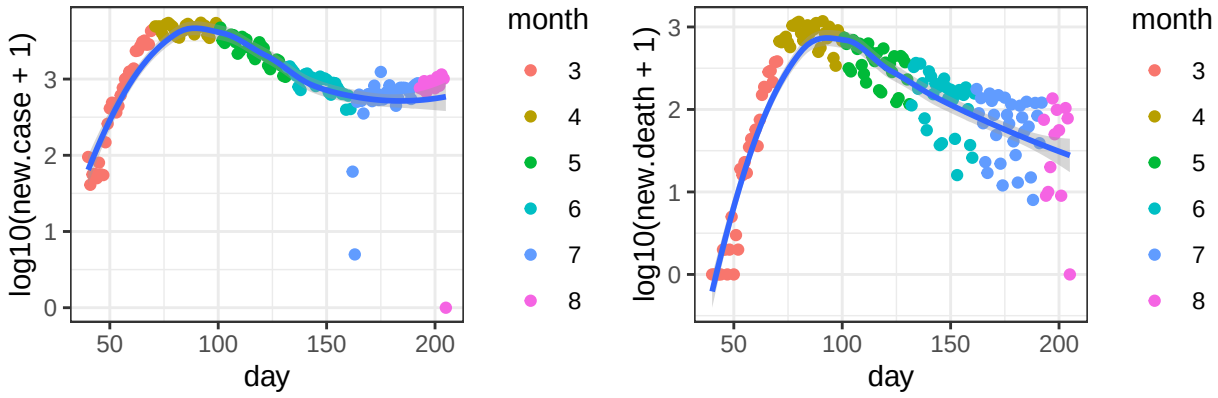
data source: https://github.com/CSSEGISandData/COVID−19

India



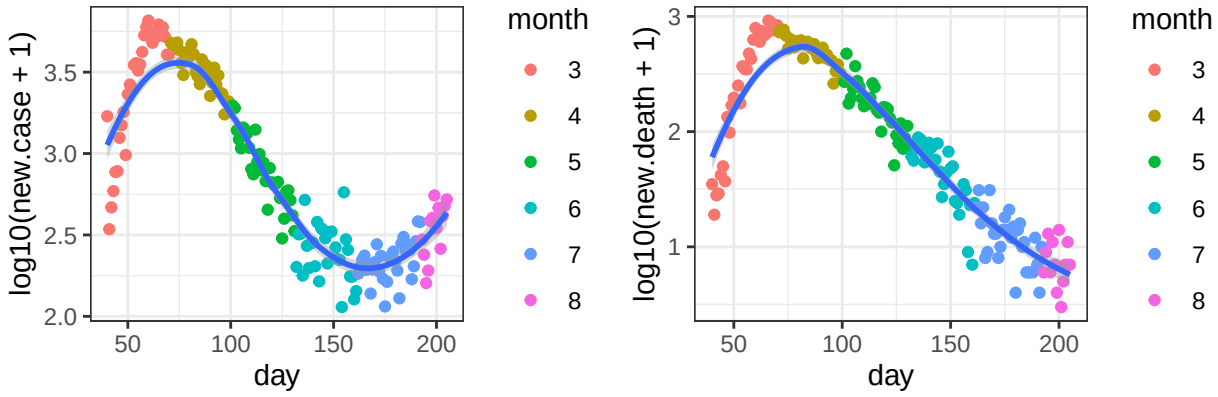
data source: https://github.com/CSSEGISandData/COVID−19

United Kingdom



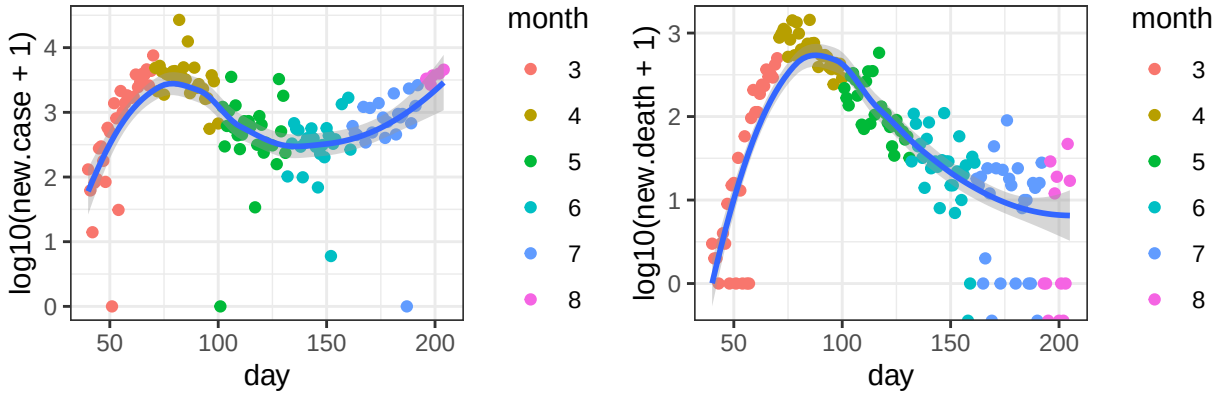
data source: https://github.com/CSSEGISandData/COVID−19

Italy



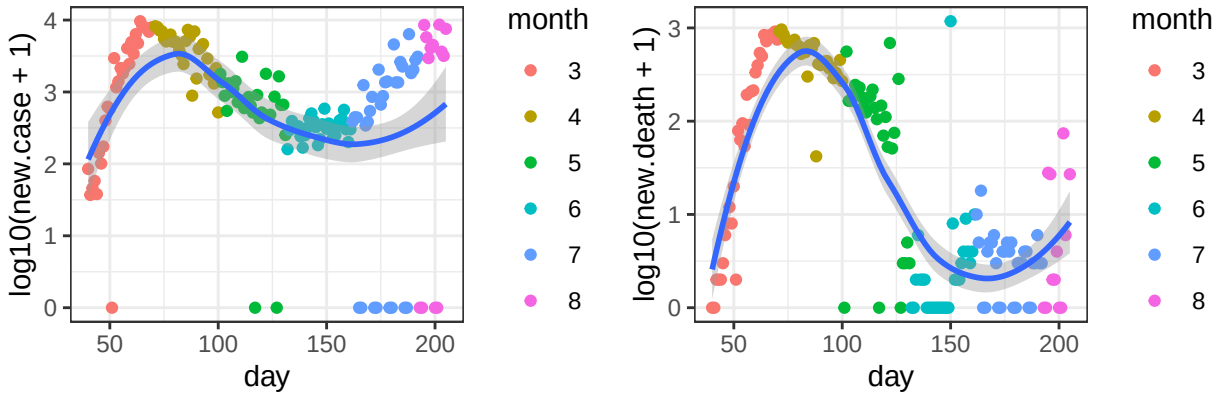
data source: https://github.com/CSSEGISandData/COVID−19

France



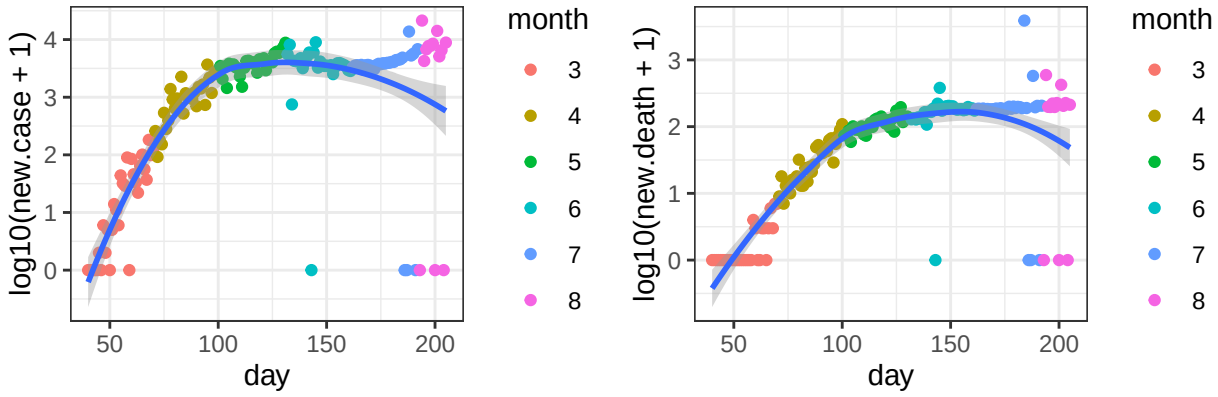
data source: https://github.com/CSSEGISandData/COVID−19

Spain



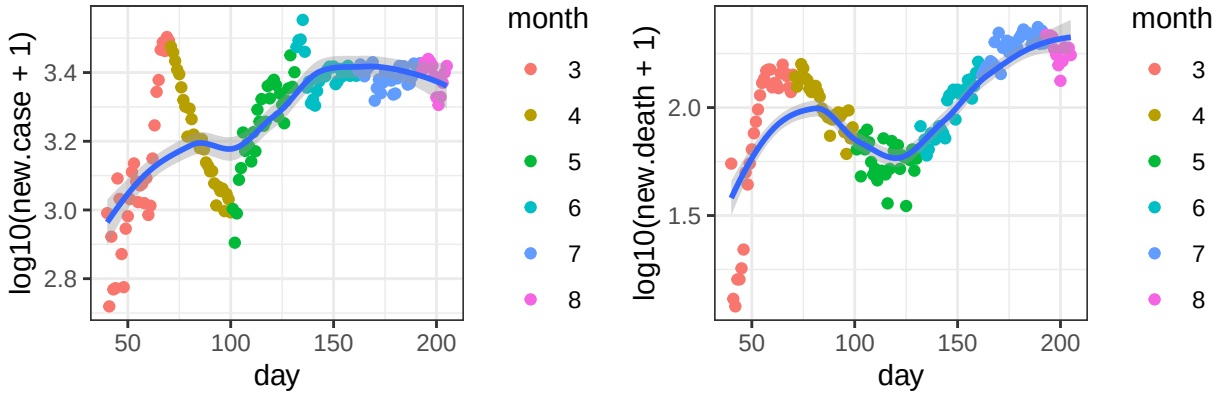
data source: https://github.com/CSSEGISandData/COVID−19

Peru



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

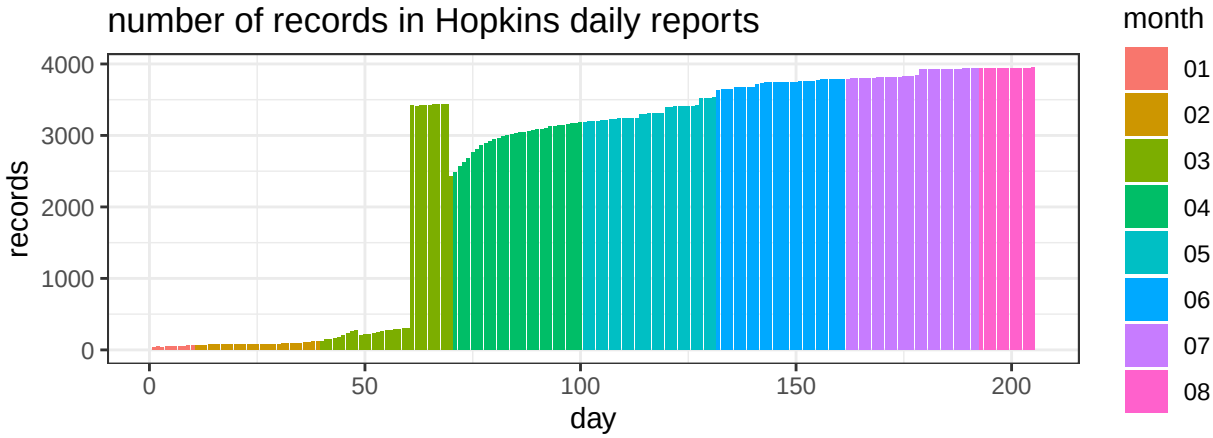
Iran



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] "2020-08-13"
```

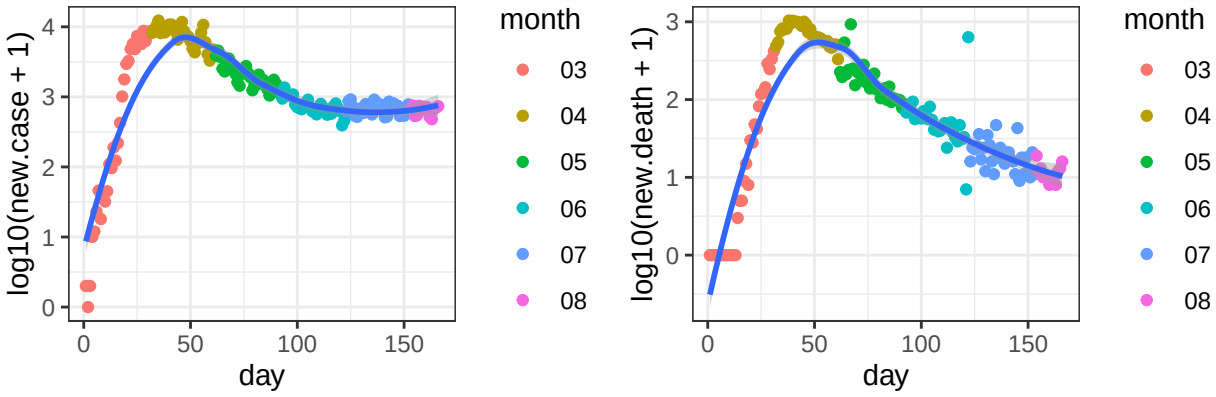
state level data

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

```
##            date          state fips  cases deaths
## 9013 2020-08-13       New York   36 428155  32399
## 9011 2020-08-13     New Jersey   34 188433  15893
## 8984 2020-08-13     California    6 603008  10995
## 9026 2020-08-13          Texas   48 536336   9878
## 8989 2020-08-13        Florida   12 557129   8912
## 9002 2020-08-13  Massachusetts   25 122423   8790
## 8994 2020-08-13       Illinois   17 202577   7934
## 9020 2020-08-13   Pennsylvania   42 126950   7474
## 9003 2020-08-13       Michigan   26  99963   6556
## 8986 2020-08-13    Connecticut    9  50782   4450
## 8990 2020-08-13        Georgia   13 212009   4440
## 8999 2020-08-13      Louisiana   22 135562   4402
## 8982 2020-08-13        Arizona    4 190850   4385
## 9017 2020-08-13           Ohio   39 105426   3755
## 9001 2020-08-13       Maryland   24  98728   3620
## 8995 2020-08-13        Indiana   18  79404   3105
## 9030 2020-08-13       Virginia   51 103622   2363
## 9014 2020-08-13 North Carolina   37 141006   2313
## 9023 2020-08-13 South Carolina   45 103909   2186
## 9005 2020-08-13    Mississippi   28  69986   2011
## 8980 2020-08-13        Alabama    1 105557   1890
## 8985 2020-08-13       Colorado    8  52242   1886
## 9031 2020-08-13     Washington   53  67862   1795
## 9004 2020-08-13      Minnesota   27  63039   1731
## 9006 2020-08-13       Missouri   29  64885   1417
## 9025 2020-08-13      Tennessee   47 125487   1300
## 9009 2020-08-13         Nevada   32  58812   1030
## 9033 2020-08-13      Wisconsin   55  67781   1029
## 9022 2020-08-13   Rhode Island   44  20240   1019
## 8996 2020-08-13           Iowa   19  50373    960
```

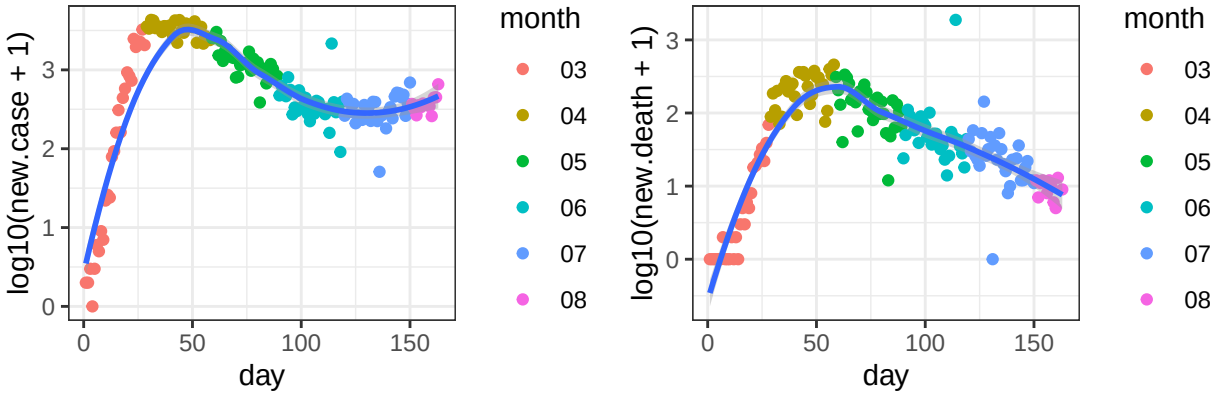
For these 30 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



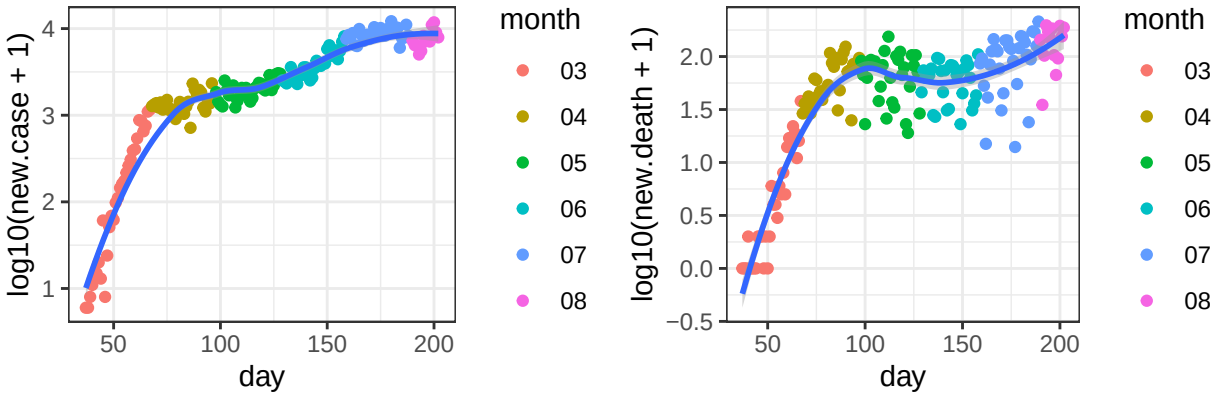
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



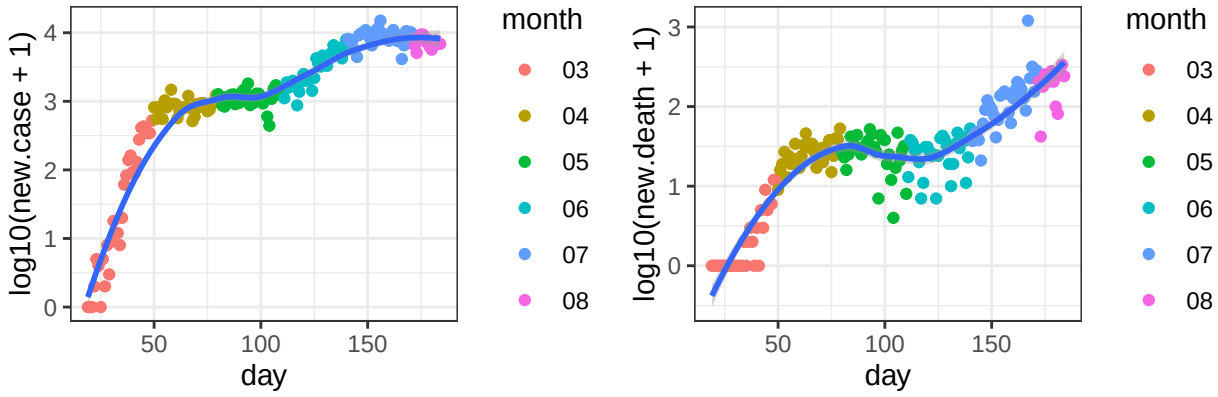
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

California



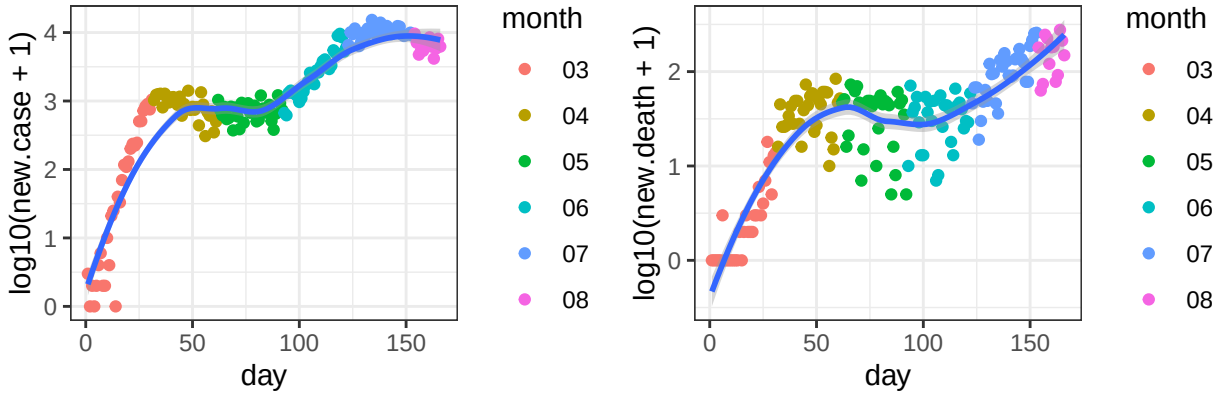
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Texas



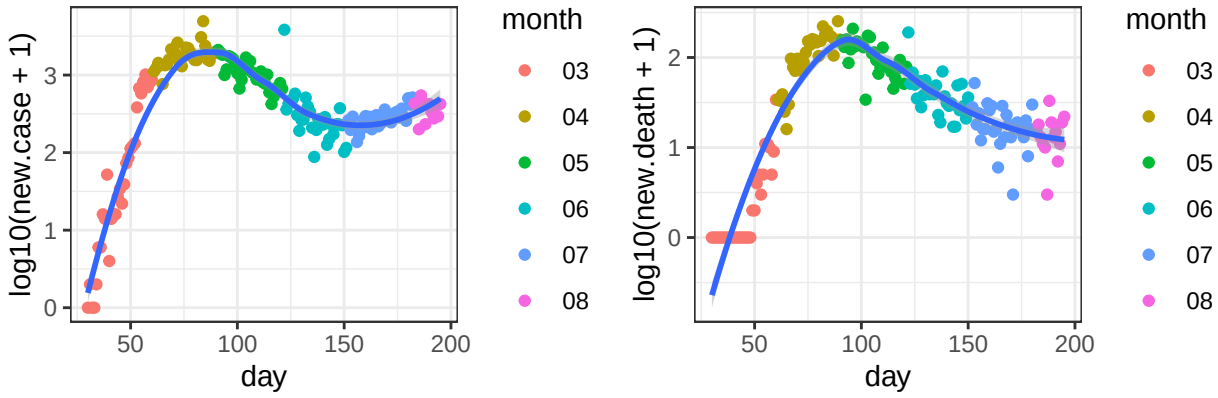
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Florida



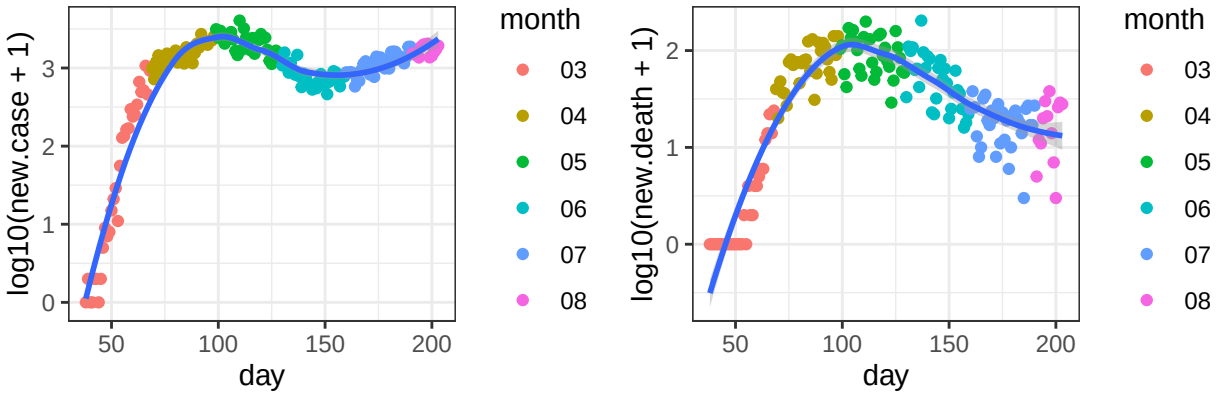
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Massachusetts



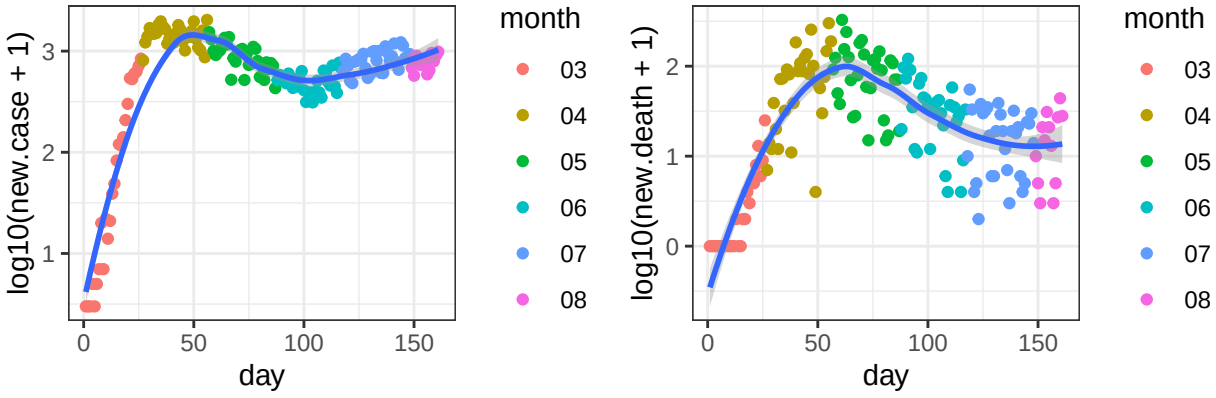
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Illinois



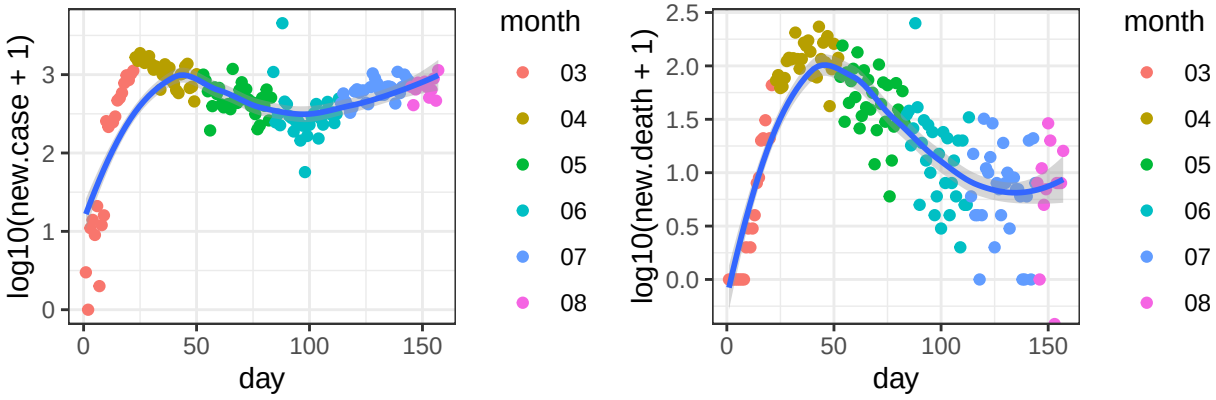
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Pennsylvania



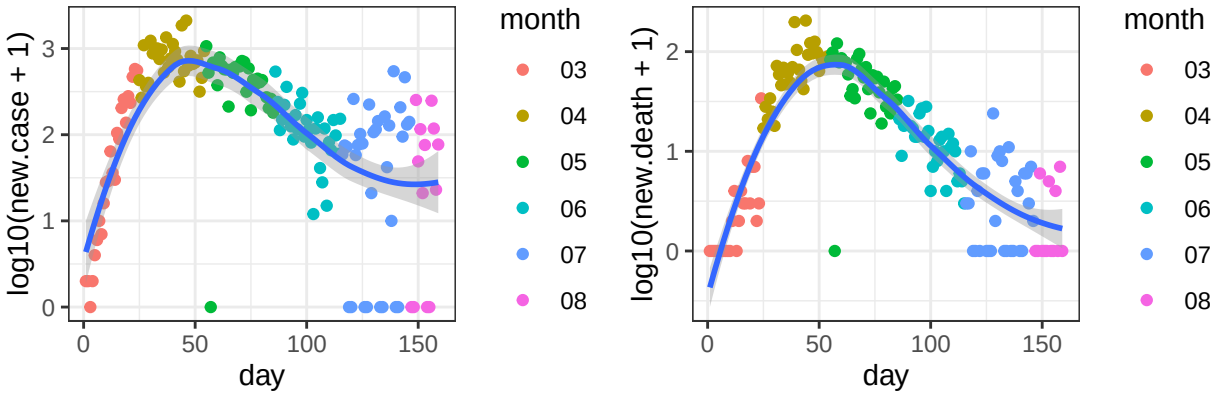
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



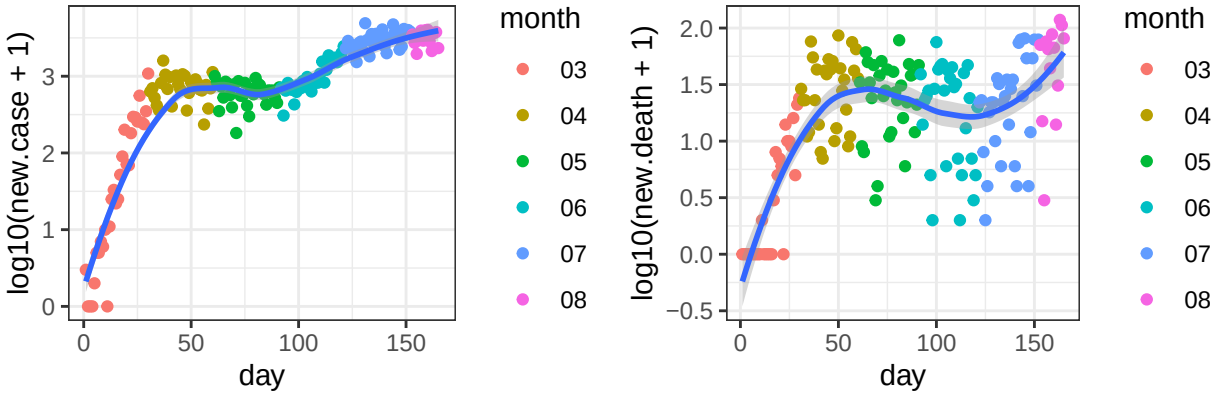
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Connecticut



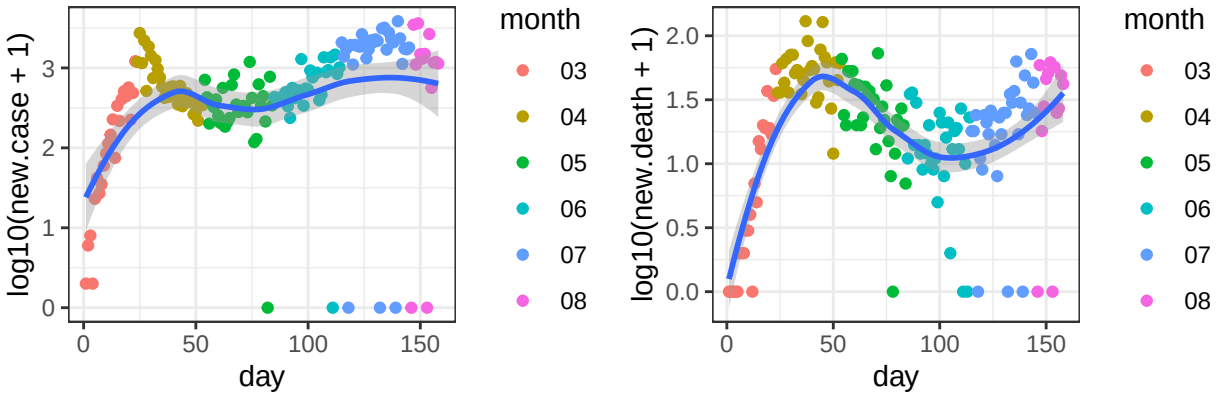
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Georgia



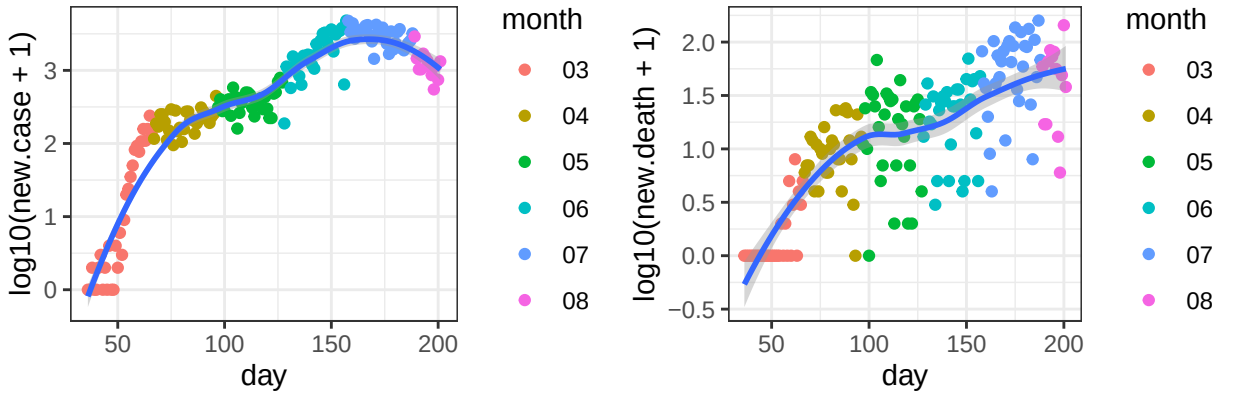
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Louisiana



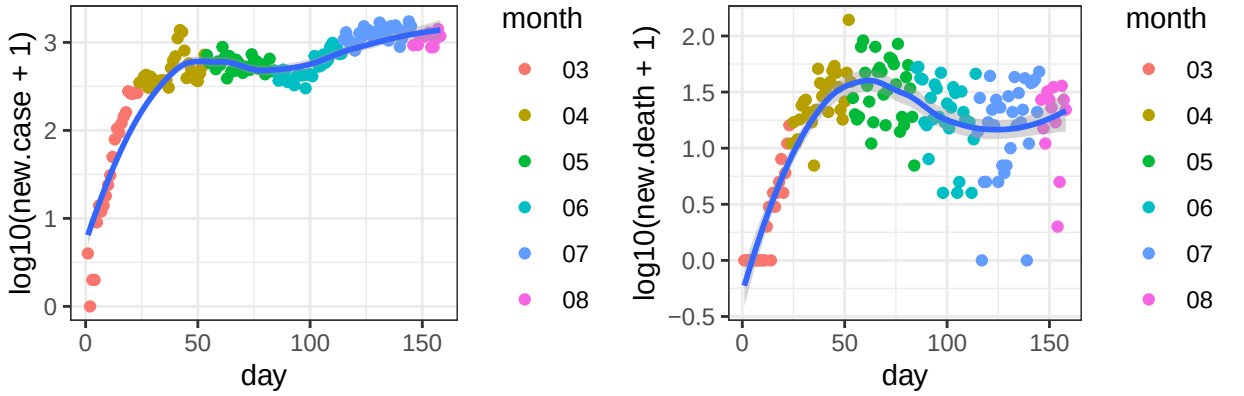
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Arizona



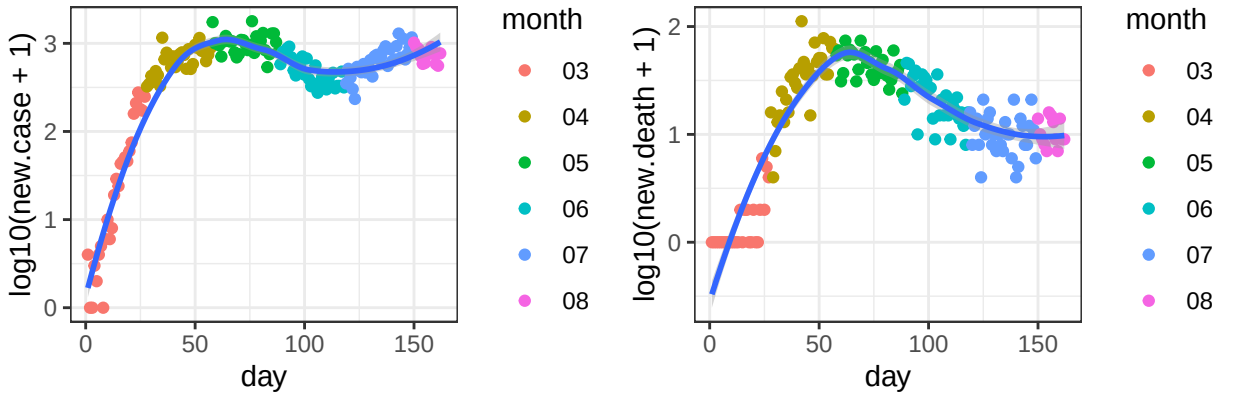
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Ohio



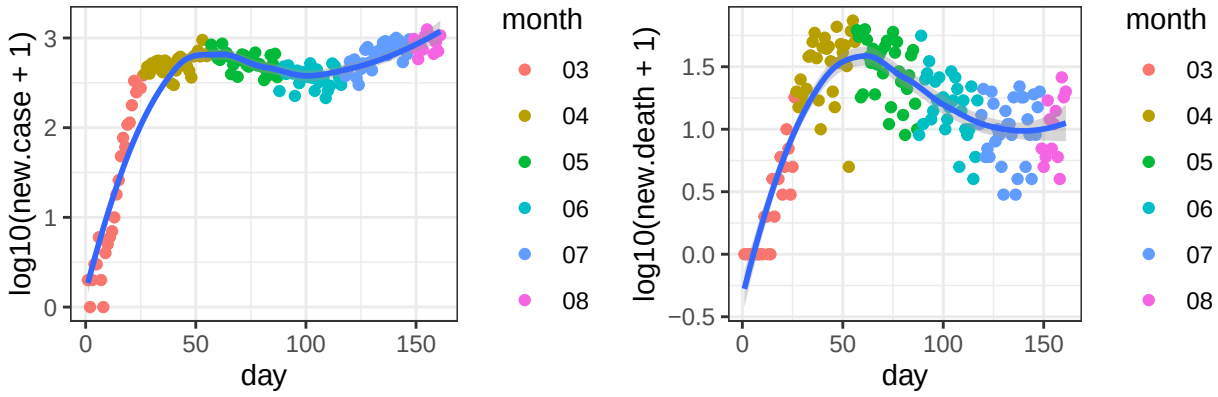
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



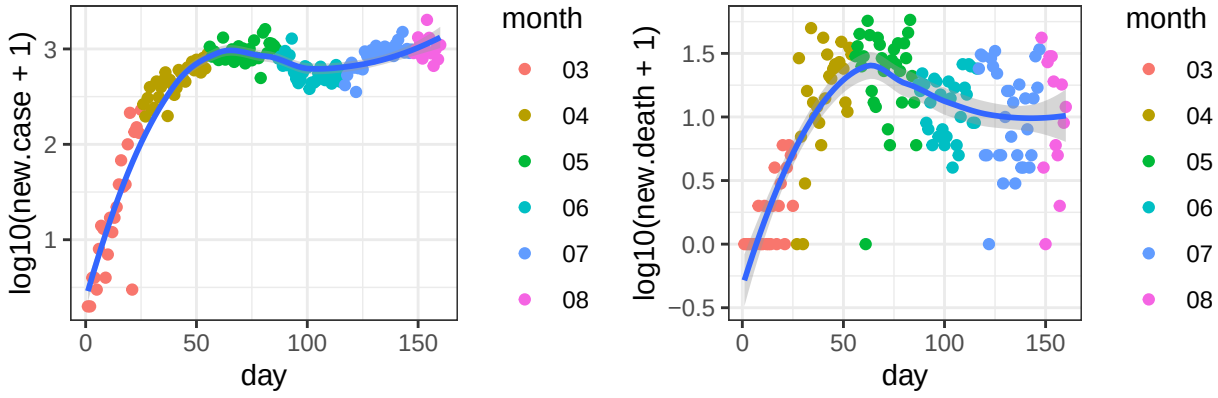
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Indiana



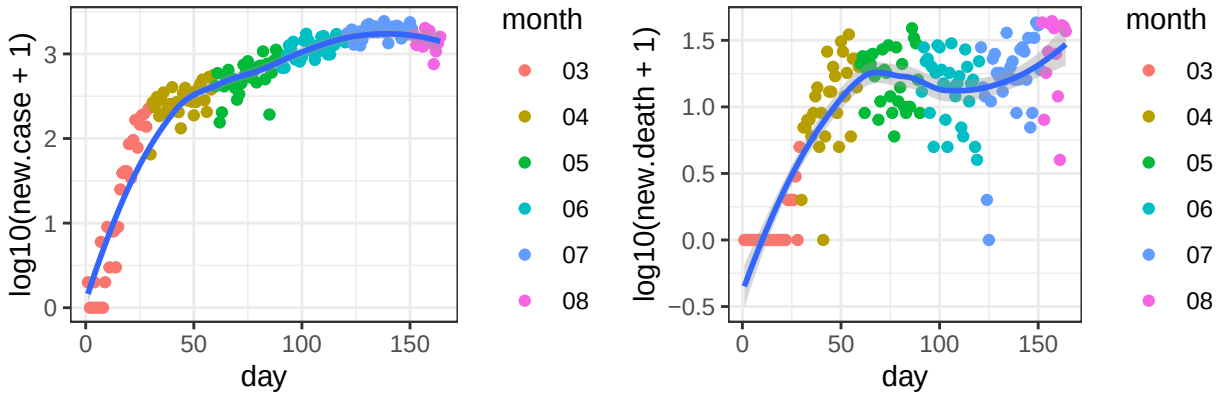
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Virginia



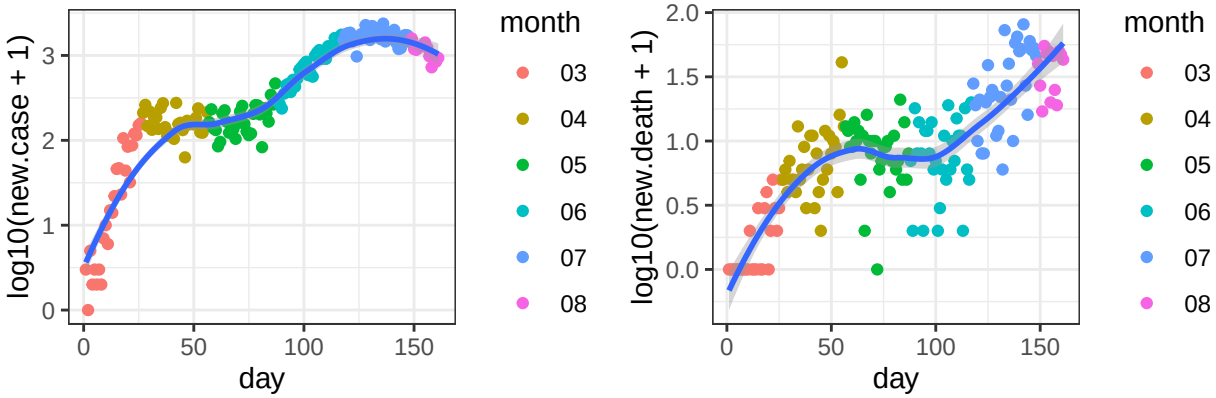
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

North Carolina



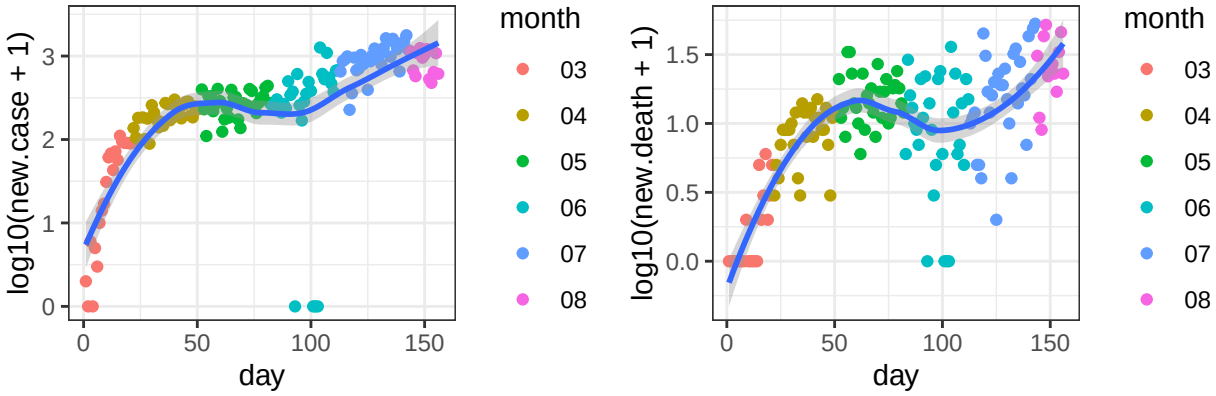
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

South Carolina



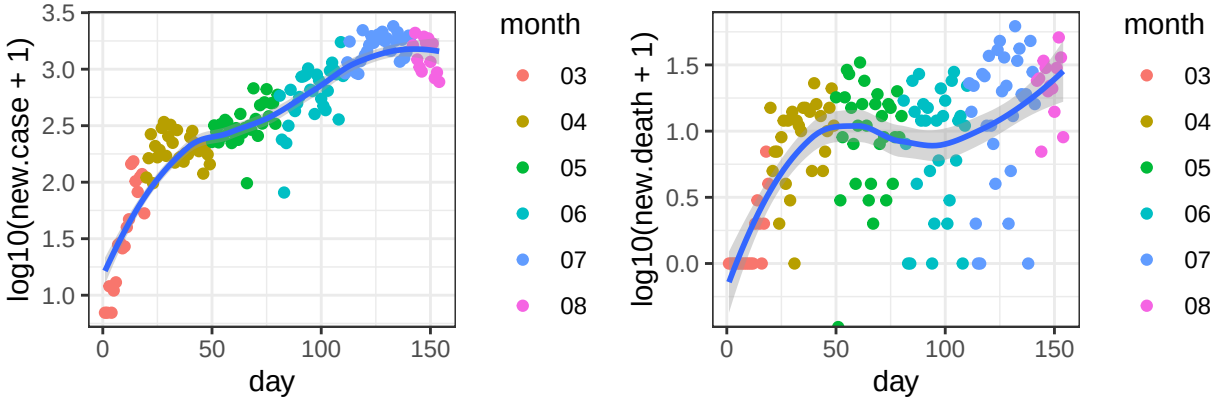
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Mississippi



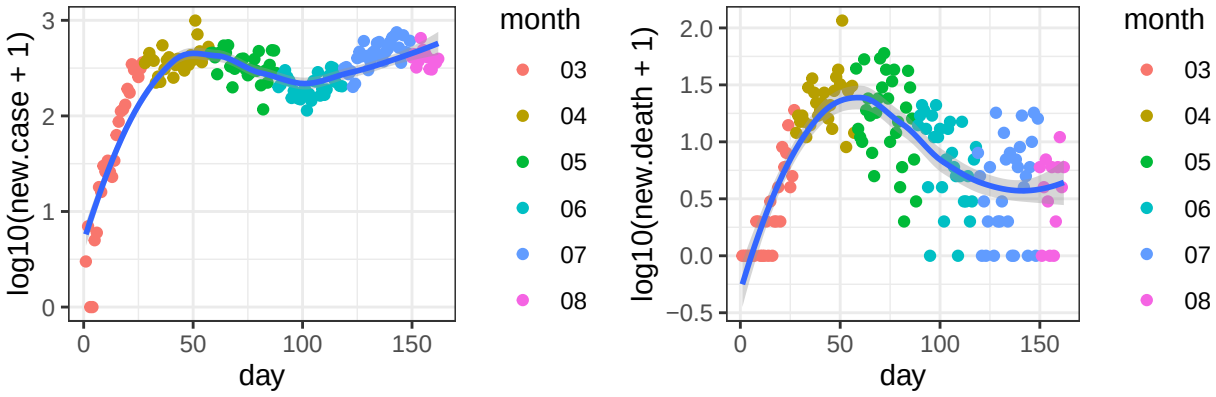
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Alabama



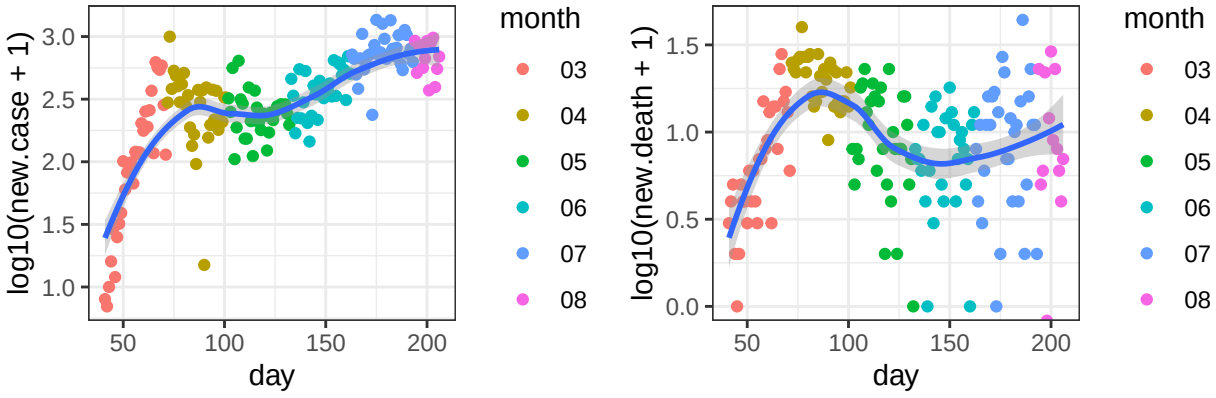
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Colorado



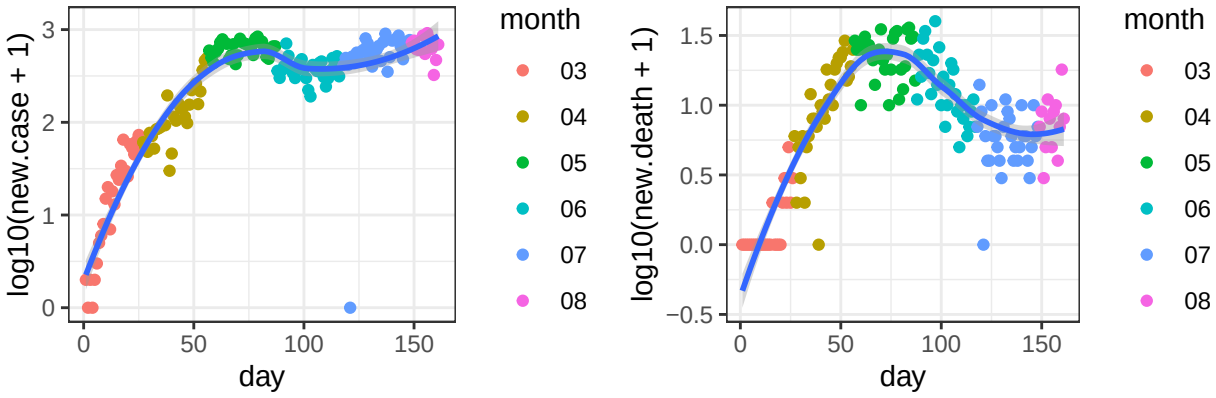
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Washington



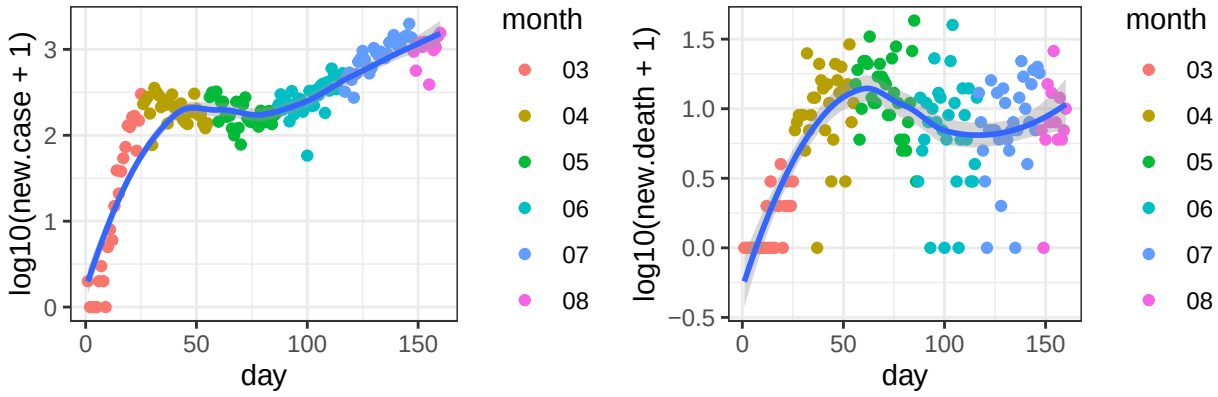
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Minnesota



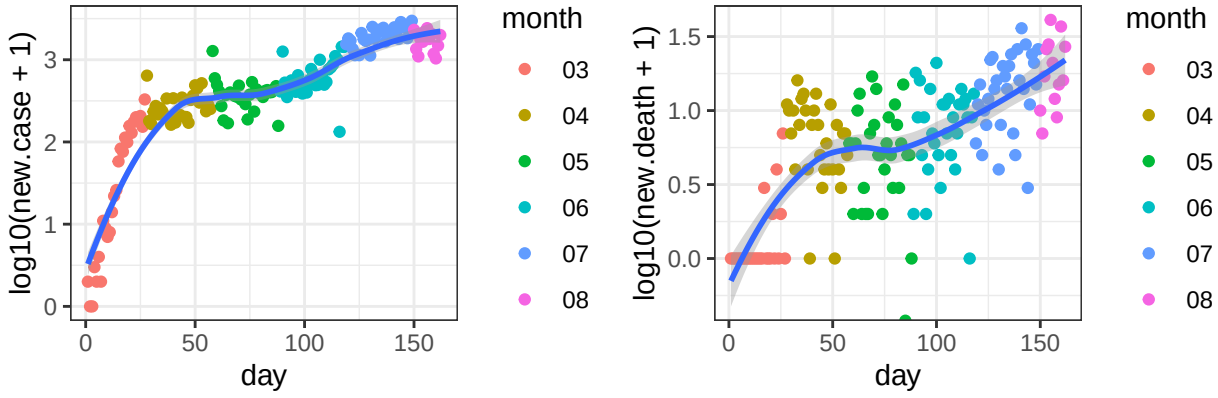
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Missouri



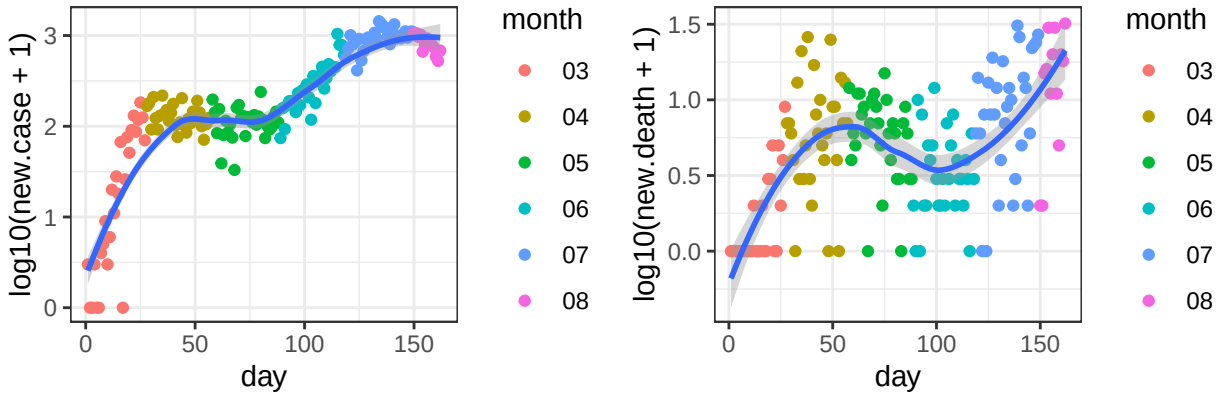
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Tennessee



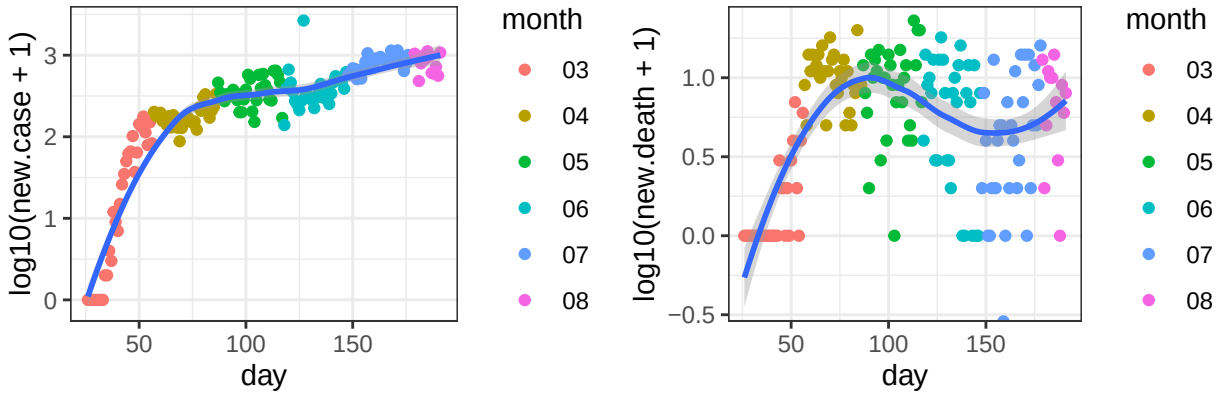
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Nevada



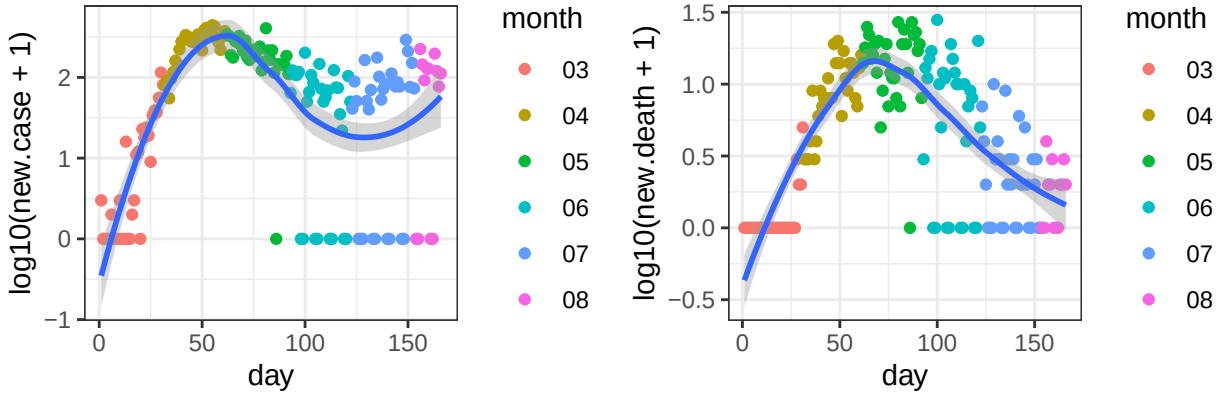
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wisconsin



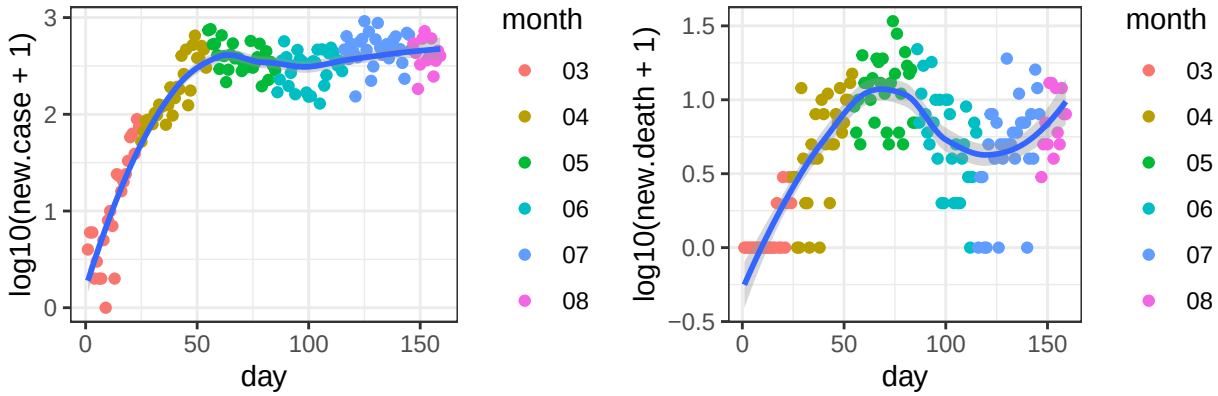
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Rhode Island



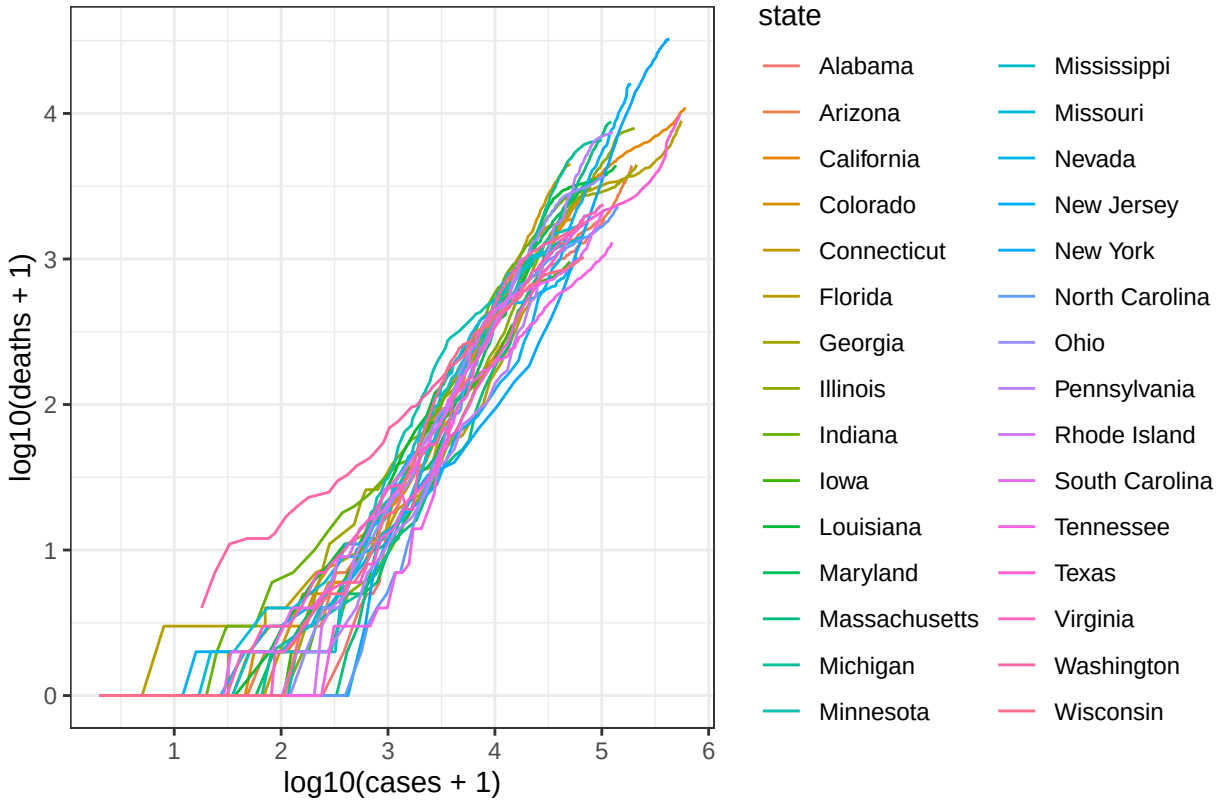
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Iowa



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

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

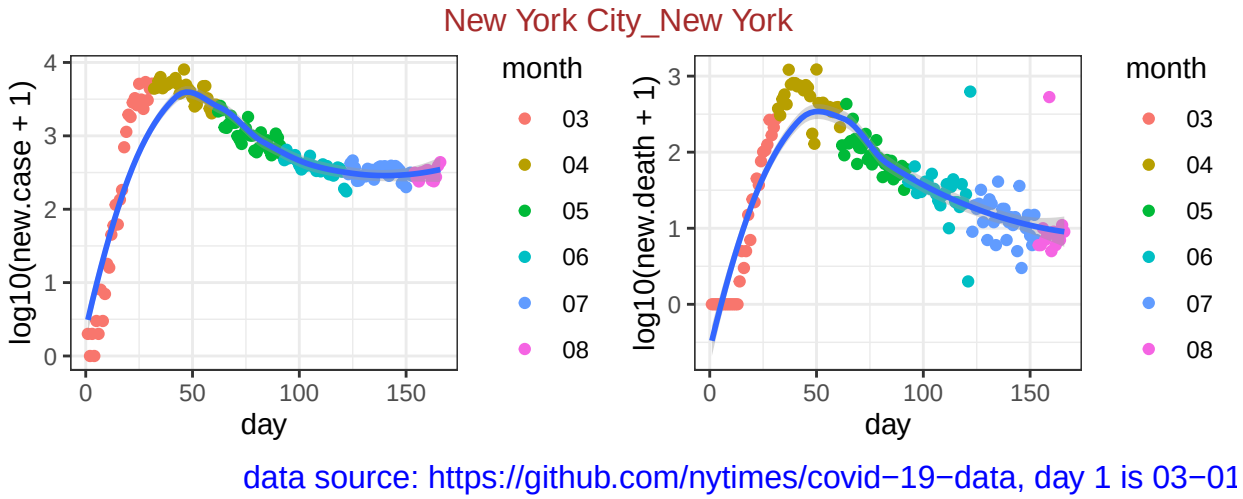
county level data

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

```
##              date          county         state  fips  cases deaths
## 429700 2020-08-13   New York City      New York    NA 233859  23610
## 428044 2020-08-13     Los Angeles    California  6037 216139   5171
## 428453 2020-08-13            Cook      Illinois 17031 113796   4943
## 429160 2020-08-13           Wayne      Michigan 26163  28715   2835
## 427942 2020-08-13         Maricopa       Arizona  4013 127768   2517
## 429699 2020-08-13          Nassau      New York 36059  43795   2195
## 429623 2020-08-13           Essex    New Jersey 34013  20074   2110
## 429618 2020-08-13          Bergen    New Jersey 34003  21188   2035
## 429071 2020-08-13       Middlesex Massachusetts 25017  26565   2006
## 429719 2020-08-13         Suffolk      New York 36103  43987   1998
## 428203 2020-08-13      Miami-Dade       Florida 12086 140983   1954
## 430136 2020-08-13    Philadelphia  Pennsylvania 42101  31910   1725
## 430546 2020-08-13          Harris         Texas 48201  89425   1713
## 429625 2020-08-13          Hudson    New Jersey 34017  19959   1508
## 429727 2020-08-13     Westchester      New York 36119  36356   1447
## 428148 2020-08-13        Hartford   Connecticut  9003  12855   1416
## 429628 2020-08-13       Middlesex    New Jersey 34023  18220   1411
## 428147 2020-08-13       Fairfield   Connecticut  9001  18126   1410
## 429636 2020-08-13           Union    New Jersey 34039  16980   1349
## 429632 2020-08-13         Passaic    New Jersey 34031  17948   1242
## 429067 2020-08-13           Essex Massachusetts 25009  17930   1194
## 429140 2020-08-13         Oakland      Michigan 26125  16083   1136
```

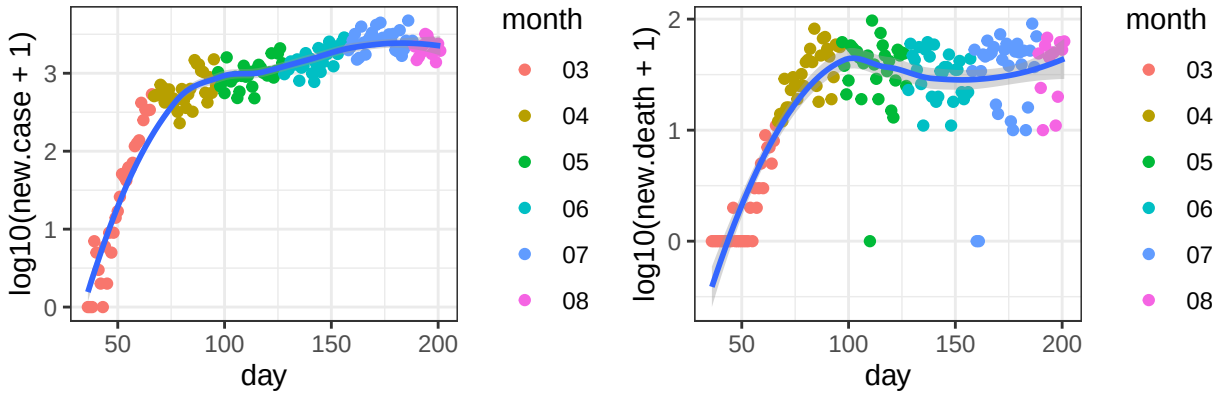
```
## 428151 2020-08-13       New Haven   Connecticut  9009  13272    1109
## 429075 2020-08-13         Suffolk Massachusetts 25025  22017    1076
## 429631 2020-08-13           Ocean    New Jersey 34029  10735    1020
## 429077 2020-08-13       Worcester Massachusetts 25027  13686    1007
## 429073 2020-08-13         Norfolk Massachusetts 25021  10682     997
## 428210 2020-08-13      Palm Beach       Florida 12099  38206     964
## 429127 2020-08-13          Macomb      Michigan 26099  11195     954
## 428166 2020-08-13         Broward       Florida 12011  64741     883
## 430553 2020-08-13         Hidalgo         Texas 48215  20767     881
## 428058 2020-08-13       Riverside    California  6065  44747     879
## 429593 2020-08-13           Clark        Nevada 32003  50569     869
## 430131 2020-08-13      Montgomery  Pennsylvania 42091  10264     859
## 429629 2020-08-13        Monmouth    New Jersey 34025  10504     856
## 430461 2020-08-13           Bexar         Texas 48029  43685     850
## 429188 2020-08-13         Hennepin     Minnesota 27053  19873     842
## 429630 2020-08-13          Morris    New Jersey 34027   7451     829
## 430235 2020-08-13      Providence  Rhode Island 44007  15644     819
## 429053 2020-08-13      Montgomery      Maryland 24031  18791     807
## 430502 2020-08-13          Dallas         Texas 48113  56428     807
## 428589 2020-08-13          Marion       Indiana 18097  16457     783
## 428055 2020-08-13          Orange    California  6059  42171     769
## 429054 2020-08-13 Prince George's      Maryland 24033  24757     766
## 430108 2020-08-13        Delaware  Pennsylvania 42045   9541     752
## 429074 2020-08-13        Plymouth Massachusetts 25023   9288     723
## 429069 2020-08-13         Hampden Massachusetts 25013   7637     709
## 430894 2020-08-13            King    Washington 53033  17223     704
## 429434 2020-08-13       St. Louis      Missouri 29189  15872     673
## 429065 2020-08-13         Bristol Massachusetts 25005   9408     637
```

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



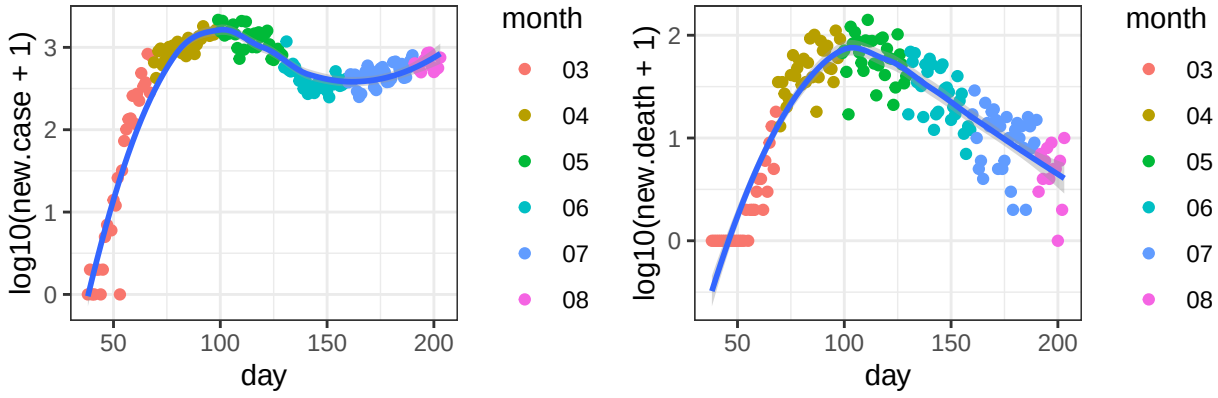
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Los Angeles_California



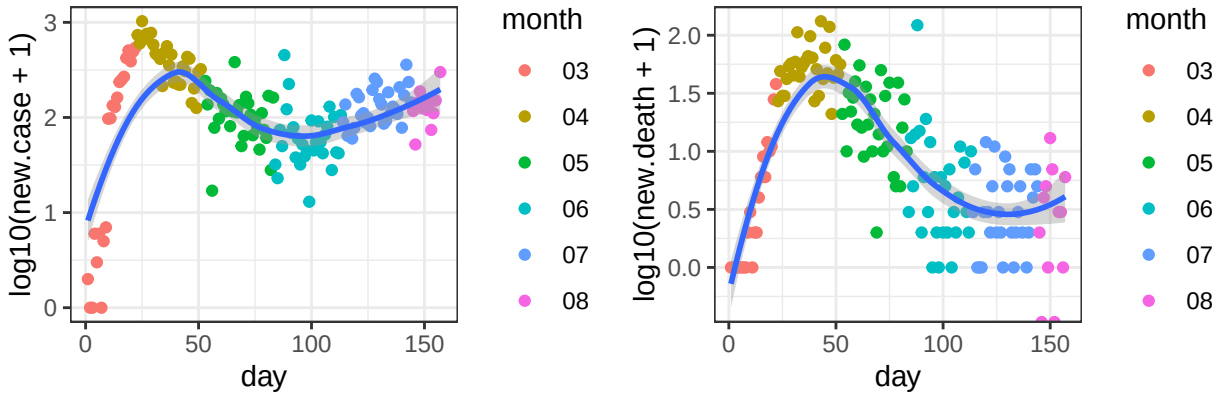
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



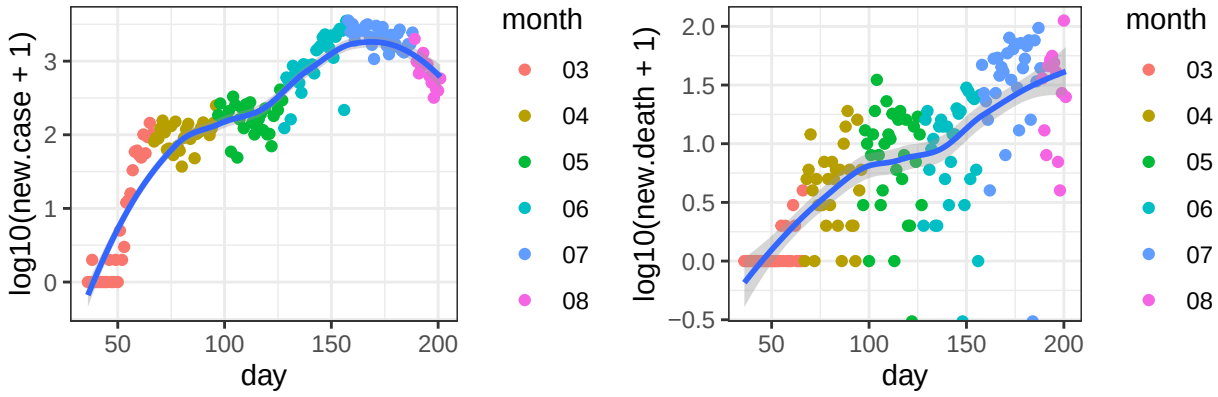
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wayne_Michigan



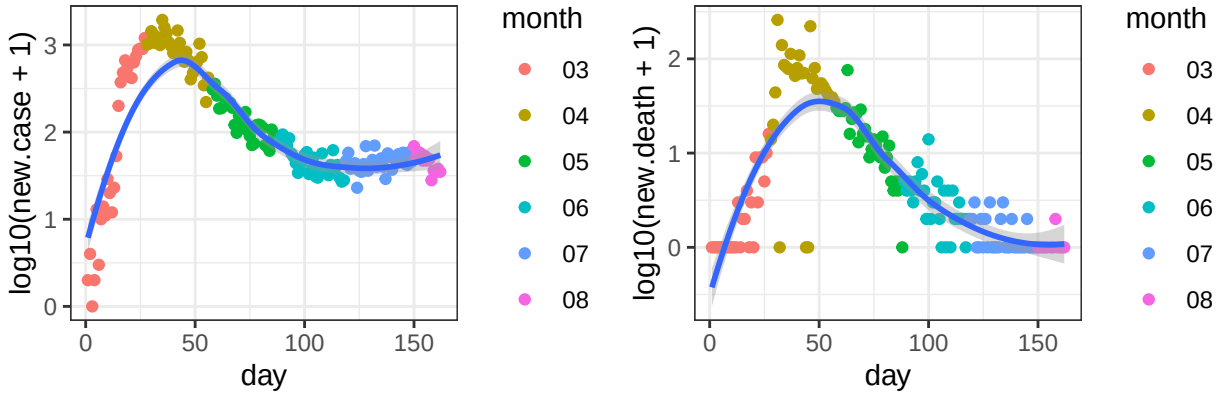
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Maricopa_Arizona



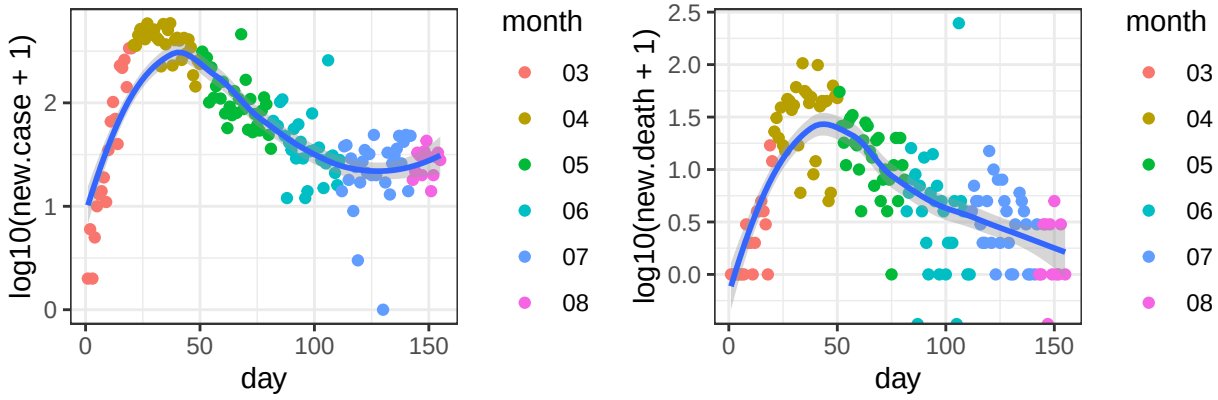
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Nassau_New York



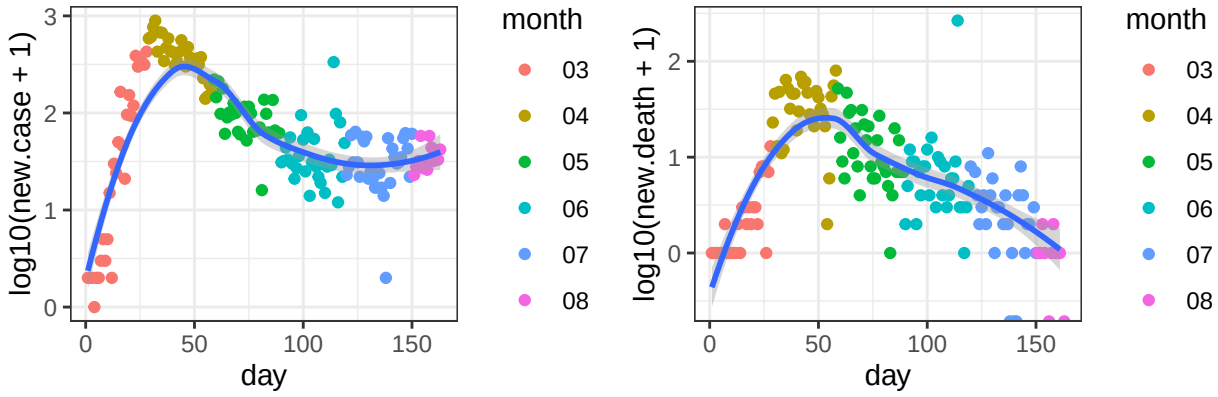
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



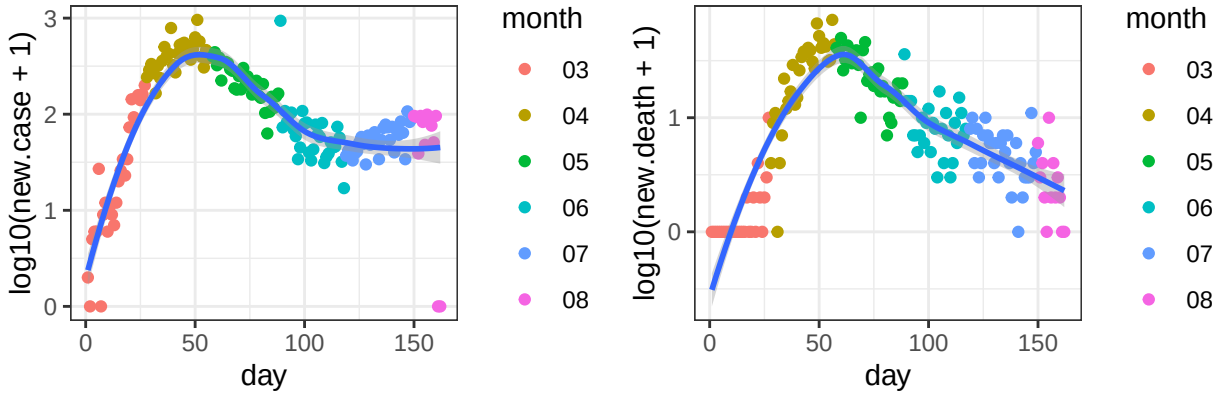
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



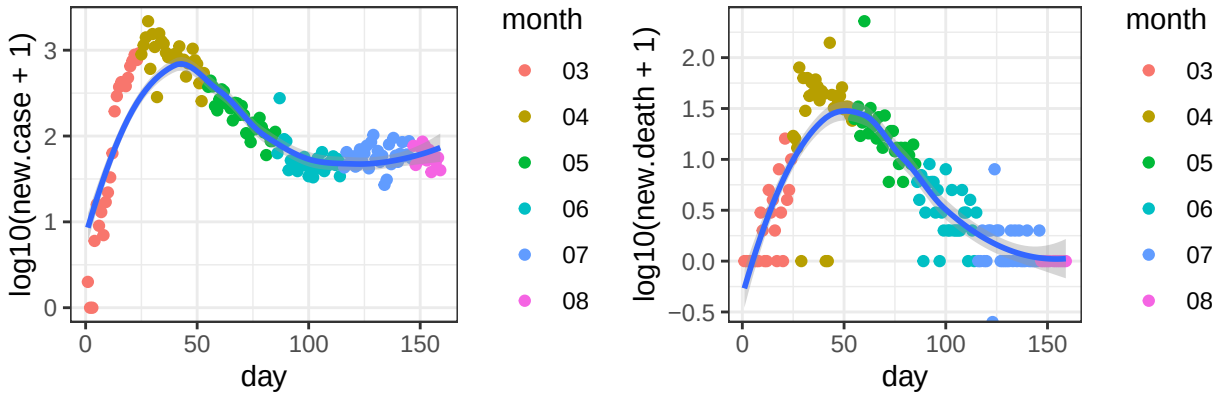
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



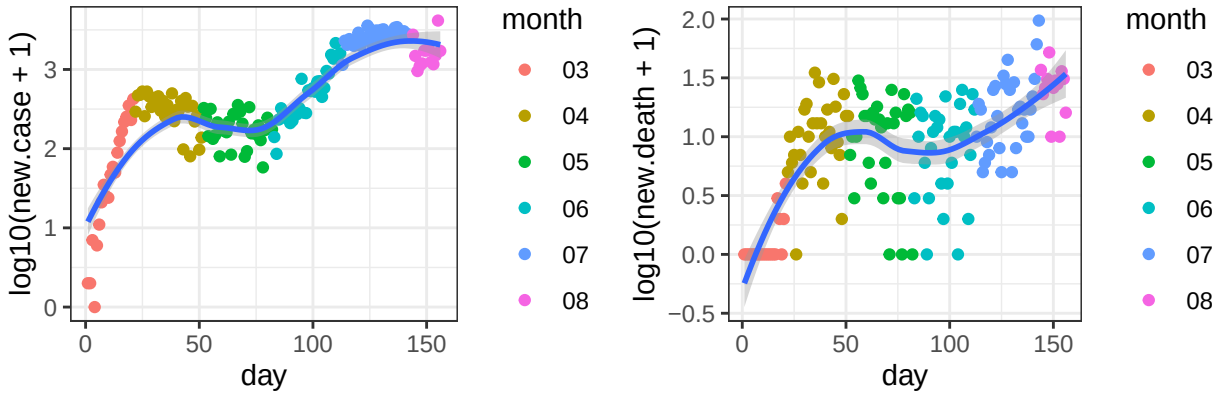
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Suffolk_New York



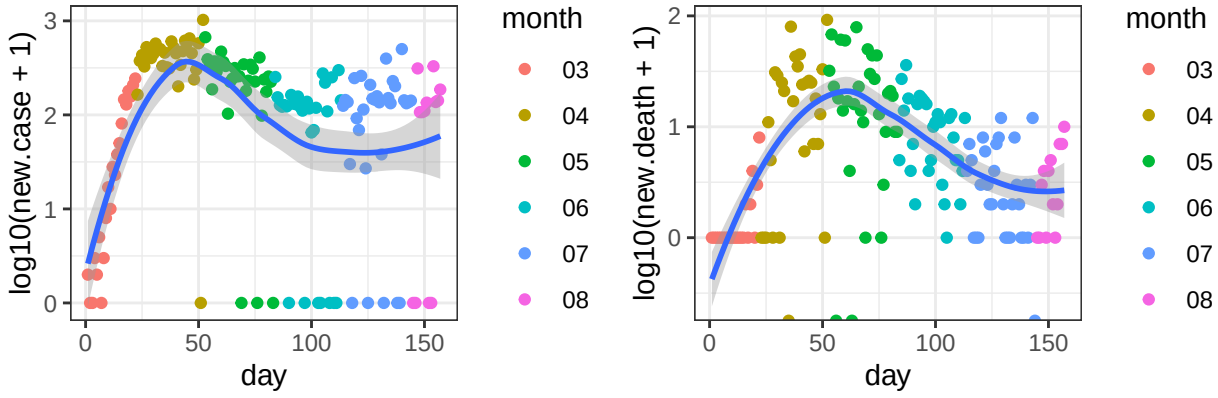
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Miami−Dade_Florida



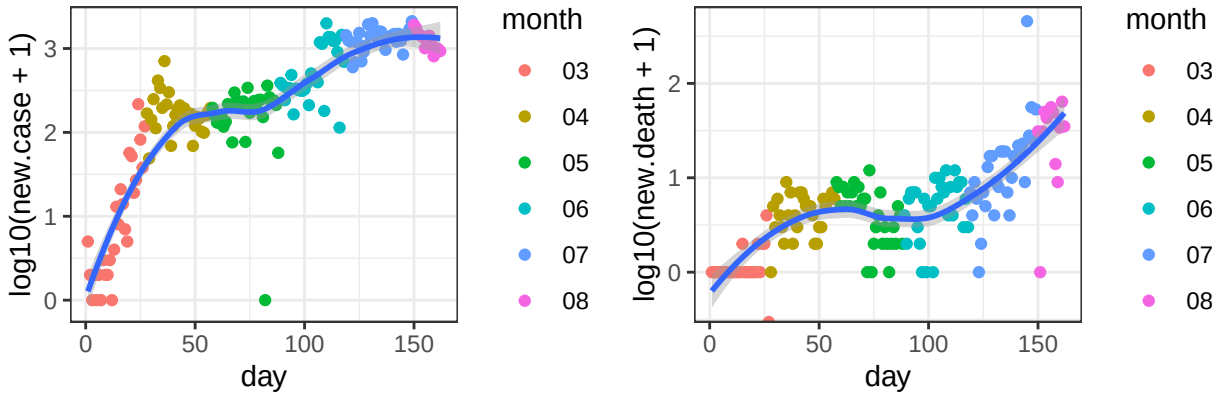
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Philadelphia_Pennsylvania



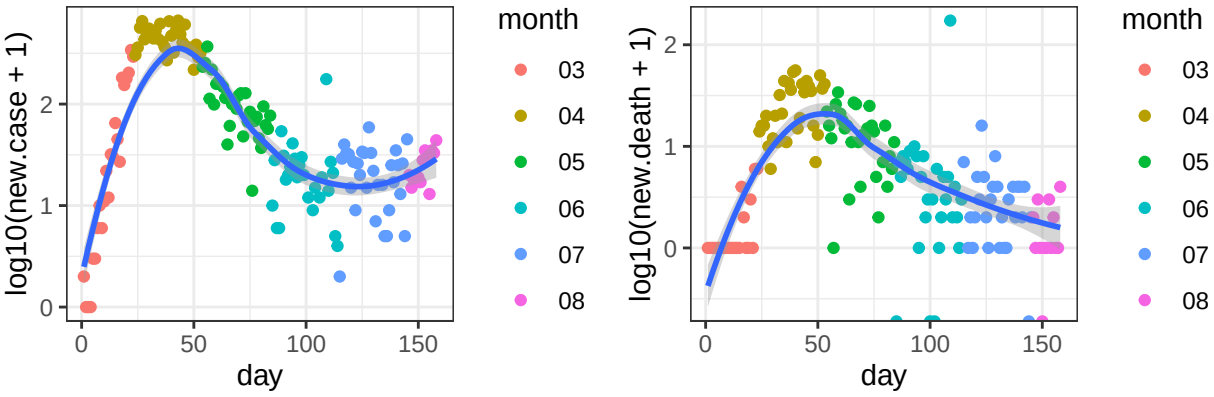
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Harris_Texas



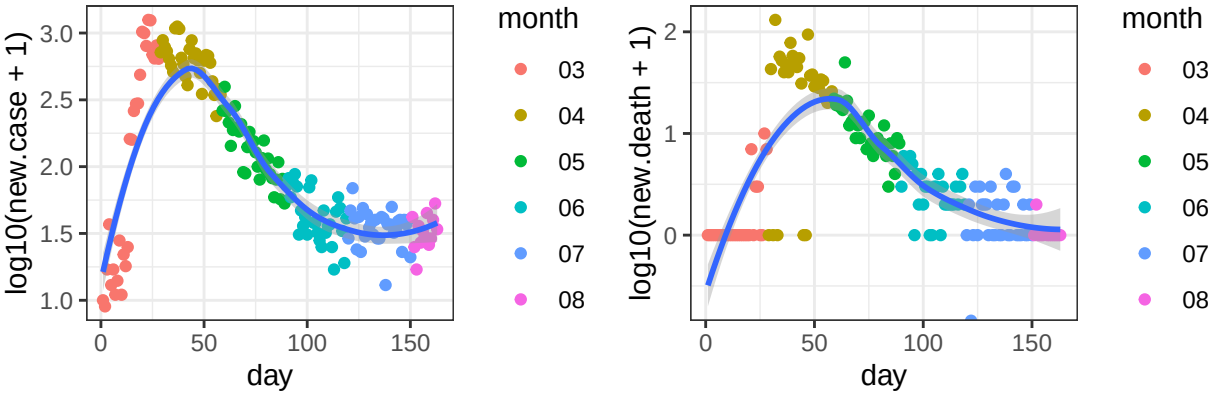
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Hudson_New Jersey



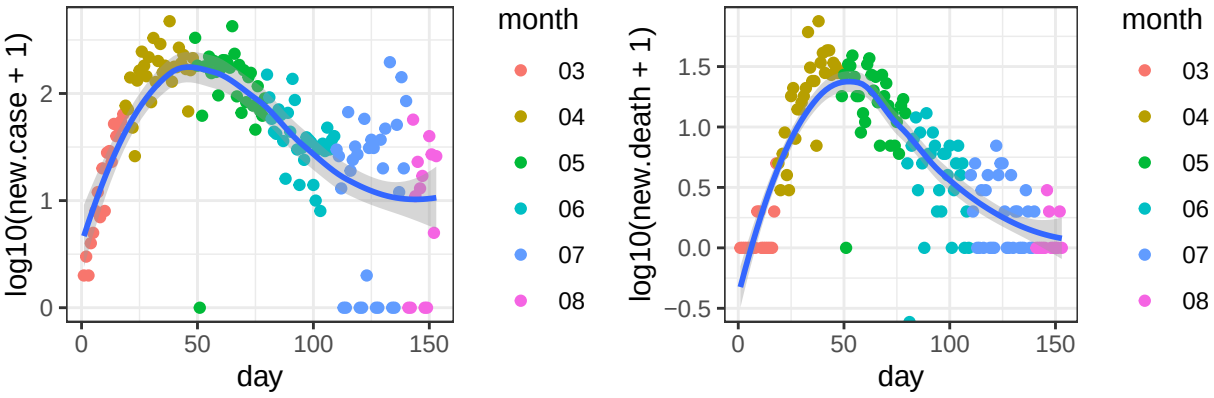
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Westchester_New York



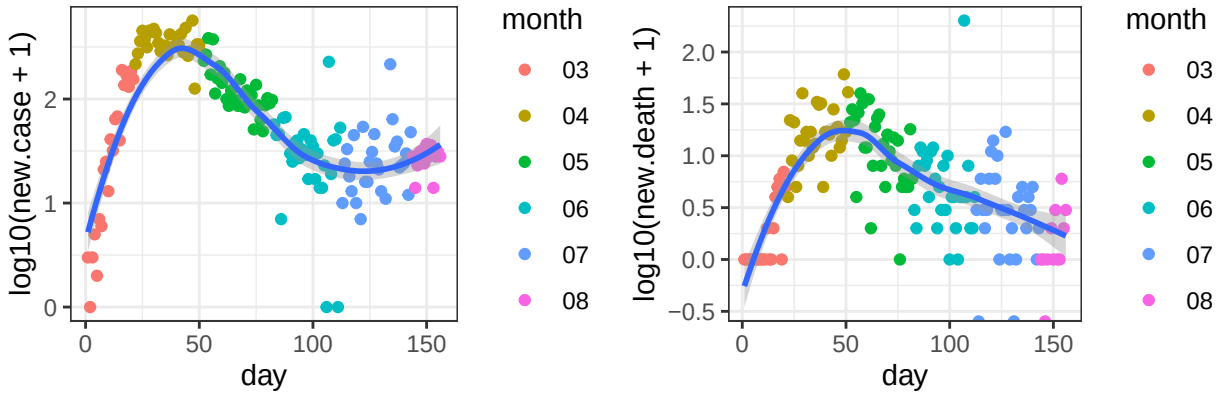
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hartford_Connecticut



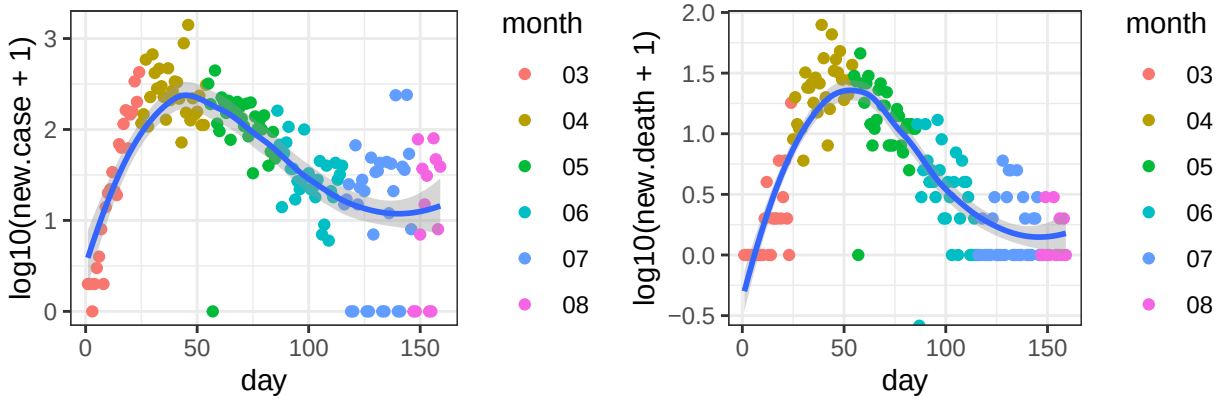
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Middlesex_New Jersey



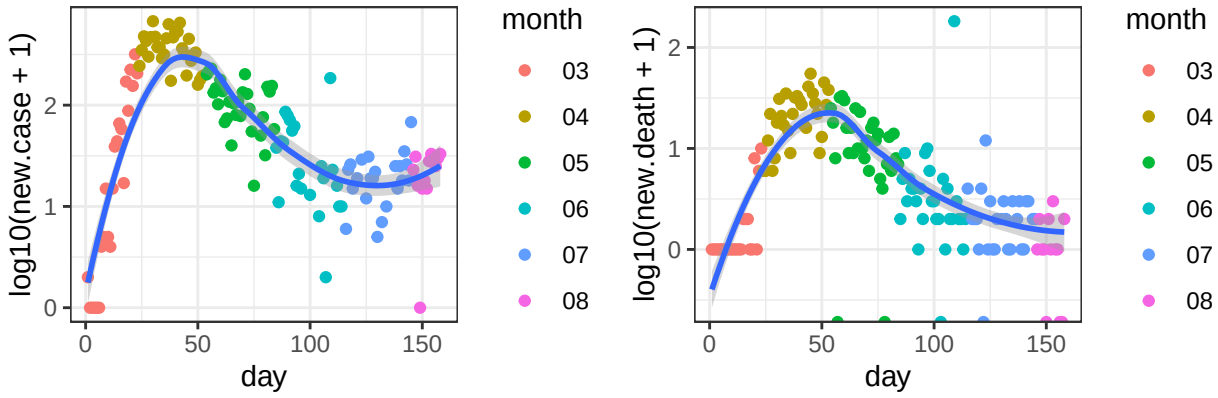
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Fairfield_Connecticut



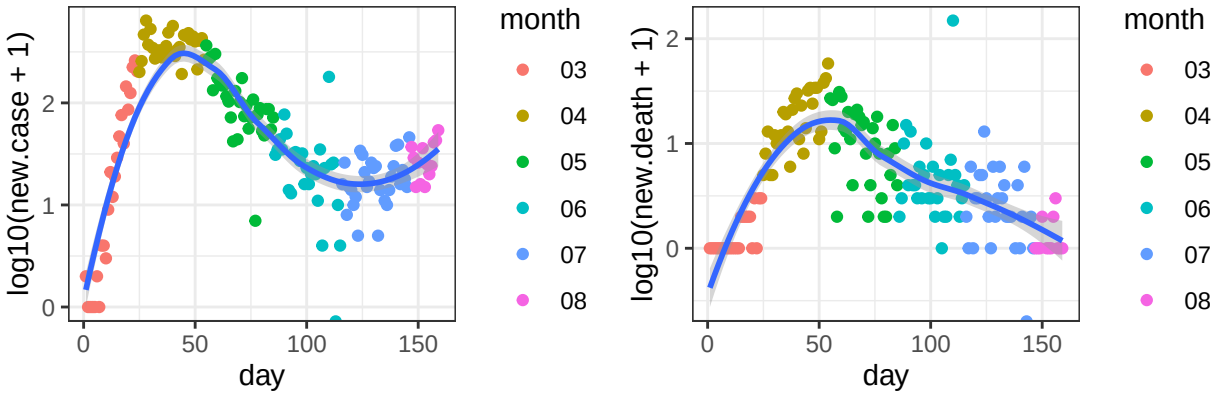
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Union_New Jersey



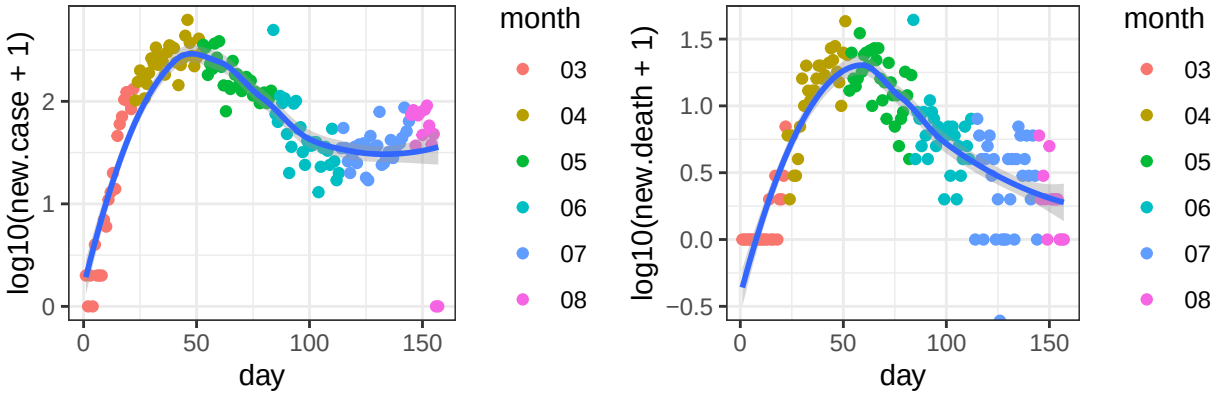
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Passaic_New Jersey



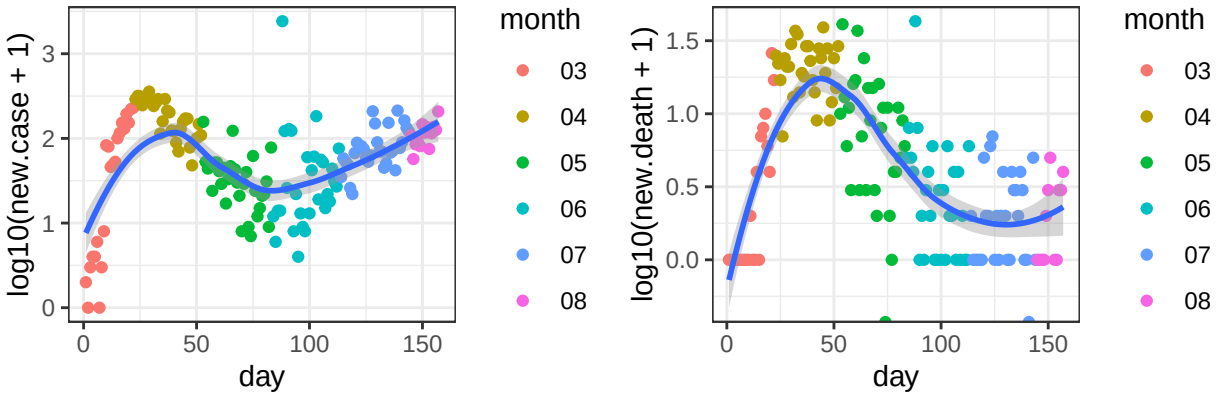
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



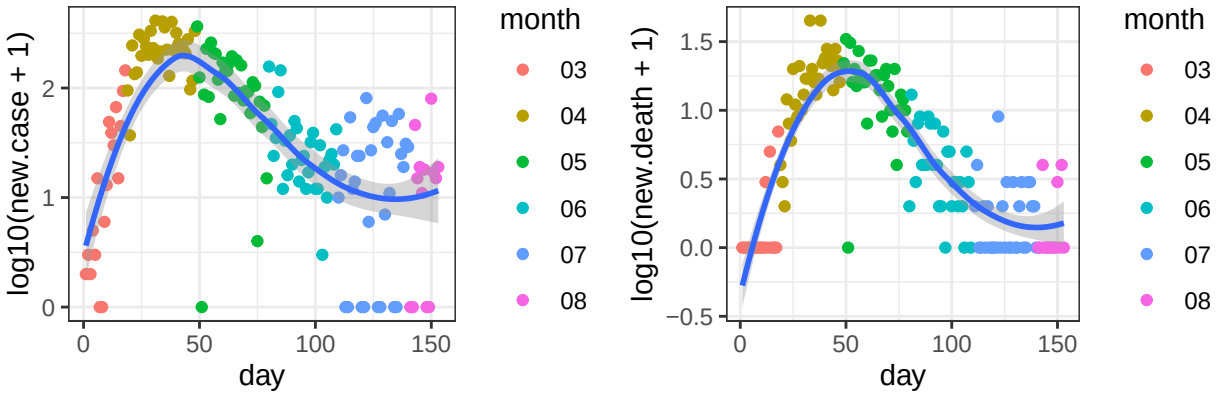
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



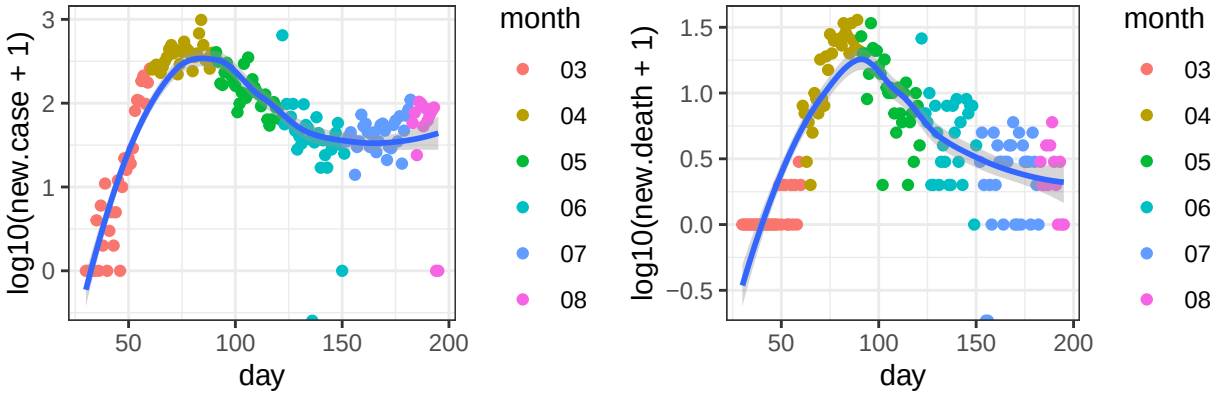
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



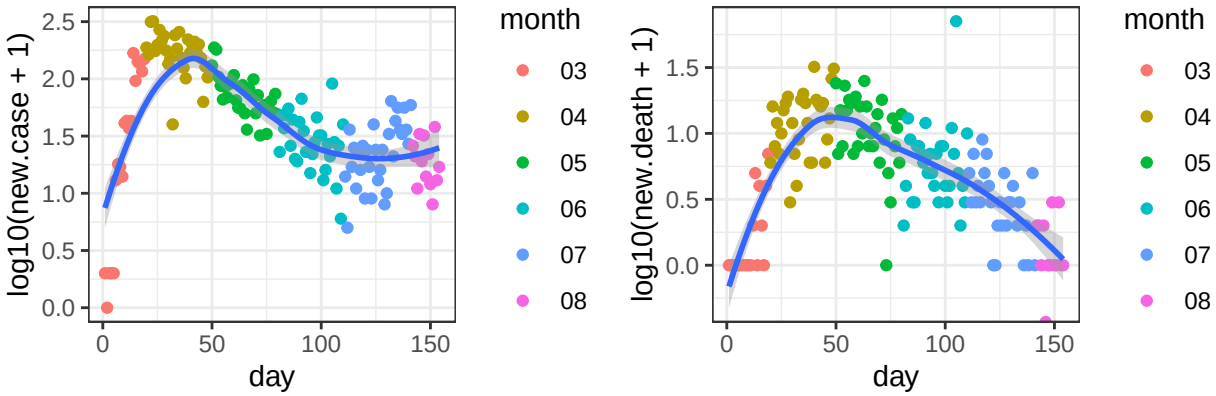
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



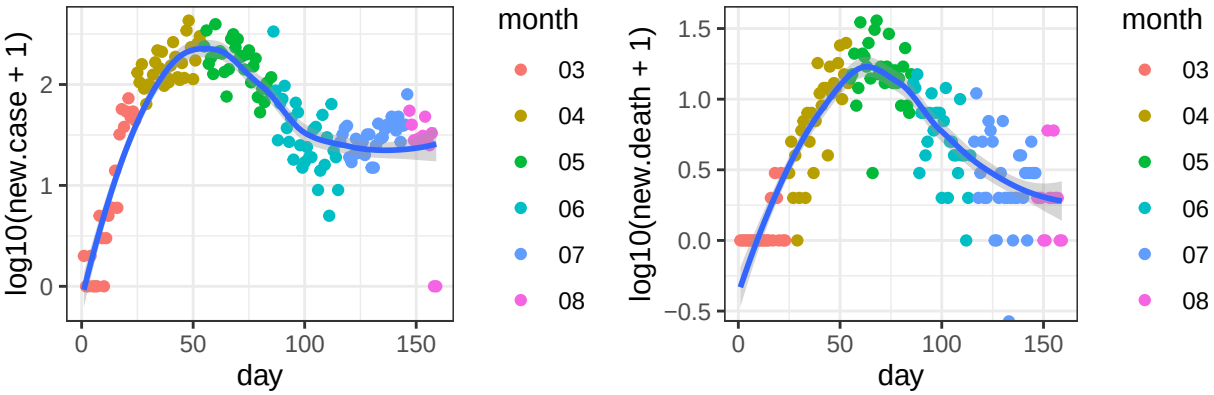
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Ocean_New Jersey



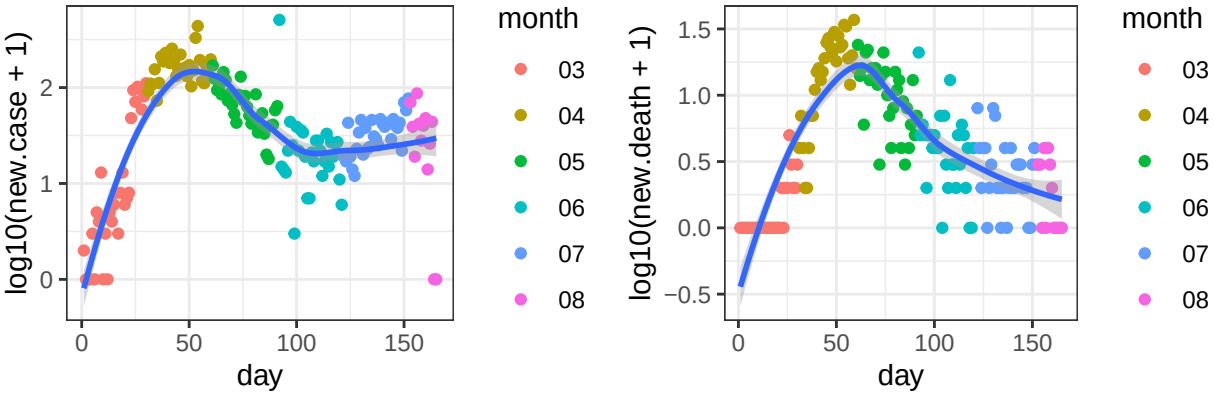
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



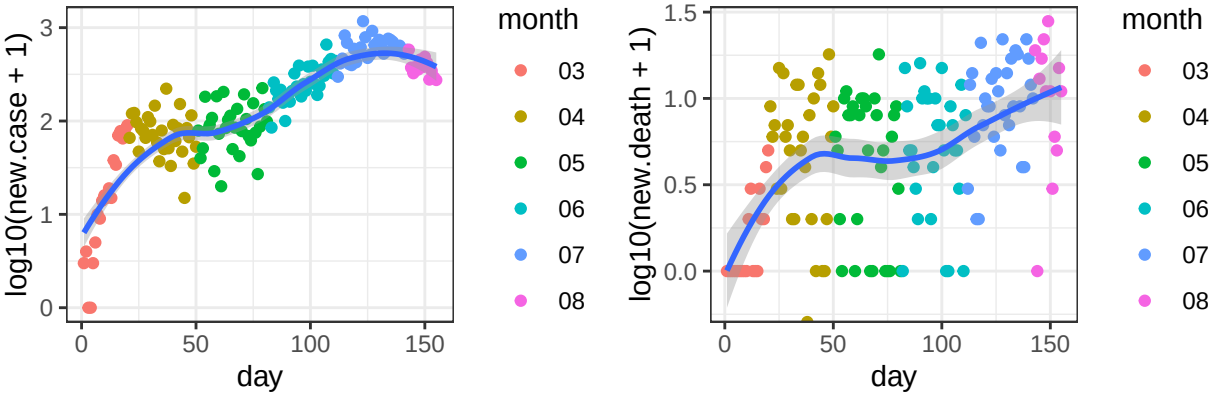
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



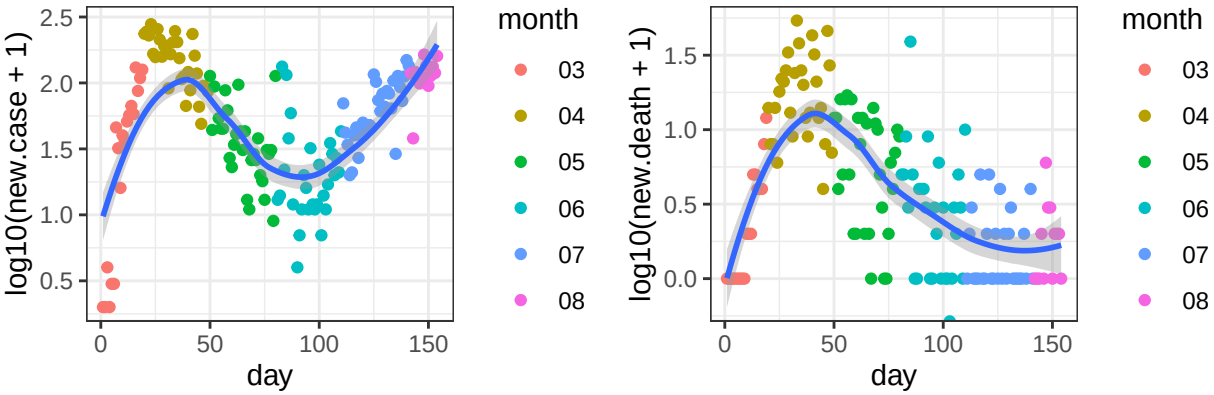
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Palm Beach_Florida



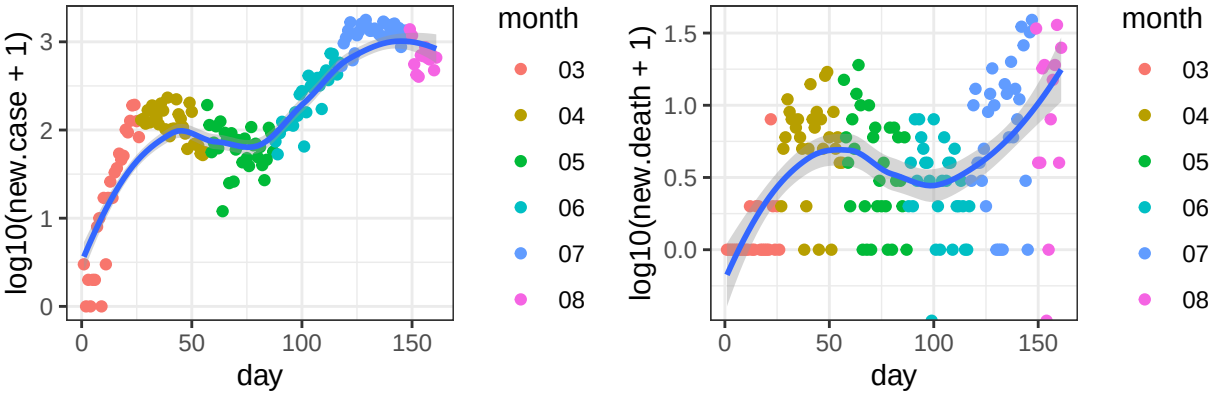
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Macomb_Michigan



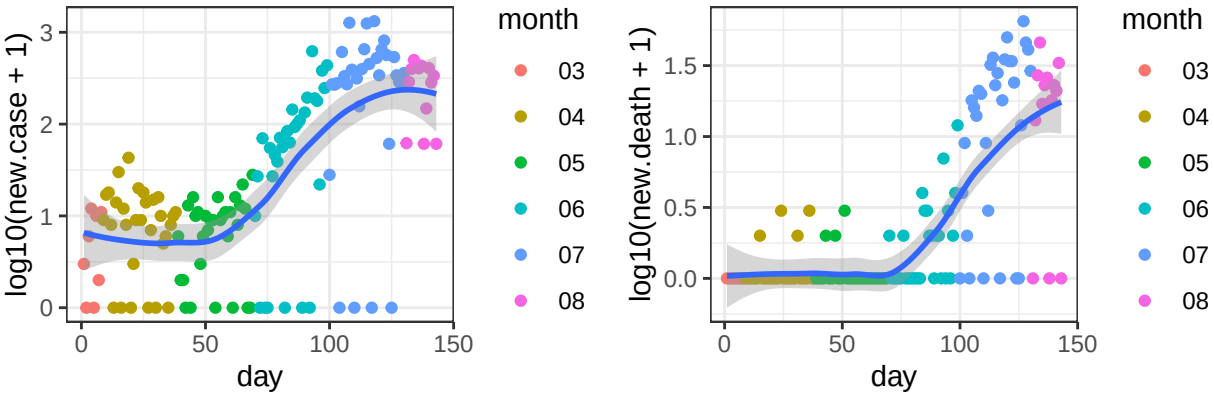
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Broward_Florida



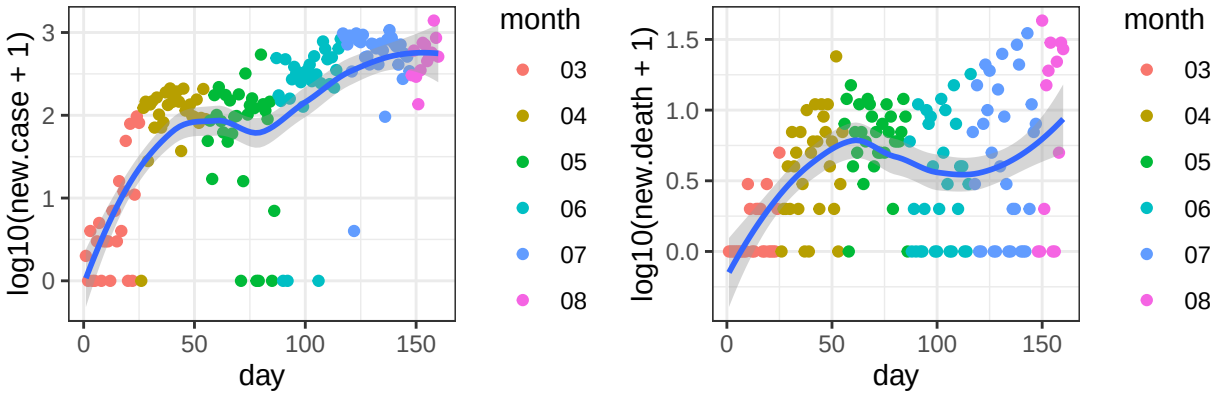
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Hidalgo_Texas



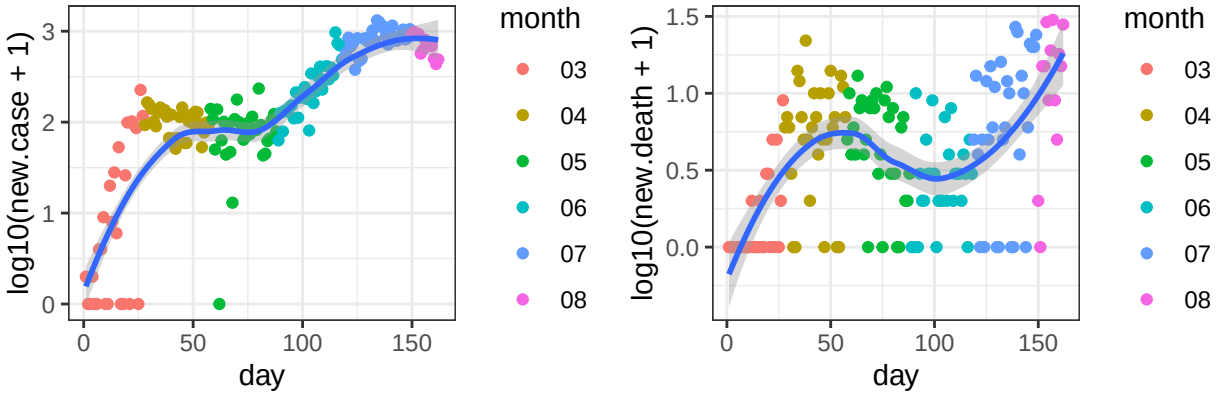
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24

Riverside_California



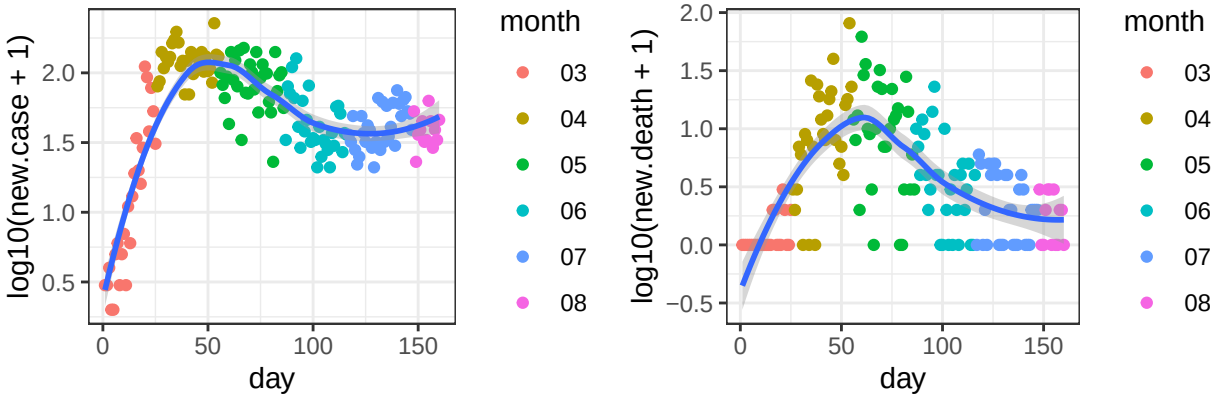
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Clark_Nevada



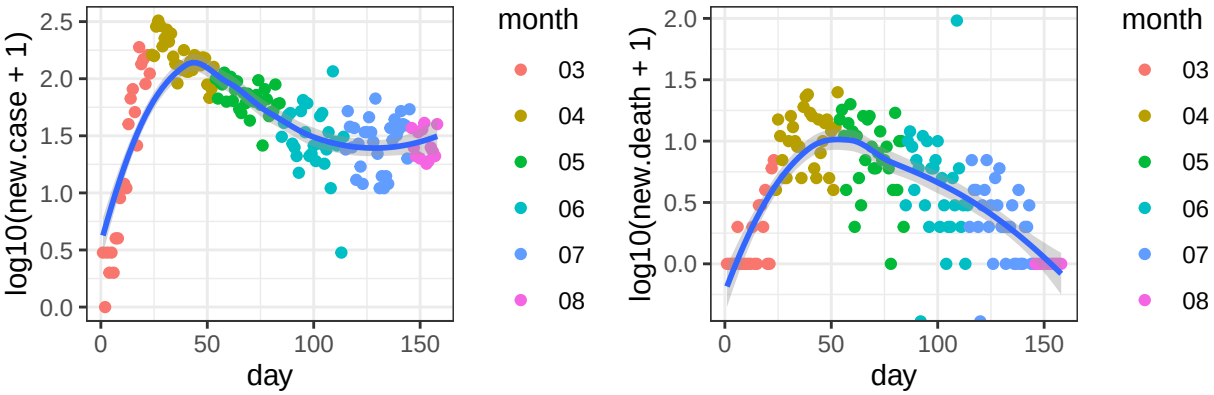
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Montgomery_Pennsylvania



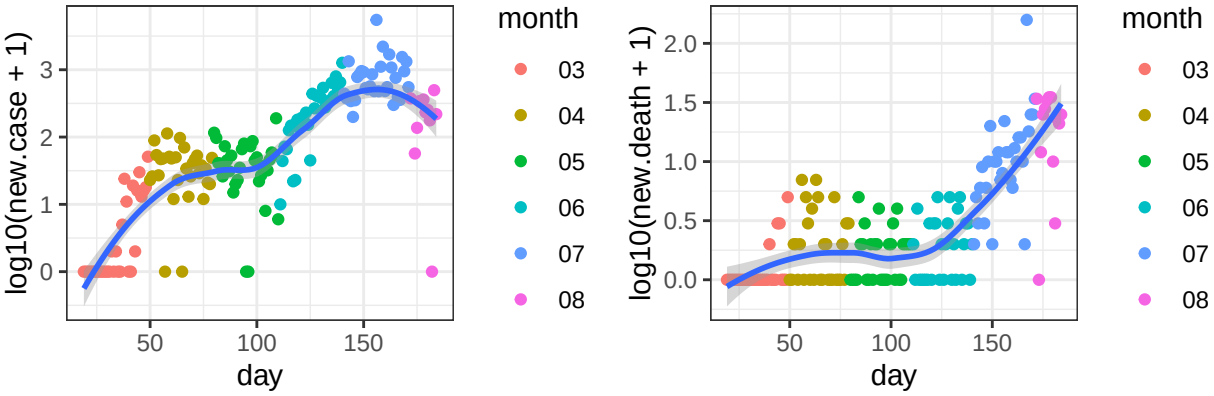
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Monmouth_New Jersey



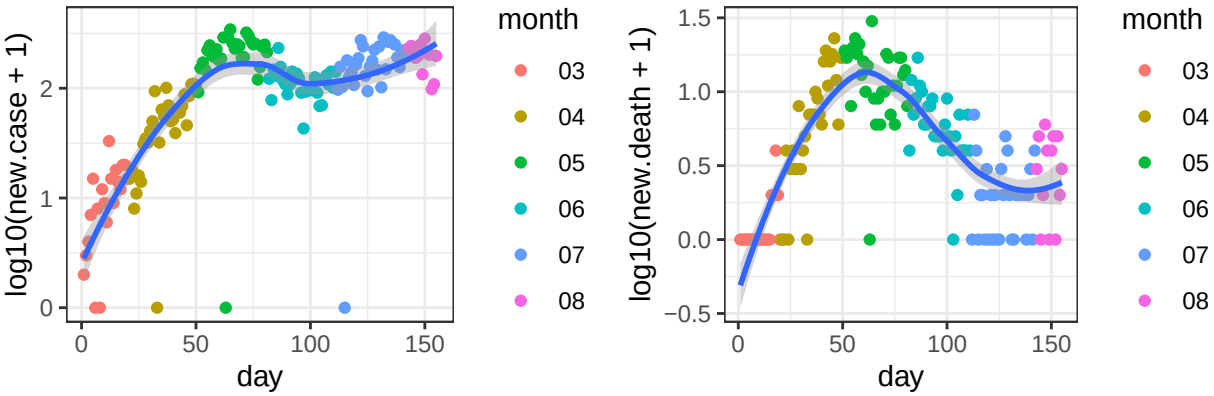
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Bexar_Texas



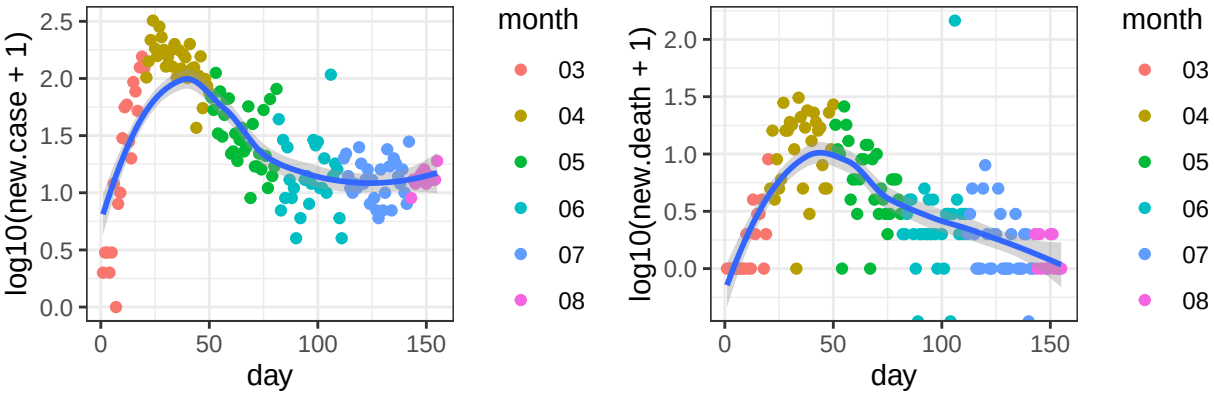
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Hennepin_Minnesota



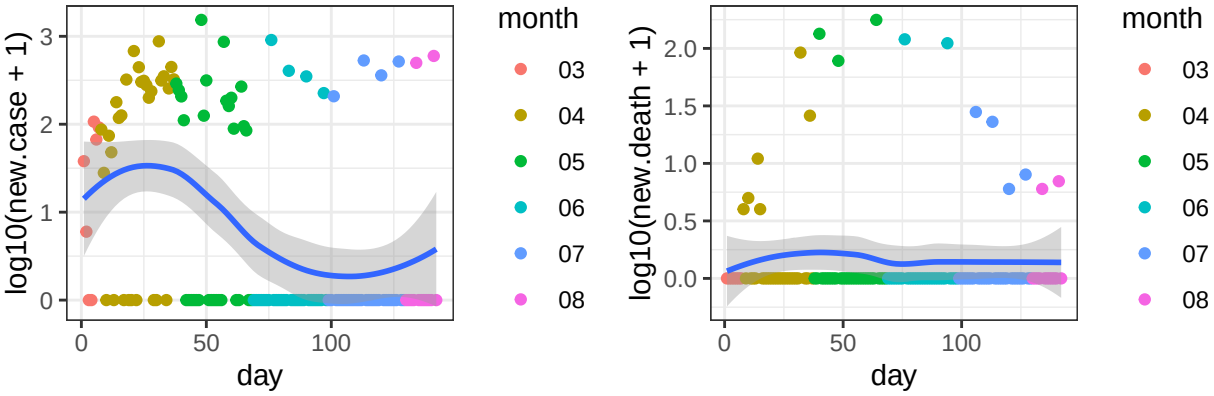
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Morris_New Jersey



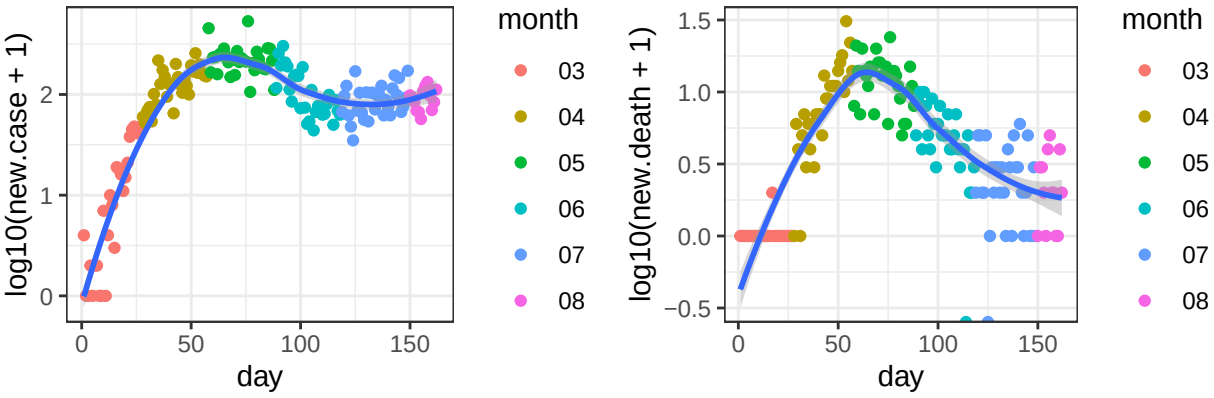
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



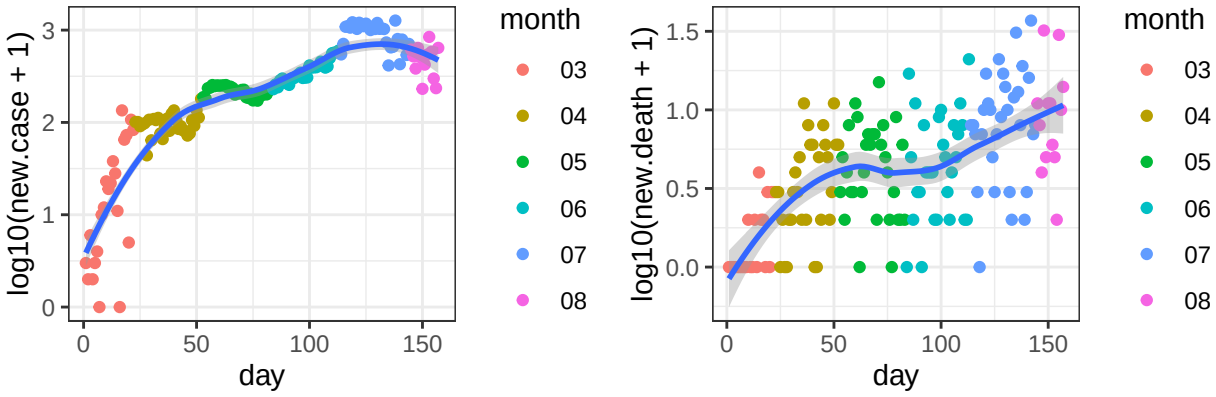
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Maryland



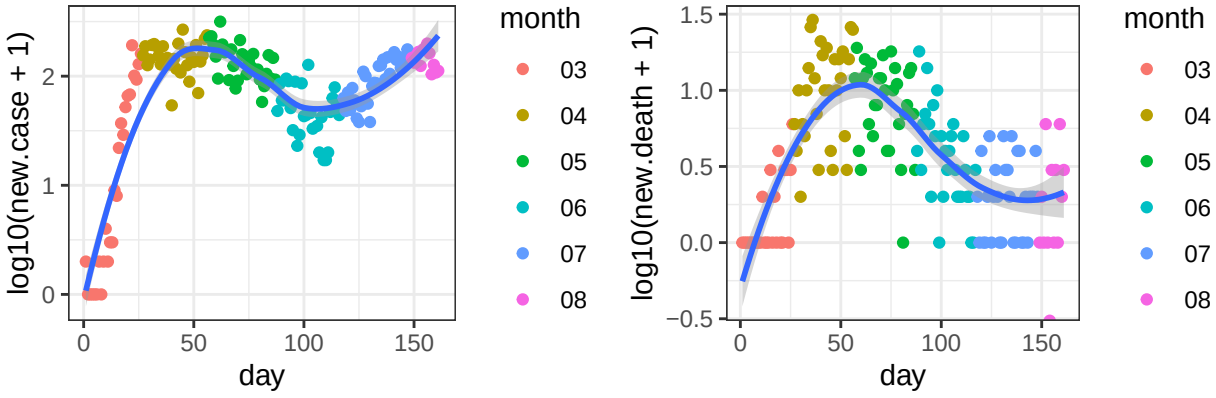
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Dallas_Texas



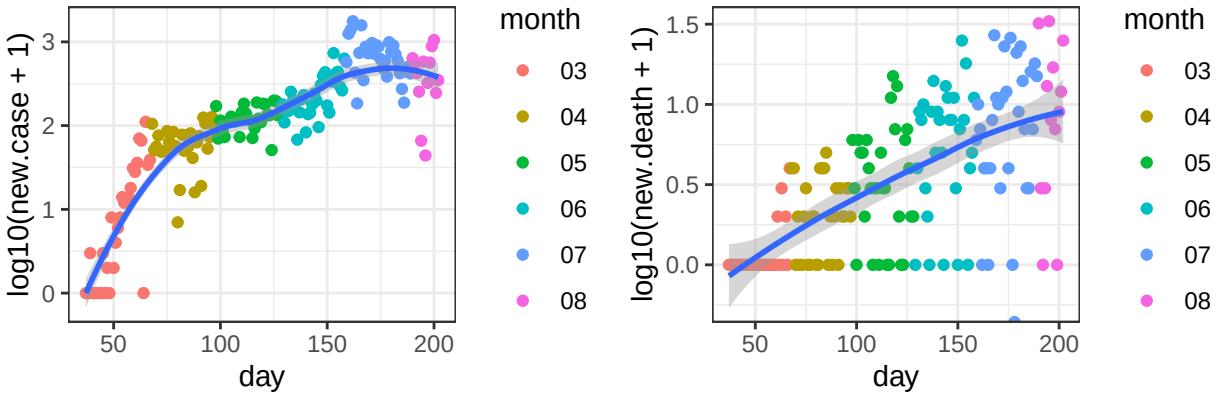
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Marion_Indiana



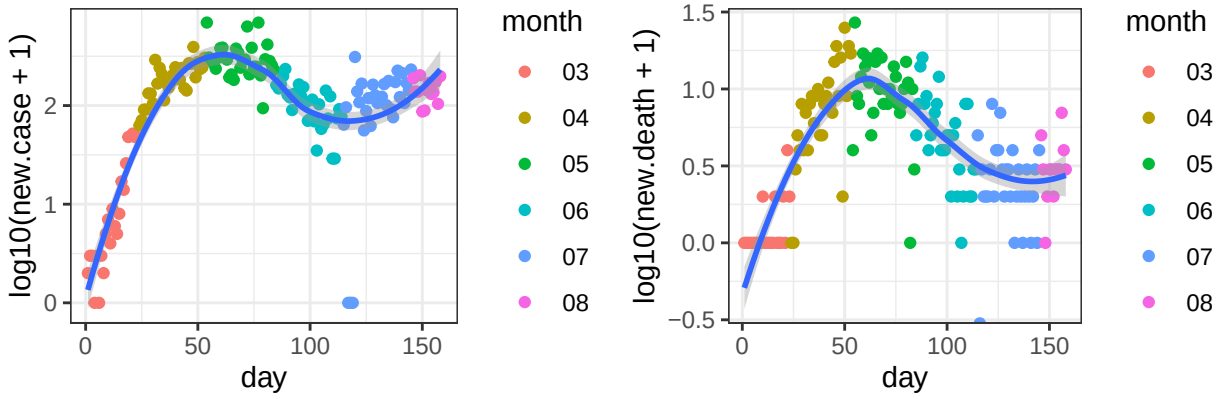
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Orange_California



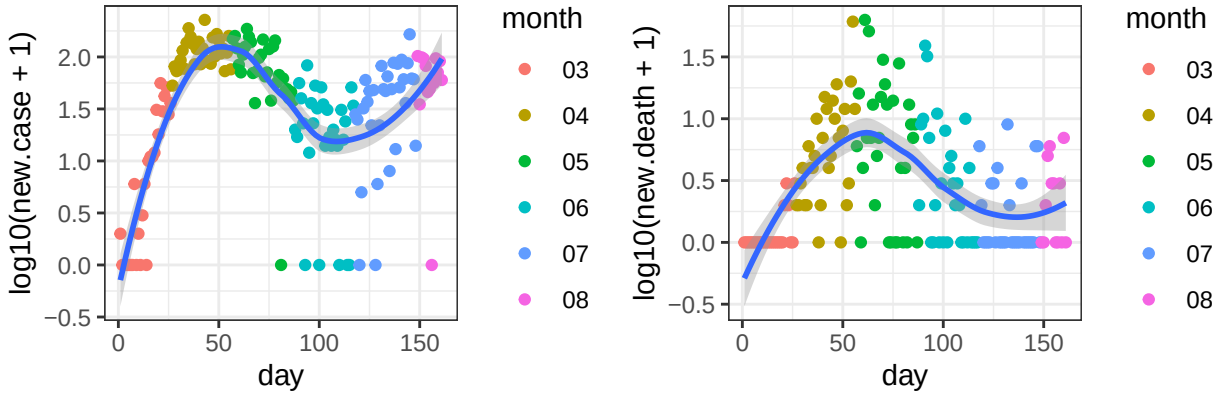
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Prince George's_Maryland



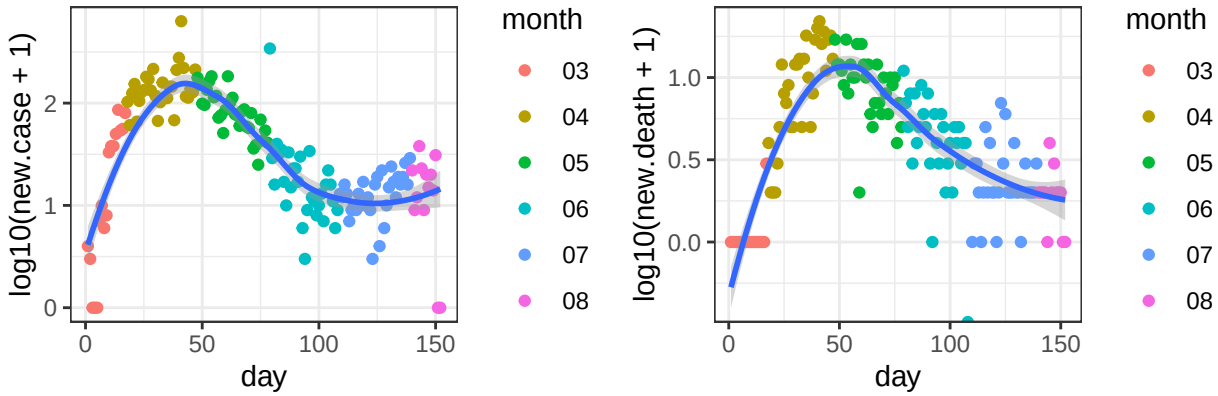
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



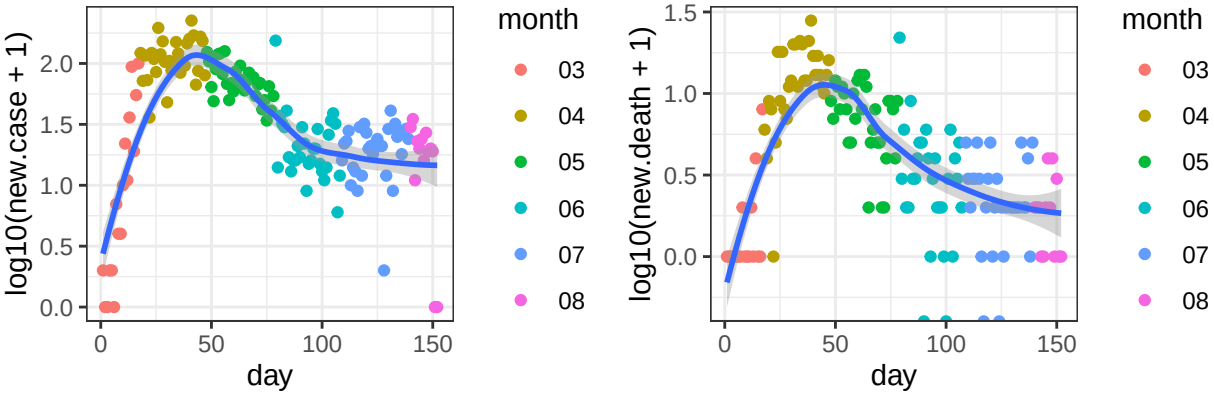
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Plymouth_Massachusetts



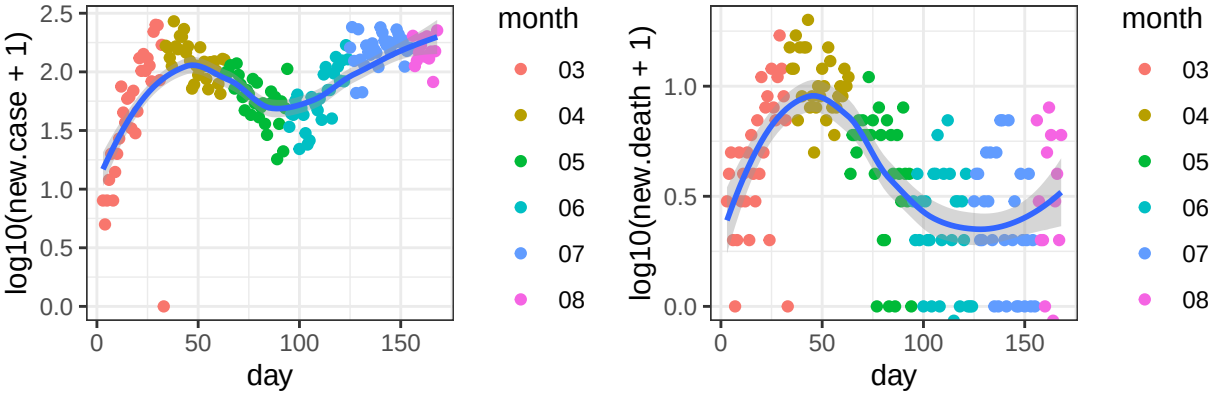
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Hampden_Massachusetts



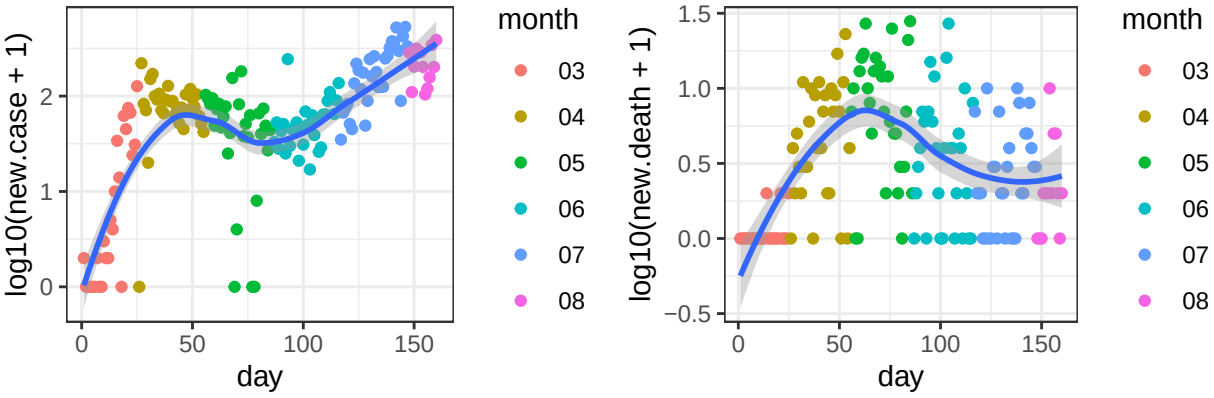
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington

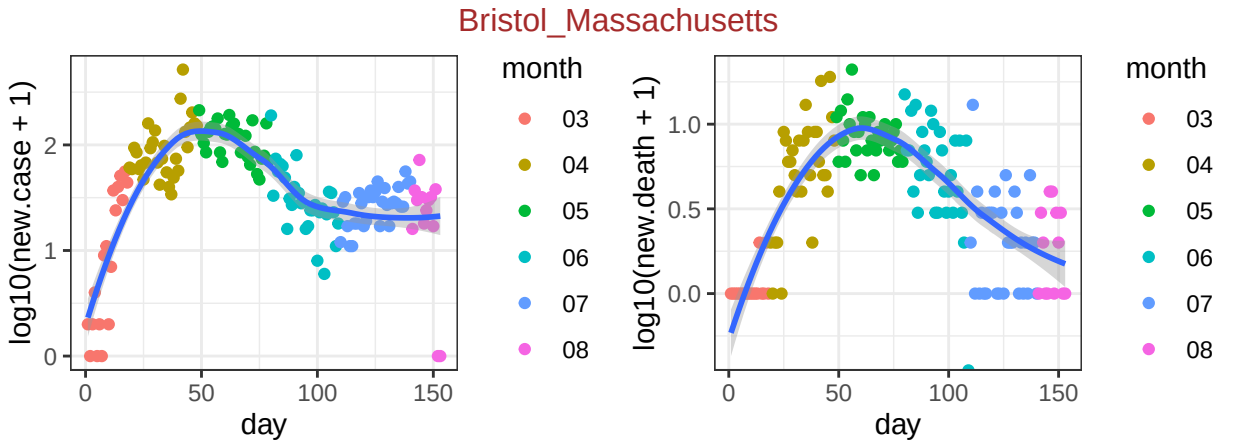


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

St. Louis_Missouri



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

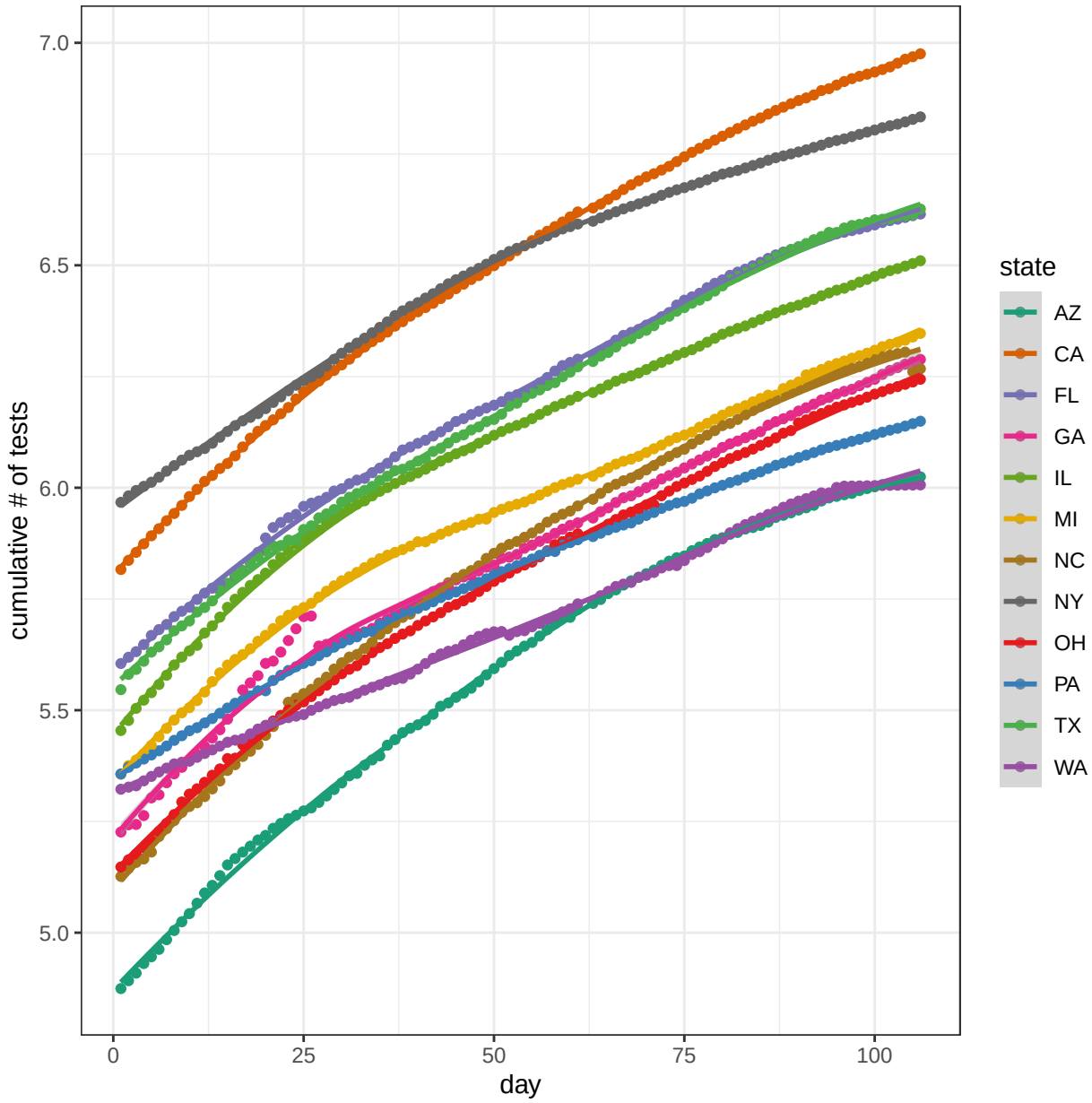


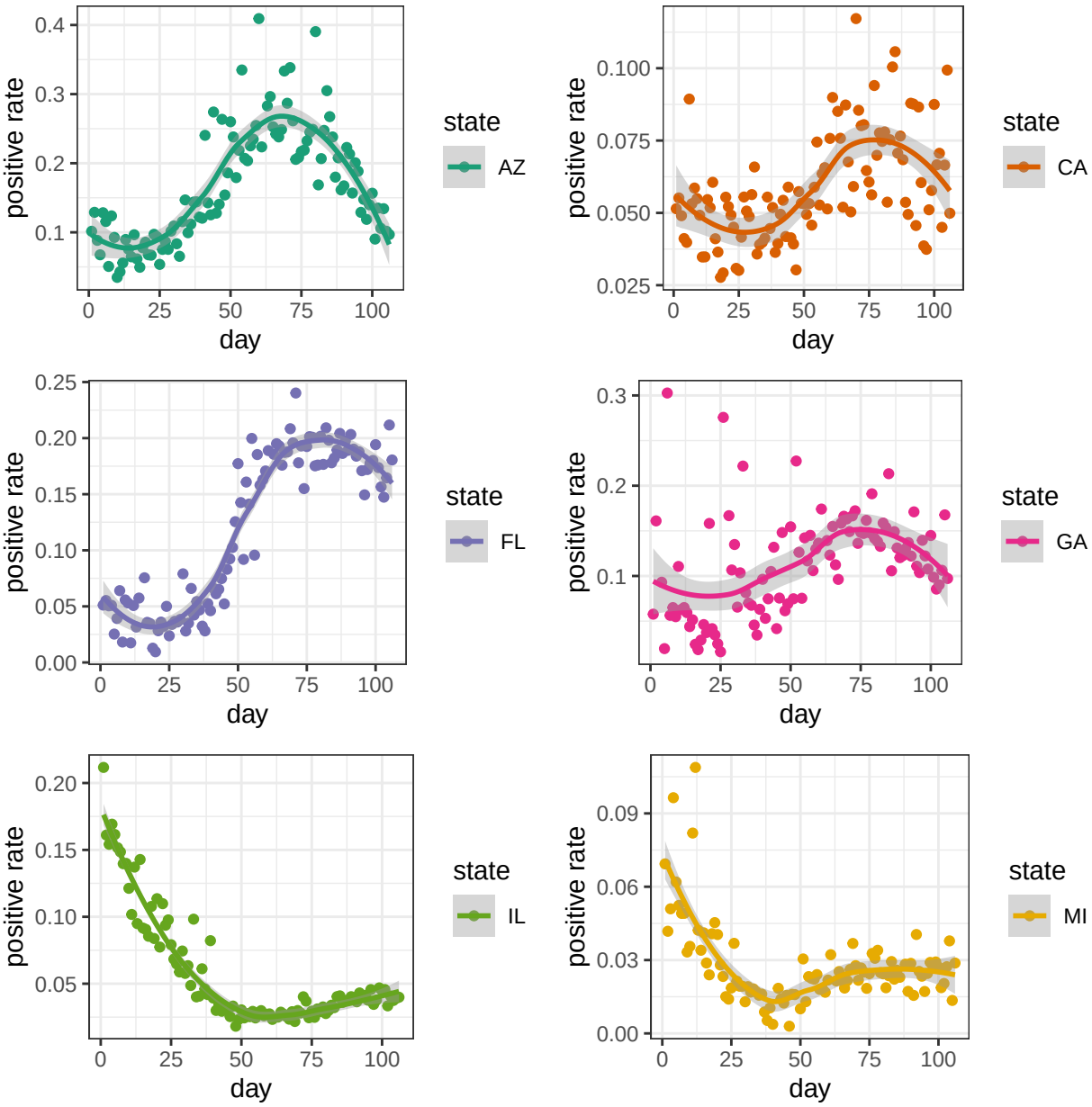
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

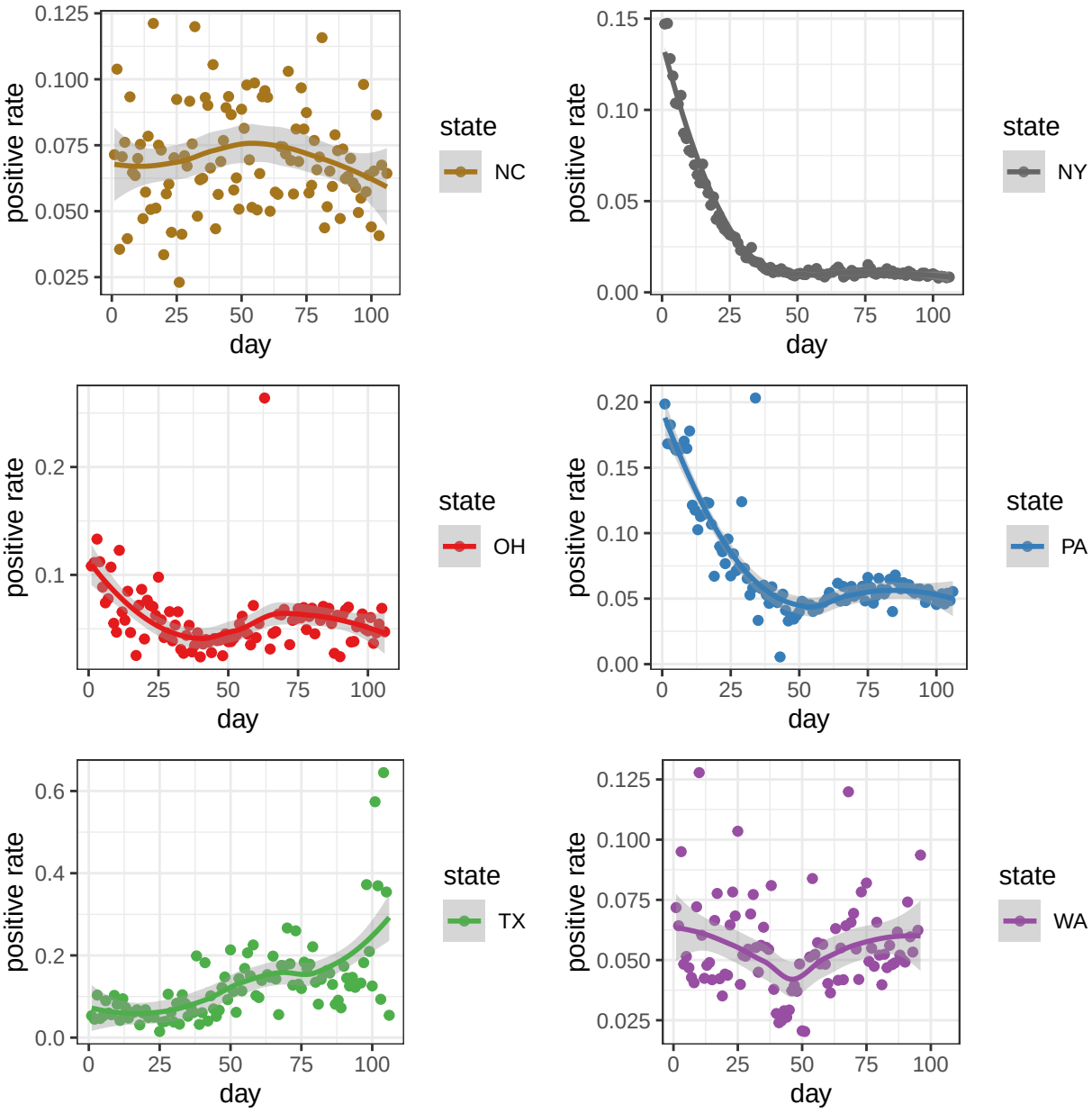
COVID Trackng

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
25
50
75
100
day
state
AZ
CA
FL
GA
IL
MI
NC
NY
OH
PA
TX
WA


positive rate
day
state
AZ
CA
FL
GA
IL
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
```