State Specific free and open epidemology analysis using R

Eric Olle

May 3, 2020

A free and open framework for the analysis of COVID-19 world wide pandemic

Background

This is the basic markdown document as part of a free and open epidemiology document. This is meant to be used free of charge and kept in the GPLv3 or equivalent to allow for ongoing analysis of the data set. This file and information may also be used for other local epidemics as needed.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

The orginal r markdown will be posted to:

https://github.com/Eric43/ID-free-reports

This is a working document and may change over time.

Required packages

The majority of the data clean up is done in the Tidyverse was created by Hadley Wickam and others. See:

https://www.tidyverse.org/

This work has affected basically every aspect of R have made R a more user friendly experience. Without this vision this report may not have been possible. If you can please support tidyverse and the work of R-studio to maintain these incredible resources for R-nerds everywhere. Hadley and all his collaborators have, in my opinion (ewo) made working in R an almost enjoyable experience. I really appreciate the Yowmans effort of turning an eclectic program based on S that I learned in the early 2000's to something that is truly remarkable for statistical and mathematical modeling.

Other packages use are lubridate, forecast, lmtest and tidycensus.

Loading the data from file

This section loads the infection data from the NY-times public data set. If needed, a csv file can be loaded using the readr read_csv() command. Make sure date is properly loaded in as a data/posix type.

```
## Parsed with column specification:
## cols(
##   date = col_date(format = ""),
##   county = col_character(),
```

```
##   state = col_character(),
##   fips = col_character(),
##   cases = col_double(),
##   deaths = col_double()
## )
```

Setting the constants that apply to the individual state

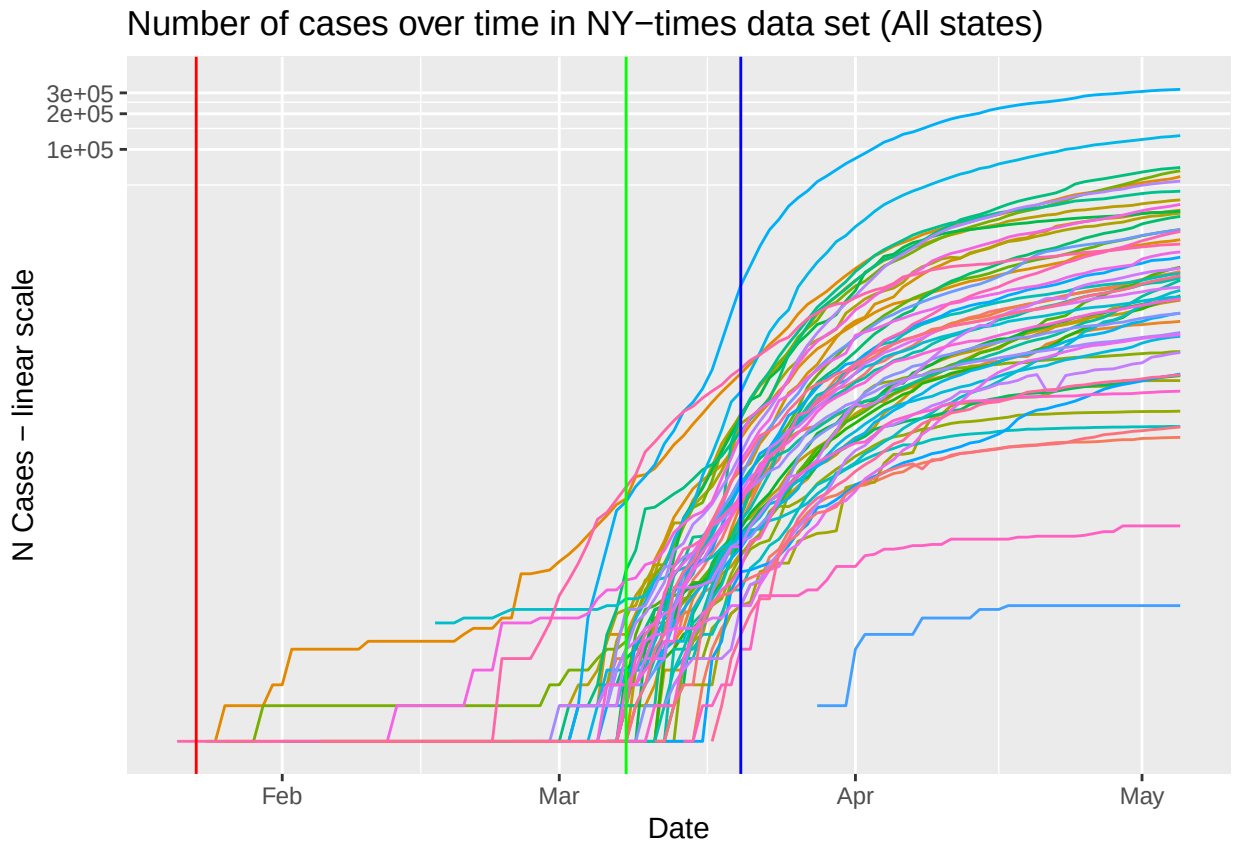
In this section to constants that will apply to the individual state(s) can be done.

```
## [1] "The state selected for analysis is: New York with lockdown of: 2020-03-20"
```

Part I. Graphing time series of infection

In this section a basic time series plot of the full US along with the selected state will be done.

A. Plotting the full state data set.



The above graph shows the different time series of infections across the USA. The red line indicates the first lock down in China of 2020-01-23. Included are the Italian lock down of 2020-03-08 in green and the New York lockdown of 2020-03-20 in blue. This should act as a way to help visualize and pinpoint different times. If additional times are needed use the geom_vline() command.

B. Selecting the state specific data from the NY Times dataset

To select a specific state use the unique state names call (above) set the state name by copy/paste or typing in with the quotation marks. The current state is set to: New York. The states lock down date is set to 2020-03-20. To change the state and/or the lock down date do so in the previous section called "Setting the constants." Finally, if the state has enough data before the lock down (i.e.5-7 days) and is still not in the lag phase of exponential growth feel free to set lock down effect analysis to "TRUE."

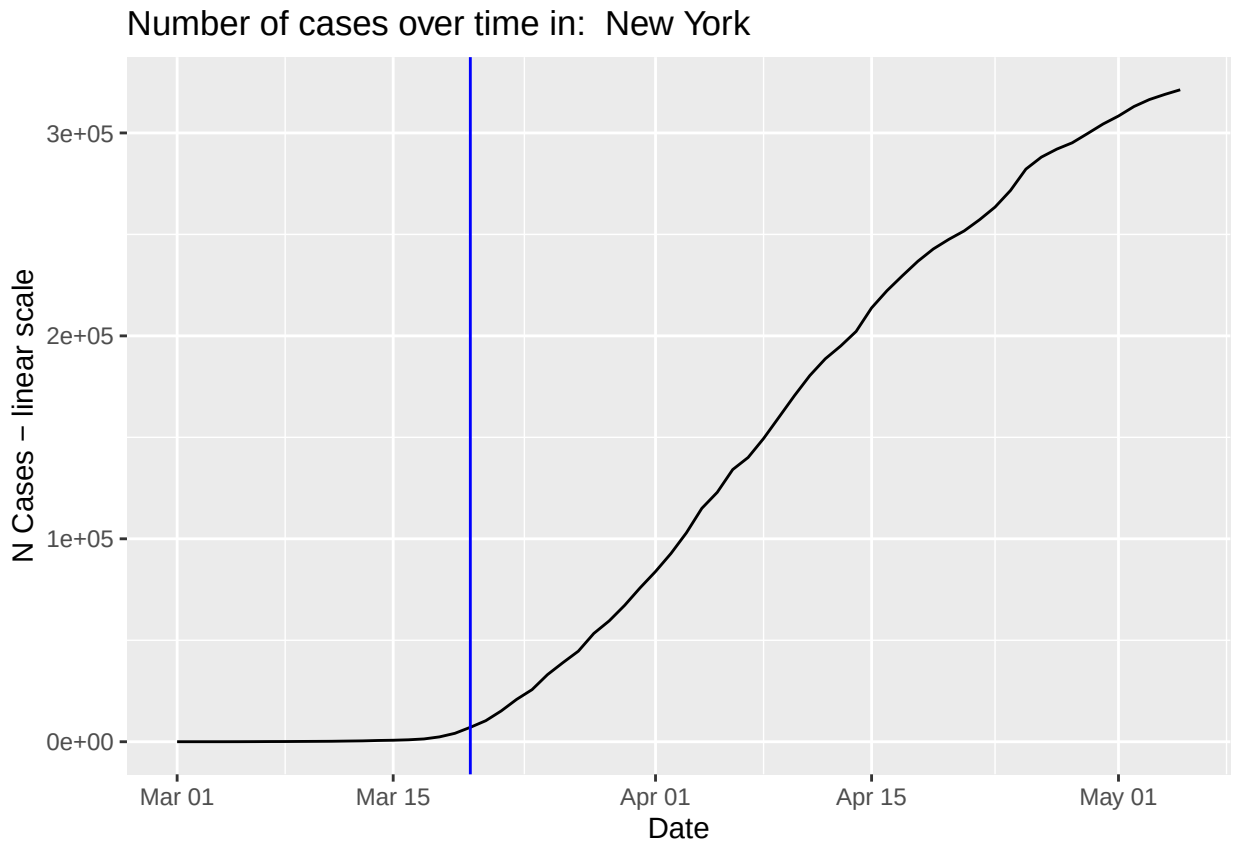
Selecting the state data from state2select var

The state data set was selected in case future geo-spatial on the rate of cases by county is needed.

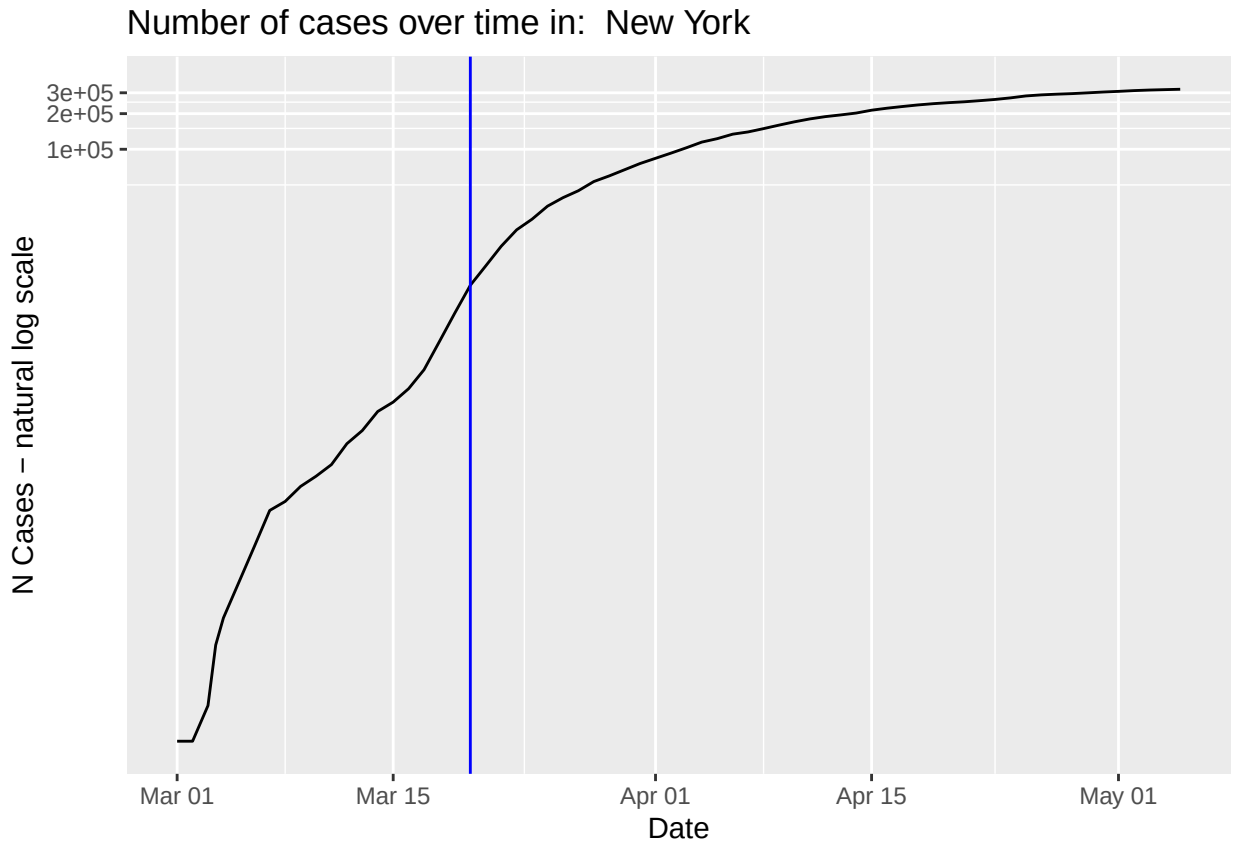
```
st <- us_counties %>%
  select(c(date, state, county, fips, cases, deaths)) %>%
  filter(state == state2select)

st_cases <- st %>% select(c(date, cases)) %>%
    group_by(date) %>%
    tally(cases)
```

C. Plotting the state specific data

Once the data is selected and grouped by state cases a general total cases per day model can be developed. Depending on the state this should not select the NA's and have different start times. Alternative methods are possible by converting to a wide format to maintain early NA data.



Depending on total case numbers(i.e.greater than 1000-5000) a log scale maybe easier to show trends and see recent trends. If under 5000 cases total this may over-represent trends (up or down) that are only part of the standard variance of the data.



Part II. Forecast models for cases load and daily differences

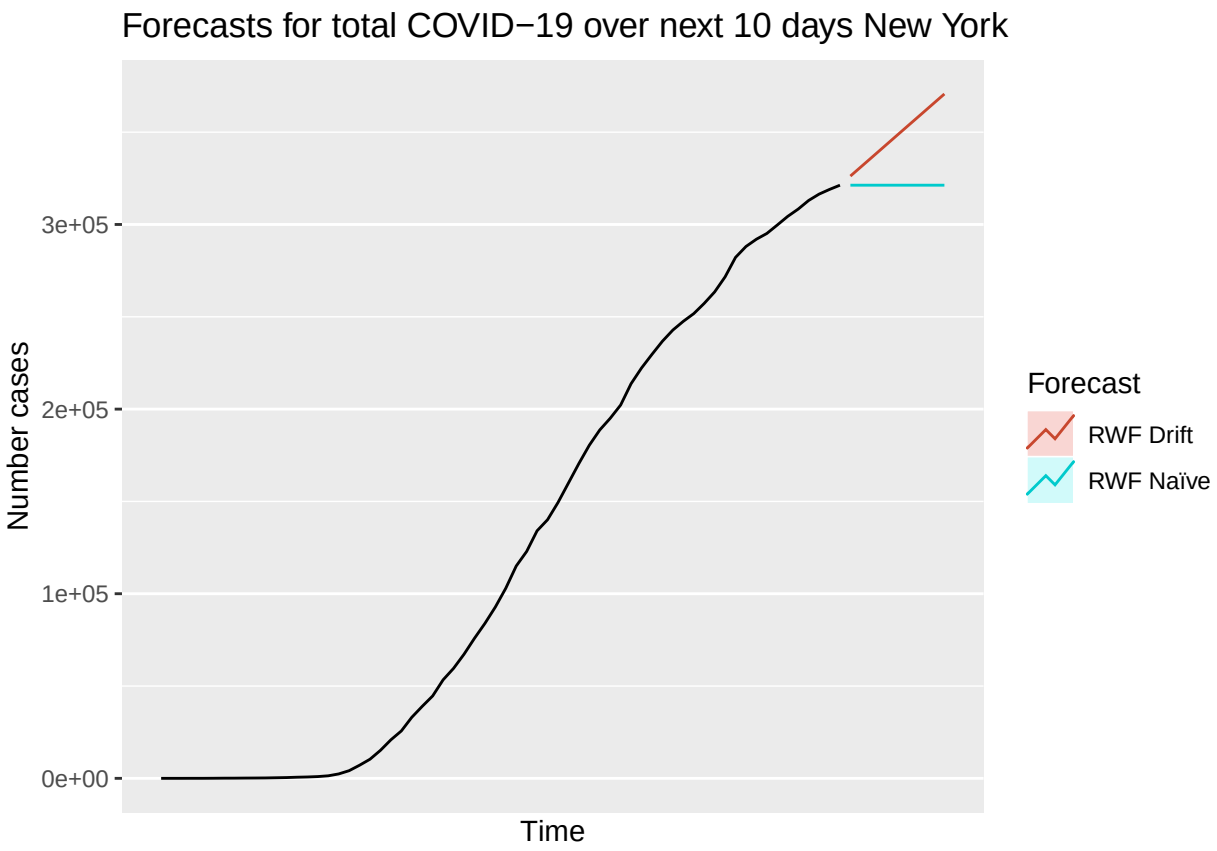
To do this you will need to convert the date data into a time-series using the Forecast and lubridate packages in R. This document is using a very basic forecasting model(s) such as: random walk, exponential time series or ARIMA. No seasonal corrections have been introduced due to the limited data set. at the time of writing. This section can provide a basic idea of where the cases may be in the next 7 to 10 days without major changes in the underlying model assumptions or variables.

The forecasting sections primarily use the forecast package by Rob J Hyndman and George Athanasopoulos it is a phenomenal package and is highly recommended that you support the continued development through purchasing of the book or maybe sending them a nice note at:

https://otexts.com/fpp2/buy-a-print-or-downloadable-version.html

The work done for this package is just amazing and between this work the seminal work on time series by Box and Jenkings (1970 & 2015) (See Hyndman et. al for current version of Box et al.) one can "see into the future." (humor intended).

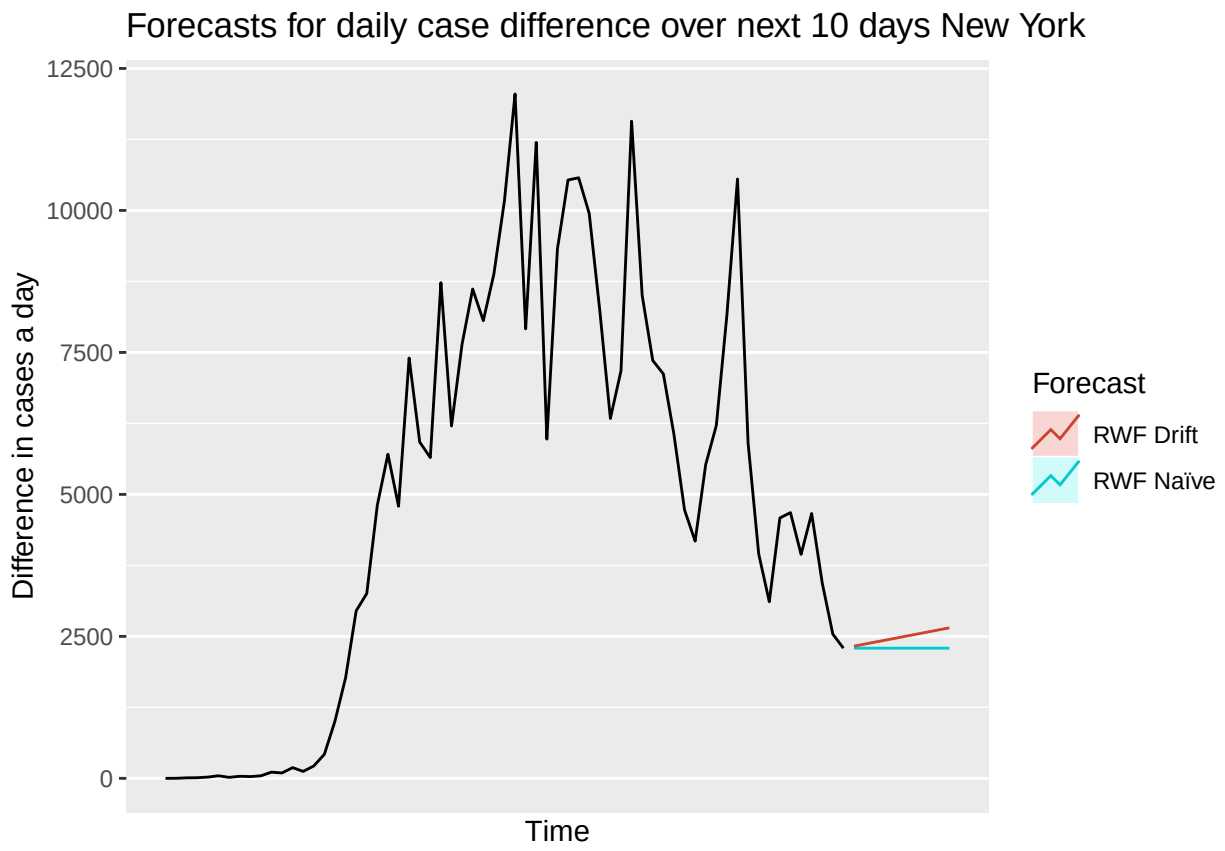
A Random walk method full data set



B. Forecasting with ETS (growth phase only)

In addition to random walk the forecast package can do exponential time series (ETS()). However there are different phases of a novel infectious agent that may not meet the underlying criteria and this method (as will the others) need to be used carefully. Make sure the system is in the appropriate growth phase and has not in lag, stationary or later phases. Additionally, it has been observed that the stay in place may affect this model and provide an unrealistic mathematical model. The exponential time series is set to FALSE. If FALSE nothing may be seen below.

C. Forecasting using a daily difference model.

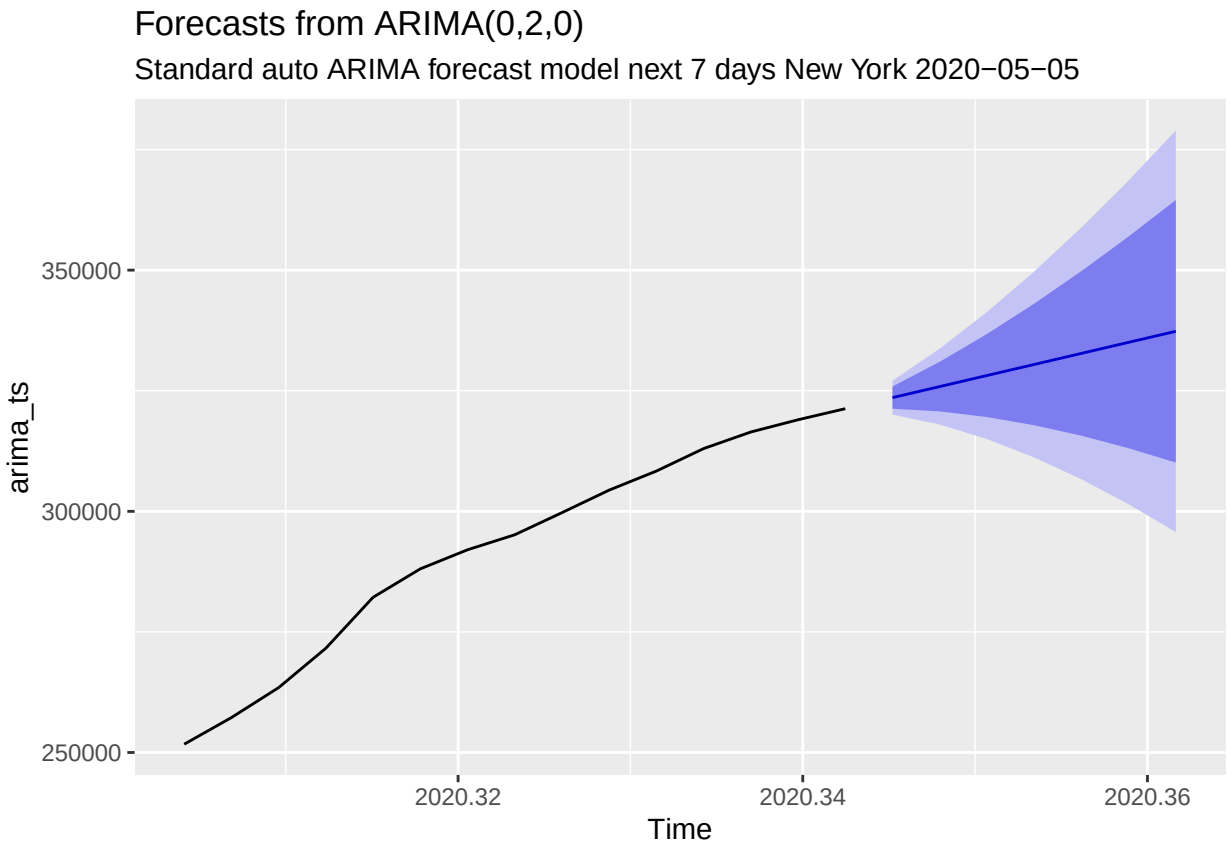


D. ARIMA modeling from two weeks

Another way to forecast the number of cases is to use ARIMA modeling (See the Forecast package citation) or:

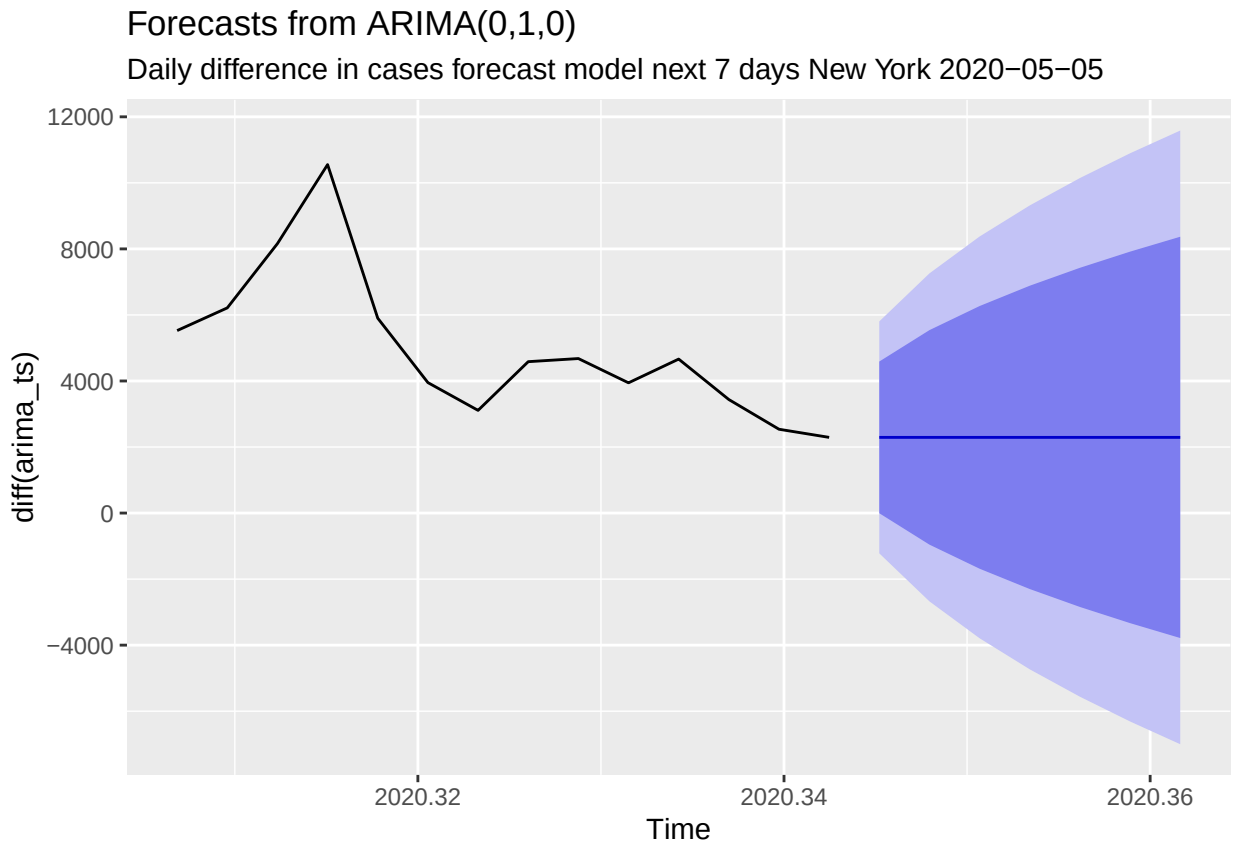
https://otexts.com/fpp2/

This and the random walk with and without drift are meant to be used together to attempt to estimate the number of cases from the previous 15 days using the auto.arima() function. There is no one "correct" answer when forecasting the future number of infections but is meant to show overall trends and help predict public health risks. The first graph is showing the log of the cases as part of a time series.



The above figure (ARIMA modeling of previous 15 days) is meant tho show an overall trend. This used 15 day window to allow for a rapid conversion to a 14 day difference model (below). It is expected that the model will show drift until full stationary phase is reached or the infection is nearly completion. This is meant to be used to predict the number of reported case over the next week.

E. ARIMA modeling difference between daily cases



The above graph is showing the last 14 days of a difference between cases/day. Depending on the stage of the infection it may show drift up or down. If during stationary phase of a public health epidemic it is expected to be a standard ARIMA model with a non-zero mean. Drift can indicate that the difference in the cases/day is changing. Therefore, it is necessary to look at the auto.arimia fit criteria for the best model indicated in the title.

Above models of a time series utilize different forecasting methods such as random walk or ARIMA modeling. These can be very useful but require a nuance approach along with multiple models to aid the forecaster (See the Hyndman et al (2020), Hyndman and Khandakar (2008)). In general these are asking the question based upon the current time series (or training data) what is the forecast number in the future. This is excellent for providing information but may lack the ability to (easily) compare the effect of of different public health measures. The next section is designed to use simple linear modeling across two time windows to see what the effect of different public health measures.

Part III. Using linear modeling in daily difference to determine the trends

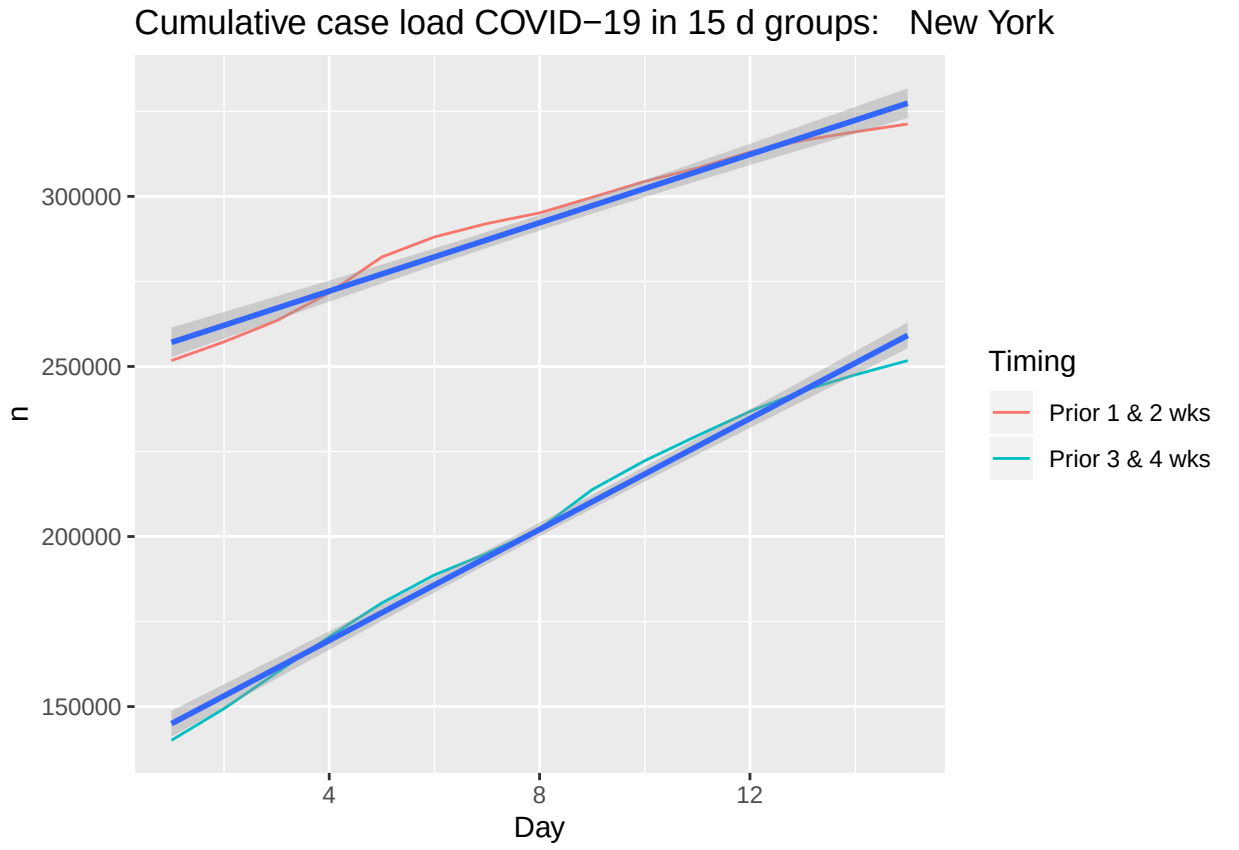
In this part it will be broken down into three sections. First section will look at the standard model for the 10-days before and after a lock-down/stay in place order. Then a standard last 10-14 days compared to previous 10-14 days and both normalized to days 1 through 10 (or 14). Second part is the start of the difference model by looking at the previous 10-14 days and comparing to the time before that. This is appropriate in states that had not transitioned from lag-phase to exponential growth phase (i.e. West Virginia). This should show a basic day-over-day trend. Third part is looking at the effect of the stay in place/lock down order to determine if it had a measurable effect. NOTE: in states with minimal/no time in exponential growth this may not be an accurate measure and recommend that Part 2, section B be used (ie. comparing two different time frames).

NOTE: This is still being worked on and needs to have the stay in place order model done.

A. Comparing the last two weeks to the previous.

The goal of this section is to align certain time frame such as 15 days in standard model or 14 days in difference model to allow for direct linear modeling and comparison. While linear modeling may have its limitations and will not be significant if the variance between the days is later, it will proved a general trend. This trend of the number of cases/day of the difference of cases/day can then be compared to see the effect of different states, stay in place or the return to work timings. The model is simple. If there is a difference between the two time frame windows the slope of the line will also be different. The line and it appearance can quickly show that a negative slop will show decreasing cases where as a horizontal line is basically stasis and a positive slope is an increase in cases. There will be limitation to using this and overlapping confidence intervals

After setting up the two different time frames in a tibble (i.e. 14 days for difference series is 15 days in standard series), the two different data sets are plotted. The first to be plotted is the standard 15 day data series with a linear model trend line from geom_smooth() command in ggplot2 can be seen below. After is the data set in the last 15 days and the before is the 15 days prior to the After data set. This was meant to coincide with the naming of the lock down (before and after the lock down) data analysis (below).

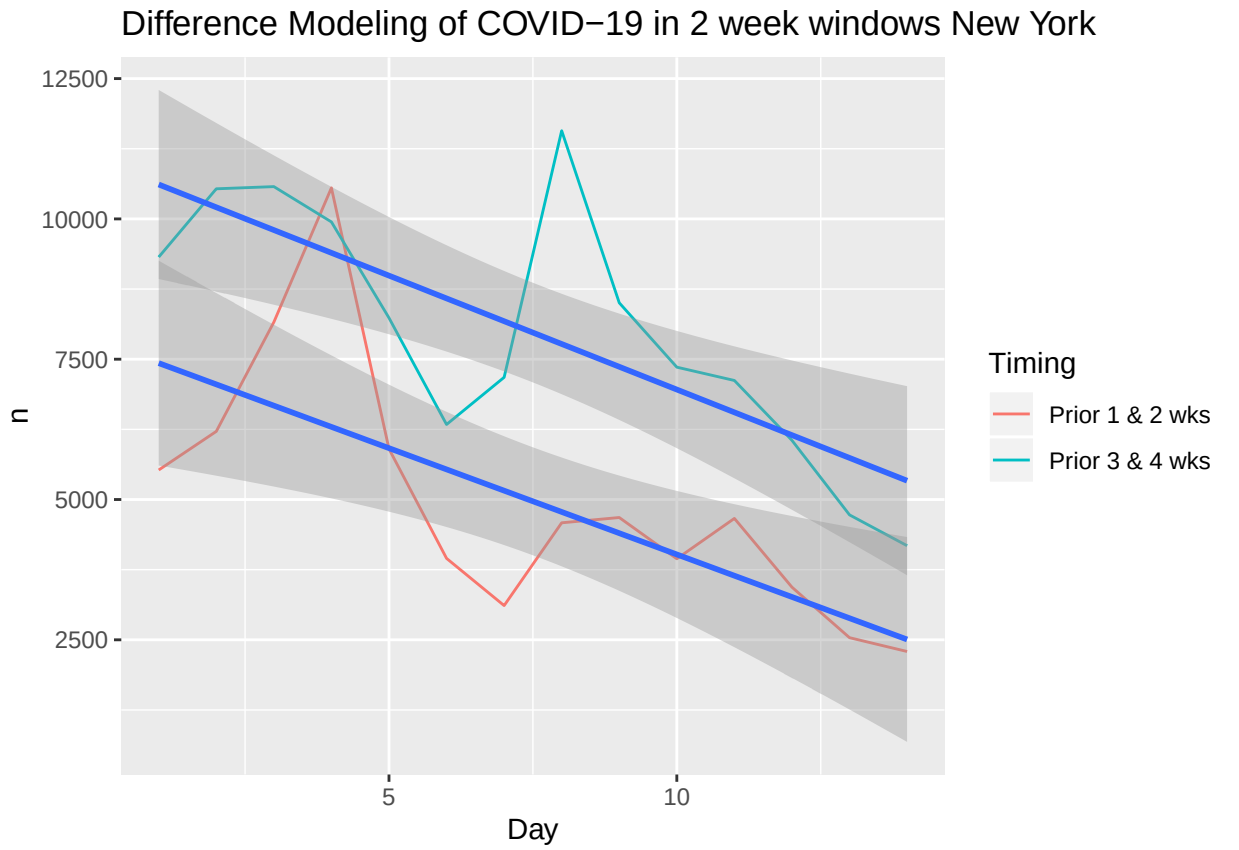


The above graph is show two different windows of time and the linear model line associated with the data. This is a standard total case model and is using 15 day windows so the data can be put through a difference model (that reduces by one day) and is comparable. The group "Prior 1 & 2 wks" is the last two weeks from the end of the NY-time data set. to 15 days ago. The group "prior 3 & 4 wks" looks at three to four weeks ago and compares. The slope of line can indicate how case numbers are changing. For example, if the last two weeks has a steep slope than three to four weeks ago it indicates that there are rate of cases had increased.

The data is showing the prior 3 to 4 weeks group from Starting date, 2020-04-07 to 2020-04-21. The prior 1 and 2 weeks group includes from Starting date, 2020-04-21 to 2020-05-05. Looking at the data set. it

is usually difficult if not impossible to see a change in the slope if the infectious disease is in the lag, exponential growth or stationary phase(es). During the death phase or transition from one phase to the next the slopes maybe different In addition, Yule-Simpson effect of big data maybe involved and the slope of the line may be unrelated or trend differently than the actual sub-grouped data. What most public health and hospital planners are interested in is a straightforward question. Does the number of new cases per day differ (i.e. lower, greater or stay the same)? Therefore, a difference model was designed to look at to determine if the difference between the day over day case numbers are changing.

B. Linear modeling difference between daily cases



The above graph "Difference modeling of COVID-19" shows the daily difference in number of cases along with a smoothed trend line and confidence interval using linear modeling. This graph provides quick comparison between the last two weeks and the two week before that (weeks 3 to 4 prior to the end of the data). Looking at the trend lines, it can be determined if more cases per day in the last two weeks are occurring when compared to the prior time frame. Additionally, the confidence interval along with the raw data lines can show how much variation occurs along with overlapping data can indicate that the daily difference is remaining basically the same. The next section analyzes the linear model to determine significance and calculated out the difference between the slopes of the line to show if daily case difference is increasing, decreasing or remaining constant.

Checking the Linear Model of the last two weeks and the last 3 to 4 weeks to compare how the number of cases differs per day. First part is to look at the LM for the previous 3 and 4 weeks. Slope of difference in cases per day is located under the Day row, Estimate column.

```
##
## Call:
```

```
## lm(formula = n ~ Day, data = diff_tib %>% filter(Timing == "Prior 3 & 4 wks"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2246.2 -1011.5   117.8   563.5  3799.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11019.1      860.7  12.802 2.34e-08 ***
## Day           -406.0      101.1  -4.016  0.00171 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1525 on 12 degrees of freedom
## Multiple R-squared:  0.5734, Adjusted R-squared:  0.5379
## F-statistic: 16.13 on 1 and 12 DF,  p-value: 0.001711
```

Second part is to look at the LM for the last two weeks. Slope of difference in cases per day is located under the Day row, Estimate column.

```
##
## Call:
## lm(formula = n ~ Day, data = diff_tib %>% filter(Timing == "Prior 1 & 2 wks"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2047.7  -714.9  -135.6   252.7  4259.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7809.2      932.6   8.374 2.35e-06 ***
## Day           -378.8      109.5  -3.458  0.00473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1652 on 12 degrees of freedom
## Multiple R-squared:  0.4992, Adjusted R-squared:  0.4574
## F-statistic: 11.96 on 1 and 12 DF,  p-value: 0.004731
```

Comparing the slopes of the line can give you an idea of how the last 14 days compare to the previous.

B.1 Calculating the differnce between the before and after groups using the difference model

The goal of this section is to use the coefficients of the linear model (i.e. differences in cases/day) to see the number of cases and if the last two week the case load is decreasing (negative number), increasing (positive number) or remaining the same (around 0 +/- number). The difference between the last two weeks and the previous two weeks is: The difference in the slope (# cases/day) is: The difference in the slope (# cases/day) is: 27.21 case-difference/day.

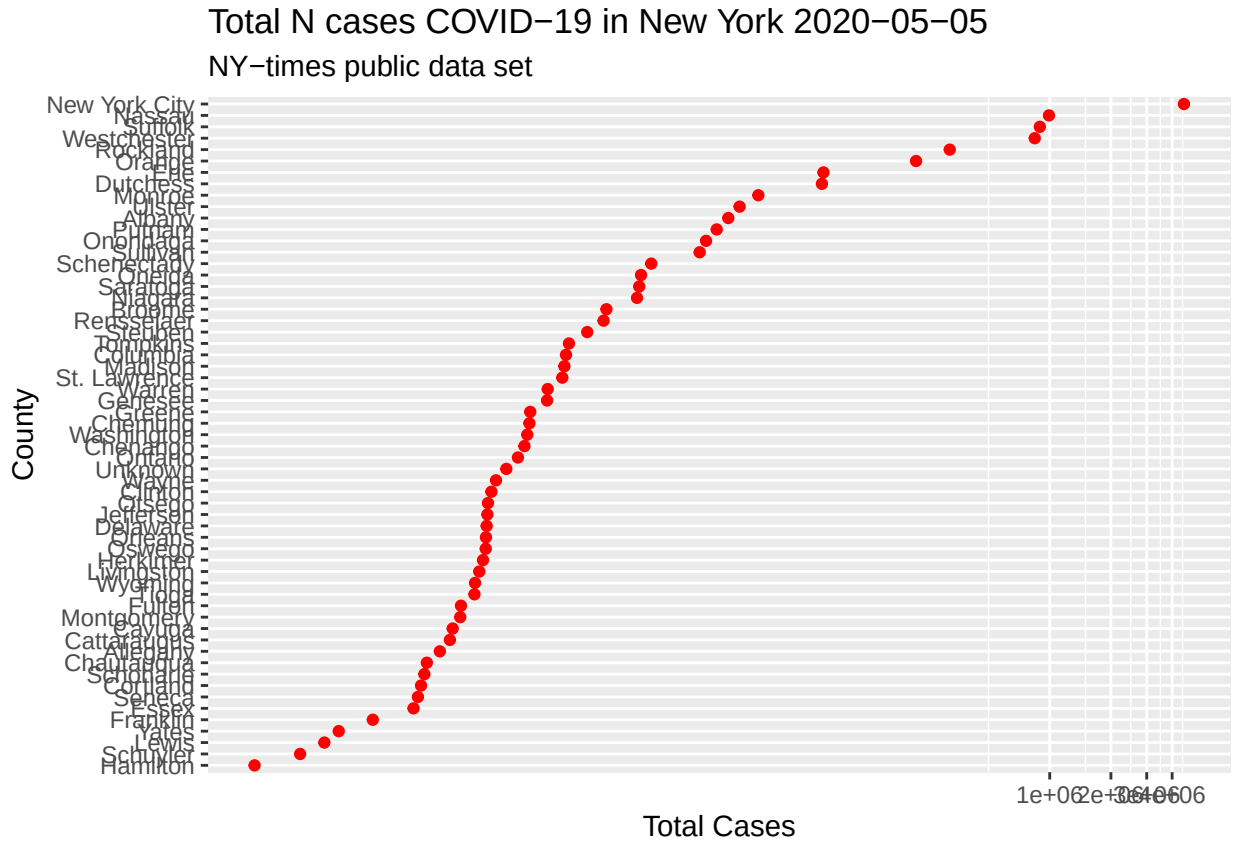
C Linear modeling to look at before and after the stay in place order

The goal of this section is to look at the effect of lock down or later on in the public health disease removal of the stay in place orders.

Lock down analysis set to FALSE.

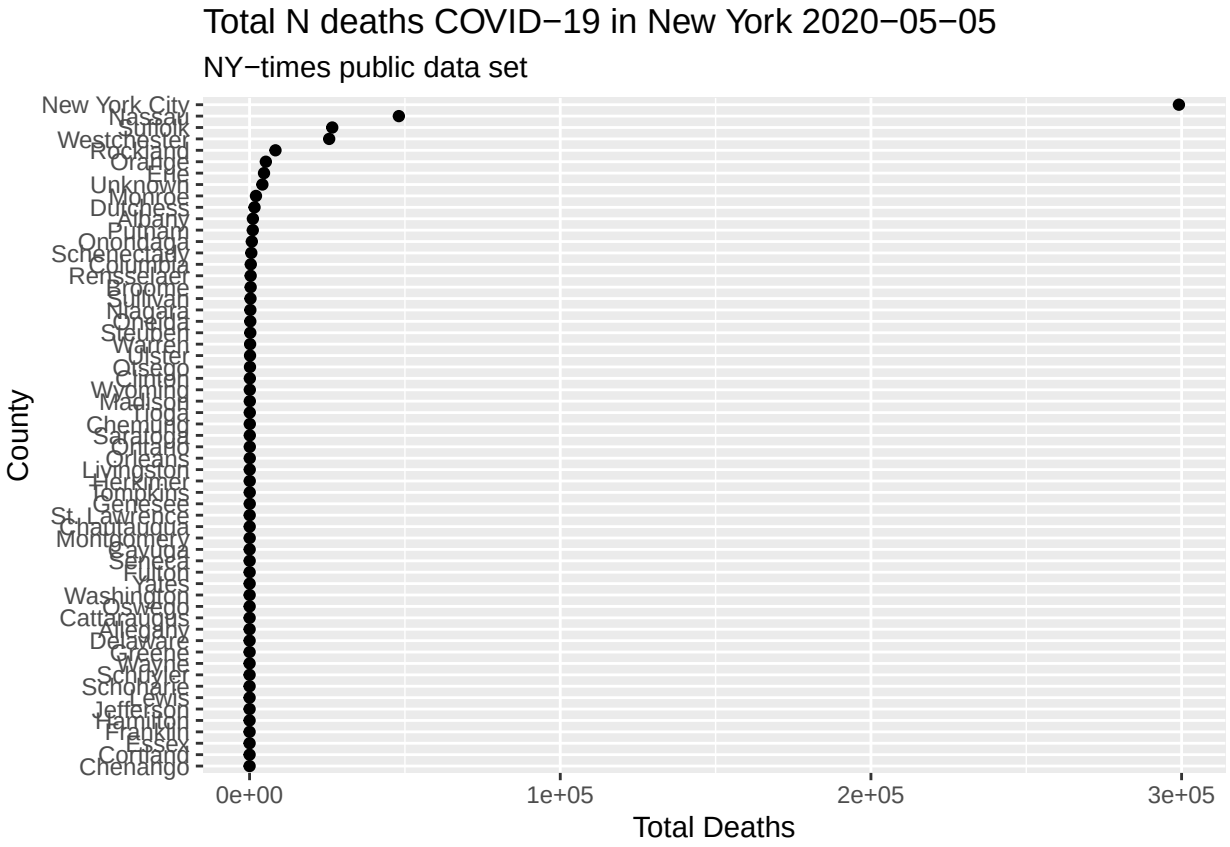
Part IV. Geospatial distrubution of cases

A. Plotting bar graph of cases by region



The above graph shows total number of cases by county. The next step is to show total mortality by county.

B. Plotting bar graph of deaths by region



C. Graphical representation of the location of cases

Collecting the population and spatial geometry data from US census.

This is a quick call to the US census using the R package tidycensus. If you do not have a census API-key please file for one on the US Census site and follow the information on registering the key using the tidycensus package manual or at:

https://walkerke.github.io/tidycensus/articles/basic-usage.html

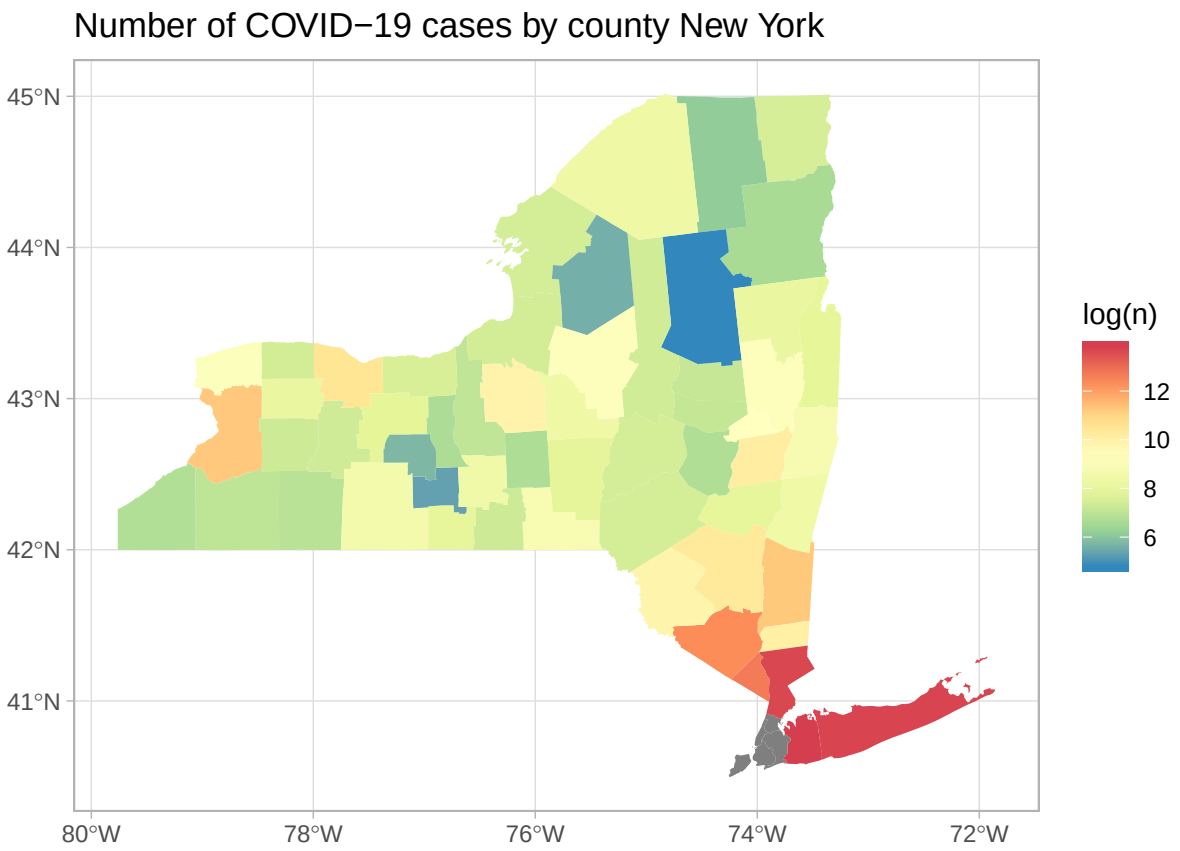
The tidycensus package is an excellent package and please let Kyle Walker know that you appreciate the work at:

https://walkerke.github.io/

```
## Getting data from the 2014-2018 5-year ACS
```

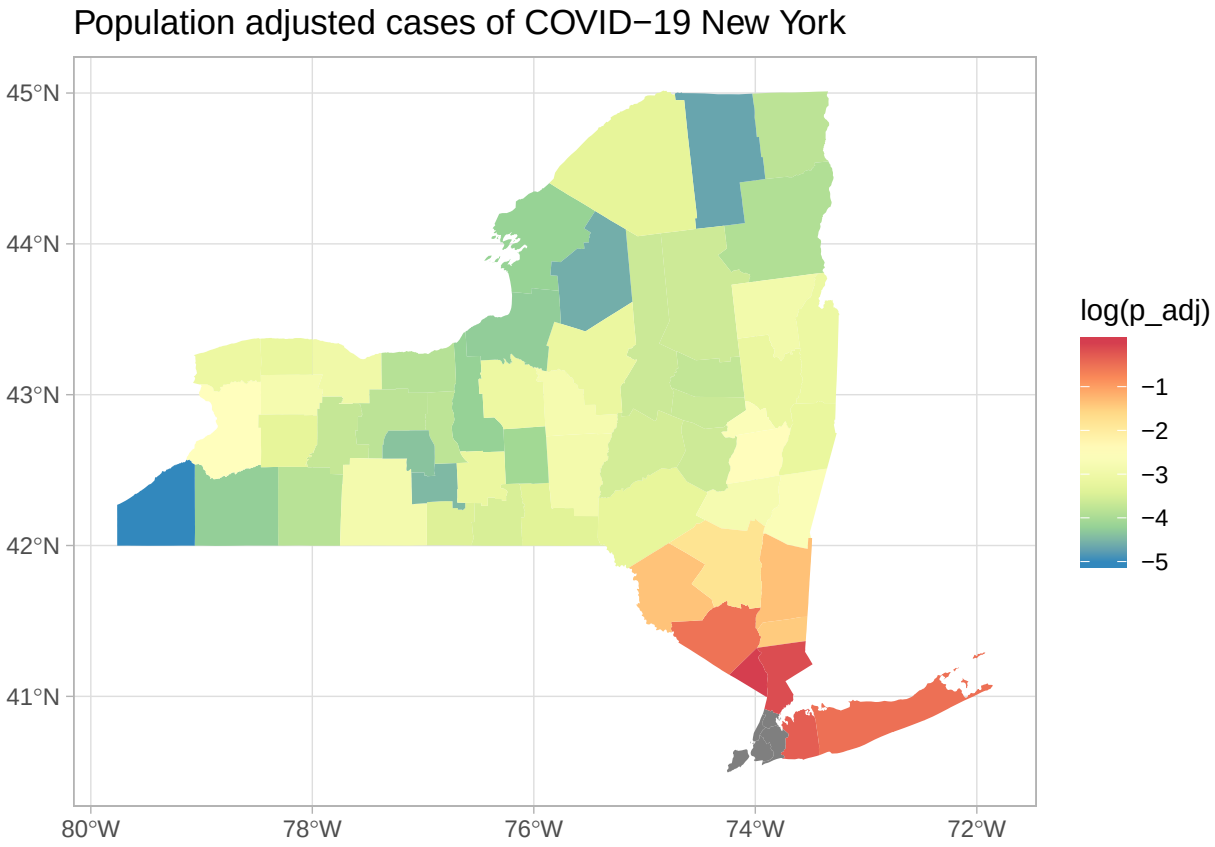
Geospatial of number of cases raw data

The goal of this plot is to show where the cases that are driving the infectious disease pan/epidemic are occurring. This is limited to the quality of the data received and due to the varied nature of different counties' public health may not be accurate. However, this should provide a population density basis for the disease. This plot should be looked at with caution because it will skew to larger urban/higher population densities and should be combined with a population adjusted geo-spatial overlay (see below) to provided a more complete picture of needs.



Geospatial overlay of cases adjusted for population

The goal of this plot it to show how the relative ratio of infections differs when adjusting for population. It is expected that the majority of the regions analyzed will remain similar, however, areas that have a higher (or lower) number of cases on a population basis should change relative color. A change higher (i.e. an average region becoming a "hot spot") could indicate a under served population or lack of hospital access. On the other hand, a region that goes from middle or higher on the spectral (log adjusted scale in this case) scale to lower could indicate access to better than average medical care. This was seen in Kanawah County, WV when unadjusted for population is a "hot spot" or red on the scale but when adjusted for population is a slightly lower risk. On the other hand the county next to Kanawah showed the opposite indicating that on a population normalized basis the public health risk was higher. One area this may prove potentially useful is in states with large urban centers (i.e. Los Angeles, CA, New York, NY or Chicago, IL to name a few). Larger urban centers may looks like the main drivers/risk populations but surrounding suburban and other outline counties may have similar risk that is not seen on the cases number plot alone.



Part V. Basic prediction of the total cases until reduction to standard health risk

In addition to number of cases and deaths there are some basic predictions that can be inferred from previous pandemics. While it seems to be fashionable to compare to the 1918 flu pandemic this is a useful number to allow a knowledge of approximate number (or percent) of cases required to decrease the infection rates below a public health risk. In 1918, the world population was around 1.8 billion and around 500 million were infected this is a 27.78% infection rate before lowering public health risk to "just the standard flu." To see how New York is on the public health risk when compared to the 1918 flu pandemic some basic calculations can be done.

Using the background of the number of population infected during the 1918 flu pandemic certain estimates can be made. This means that when the population infected is roughly equal to the the percentage of those infected with the flu the public health risks is low enough to begin reducing the need for public masks/sanitzation etc. This would mean that for this state to achieve similar percent infected as the 1918 flu around 5449570 are needed. Currently, there are approximately 8278006. this is around 42.19%. Currently the total mortality in the state is 432300 and this accounts for 5.22% of the total verified cases.

Part VI. References

A. R - The program

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

B. R - Packages

Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2020). *forecast: Forecasting functions for time series and linear models*. R package version 8.11, <URL: http://pkg.robjhyndman.com/forecast>.

Hyndman RJ, Khandakar Y (2008). "Automatic time series forecasting: the forecast package for R." *Journal of Statistical Software*, *26*(3), 1-22. <URL: http://www.jstatsoft.org/article/view/v027i03>.

Kyle Walker (2019). tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R package version 0.9.2. https://CRAN.R-project.org/package=tidycensus

Part VII. Authors and contributors

A. Authors.

Eric W. Olle. Author and creator of this document

B. Contributors.