

State Specific free and open epidemiology analysis using R

Eric Olle

May 4, 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>.

This is a rough draft and has not been checked for spelling or grammar!

This is a working document that is part of an ongoing R markdown.

The original r markdown will be posted to:

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

Need to do the references at the end.

Required packages

The majority of the data clean up is done in the Tidyverse created by Hadley Wickam and others. Hadley and his work on ggplot along with 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.

Loading the data from file

This section loads the infection data from the NY-times public dataset. 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.

```
us_counties <- read_csv("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv")
```

```
## 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 the constants that will apply to the individual state(s) can be done.

```
# Which specific state?

state2select <- "West Virginia"

# What is the state specific lockdown?
# (follow the correct date format)

### Need a look up table to autopopulate lockdown
lockdown_st <- as.Date("2020-03-20")

lockdown_effect <- FALSE

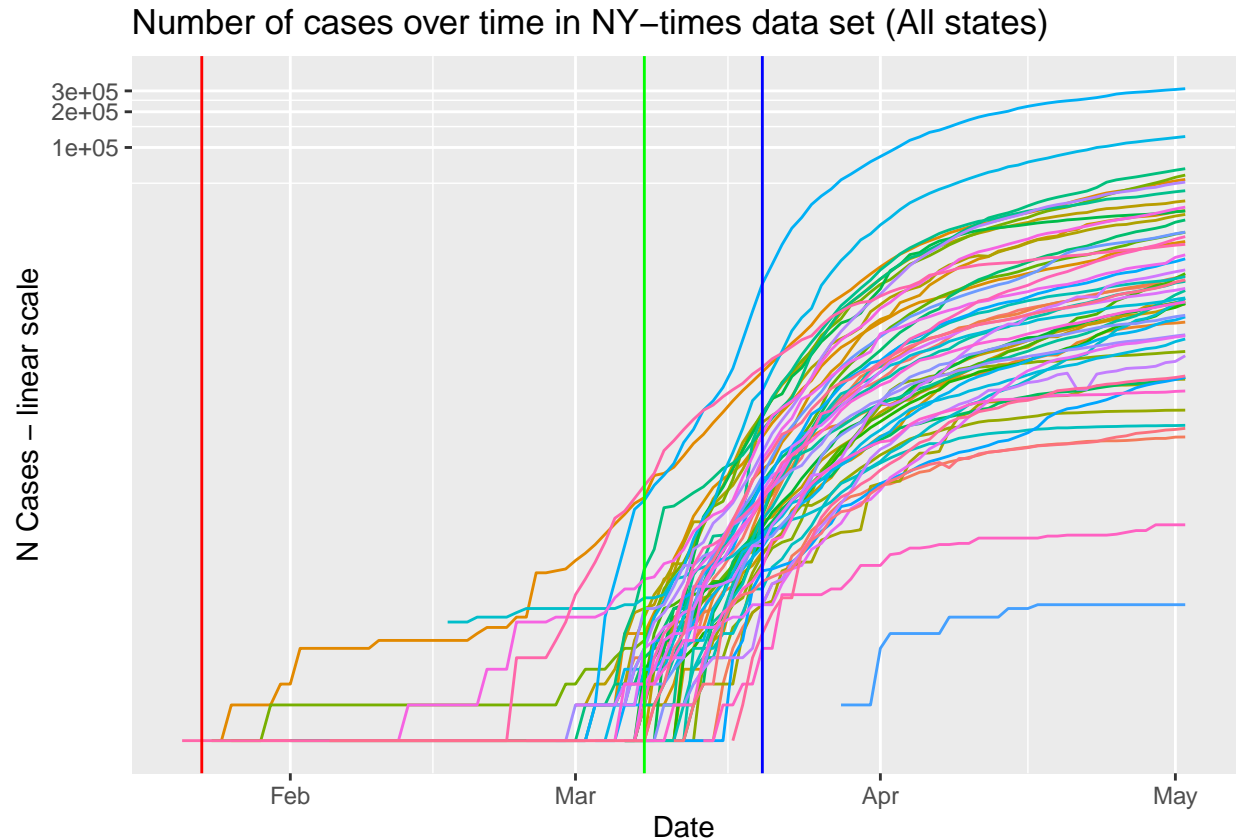
# lockdown data

## Peoples Republic of China
lockdownprc <- as.Date("2020-01-23")
## Italy
lockdownitl <- as.Date("2020-03-08")
## USA - New York
lockdownnyc <- as.Date("2020-03-20")
```

Part 1. Time series analysis

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

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 lockdown in China of 2020-01-23. Included are the Italian lockdown 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.

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: West Virginia. The states lockdown date is set to 'r lockdown_st'. To change the state and/or the lockdown date do so in the previous section called "Setting the constants." Finally, if the state has enough data before the lockdown (i.e. 5-7 days) and is still not in the lag phase of exponential growth feel free to set lockdown effect analysis to "TRUE."

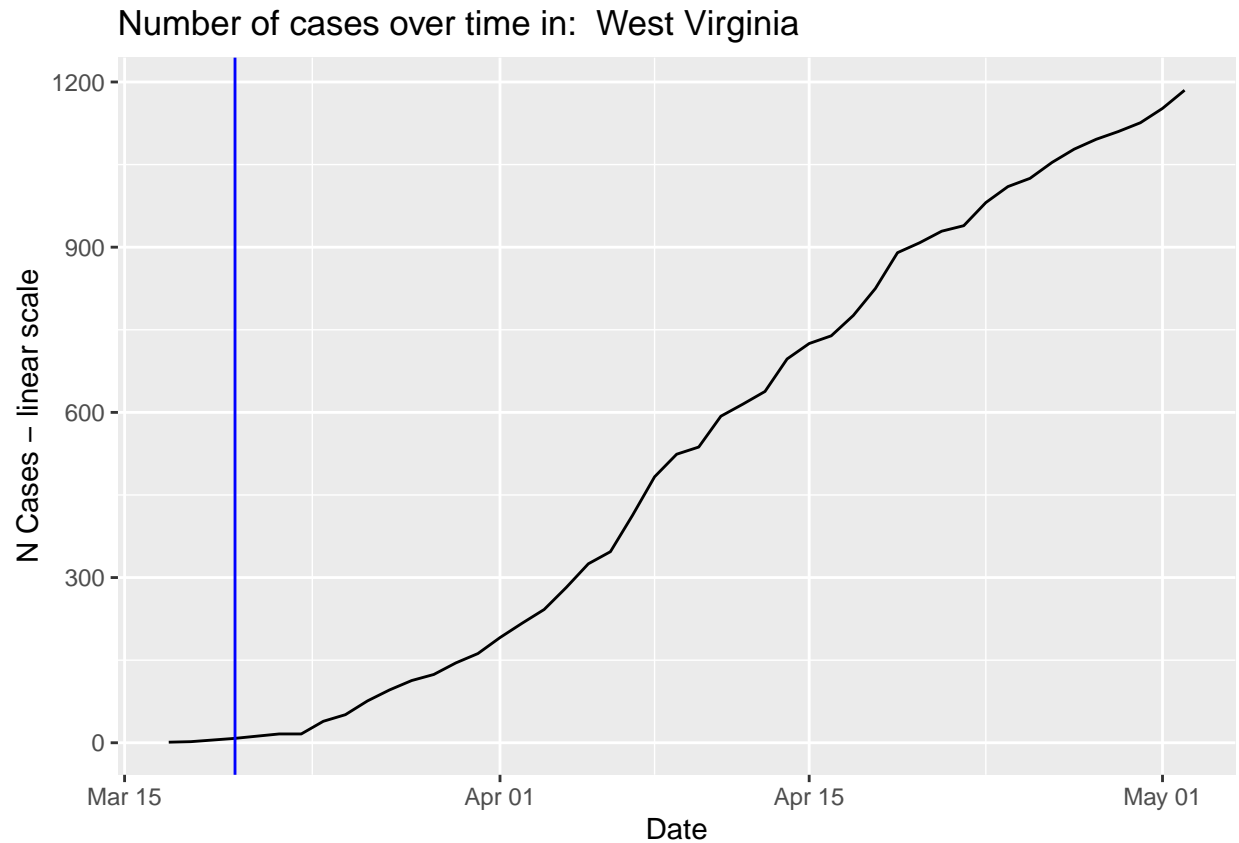
Selecting the state data from state2select var

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

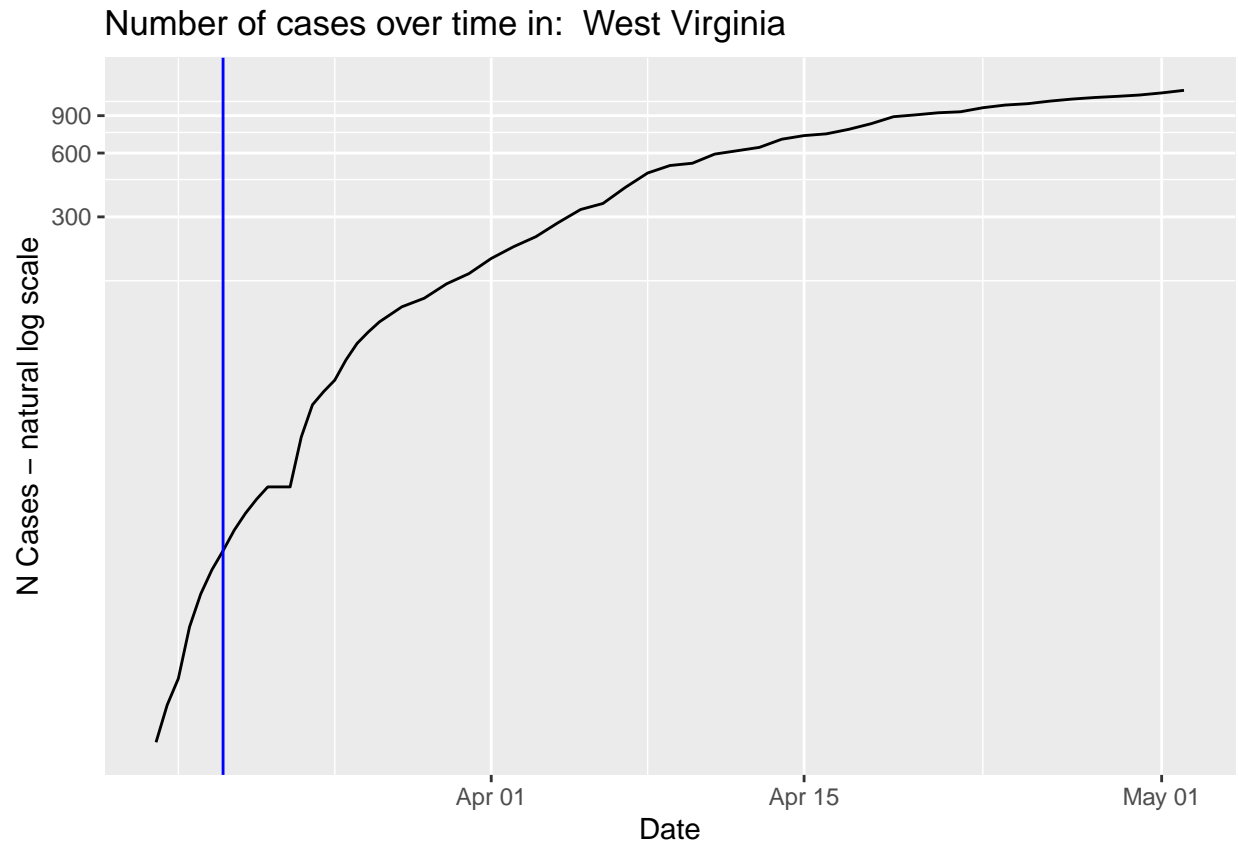
```
st <- us_counties %>%  
  select(c(date, state, county, fips, cases, deaths)) %>%  
  filter(state == state2select)  
  
st_cases <- st %>% select(c(date, cases)) %>%  
  group_by(date) %>%  
  tally(cases)
```

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.



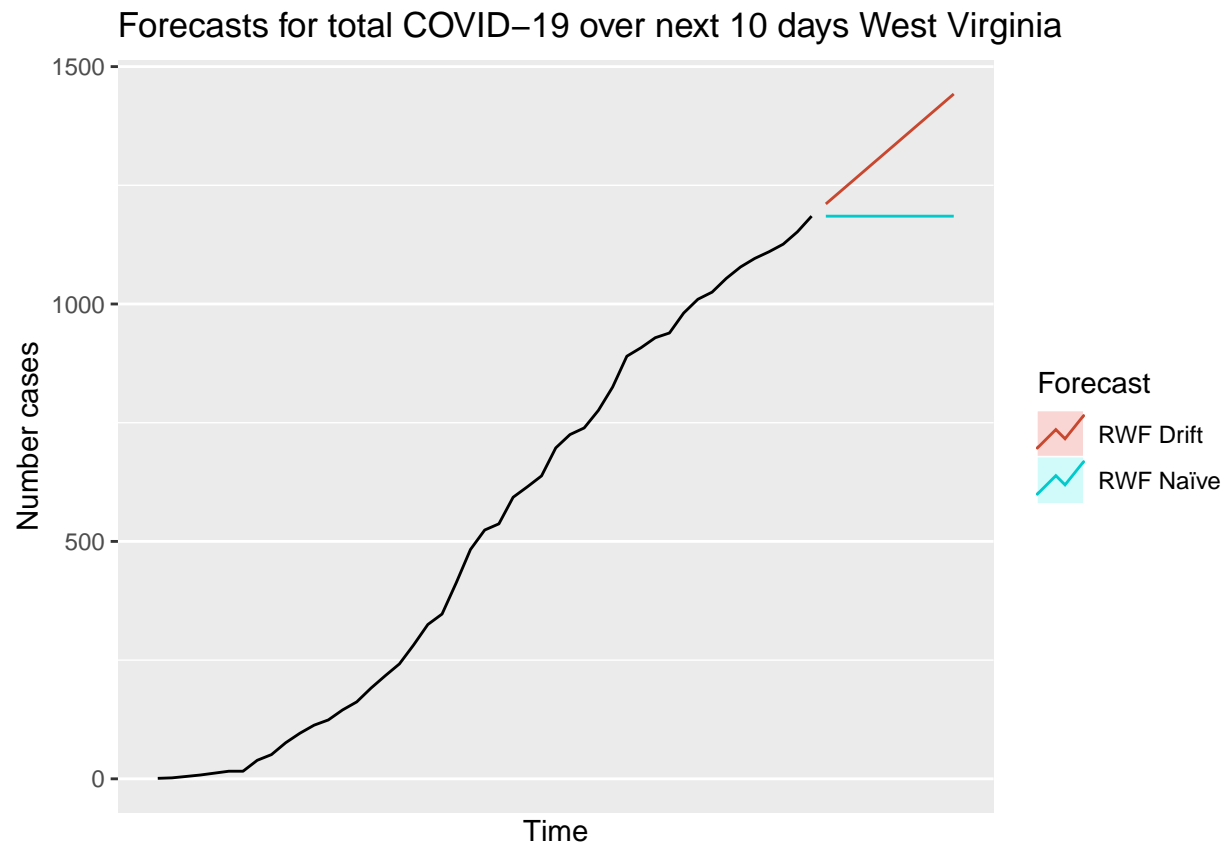
Basic forecast model for the next 10 days.

To do this you will need to convert the date data into a time-series using the Forecast and lubridate packages in R. This is a very basic forecasting model and just uses random walk with or without drift. Depending on the phase of the infectious disease this may need to be done using ARIMA modeling (later sections). However, this can provide a basic idea of where the cases may be in 10 days without major changes in the outlying 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:

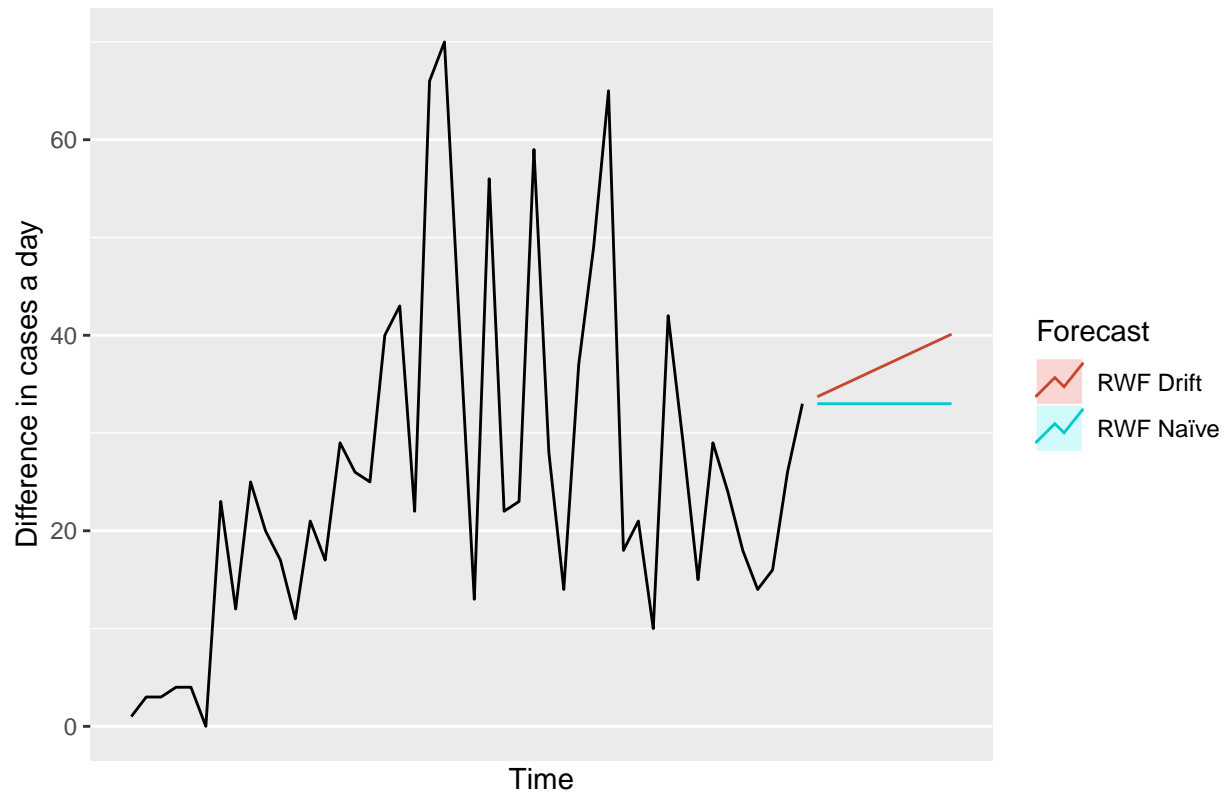
<https://otexts.com/fpp2/buy-a-print-or-downloadable-version.html>

The work done for this package is just amazing.



Doing a basic difference model with forecasting.

Forecasts for daily case difference over next 10 days West Virginia



ARIMA modeling prior 14 (15 for non differenc model) days in a difference model

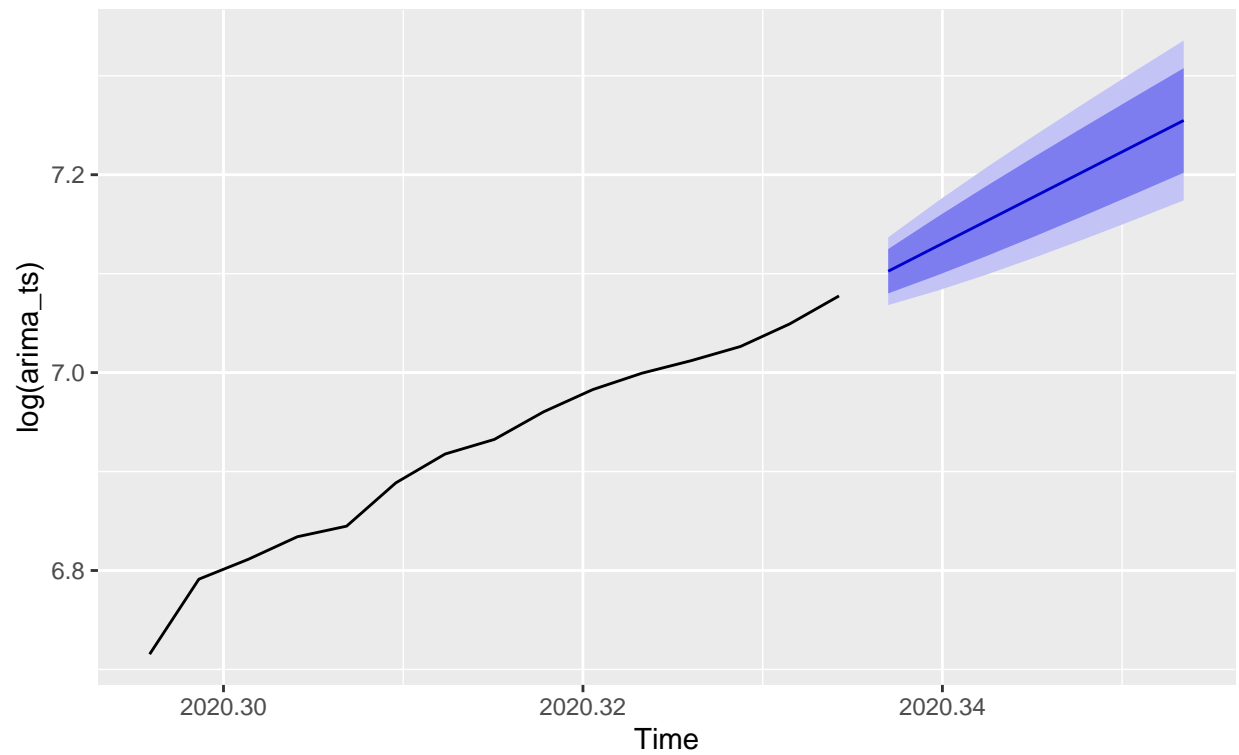
Another way to forecast the number of cases is to use ARIMA modeling (See the Forecast package citation) or:

<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 nubmer of infections but is meant to show overall treands and help predict public health risks. The first graph is showing the log of the cases as part of a time series.

Forecasts from ARIMA(1,1,0) with drift

Standard log(cases) forecast model next 7 days West Virginia 2020-05-02



The above figure (ARIMA modeling of previous 15 days) is meant to show an overall trend. This used 15 day window to allow for a rapid conversion to a 14 day difference model (below).

Forecasts from ARIMA(1,0,0) with non-zero mean

Daily difference in cases forecast model next 7 days West Virginia 2020-05-02



Part 2. 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/lockdown 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.

Part 2. Section A Comparing the last two weeks to the previous.

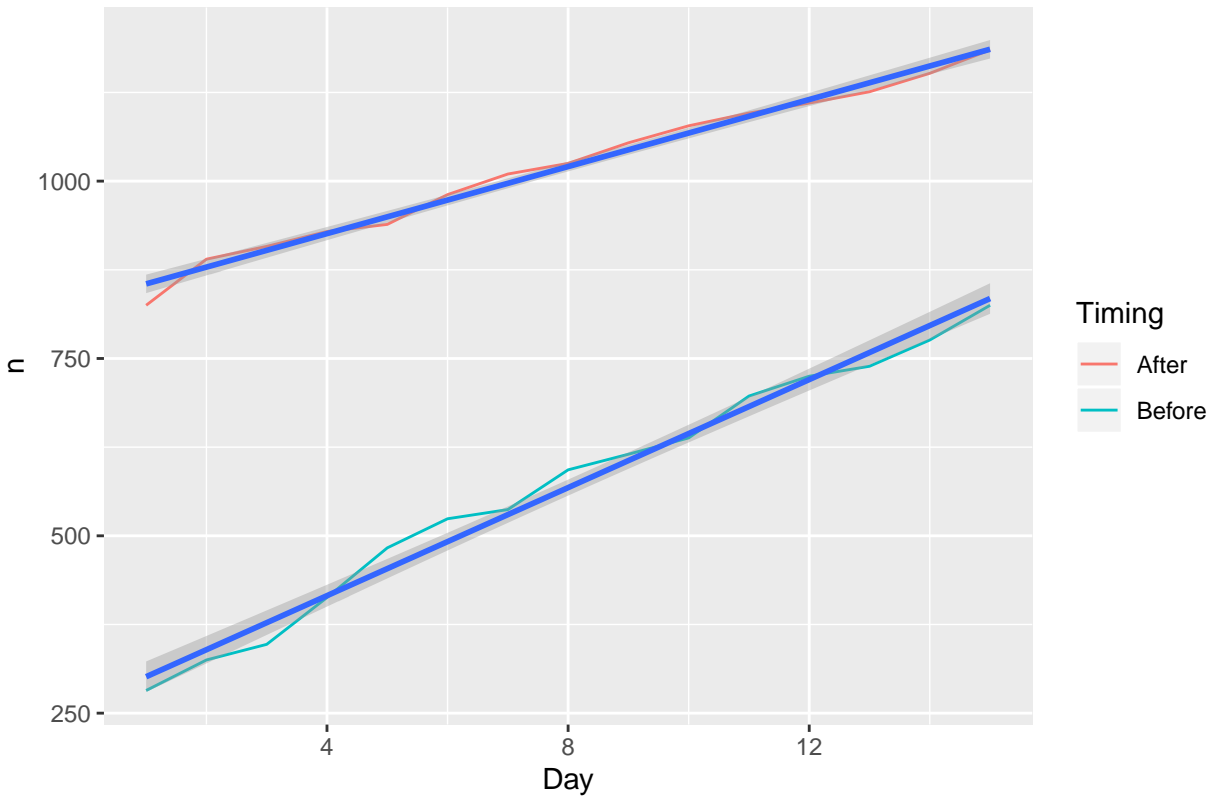
Comparing the last n days versus previous time frame using standard case number and difference modeling.

Above shows a basic difference model of a time series with a random walk forecast. If it appears that a form of stasis (i.e. random variation around an estimated mean) was achieved then it may be possible to use ARIMA modeling for a better determination of actual Difference(cases).

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

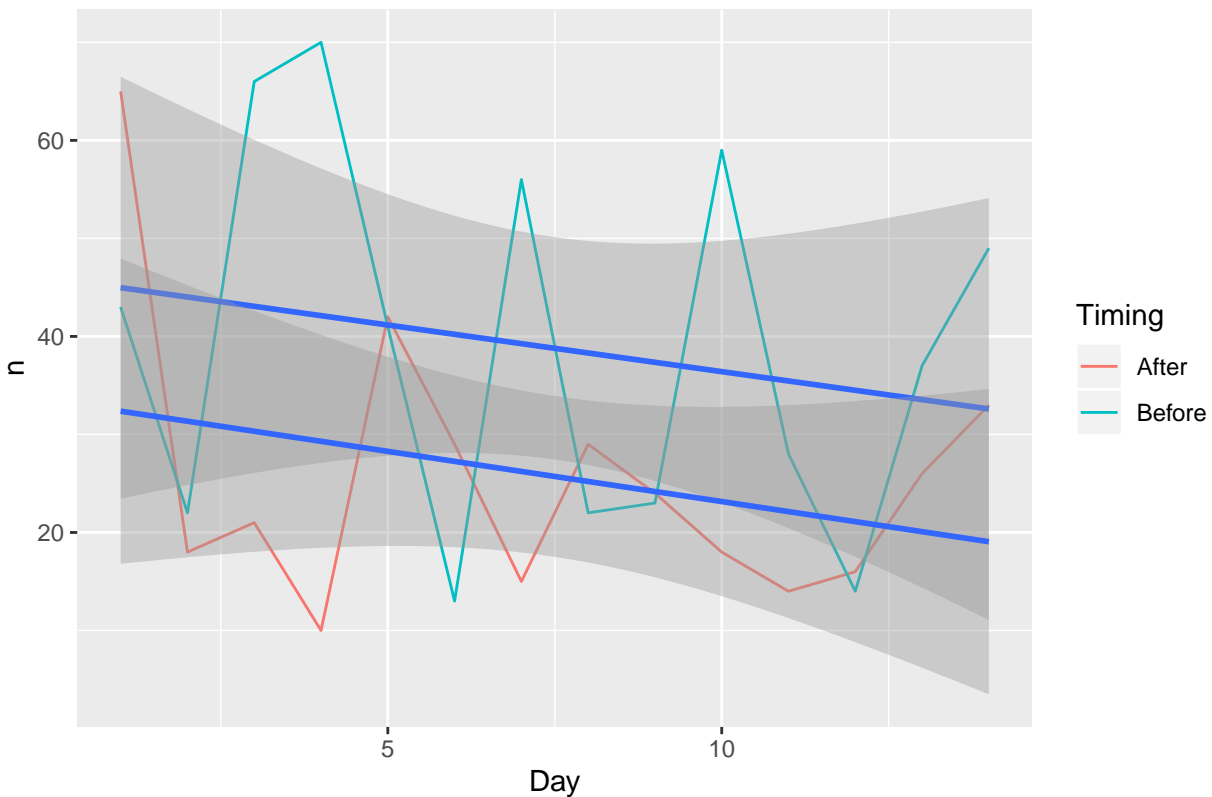
the data set in the last 15 days and the before is the 15 days prior to the After data set. This was meant to coincide with the naming of the lockdown (before and after the lockdown) data analysis (below).

Cumulative case load COVID-19 in 15 d groups: West Virginia



The above data is showing the before group from Starting date, 2020-04-04 to 2020-04-18. The after group includes from Starting date, 2020-04-18 to 2020-05-02. Looking at the dataset 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 may be different. In addition, Yule-Simpson effect of big data may be involved and the slope of the line may be unrelated or trend differently than the actual sub-grouped data. What we are really 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.

Difference Modeling of COVID-19 in 2 week windows West Virginia



Checking the Linear Model of the before and after.

First part is to look at the LM for the before group. 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 == "Before"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.213 -15.822  -1.068   16.654   27.884
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.9231    10.9952   4.177  0.00128 **
## Day         -0.9516     1.2913  -0.737  0.47531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.48 on 12 degrees of freedom
## Multiple R-squared:  0.0433, Adjusted R-squared:  -0.03643
## F-statistic: 0.5431 on 1 and 12 DF, p-value: 0.4753
```

Second part is to look at the LM for the After group. 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 == "After"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.299  -9.025  -2.642   5.388  32.629
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.3956     7.9539   4.199  0.00123 **
## Day         -1.0242     0.9341  -1.096  0.29442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.09 on 12 degrees of freedom
## Multiple R-squared:  0.09105,    Adjusted R-squared:  0.01531
## F-statistic: 1.202 on 1 and 12 DF,  p-value: 0.2944
```

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

Calculating the difference between the before and after groups using the difference model

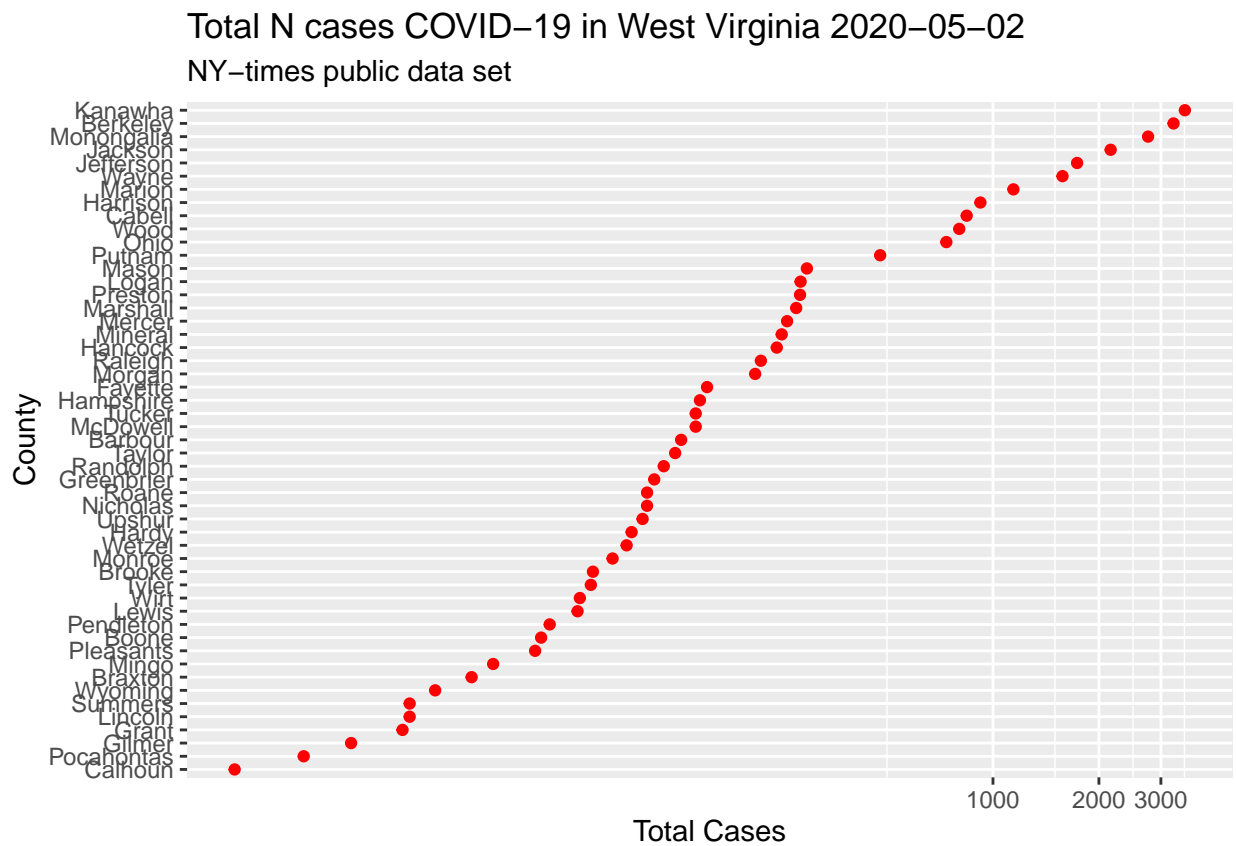
The goal of this section is to use the coefficients of the linear model (i.e. differences in cases/day) to see the number of cases and if the last two week the case load is decreasing (negative number), increasing (positive number) or remaining the same (around 0 +/- number). The difference between the last two weeks and the previous two weeks is: The difference in the slope (# cases/day) is: -0.07 case-difference/day.

Part 2, Section B Linear modeling to look at before and after the stay in place order

Lockdown analysis set to FALSE.

Part 3. Geospatial distrubution of cases

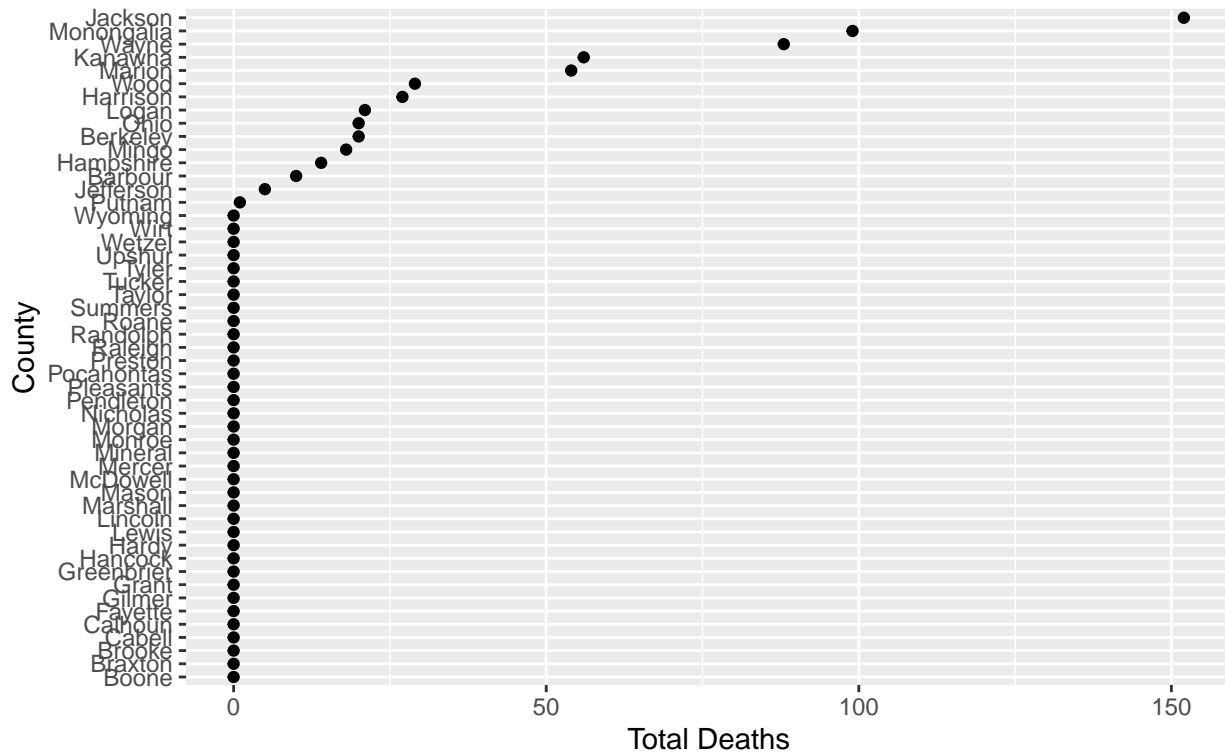
Part 3 Section A Plotting bar graph of cases



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

Total N deaths COVID-19 in West Virginia 2020-05-02

NY-times public data set



Part 3 Section B plotting using state shape files

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:

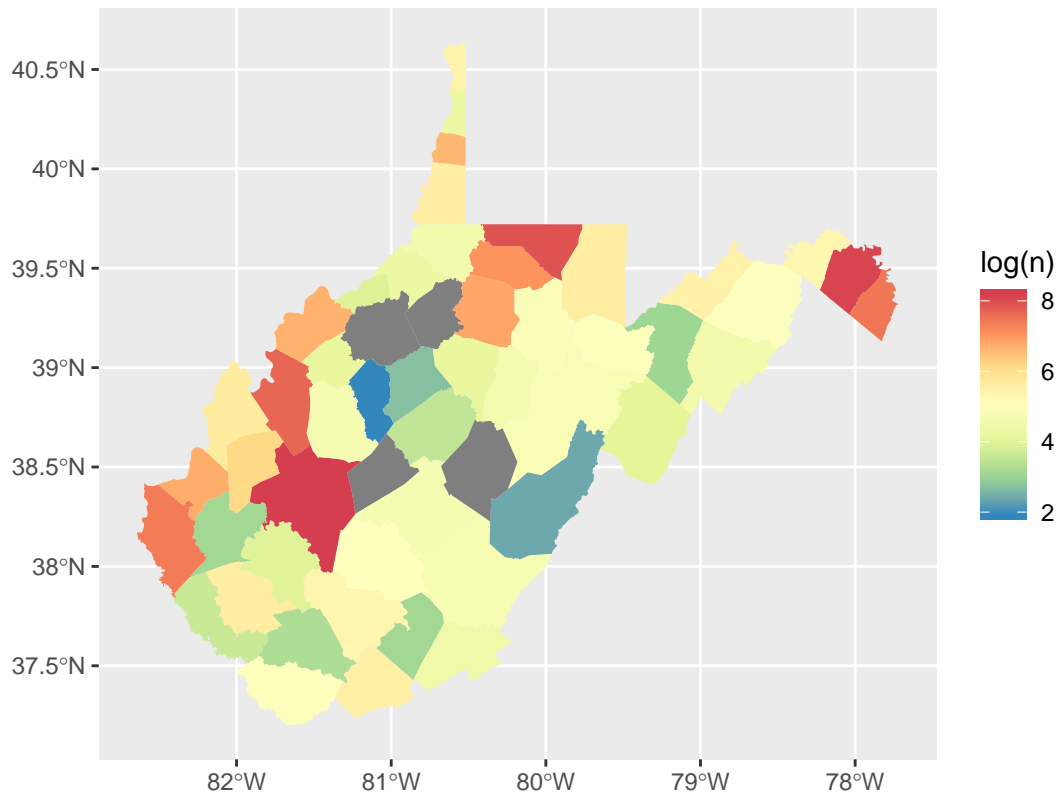
<https://walkerke.github.io/>

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

Geospatial of number of cases

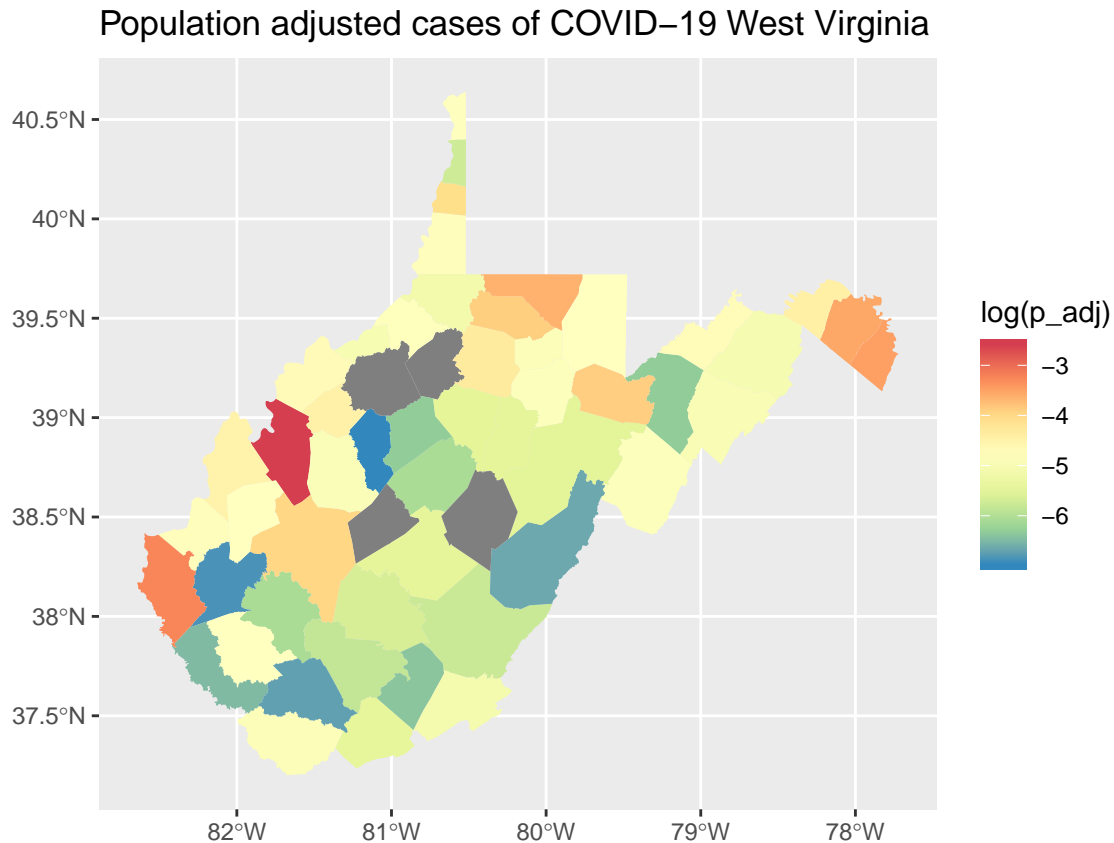
The goal of this plot is to show where the cases that are driving the infectious disease pan/epidemic are occurring. This is limited to the quality of the data recieved and due to the varied nature of different counties' public health may not be accurate. However, this should provide a population densisty 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 provied a more complete picture of needs.

Number of COVID–19 cases by county West Virginia



Geospatial overlay of probability of case given its population

The goal of this plot is 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 “hotspot”) could indicate a underserved 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 “hotspot” 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 oppisite 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 surronding suburban and other outline counties may have similar risk that is not seen on the cases number plot alone.



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 West Virginia 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/santization etc. This would mean that for this state to achieve similar percent infected as the 1918 flu around 508071 is needed. Currently, there are approximately 24518. this is around 1.34%. Currently the total mortality in the state is 614 and this accounts for 2.5% of the total verified cases.