

Here's a chart of the trends in COVID-19 cases in Texas, with and without being adjusted by a multiplier of the square root of the test positivity rate. The vertical scale has been removed because we don't have a way of translating the red adjusted line into actual numbers of cases. For this chart, I've converted both lines to indexes that come together at the end of the period by design. A good estimate of actual absolute case numbers, allowing me to put the scale back on the vertical axis, would certainly involve the red line being shifted upwards by some additional and unknown multiplier. So let's just focus on trends.

I've used a smoothed version of the test positivity rate after modelling it with a generalized additive model, to handle data problems relating to test numbers; and seven day moving averages of both series to deal with the weekly 'seasonality' of the data. Code is at the bottom of the post.

We can see that @SeanTrende is at least partly justified. If you adjust the confirmed cases per day this way, the latest values, while worrying, are not 'records' exceeding the high point in mid April.

But they are still going up, which means that COVID-19 cases do seem to be accelerating in Texas even when we take into account the higher number of tests being undertaken.

To get that red line to level out you need to use the most-maximalist version of adjustment possible and multiply the number of cases by the test positivity rate itself (rather than its square root). This would be equivalent to treating the people being tested as a random sample representative of the overall Texas population (not self-selecting for sicker people at all), which is not plausible.

Here's a similar chart for the 12 US states with the most COVID-19 cases:

There's some interesting patterns there. I'm pretty much satisfied the adjusted values are more accurate pictures of the incidence trends in these states than the original case numbers.

The R code for doing this is below. Comments welcome.

```
library(tidyverse)
library(scales)
library(janitor)
library(ggseas)
library(mgcv)

#-----Data import and tidying-----

states_orig <- read_csv("https://covidtracking.com/api/v1/states/daily.csv")
states_info <- read_csv("https://covidtracking.com/api/v1/states/info.csv")

states <- states_orig %>%
  mutate(date = as.Date(as.character(date), format = "%Y%m%d")) %>%
  clean_names() %>%
  # force total number of tests to be at least as many as the number
  of positives:
  mutate(total_test_results_increase = pmax(positive_increase,
total_test_results_increase)) %>%
  mutate(pos_rate = positive_increase / total_test_results_increase)
%>%
  arrange(date) %>%
  mutate(date_n = as.numeric(date)) %>%
  left_join(select(states_info, state, state_name = name), by =
"state")

# Just the 12 biggest states
states12 <- states %>%
  group_by(state) %>%
  summarise(max_pos = max(positive)) %>%
```

```

    arrange(desc(max_pos)) %>%
    slice(1:12) %>%
    inner_join(states, by = "state") %>%
    # state has to be a factor for use in mgcv::gam:
    mutate(state_name = fct_reorder(state_name, positive, .fun = sum))
%>%
    arrange(date) %>%
    ungroup()

#-----Smooth the positive test rates-----
mod <- gam(pos_rate ~ state_name + s(date_n, by = state_name),
          data = states12,
          family = quasibinomial,
          weights = total_test_results_increase)

states12$pos_rate_smoothed <- predict(mod, newdata = states12, type =
"response")

d <- states12 %>%
  mutate(adj_pos = positive_increase * sqrt(pos_rate_smoothed)) %>%
  select(date, state_name, positive_increase, adj_pos) %>%
  gather(variable, value, -date, -state_name) %>%
  mutate(variable = if_else(variable == "adj_pos",
                           true = "Adjusted for test positivity
rate",
                           false = "Original")) %>%
  group_by(state_name, variable) %>%
  arrange(date) %>%
  mutate(value = value / value[n()] * 100)

#-----Common themes and labels-----

the_theme <- theme(axis.text.y = element_blank(),
                  panel.grid.major.x = element_blank(),
                  panel.grid.minor.x = element_blank(),
                  panel.grid.minor.y = element_blank())

the_labs <- labs(x = "",
               colour = "",
               y = "New daily confirmed cases",
               caption = "Source: data from covidtracking.com,
positivity adjustment by http://freerangestats.info")

#-----Plots-----

# Top 12 states
d %>%
  ggplot(aes(x = date, y = value, colour = variable)) +
  facet_wrap(~state_name, scale = "free_y") +
  stat_rollaplyr(index.ref = 60, width = 7) +
  the_theme +
  the_labs +
  scale_colour_brewer(palette = "Set1") +
  ggtitle("Trends in daily COVID-19 cases (rolling seven-day average,
scale-free index)",
         "With and without adjustment for proportion of tests that
return positives, suggesting relatively more unknown cases in March

```

and April.")

Texas:

```
d %>%
  filter(state_name == "Texas") %>%
  ggplot(aes(x = date, y = value, colour = variable)) +
  stat_rollapplyr(index.ref = 70, width = 7) +
  theme +
  theme_labs +
  scale_colour_brewer(palette = "Set1") +
  ggtitle("Trends in daily COVID-19 cases in Texas (rolling seven-day
average, scale-free index)",
          "After adjustment for test-positivity, new cases are still
accelerating.")
```