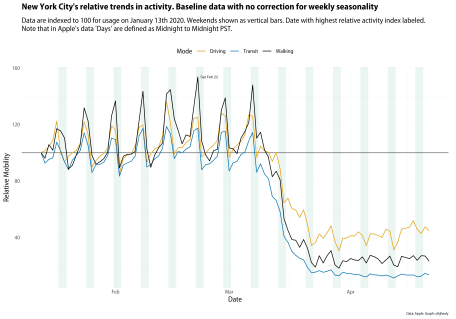
Here’s what the data look like:

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18 | > apple\_mobility  # A tibble: 39,500 x 5  geo\_type region transportation\_type date index  <chr> <chr> <chr> <date> <dbl>  1 country/region Albania driving 2020-01-13 100  2 country/region Albania driving 2020-01-14 95.3  3 country/region Albania driving 2020-01-15 101.  4 country/region Albania driving 2020-01-16 97.2  5 country/region Albania driving 2020-01-17 104.  6 country/region Albania driving 2020-01-18 113.  7 country/region Albania driving 2020-01-19 105.  8 country/region Albania driving 2020-01-20 94.4  9 country/region Albania driving 2020-01-21 94.1  10 country/region Albania driving 2020-01-22 93.5  # … with 39,490 more rows |

The index is the measured outcome, tracking relative usage of directions for each mode of transportation. Let’s take a look at the data for New York.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29 | raw\_ny <- apple\_mobility %>%  filter(region == "New York City") %>%  select(region:index) %>%  rename(mode = transportation\_type) %>%  mutate(mode = tools::toTitleCase(mode),  weekend = isWeekend(date),  holiday = isHoliday(as.timeDate(date), listHolidays())) %>%  mutate(max\_day = ifelse(is\_max(index), date, NA),  max\_day = as\_date(max\_day))  p\_raw\_ny <- ggplot(raw\_ny, mapping = aes(x = date, y = index,  group = mode, color = mode)) +  geom\_vline(data = subset(raw\_ny, holiday == TRUE),  mapping = aes(xintercept = date),  color = my.colors("bly")[5], size = 2.9, alpha = 0.1) +  geom\_hline(yintercept = 100, color = "gray40") +  geom\_line() +  geom\_text\_repel(aes(label = format(max\_day, format = "%a %b %d")),  size = rel(2), nudge\_x = 1, show.legend = FALSE) +  scale\_color\_manual(values = my.colors("bly")) +  labs(x = "Date", y = "Relative Mobility",  color = "Mode",  title = "New York City's relative trends in activity. Baseline data with no correction for weekly seasonality",  subtitle = "Data are indexed to 100 for usage on January 13th 2020. Weekends shown as vertical bars. Date with highest relative activity index labeled.\nNote that in Apple's data 'Days' are defined as Midnight to Midnight PST.",  caption = "Data: Apple. Graph: @kjhealy") +  theme(legend.position = "top")  p\_raw\_ny |



Relative Mobility in New York City. Touch or click to zoom.

As you can see, we have three series. The weekly pulse of activity is immediately visible as people do more or less walking, driving, and taking the subway depending on what day it is. Remember that the data is based on requests for directions. So on the one hand, taxis and Ubers might be making that sort of request every trip. But people living in New York do not require turn-by-turn or step-by-step directions in order to get to work. They already know how to get to work. Even if overall activity is down at the weekends, requests for directions go up as people figure out how to get to restaurants, social events, or other destinations. On the graph here I’ve marked the highest relative value of requests for directions, which is for foot-traffic on February 22nd. I’m not interested in that particular date for New York, but when we look at more than one city it might be useful to see how the maximum values vary.

The big COVID-related drop-off in mobility clearly comes in mid-March. We might want to see just that trend, removing the “noise” of daily variation. When looking at time series, we often want to decompose the series into components, in order to see some underlying trend. There are many ways to do this, and many decisions to be made if we’re going to be making any strong inferences from the data. Here I’ll just keep it straightforward and use some of the very handy tools provided by the [tidyverts](https://tidyverts.org/) (sic) packages for time-series analysis. We’ll use an [STL decomposition](https://feasts.tidyverts.org/reference/STL.html) to decompose the series into *trend*, *seasonal*, and *remainder* components. In this case the “season” is a week rather than a month or a calendar quarter. The trend is a locally-weighted regression fitted to the data, net of seasonality. The remainder is the residual left over on any given day once the underlying trend and “normal” daily fluctuations have been accounted for. Here’s the trend for New York.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30 | resids\_ny <- apple\_mobility %>%  filter(region == "New York City") %>%  select(region:index) %>%  rename(mode = transportation\_type) %>%  mutate(mode = tools::toTitleCase(mode)) %>%  as\_tsibble(key = c(region, mode)) %>%  model(STL(index)) %>%  components() %>%  mutate(weekend = isWeekend(date),  holiday = isHoliday(as.timeDate(date), listHolidays())) %>%  as\_tibble() %>%  mutate(max\_day = ifelse(is\_max(remainder), date, NA),  max\_day = as\_date(max\_day))  p\_resid\_ny <- ggplot(resids\_ny, aes(x = date, y = remainder, group = mode, color = mode)) +  geom\_vline(data = subset(resids, holiday == TRUE),  mapping = aes(xintercept = date),  color = my.colors("bly")[5], size = 2.9, alpha = 0.1) +  geom\_line(size = 0.5) +  geom\_text\_repel(aes(label = format(max\_day, format = "%a %b %d")),  size = rel(2), nudge\_x = 1, show.legend = FALSE) +  scale\_color\_manual(values = my.colors("bly")) +  labs(x = "Date", y = "Remainder", color = "Mode",  title = "New York City, Remainder component for activity data",  subtitle = "Weekends shown as vertical bars. Date with highest remainder component labeled.\nNote that in Apple's data 'Days' are defined as Midnight to Midnight PST.",  caption = "Data: Apple. Graph: @kjhealy") +  theme(legend.position = "top")    p\_resid\_ny |

Trend component of the New York series. Touch or click to zoom.

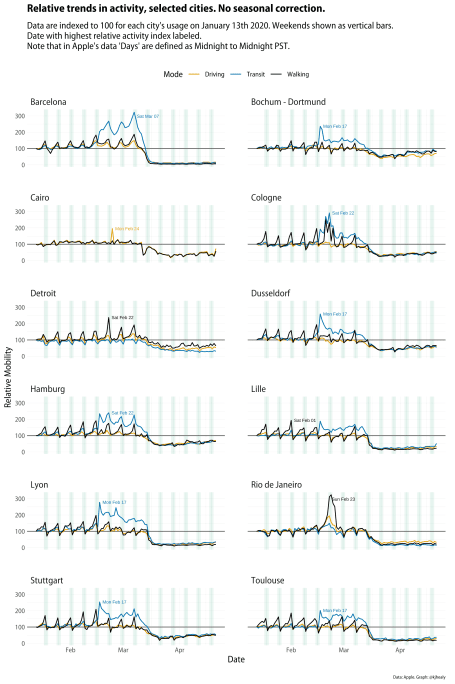
We can make small multiple showing the raw data (or the components, as we please) for all the cities in the dataset if we like:

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16 | p\_base\_all <- apple\_mobility %>%  filter(geo\_type == "city") %>%  select(region:index) %>%  rename(mode = transportation\_type) %>%  ggplot(aes(x = date, y = index, group = mode, color = mode)) +  geom\_line(size = 0.5) +  scale\_color\_manual(values = my.colors("bly")) +  facet\_wrap(~ region, ncol = 8) +  labs(x = "Date", y = "Trend",  color = "Mode",  title = "All Modes, All Cities, Base Data",  caption = "Data: Apple. Graph: @kjhealy") +  theme(legend.position = "top")  p\_base\_all |

Data for all cities. Touch or click to zoom.

This isn’t the sort of graph that’s going to look great on your phone, but it’s useful for getting some overall sense of the trends. Beyond the sharp declines everywhere—with slightly different timings, something that’d be worth looking at separately—a few other things pop out. There’s a fair amount of variation across cities by mode of transport and also by the intensity of the seasonal component. No-one is walking anywhere in Dubai. Some sharp spikes are evident, too, not always on the same day or by the same mode of transport. We can take a closer look at some of the cities of interest on this front.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35 | focus\_on <- c("Rio de Janeiro", "Lyon", "Bochum - Dortmund", "Dusseldorf",  "Barcelona", "Detroit", "Toulouse", "Stuttgart",  "Cologne", "Hamburg", "Cairo", "Lille")  raw\_ts <- apple\_mobility %>%  filter(geo\_type == "city") %>%  select(region:index) %>%  rename(mode = transportation\_type) %>%  mutate(mode = tools::toTitleCase(mode),  weekend = isWeekend(date),  holiday = isHoliday(as.timeDate(date), listHolidays())) %>%  filter(region %in% focus\_on) %>%  group\_by(region) %>%  mutate(max\_day = ifelse(is\_max(index), date, NA),  max\_day = as\_date(max\_day))    ggplot(raw\_ts, mapping = aes(x = date, y = index,  group = mode, color = mode)) +  geom\_vline(data = subset(raw\_ts, holiday == TRUE),  mapping = aes(xintercept = date),  color = my.colors("bly")[5], size = 1.5, alpha = 0.1) +  geom\_hline(yintercept = 100, color = "gray40") +  geom\_line() +  geom\_text\_repel(aes(label = format(max\_day, format = "%a %b %d")),  size = rel(2), nudge\_x = 1, show.legend = FALSE) +  scale\_color\_manual(values = my.colors("bly")) +  facet\_wrap(~ region, ncol = 2) +  labs(x = "Date", y = "Relative Mobility",  color = "Mode",  title = "Relative trends in activity, selected cities. No seasonal correction.",  subtitle = "Data are indexed to 100 for each city's usage on January 13th 2020. Weekends shown as vertical bars.\nDate with highest relative activity index labeled.\nNote that in Apple's data 'Days' are defined as Midnight to Midnight PST.",  caption = "Data: Apple. Graph: @kjhealy") +  theme(legend.position = "top") |



Selected cities only. Touch or click to zoom.

Look at all those transit peaks on February 17th. What’s going on here? At this point, it might useful to take a look at the residual or remainder component of the series rather than looking at the raw data, so we can see if something interesting is happening.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31 | resids <- apple\_mobility %>%  filter(geo\_type == "city") %>%  select(region:index) %>%  rename(mode = transportation\_type) %>%  mutate(mode = tools::toTitleCase(mode)) %>%  filter(region %in% focus\_on) %>%  as\_tsibble(key = c(region, mode)) %>%  model(STL(index)) %>%  components() %>%  mutate(weekend = isWeekend(date),  holiday = isHoliday(as.timeDate(date), listHolidays())) %>%  as\_tibble() %>%  group\_by(region) %>%  mutate(max\_day = ifelse(is\_max(remainder), date, NA),  max\_day = as\_date(max\_day))    ggplot(resids, aes(x = date, y = remainder, group = mode, color = mode)) +  geom\_vline(data = subset(resids, holiday == TRUE),  mapping = aes(xintercept = date),  color = my.colors("bly")[5], size = 1.5, alpha = 0.1) +  geom\_line(size = 0.5) +  geom\_text\_repel(aes(label = format(max\_day, format = "%a %b %d")),  size = rel(2), nudge\_x = 1, show.legend = FALSE) +  scale\_color\_manual(values = my.colors("bly")) +  facet\_wrap(~ region, ncol = 2) +  labs(x = "Date", y = "Remainder", color = "Mode",  title = "Remainder component for activity data (after trend and weekly components removed)",  subtitle = "Weekends shown as vertical bars. Date with highest remainder component labeled.\nNote that in Apple's data 'Days' are defined as Midnight to Midnight PST.",  caption = "Data: Apple. Graph: @kjhealy") +  theme(legend.position = "top") |

Remainder components only. Touch or click to zoom.

As a rule, when you see a sharp change in a long-running time-series, you should always check to see if some aspect of the data-generating process changed—such as the measurement device or the criteria for inclusion in the dataset—before coming up with any substantive stories about what happened and why. This is especially the case for something susceptible to change over time, but not to extremely rapid fluctuations. … As Tom Smith, the director of the General Social Survey, likes to say, if you want to measure change, you can’t change the measure.

In this case, there’s a further wrinkle. I probably would have been quicker to twig what was going on had I looked a little harder at the raw data rather than moving to the remainder component of the time series decomposition. Having had my eye caught by Rio’s big Carnival spike I went to look at the remainder component for all these cities and so ended up focusing on that. But if you look again at the raw city trends you can see that the transit data series (the blue line) spikes up but then *sticks around* afterwards, settling in to a regular presence, at quite a high relative level in comparison to its previous non-existence. And this of course is because people have begun to use this new feature regularly. If we’d had raw data on the absolute levels of usage in transit directions this would likely have been clearly more quickly.

The tendency to launch right into what social scientists call the “Storytime!” phase of data analysis when looking at some graph or table of results is really strong. We already know from other COVID-related analysis how tricky and indeed dangerous it can be to mistakenly infer too much from what you think you see in the data. Taking care to understand what your measurement instrument is doing really does matter. In this case, I think, it’s all the more important because with data of the sort that Apple have released, it’s fun to just jump into it and start speculating. That’s because we don’t often get to play with even highly aggregated data from sources like this. I wonder if, in the next year or so, someone doing an ecological, city-level analysis of social response to COVID-19 will inadvertently get caught out by the change in the measure lurking in this dataset.