Details on [The Covid19 Commerce Patterns Dashboard]() from SafeGraph

Ryan Fox Squire, Data Scientist @ SafeGraph
3/25/2020

Introduction

SafeGraph provides open access to clean and accurate data on millions of physical places across the United States and Canada. This empowers analysts, data scientists, and developers to deliver actionable business analysis and meaningful socio-economic research that drives noticeable, real-world impact.

Our goal for The Covid19 Commerce Patterns Dashboard is to show changes in day-to-day foot-traffic to different sectors of the economy in response to Covid19. Historically many areas of commerce exhibit significant seasonal variability (i.e., foot-traffic to airports in March is always larger than in February). So our goal is also to contextualize behavior in 2020 against historic seasonal patterns.

This document briefly summarizes how these visualizations are created so that these analyses can be easily reproduced and extended by the broader community of researchers, academics, governments, and other organizations for the public good currently working with SafeGraph data. Data is available at no cost to these organizations.

What is the data?

SafeGraph Patterns is aggregated from a panel (sample of population that is measured longitudinally) of millions of opt-in mobile devices (e.g., smart phones) in the USA, provided by Veraset. SafeGraph uses geospatial point-of-interest polygons and time-clustering methods to summarize GPS data into aggregated visit counts for individual points-of-interest (e.g., a Walmart in San Leandro, CA). The data in the analysis is anonymized and aggregated and we take extra steps to preserve privacy in any data we share.

How are the visualizations built?

We start with aggregated daily counts of visits to individual POI, across the USA (i.e., the SafeGraph Patterns column 'visits_by_day').

We then group by year, month, day, and a particular brand (or a particular naics_code), e.g.

```
SELECT brand, year, month, day, SUM(visits_by_day) as raw_visit_sum
FROM safegraph_patterns_daily
```

```
GROUP BY brand, year, month, day
WHERE brand = 'Walmart'
```

Due to the nature of sampling from a dynamic and unstable panel, comparisons across long time periods are methodologically challenging. We employ several data transformations to make the data easier to interpret.

Different parts of the dashboard group by NAICS or by Brand. The "industry" sections are grouped by NAICS code. Here are the naics codes we are using:
'512131' : "Movie Theaters",
'445110' : "Supermarkets",
'622110' : "Hospitals",
'531120' : "Malls",
'722511' : "Full Service Restaurants",
'452319' : "General Merchandise Stores",
'722513' : "Limited Service Restaurants",
'721110' : "Hotels",
'722515' : 'Snack and Nonalcoholic Beverage Bars',
'722410' : 'Drinking Places (Alcoholic Beverages)',
'488119' : "Top Airports", ** For Airports we are not using all airports, but we are using a hand-curated list of airport SGPIDS representing the top USA airports.

Normalization

To correct for short-term (day-to-day and month-to-month) changes in the panel, we divide the daily raw_visit_sum by a daily aggregated count of unique visitors (to all industry, brand, or census areas). This is the **total_devices_seen** in Normalization Stats. This normalization helps control for short-term variability in the panel day-to-day. For example, if the panel population grew 20% overall, but the frequency of people visiting Walmart did not change, then the **normalized daily count** would be constant. There are many ways to implement this sort of normalization; in general we do not find that different methods significantly alter the overall results.

Smoothing

Many industries and brands exhibit significant within-week variation (e.g., most retail shows larger foot-traffic on saturday versus a weekday). We smooth over a seven day window to focus the analysis on the overall trend in a category or brand across many weeks, while still capturing important day-by-day changes. Specifically we use the python pandas method rolling(window=7, center=False).mean(). So, the data point associated with March 8th, is the mean of all 7 data points from March 2nd - March 8th (inclusive).

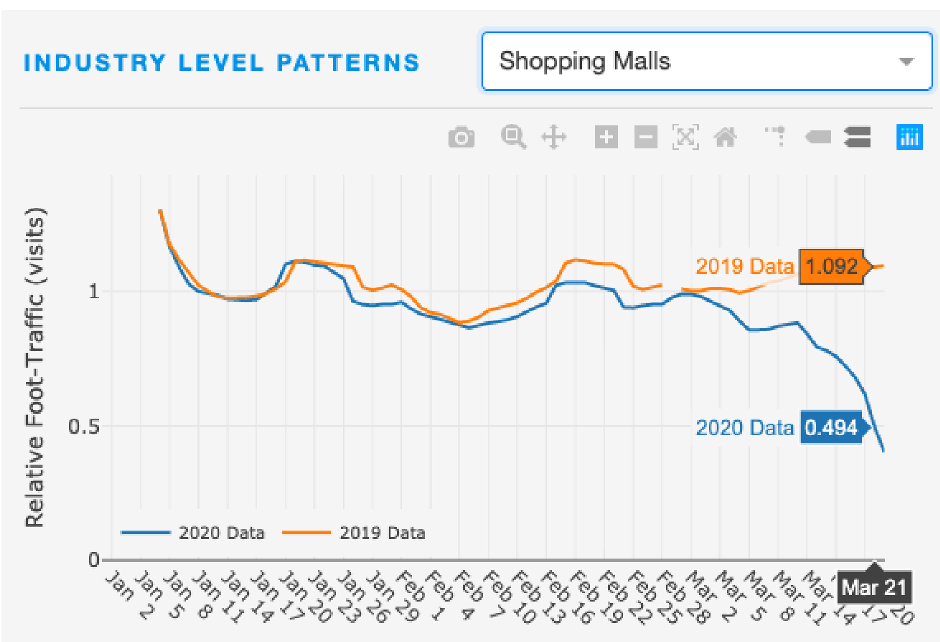
Year-over-year standardization

Finally, long term periods (years) are associated with larger changes in the panel population that create additional sources of sampling bias. These changes are not only to the size of the panel population, but also to the ability to measure and detect visits in general. Therefore, for the purposes of visualization, we make a strong assumption: foot-traffic in early January is set == 1 in both 2019 and 2020. We call this **standardization**. This allows one to easily compare how the timeseries evolve day-by-day within each year, but it removes any signal about the absolute difference between 2019 and 2020. If you have data or a strong prior belief about the difference between 2019 and 2020, this is easily incorporated as a fixed offset to the 2020 data series. For our purposes we have no such data or beliefs and therefore all 2020 and 2019 time series are set to be equal in early January as a conservative standardization.

The specific algorithm we use for **standardization** is as follows. We collect the 7 data points between the 2nd and 3rd Wednesday of January, and compute the median for those 7 values (this is the **standardization factor**). Then we divide the entire time series for that year by the constant **standardization factor**. The dates used are: [2019-01-09, 2019-01-01-16) and [2020-01-08, 2020-01-01-15).

Interpreting results & quantifying uncertainty

Let's take an example from the dashboard for Shopping Malls on Mar 21st.



The 2020 value is 0.494 and the 2019 value is 1.092.

Here are some statements you can make about this data:

- Foot-traffic to Malls on March 21st 2020 is down 50.6% from early January of that year.
- Foot-traffic to Malls on March 21st 2019 was up 9.2% from early January of that year.
- On March 21st, 2020 foot-traffic to Malls is down 54.8% (0.494-1.092)/1.092*100 from where it **would** have been if 2020 YTD growth and 2019 YTD growth had been identical.
- If we assume 0% YoY growth in January, foot-traffic to Malls is down 54.8% YoY compared to the same time in 2019.

This analysis and visualization does not quantify the uncertainty of these measurements, and therefore, strictly speaking, you cannot make statistical statements about the certainty of these estimates. However, we've written about quantifying uncertainty in national-level analysis with patterns data before, in a different context. The panel is a very large sample of the population (covering roughly ~ 10% of the USA population), and is well-balanced across geographies. We've found estimates are very precise when estimating behavior at the national level. And from working with retailers, we've found our foot traffic counts to closely correlate with foot traffic measured by physical counters and transaction data. We welcome your feedback on how to best incorporate uncertainty into these specific visualizations.

Contact

Questions about this methodology?

- If you are a member of our public-good research consortium (or want to be), please ask in our slack community.
- If you are a journalist or news organization, please email press@safegraph.com