Norria Tian

Michael Russo

Prof. Christopher Musco

Introduction to Machine Learning

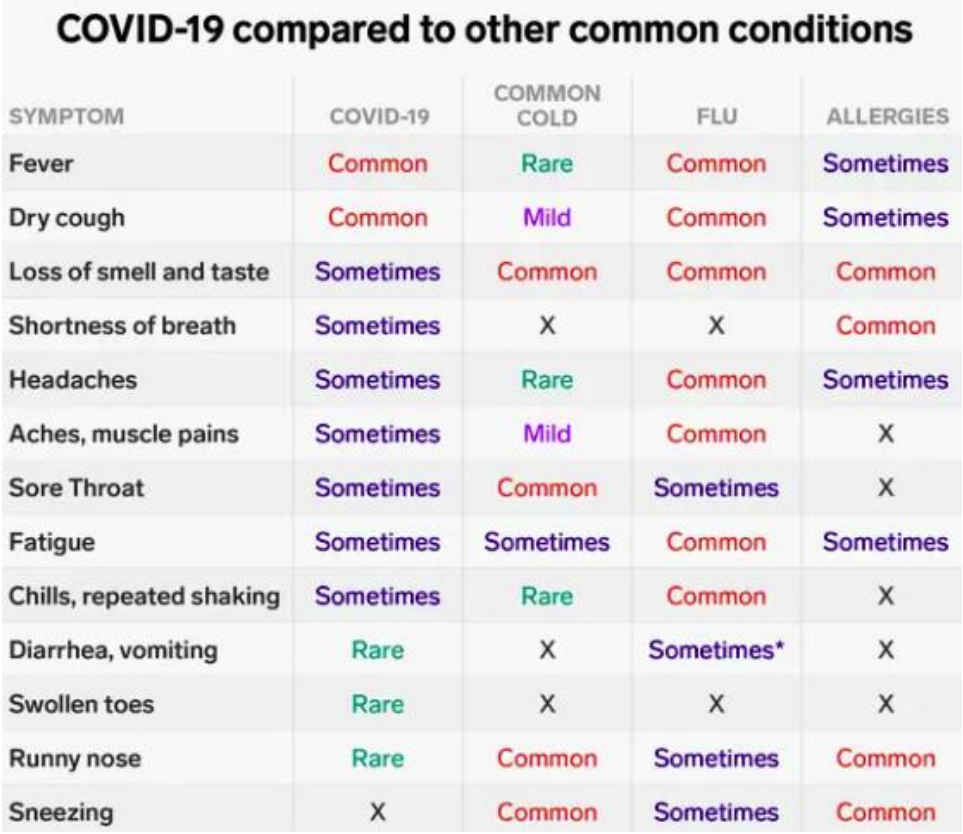
May 05, 2020

Introduction

Many people are making different models of the coronavirus infection rate based on information about the disease itself and the actions people take such as travel restrictions, population density, local social distancing measures, testing availability, and estimated infectivity. After reading a New York Times article about how Google search data may be useful for predicting the infection rate of the virus, we decided to make our project on how the current pandemic has changed people's search trends. We are using publicly available Google search trends data in New York state.

The underlying idea behind the news article is that if there is a sudden increase in searches for key words such as "coughing," "fever," or other words/symptoms associated with the virus, could it mean that there will be an impending increase in coronavirus cases? Our goal is to find out just how much the search trends have changed since Jan 1, 2020. We used the pytrends API to collect search data for key words for every year since 2009. We then use multiple linear regression to find how well the search rate for coronavirus-related words before 2020 matches the real search data. We focus more on the rates rather than the number of searches or number of infected because the search data provided by Google is automatically normalized on a scale from 0 – 100. Google does provide more detailed search data that is not normalized; however, it is either available for a fee or available for healthcare researchers.

We conduct a correlation and regression analysis based on search data of key word clusters (e.g., cough, coughing). We do this by creating multiple linear regression models based on past search data to predict what the trends for 2020 could have been without the pandemic and media influence. Using the loss from those models, we can determine how much search data is linked to the pandemic. We preprocess this data by normalizing each keyword cluster to be on a scale from 0 – 1 based on how popular the searches were. The final scores of 0 – 1 for each cluster become the features used for multiple linear regression. Another small thing we note is the possible influence of the H1N1 pandemic, which lasted from April 17, 2009 to August 11, 2010, on our early search data. The figure below shows the list of symptoms that are associated with the coronavirus. Our list of key words all came from this chart from businessinsider.com which is intended to inform people which symptoms are closer to COVID and which are not. We still include sneezing related words with our other features to observe the search trends of a commonly misconceived symptom of COVID.



COVID-19 compared to other common conditions

SYMPTOM	COVID-19	COMMON COLD	FLU	ALLERGIES
Fever	Common	Rare	Common	Sometimes
Dry cough	Common	Mild	Common	Sometimes
Loss of smell and taste	Sometimes	Common	Common	Common
Shortness of breath	Sometimes	X	X	Common
Headaches	Sometimes	Rare	Common	Sometimes
Aches, muscle pains	Sometimes	Mild	Common	X
Sore Throat	Sometimes	Common	Sometimes	X
Fatigue	Sometimes	Sometimes	Common	Sometimes
Chills, repeated shaking	Sometimes	Rare	Common	X
Diarrhea, vomiting	Rare	X	Sometimes*	X
Swollen toes	Rare	X	X	X
Runny nose	Rare	Common	Sometimes	Common
Sneezing	X	Common	Sometimes	Common

Figure 1: Common Symptoms (from businessinsider.com)

Methodology & Discussion

The unofficial API pytrends is used to interface with Google Trends and query for search entry data used in this project.

A series of symptoms commonly associated with the COVID-19, influenza, common colds, and seasonal allergies are investigated. They are broadly divided into these categories: Cough, Sneeze, Shortness of Breath, Sore Throat, Runny Nose, Fever, Muscle Pain, Fatigue, Anosmia, Chills, and Swollen Toes (included in the kw_lists).

The Google Trends data for daily normalized scores of each of the entries between January 1 and May 1 of each year from 2009 to 2020 are included. The leap day February 29 is dropped from the leap years 2012, 2016, and 2020 when performing a multiple linear regression with the other years.

Google Trends does not provide direct search volume for an entry. “Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity” to give a score normalized on the range of 0 to 100. It represents the proportion of a search entry to all search entries. A single normalized score for each symptom category (containing multiple search entries, also known as keywords) each day is obtained by dividing the sum of all the scores in the same time period (day) in the same geographical location (NY), by the number of keywords. This is essentially the mean of the scores in the category. Such

means are tabulated in tables like the one in Figure 2. The rows represent the date (how many days away from January 1) and the columns are the years from 2009 to 2020. Taking the mean does present issues such as: 1) Are the scores additive? It is assumed they are since they each represent the proportion out of the same total daily volume in the same region; 2) Is weighing them equally (mean) in calculating the combined normalized score a good idea? This is discussed next.

Normalized daily Google Trends scores for the symptom COUGH 01/01-01/05 from 2009 to 2020

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	0.1250	0.1925	0.3100	0.2450	0.2450	0.2725	0.3000	0.2125	0.2450	0.2625	0.3425	0.1950
1	0.1275	0.1375	0.3200	0.3000	0.2625	0.2725	0.3425	0.2875	0.2475	0.2900	0.3100	0.1475
2	0.2625	0.1100	0.2150	0.2175	0.2600	0.3025	0.2950	0.2975	0.2625	0.2750	0.3400	0.1725
3	0.1500	0.2550	0.1875	0.1800	0.2475	0.3575	0.2425	0.2775	0.3150	0.2075	0.2775	0.1500
4	0.2225	0.1450	0.1425	0.1475	0.3125	0.2925	0.2475	0.2675	0.2750	0.2425	0.3250	0.1475
...	...	...	...	...	...	...	...	...	...	...	...	...
116	0.1300	0.0825	0.1450	0.1500	0.2275	0.1550	0.1875	0.1825	0.1700	0.1700	0.2025	0.1125
117	0.2025	0.1275	0.1300	0.1375	0.1075	0.2100	0.1525	0.1650	0.1350	0.1550	0.1675	0.0850
118	0.2675	0.1750	0.1850	0.1275	0.1000	0.2200	0.1600	0.1250	0.1675	0.2125	0.1800	0.0975
119	0.1925	0.1750	0.1350	0.1575	0.1700	0.2075	0.1800	0.1875	0.1925	0.1175	0.2075	0.0975
120	0.2200	0.0975	0.2200	0.1150	0.1600	0.1650	0.1250	0.1700	0.2025	0.1700	0.1925	0.0725

121 rows × 12 columns

Figure 2

The selection of search entries (keywords) to represent each symptom is critical and inevitably introduces bias. There is an inherent bias for users to textualize the symptoms using certain preferred words. This bias is compounded by the choice in this project to select from all the related queries to the symptoms investigated. The keyword lists (kw_list) have taken this into account and include only the most popular search entries grouped by their relatedness. There is the possibility to introduce associated weight to each search entry in the calculation of the normalized score for each symptom; However, the assignment of the weight ratio would contain bias. Perhaps the assignment could reference potential research done on the tendency of native English speakers to gravitate towards certain words more than others; But this effect should be negligible and could introduce a new set of complications. It is presumed that such preference is constant through the years and impacts the scores from each year in the same manner that the effect becomes negligible. Hence, such weighted normalization is avoided in this project.

The normalized Google Trends score for each symptom category is then plotted against the date (Day) for each year to give an overview of any unusual behavior. An example is shown in Figure 3 where the normalized Google Trends score for the symptom category COUGH from Day 0 to Day 120 (from January 1) is plotted for each year. Black represents the years 2011 through 2019, red represents the current year 2020 (COVID-19); Blue and yellow represent 2009

and 2010 respectively (H1N1). The red line has an obvious peak around Day 50 – 90 (Feb 19 – Mar 30).

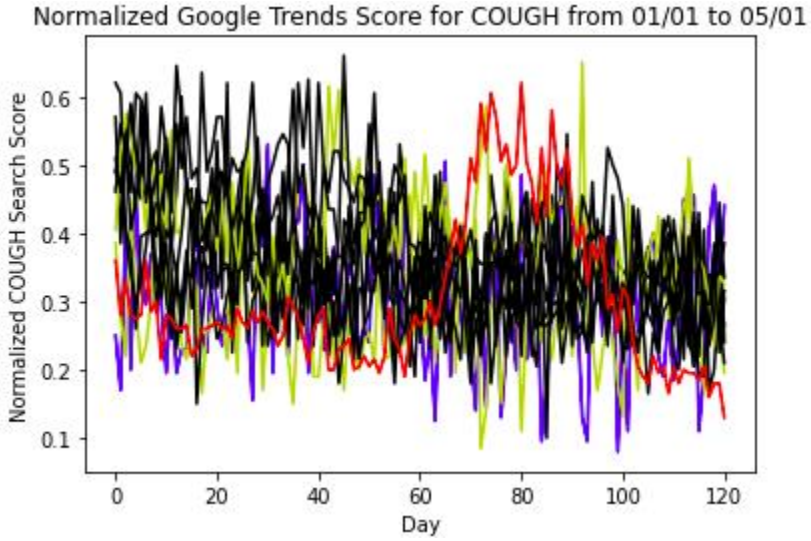


Figure 3

Multiple linear regression (MLR) is performed for each symptom category, using the years 2011 through 2019 as the training data and 2009-2010 and 2020 as separate tests. The results are plotted in Figures 4-6 and the red line represents the test data (actual scores of COUGH in 2009, 2010, and 2020), and the black represents the scores predicted using the data from years 2011 through 2019. The red line shows a drastic peak deviating from the predicted black line, indicating an unusual surge in interest in the entries relating to the symptom COUGH in 2020. The 2009 and 2010 plots do not show any particular interest. For an even clearer display, the COUGH trendline can be centralized by subtracting the average of the scores of the years 2011-2019, shown in Figure 7.

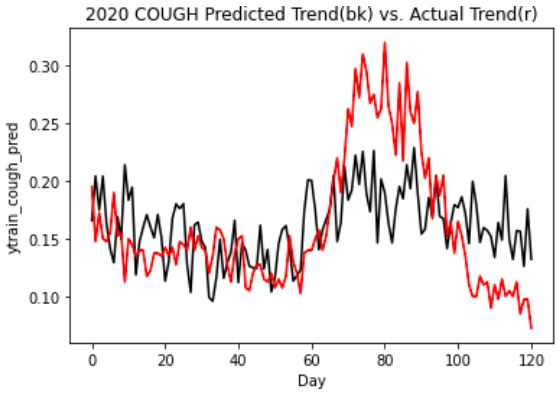


Figure 4: COVID-19

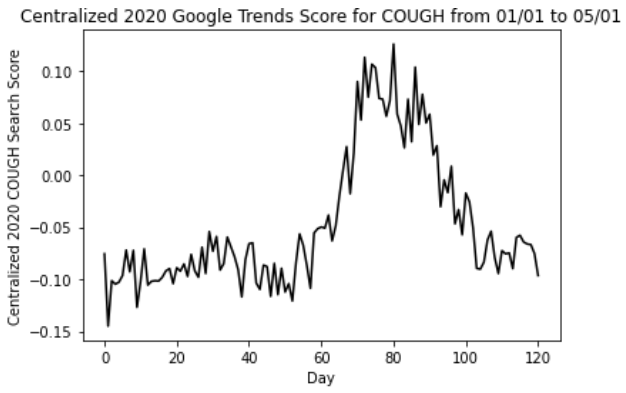


Figure 7: COUGH Centralized

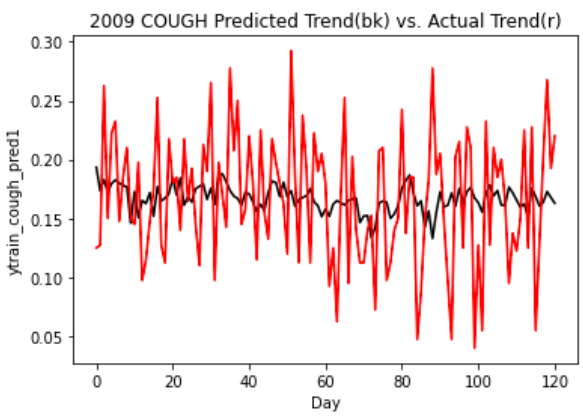


Figure 5: H1N1

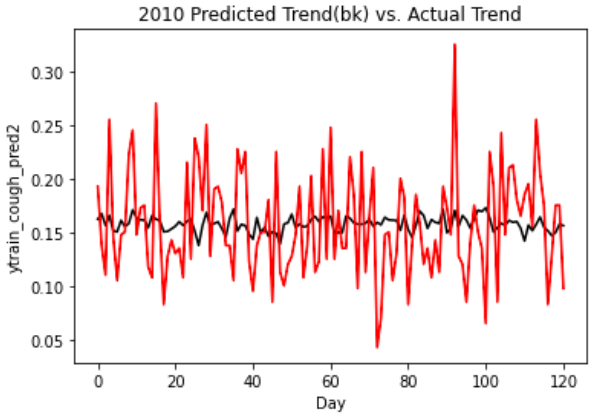


Figure 6: H1N1 Continued

Results

The results for the symptom category are showcased as an example in Methodology; the results afor the other categories are included below.

The normalized trendline for the symptom SNEEZE (Plot 1) appears to have a peak around Day 50 – 90 (Feb 19 – Mar 29) like COUGH. It blends in with rest of the years and the peak is not very prominent. Sneezing is in fact not a symptom of the COVID-19 but one frequent to the common cold and allergies. It could have been mistaken for a symptom of the COVID-19, which explains its sudden rise around the time period, unlike the constant up and down shape of the other years.

The symptom BREATH (which stands for short of breath) is common to allergies and less common to COVID-19 than FEVER AND COUGH. Its normalized trendline (Plot 2) takes a similar shape with a peak centralized around Day 80 as other 2020 trendlines, but otherwise blends in with the data from the past decade. In fact, the scores appear to be lower than usual.

A SORE THROAT is frequent for the common cold but occurs only sometimes for COVID-19 and influenza. Its normalized trendline (Plot 3) shows the usual peak but otherwise does not overwhelm the data from the previous decade.

A RUNNY NOSE is rare for COVID-19 but frequent for the common cold and the allergies of which peak coincides with the coronavirus pandemic. The normalized trendline (Plot 4) displays the characteristic 2020 peak prominently, perhaps due to the natural rise in frequency of this symptom around spring time and panic caused by COVID-19.

The symptom FEVER is common to COVID-19 and the flu. Its trendline in Plot 5 takes the usual 2020 shape with a peak centralized around Day 80 but is otherwise no more significant in popularity compared to data from the previous decade. This could perhaps be due to the sheer frequency of fever as a common illness.

The trendline (Plot 6) for PAIN (which stands for muscle pain or ache) is indistinguishable from the other trendlines. This is no surprise because it only occurs sometimes for COVID-19 and is an incredibly frequent symptom for many other afflictions. It is also one of

the less known symptoms of COVID-19. The actual trendline in Plot 19 show very little deviation from the predicted PAIN trendline in terms of center.

Plot 7 shows the normalized trend for FATIGUE. There is no obvious spike in the 2020 trendline that shows for most of the other symptoms. When compared to the predicted trendline in Plot 20, there is no significant stray from the predicted trendline. This is expected since FATIGUE only occurs sometimes for COVID-19 and is characteristic of a large myriad of illnesses physiological and psychological (e.g. diabetes, depression etc.). Like PAIN, FATIGUE trendline shows little deviation from its predicted trendline in Plot 20.

Plot 8 shows the normalized trend for ANOSMIA clustered with AGEUSIA (both of which are included in the kw_list), symptoms which means loss of smell and taste respectively. Although they are not common, it shows a very clear spike in searches around day 80. This symptom was not discovered in the beginning portion of the pandemic timeline. Although they were always symptoms, the searches before day 80 were extremely few, around 0.0 as shown in Plot 21. This is a good example of how search data can be influenced by media reporting. The fact that it went from no searches at all to a sudden peak may indicate that people were googling to know what these uncommon words mean instead of checking their symptoms online. The American Academy of Otolaryngology called for ANOSMIA to be added to a list of screening tools for possible COVID-19 infections on Mar 22, 2020. ANOSMIA as a symptom was subsequently reported widely in media. The trendline peak coincides exactly with that date.

Plot 9 shows the normalized trend for CHILLS. There is a noticeable peak between days 70 and 80 that may be easier to see in Plot 25. The 2020 trendline follows the predicted trendline closely in Plot 25 until it reaches around day 70, where the spikes begin. This is an uncommon symptom of the virus, but there is still a greater number of searches compared to the predicted trend.

Plot 10 shows the normalized trend for SWOLLEN TOES. This trend inconsistently spikes and returns to 0, which is better seen in Plot 22. The peaks become noticeably higher after day 80 which seems to be the common point of interest for most of these graphs. Swollen toes are considered a rare symptom, yet the peaks are still higher than the predicted trend for 2020.

Plot 11 shows the normalized trend for COVID itself. This cluster includes searches of “covid symptoms,” “coronavirus symptoms,” and “severe acute respiratory syndrome.” Even this cluster, which wouldn’t have had significant searches in previous years, has the same signature peak at the same time as the other graphs.

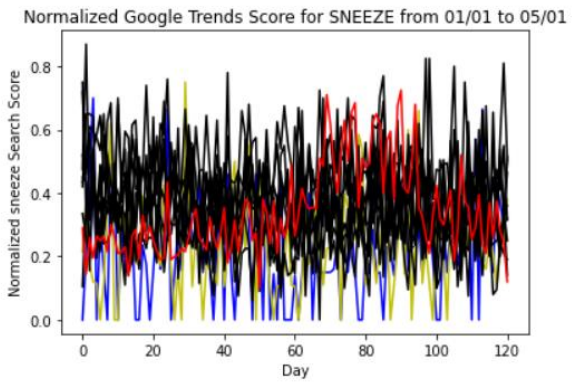
Plot 12 shows the normalized trend for FLU symptoms searches. The flu trendline does have the same characteristic peak, but it also follows the predicted line very closely in the beginning of Plot 24. Then after the peak around day 70, the flu symptom searches actually went below the predicted line in the most recent weeks. The peak above the line at the same time as the other symptoms does indicate a link to the virus, but the recent dip may indicate something about the social distancing measures. People who are staying at home across the state are less likely to catch the flu because it transmits similarly to the virus.

A separate category (kw_list14) which contains the keyword entries “covid test sites,” “coronavirus test sites” and “coronavirus testing” is included. Its normalized trendline is plotted in Plot 13 which shows the peak around Day 75-80 and a subsequent decline till present (the decline is briefly interrupted by a small local maximus around Day 110).

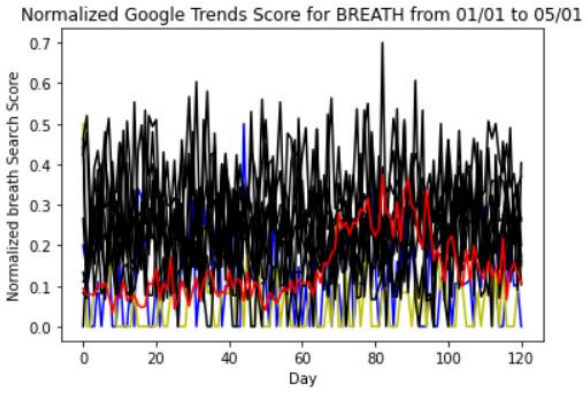
As a side note, Plots 26 and 27 show the predicted FLU cluster for the years 2009 and 2010 respectively. After around 110 days into 2009, there is a sudden spike in flu searches that go beyond the predicted value based on the subsequent years. This coincides with the H1N1 pandemic which was declared a national emergency by the WHO on April 25, 2009. This continues with peaks above the predicted trendline in the beginning of 2010. The actual flu searches in 2010 stabilize back to the predicted trendline in the middle of the year, when the WHO announced the end of the pandemic on August 11.

Overall, we were able to see a clear increase in number of searches for most of the word clusters we had chosen. The peaks on each graph appeared in the same time period of the first 70-80 days of this year and were significantly higher than the predicted trend. These results indicate that the pandemic has had a measurable effect on search data which makes it a potential feature to consider when modeling the infection rate of the virus. Some of the peaks investigated also directly correspond to the dates of media reports; It is clear that journalistic verbiage has a significant impact on specific search entries. Our project is also limited by the lack of access to nonpublic health data from both Google and elsewhere. If we had more time, we could have tried this method for more states, and perhaps regions in other countries. We can see how search data is affected in NY state, but how would it be affected in regions with limited access to the internet? The next natural question to ask is how well search data can predict infectivity, if at all? This would require datasets about the virus itself, such as the New York Times public dataset which provides daily confirmed infection data around the world.

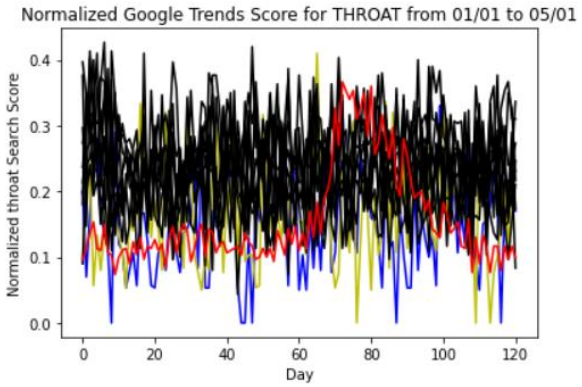
Appendix



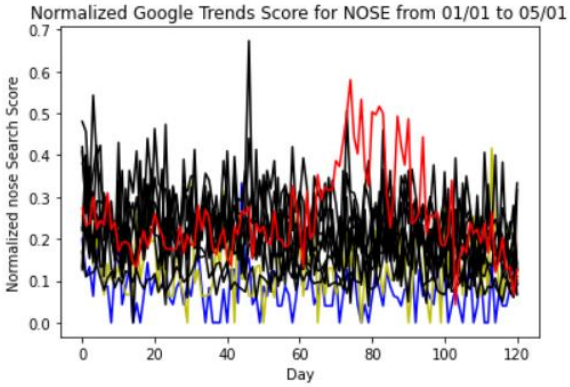
Plot 1



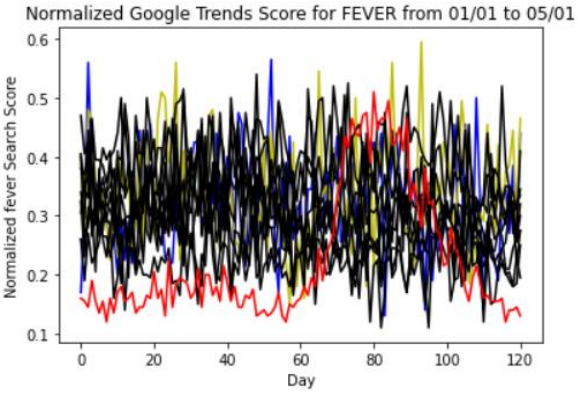
Plot 2



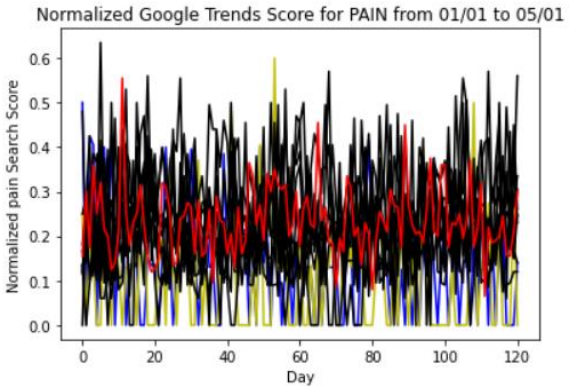
Plot 3



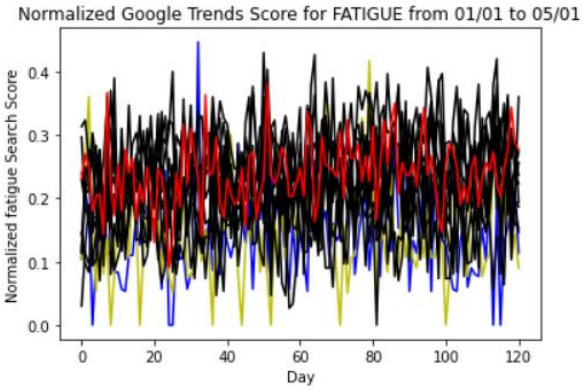
Plot 4



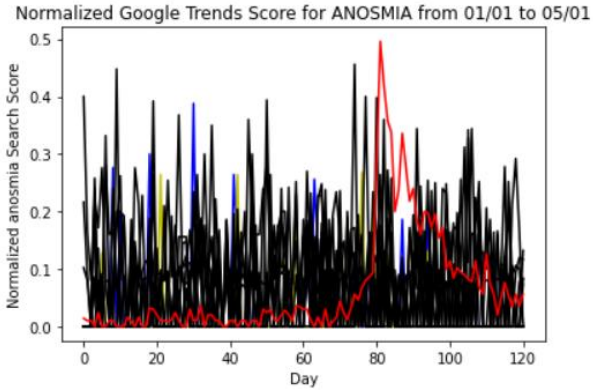
Plot 5



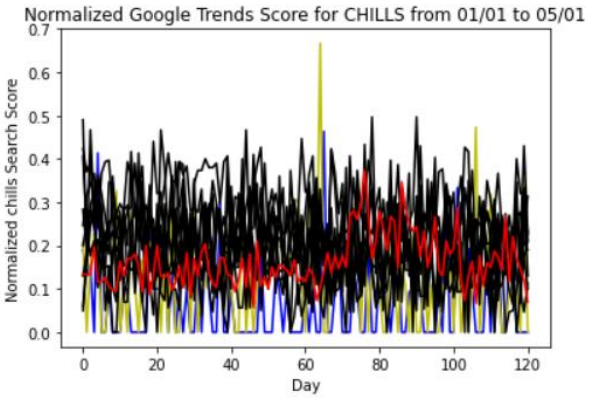
Plot 6



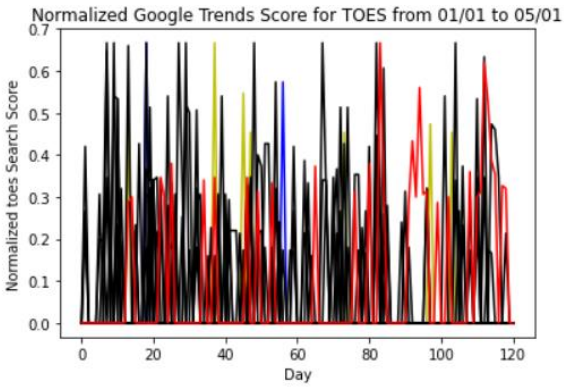
Plot 7



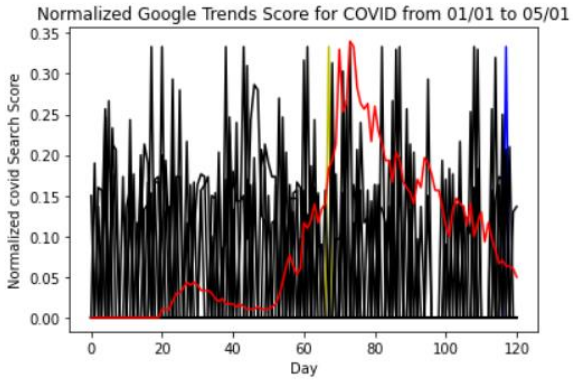
Plot 8



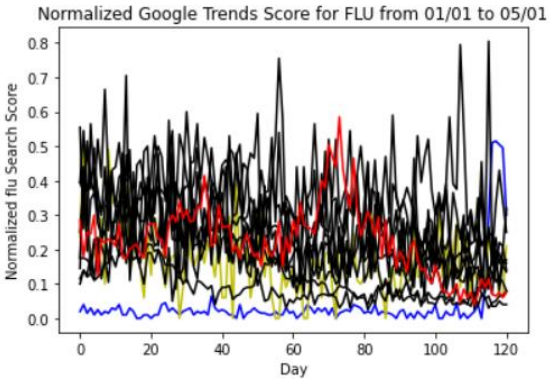
Plot 9



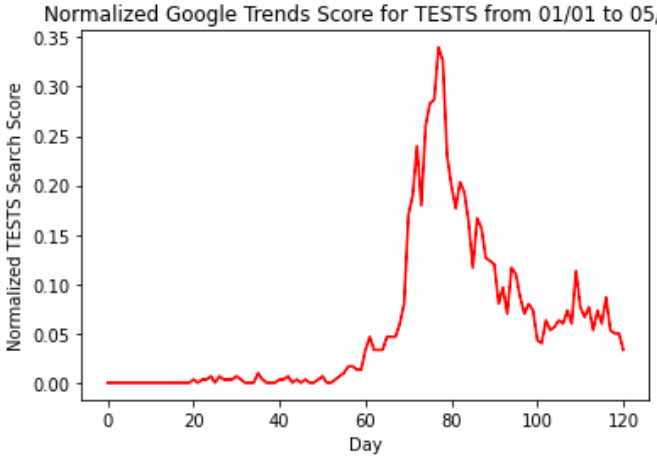
Plot 10



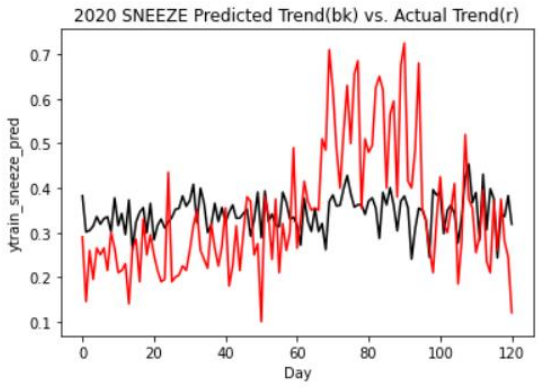
Plot 11



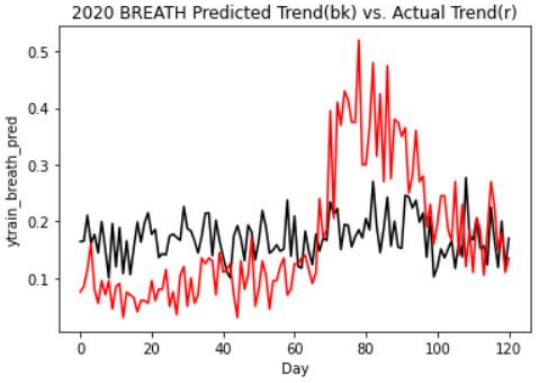
Plot 12



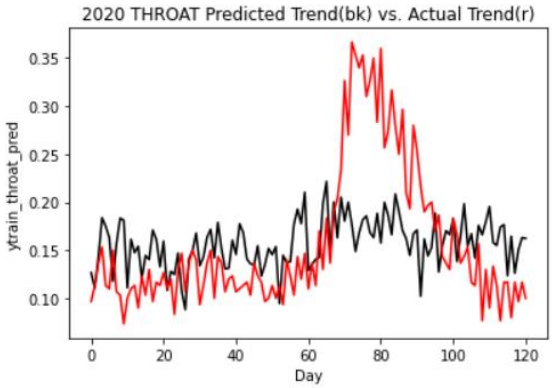
Plot 13



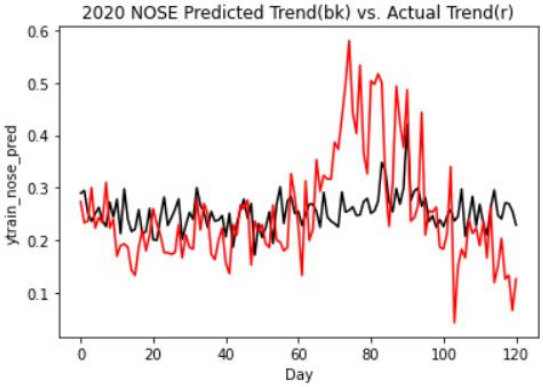
Plot 14: SNEEZE MLR



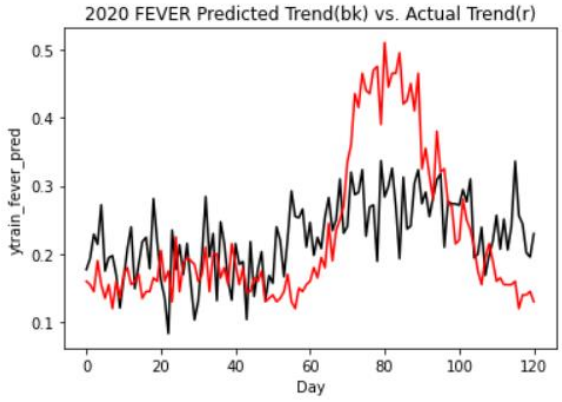
Plot 15: BREATH MLR

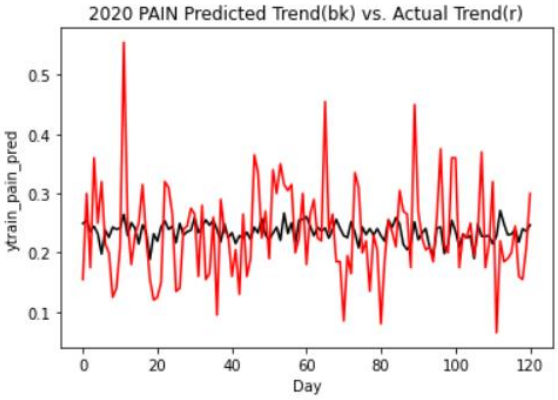


Plot 16: THROAT MLR

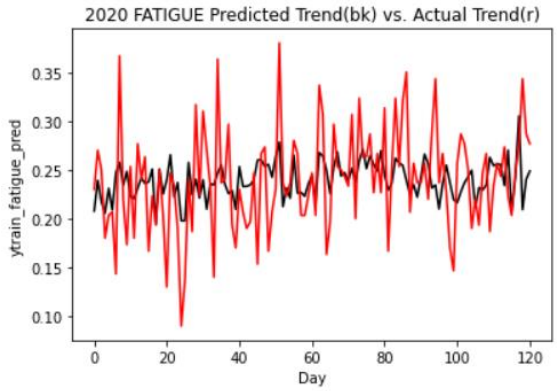


Plot 17: NOSE MLR

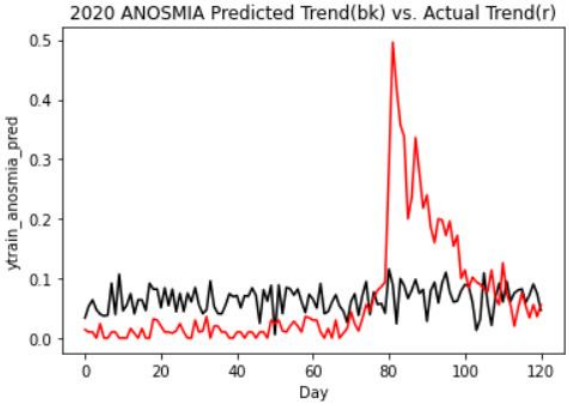




Plot 18: FEVER MLR

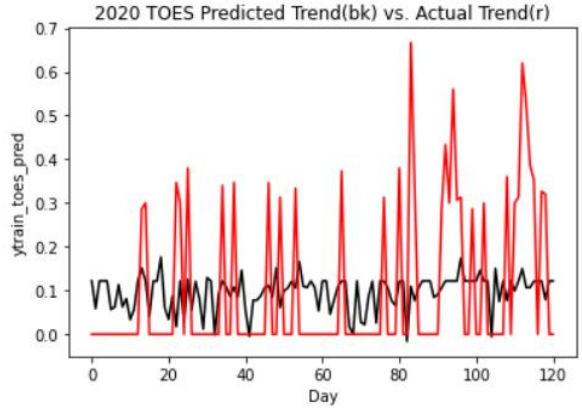


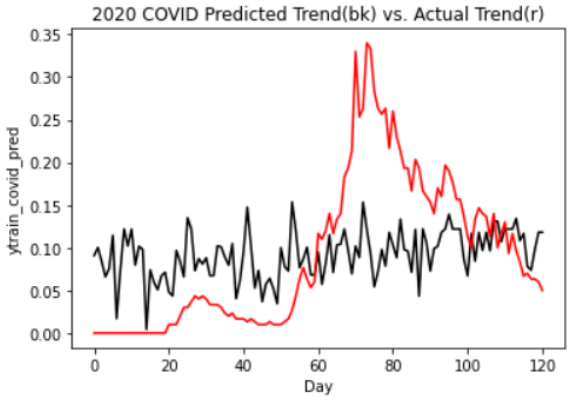
Plot 19: PAIN MLR



Plot 20: FATIGUE MLR

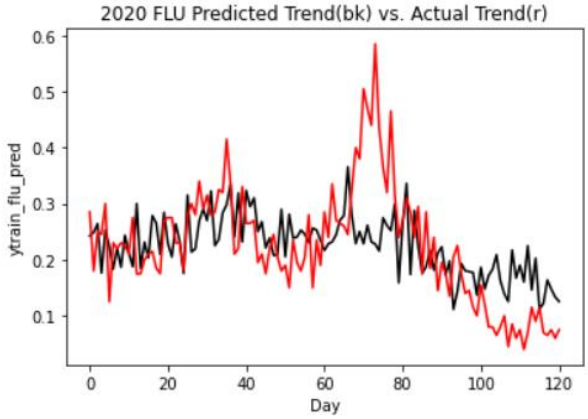
Plot 21: ANOSMIA MLR

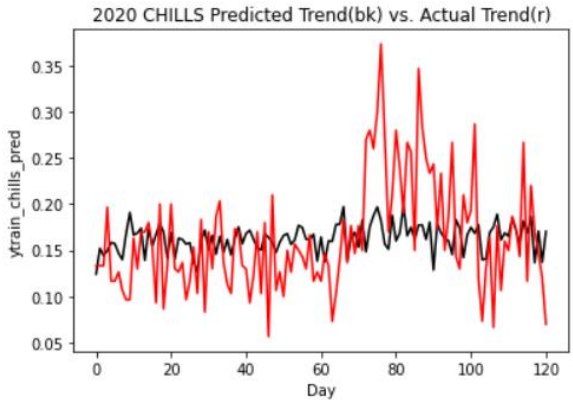




Plot 22: TOESMLR

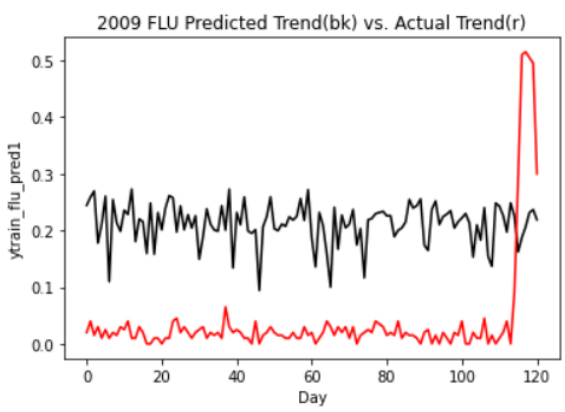
Plot 23: COVID MLR



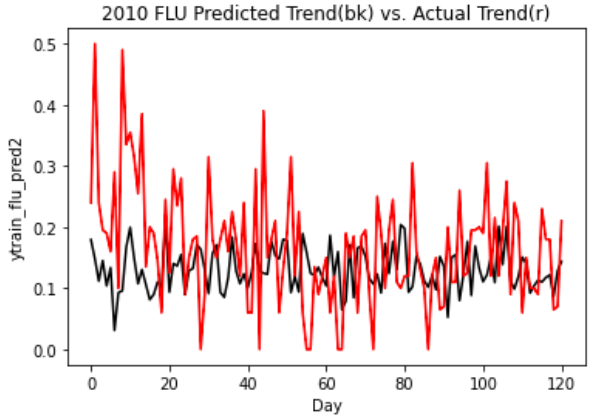


Plot 24: FLU MLR

Plot 25: CHILLS ML



Plot 26: FLU 2009



Plot 27: FLU 2010

	Year	L2 Loss		Year	L2 Loss
Cough	2020	61.1307	Fatigue	2020	46.1163
	2009	45.6124		2009	87.558
	2010	37.3042		2010	102.2788
Sneeze	2020	312.0063	Anosmia	2020	137.2568
	2009	412.9398		2009	88.1137
	2010	383.4195		2010	46.2619
Breath	2020	218.1355	Chills	2020	55.837
	2009	321.6851		2009	143.3933
	2010	323.3119		2010	213.7566
Throat	2020	88.0176	Toes	2020	435.6777
	2009	87.7033		2009	111.8567
	2010	100.7921		2010	215.5508
Nose	2020	173.3558	COVID	2020	127.3174
	2009	50.5646		2009	14.6446
	2010	106.2372		2010	13.6823
Fever	2020	214.9768	FLU	2020	193.9647
	2009	117.6874		2009	154.8028
	2010	129.3189		2010	162.5074
Pain	2020	93.9891			
	2009	236.1354			
	2010	273.3966			

Figure 8