# Study of inter-regional relations for various disease symptoms generated by Search Trends

Anshuman Sinha
Georgia Institute of Technology, Atlanta, US

Bhavay Aggarwal
Georgia Institute of Technology, Atlanta, US

## ABSTRACT

The use of hospital resources and the development of management plans to best manage infected patients depend on accurate forecasting of COVID-19 cases. However, monitoring sensors is an excellent method to measure the spread of the disease; it is an expensive task and has various privacy and ethical issues. A low-cost alternative to sensors in monitoring the sensors can be leveraging search trends on disease symptoms. It is not far-fetched to think that people with such symptoms would google them to get more information and potential remedies or cures. If this assumption is experimentally validated, it becomes possible to forecast the disease by forecasting symptoms related to Covid-19. In this work, we look at how the disease forecasting problem has been previously approached and how search trends data can be incorporated into it. We also look at predicting pharmacological demands such as vaccines, hospital beds, and medical requirements. Moreover, we also look at the possibilities of including misinformation from media on search trends in order to improve our model.

## KEYWORDS

Covid-19, Google search trends, Epidemiology, Deep learning, Time-series, Disease forecasting

**ACM Reference Format:**
Anshuman Sinha and Bhavay Aggarwal. 2022. Study of inter-regional relations for various disease symptoms generated by Search Trends. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 WORK DISTRIBUTION

- Conceptualisation: Equal btw Anshuman S. and Bhavay A.
- Writing: Equal between Anshuman S and Bhavay A.
- Data Collection: Equal between Anshuman S and Bhavay A.
- Work: Initial Findings and Statistical summary, Trends analysis, IMF analysis: Bhavay A.
- Work: Embedding generation and Random projection, Lead time analysis, Spatio-temporal study: Anshuman S.

## 2 INTRODUCTION

The prediction, modelling and containment of the spread of diseases is one of the most important and challenging problems of the modern world. In our previously submitted project proposal, we formulated our problem of 'Studying the inter-regional relations for various disease symptoms generated by search trends data in the context of Covid-19'. The overall study is divided into the following sub-parts:

- **Studying correlations:** Finding the top-k most correlated symptoms from our trends data. (details of the process given in section 6 and 7) And further expanding the study to inter-regional correlation of symptoms and cases.
- **Inference study:** Based on the study of co-relation we observe the various features of the pandemic such as:
  - i **Phase changes** and its spatial correlation between neighbouring states. We will study and compare the trends in search data and Covid cases; And also study the effect of changing trends on neighbouring states.
  - ii Study of **lead time** and inferring how search data from a better connected state can be used to restrict surge in cases of the neighbouring states.And extending this to find lead time over surge in vaccine demands of the state.
  - iii Correlation between search trends and **targets** like, hospitalisation, severe infections, mortality etc.
- **Predictions:** In this work, we will perform Covid-19 prediction in the following cases.
  - i Prediction of Covid-19 cases with the help of trends of most correlated symptoms of the state and extending it to prediction based on neighbouring states.
  - ii Comparing the results of lead time with our predicted cases and providing a validation to our model.
  - iii Predicting the demand of pharmacological amenities such as vaccines, drugs based on search trends; such that the disease can be better managed.
- **Model improvement:** Study the impact of media influence (such as twitter trends) on search trend's shortcoming of over-predict the related diseases and eventually its effects on estimates of Covid-19 (In sense, studying the impact of miss-information).

We will be working on the above mentioned topics following the same order as listed above. As of now, we have started working on Inference study (finished most of the parts, results shown later in the draft).

## 3 RESPONSE TO INITIAL COMMENTS

- ***'Try to get papers from well known journals and improve the quality of work'***: We have read and implemented research papers from better journals and conferences in various sections of our work. Such as exploring the possibilities of using other data-sets[13] [6] [10] , Implementing algorithms to better our current model [12] [11] [3], Improving performance of our current model by implementing better feature engineering.[2] [8]
- ***'Make sure you compute correlations and predictions in a real-time..'***: This is what we had initially planned to do but did not mention it explicitly in our proposal. Our LSTM

*Conference'17, July 2017, Washington, DC, USA*
2022. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

model defines a time-steps that the model we look in the past while predicting any future value. Somewhat like a sliding window over the data-set.

- ***'Project needs to be expanded a bit..'***: We had initially planned on predicting Covid cases of a selected state from the Google trends data of its neighboring states and also to study the effect of miss information like media trends affect the trends in symptoms search data. Now, we will be extending our work to
  - i Trends study: Phase changes and its effects.
  - ii Lead time analysis: Predicting lead time through search trends.
  - iii Targets study: Hospitalisation, Vaccine requirements etc.
  - iv Anomaly detection: Anomalous data based on trends of neighbouring states.
  - v Spatio-temporal study: Studying the spatial effect of all the above mentioned points.
- ***'There might be a CS problem here on finding top-k correlated time-series efficiently'***: Although there can be many approach towards solving this problem, we have decided not use any predefined library but to write our own method. In this work are using Random projection method as mentioned in the Query stream search paper by Chien et al. [2] (Which was mentioned in our comments)

## 4 REVIEW OF OUR PREVIOUS WORK

In the proposal submission, we had formulated the problem and defined our approach towards the solution of our problem (Refer Problem formulation and Approach section of Proposal).

We had also simulated a basic LSTM model for predicting Covid 19 cases in United States based on Google trends data of keyword "Covid" from the start of year 2022 to 50 weeks into the year. The model seems to match the trends plots of Covid 19 daily cases. The preliminary analysis shows decent results due to the good correlation between GT trends and original Covid 19 cases. Figure 1 and 2 shows our work from the initial model.

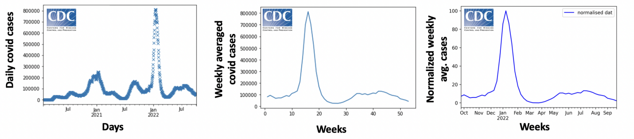

**Figure 1: Number of daily Covid case data obtained from CDCs website.**

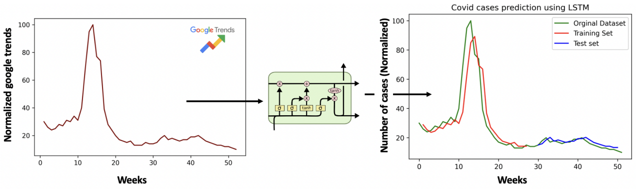

**Figure 2: Prediction from a basic LSTM model.**

## 5 DATA COLLECTION APPROACH

In the current work will be implementing data from various search sources mentioned in table 1. These datasets are implemented according to the model, such as trends data from Google and Facebook are used as symptomatic predictions, Hospital capacity data for studying the measure of targets , The media coverage data in improving the model etc. The weekly covid cases data was downloaded from CDC. Daily google trends data was downloaded from 2020 to 2022 and resampled into a weekly format. As detailed by the authors in [1], the search trend is defined as the ratio between the normalized search volume for a given symptom in a given region during a given week and the median search volume for that region-symptom pair. This is done to ensure that the google search are comparable between regions.The timeframe of the data was restricted to start from $2020 - 01 - 26$ and end at $2022 - 10 - 13$. These dates coincide with the first week of covid cases and the last week of reported covid cases available on CDC and consists of a total of 141 weeks. We also used the daily cases in prediction of lead time, by converting the data to lower a dimension first and then applying the later algorithms. We plan to begin incorporating the other mentioned datasets in the future.

## 6 DESCRIPTION OF INITIAL FINDING AND STATISTICAL SUMMARY OF DATASET:

Examining the relationship between Google Trends and the occurrence of COVID-19 is the first step in determining the usability of Google Trends data in the predictability of COVID-19. The Pearson correlation coefficients (r) between the ratio (COVID19 deaths)/(COVID-19 cases) and Google Trends data are calculated since Pearson correlation analysis is the benchmark analysis in this kind of methodology. We perform our experiments on the national level and also for the states - California, New York, Texas, Alaska, Mississippi and Georgia.

We first checked the PCC (similar to the work of [9] and [5]) of the weekly covid cases with the search trends for all the symptoms. The top-10 symtoms and the corresponding PCC values are listed in table 2 This approach however has a major drawback, it assumes that the variance in the data is homogeneous across the data range.

We observed that many unrelated symptoms ended up having a high correlation value. Although, these symptoms can be manually filtered, we wanted to look at how correlated the symptoms in smaller windows over the entire time frame. We created rolling windows of 4 weeks for both the datasets and performed Pearson's correlation on the symptom window with the corresponding cases window (table 3). The table shows the symptoms which appeared the most in the top-15 most correlated symptoms. 50% indicates the median PCC value across all windows and is the metric we focus on(symptom with maximum value highlighted for each region). For the selected regions, Ageusia has the highest median PCC value majority of the times, so it is a symptom we can explore a bit more. The disease "ageusia" refers to the loss of sense of taste and is prominent symptom among people infected with covid. Moreover, the other top symptoms are also prominent symptoms of covid. This result shows the rolling window/local PCC is an effective method

| Application | Data type | Publisher | Link |
|---|---|---|---|
| Covid 19 cases data | COVID-19 statistics | Johns Hopkins University | github/CSSEGISandData/COVID-19 |
| Covid 19 cases data | COVID-19 statistics | CDC | cdc/2019-ncov |
| Measuring emotions | Textual/Embeded data | COVID-19 Real World Worry | github/ben-aaron188/covid19worry |
| Media Coverage | National and International news article and interpretation | Humanities and social sciences communication [nature] | nature/News-media-coverage-of-COVID-19 |
| Search trends data | Normalised keyword search data | Google | google/search-trends |
| COVID-19 Trends and Impact Survey | Symptoms and behaviour data | Facebook and Delphi CMU | cmu-delphi/symptom-survey |
| Hospital Capacity Data | Number of beds occupied and available state wise data | U.S. Department of Health & Human Services | healthdata/capacity-bed-data |

**Table 1: Data Sources**

| Symptom | PCC |
|---|---|
| Hypoxemia | 0.643161 |
| Eye pain | 0.635909 |
| Night sweats | 0.624184 |
| Nasal congestion | 0.582434 |
| Sinusitis | 0.545885 |
| Bradycardia | 0.525026 |
| Throat irritation | 0.515732 |
| Ageusia | 0.500023 |
| Post-nasal drip | 0.491759 |
| Anosmia | 0.483690 |

**Table 2: Global Person’s Correlation between symptoms and covid cases**

to extract the most correlated symptoms.

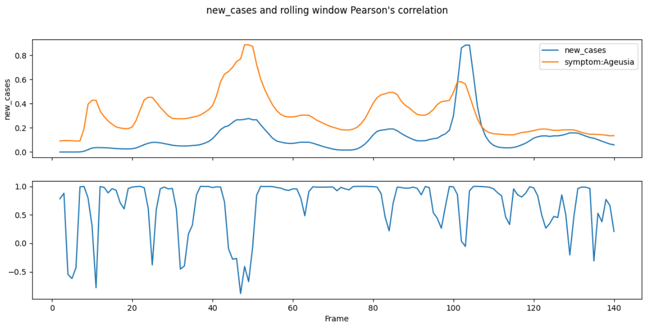

**Figure 3: Plot of local window search trends of Ageusia along with the weekly covid cases for the entire nation. The bottom plot shows the PCC for the rolling window.**

Next up, we perform the EMD analysis as decribed by [14], [4] and [7] to determine the statistical significance of the input data. The benefit of using the EMD method is that it can eliminate noise and non-stationary oscillations that are unrelated to features of the trend (for instance, informative searches and influential search) that do not correspond to the actual disease (Fever, Asthma etc.) or to the related disease (Covid 19). We removed the lower order IMF’s and combined the first and second order IMF to form the new curve for each symptom’s google trends. We again performed local PCC on the new symptom trend curves with the covid cases. We compared the correlation obtained this way with the original local PCC and observed that the correlation was comparatively lower. This suggests that the assumption we made that the oscillations were noise in the trends data was incorrect and these oscillations are important trends. Our next step to limit the overrepresentation of symptom search trends induced by media/social media would be to incorporate the media coverage data listed in table 1.

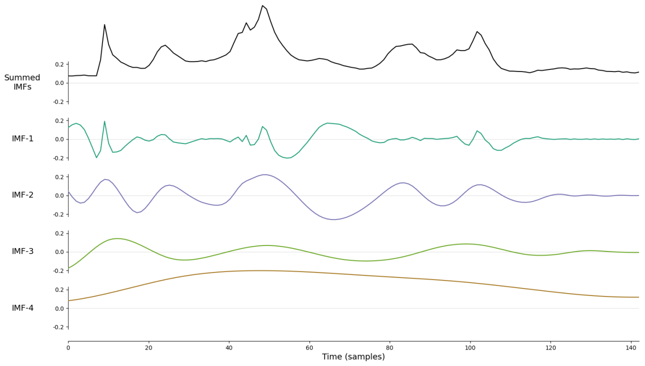

**Figure 4: Google search trends for "Ageusia" divided into its IMF components**

## 7 DESCRIPTION OF MATHEMATICAL BACKGROUND NECESSARY FOR THE PROBLEM:

A background knowledge of Statistics, Probability and Linear algebra is required to understand the working of our models. We briefly

| | | Low-grade fever | Fever | Pneumonia | Ageusia | Hypoxemia | Dysgeusia | Anosmia | Shortness of breath |
|---|---|---|---|---|---|---|---|---|---|
| National | mean | 0.55 | 0.47 | 0.41 | 0.65 | 0.56 | 0.49 | 0.46 | 0.36 |
| | 25% | 0.38 | 0.02 | -0.01 | 0.46 | 0.22 | 0.19 | 0.00 | -0.02 |
| | 50% | 0.86 | 0.82 | 0.76 | **0.92** | 0.85 | 0.73 | 0.72 | 0.51 |
| | 75% | 0.96 | 0.96 | 0.95 | 0.98 | 0.97 | 0.93 | 0.97 | 0.93 |
| California | mean | 0.42 | 0.39 | 0.28 | 0.57 | 0.52 | 0.43 | 0.40 | 0.30 |
| | 25% | 0.01 | -0.02 | -0.32 | 0.37 | 0.28 | 0.02 | -0.02 | -0.11 |
| | 50% | 0.62 | 0.61 | 0.52 | **0.81** | **0.81** | 0.62 | 0.61 | 0.42 |
| | 75% | 0.95 | 0.92 | 0.91 | 0.96 | 0.96 | 0.93 | 0.92 | 0.89 |
| Texas | mean | 0.35 | 0.30 | 0.45 | 0.45 | 0.49 | 0.36 | 0.28 | 0.30 |
| | 25% | -0.07 | -0.17 | 0.12 | 0.12 | 0.00 | 0.00 | -0.23 | -0.12 |
| | 50% | 0.57 | 0.57 | 0.68 | 0.68 | **0.76** | 0.54 | 0.48 | 0.49 |
| | 75% | 0.87 | 0.89 | 0.94 | 0.94 | 0.94 | 0.87 | 0.89 | 0.86 |
| New York | mean | 0.62 | 0.49 | 0.35 | 0.62 | 0.62 | 0.42 | 0.50 | 0.43 |
| | 25% | 0.38 | 0.17 | -0.03 | 0.40 | 0.38 | 0.03 | 0.23 | 0.05 |
| | 50% | 0.85 | 0.76 | 0.65 | **0.86** | 0.85 | 0.66 | 0.73 | 0.65 |
| | 75% | 0.95 | 0.93 | 0.93 | 0.97 | 0.95 | 0.88 | 0.94 | 0.91 |
| Alaska | mean | 0.13 | 0.31 | 0.26 | N/A | N/A | N/A | N/A | 0.14 |
| | 25% | -0.58 | -0.06 | -0.18 | N/A | N/A | N/A | N/A | -0.31 |
| | 50% | 0.32 | **0.57** | 0.35 | N/A | N/A | N/A | N/A | 0.24 |
| | 75% | 0.80 | 0.84 | 0.84 | N/A | N/A | N/A | N/A | 0.60 |
| Mississippi | mean | 0.33 | 0.36 | 0.21 | 0.24 | 0.30 | 0.16 | 0.25 | N/A |
| | 25% | 0.00 | -0.01 | -0.28 | -0.25 | -0.12 | -0.30 | -0.24 | N/A |
| | 50% | 0.48 | **0.62** | 0.26 | 0.43 | 0.51 | 0.21 | 0.40 | N/A |
| | 75% | 0.87 | 0.91 | 0.80 | 0.86 | 0.87 | 0.67 | 0.86 | N/A |
| Georgia | mean | 0.45 | 0.40 | 0.33 | 0.53 | 0.50 | 0.33 | 0.46 | N/A |
| | 25% | 0.01 | -0.04 | -0.08 | 0.18 | 0.02 | -0.03 | 0.02 | N/A |
| | 50% | 0.76 | 0.70 | 0.62 | 0.71 | **0.81** | 0.43 | 0.70 | N/A |
| | 75% | 0.94 | 0.94 | 0.92 | 0.97 | 0.97 | 0.85 | 0.92 | N/A |

**Table 3: Local Person's Correlation between symptoms and covid cases**

point the highlights of the fundamentals which we have used in our current work:

## 7.1 Random projections

Since, our time-series data was of very high dimension (e.g. Dim 52 for Weekly data and Dim 365 for Daily data) we are required to lower the dimension of the data as well as transform it such that it takes less space.

Random projection is used to project the time-series vector on a random hyperplanes, such that we can evaluate which side of the hyperplane these vectors are lying. Further, we randomly pick a sufficient number of hyperplanes from a Gaussian distribution to remove the stochasticness involved in the process. Figure 4, shows the schematic of this process.

## 7.2 Johnson-Lindenstrauss Lemma

In order to understand that vector representation on the low dimensional subspace (or on the hyperplane) is similar to the original distance between these two vectors we need the Johnson-Lindenstrauss Lemma.

Informally, the lemma states that we can map N points into a much smaller Euclidean space, specifically one of logarithmic dimension, while maintaining the pairwise Euclidean distances of the given points up to a multiplicative factor of 1 +. N points are defined as being in the N-dimensional Euclidean space.

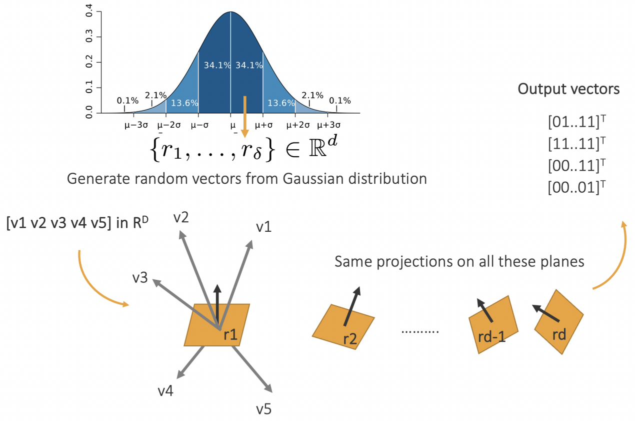

**Figure 5: A pictorial representation of the main question that the Johnson-Lindenstrauss Lemma resolves.**

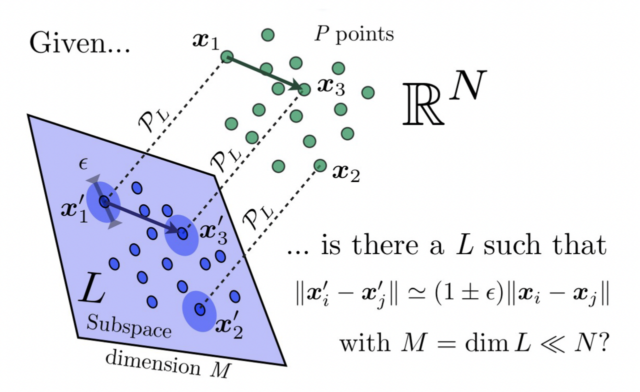

**Figure 6: A pictorial representation of the main question that the Johnson-Lindenstrauss Lemma resolves.**

## 7.3 Statistical correlation

In order to find how similar the google search trends is similar to the weekly covid cases, we are using Pearson's correlation. Pearson's correlation measures the statistical relationship between two continuous variables and outputs a number between −1 and 1 that measures the strength and direction of the relationship between two variables.

$$\textit{Pearson's correlation coefficient}(\rho) =$$
$$covariance(X, Y)/(stdv(X) * stdv(Y))$$
$$covariance(X, Y) = (sum(x - mean(X)) * (y - mean(Y))) * 1/(n - 1)$$

# 8 FORMAL DESCRIPTION OF IMPORTANT ALGORITHMS USED

## 8.1 Binary Embedding using LSH

- Make k random vectors of length d each, where d is the feature vector's dimension and k is the size of the bitwise hash value.
- Calculate the dot product of the random vector and the observation for each random vector. If the dot product result is positive, set the bit value to 1; otherwise, set it to 0.
- Concatenate all of the bit values that were calculated for the k-dot products.

## 8.2 Spatial effects on lead time

- Find the top-k most correlated symptoms of a state.
- Find the top-k most correlated symptoms of the neighbouring states.
- Compute the lead time of each neighbouring state from their symptoms.
- Re-compute the lead time of each neighboring state from the symptom of the original state.
- Evaluate the gain in lead time for each neighbouring state.

| National data | Top k symptoms | Match | Lead time (t*) |
|---|---|---|---|
| 1 | Anosmia | 18 | 271 |
| 2 | Acute bronchitis | 17 | 267 |
| 3 | Ageusia | 17 | 271 |
| 4 | Bronchitis | 17 | 267 |
| 5 | Chills | 17 | 271 |
| 6 | Common cold | 17 | 267 |
| 7 | Cough | 17 | 267 |
| 8 | Dysgeusia | 17 | 270 |
| 9 | Fever | 17 | 14 |
| 10 | Hemoptysis | 17 | 270 |
| 11 | Hypoxemia | 17 | 271 |
| 12 | Low-grade fever | 17 | 7 |
| 13 | Nasal congestion | 17 | 271 |
| 14 | Pneumonia | 17 | 6 |
| 15 | Post-nasal drip | 17 | 15 |

**Table 4: Lead Time of symptom search trends on national weekly covid cases.**

# 9 SOME INITIAL RESULTS

## 9.1 Lead time analysis

The lead time analysis was performed on the National covid (daily cases) data with respect to the symptoms data (for year 2022, Jan to Oct; 274 dimesnion vector). We first found the most correlated symptom with the help of top-15 symptom search (using dimension reduction) and then analyse the lead time on the original data wrt. to these most correlated symptoms. Table 4 shows the result which shows logical lead time wrt. some important symptoms like 'Fever', 'Low-grade Fever' etc and many of these search trends tend not to align well with the original data (hence, we get very high lead times)

## 9.2 Spatio-temporal study

We also studied the most correlated symptom of neighbouring states and whether they give us better lead time than the original state or not! For this study we took neighbours of Georgia (Alabama and South Carolina) and studied the lead time of these neighbouring states wrt. the symptoms searches in Georgia. The results are detailed in tables 5 and 6 for Alabama and South Carolina respectively.

Here, we observe that we get a better lead time for the neighbors with the most correlated symptoms of Georgia than from their own symptoms search trend. These results show the possibility of spread of covid from Georgia in these states, and also tells that we can get a better insight (lead time) with a neighbouring state (which maybe better connected to other states of the country) and make use of this gain in lead time to address the forthcoming situation in the state.

# 10 GENERAL DIFFICULTIES IN OUR WORK

- Top k searches : We were getting very high dimensional data (especially when considering daily cases).

| Top k symptoms | Match | $t^*$ | $t^*$ wrt Georgia | Gain in $t^*$ |
|---|---|---|---|---|
| Ageusia | 27 | 10 | 11 | 1 |
| Hypoxemia | 27 | 8 | 8 | 0 |
| Shortness of breath | 27 | 49 | 47 | -2 |
| Anosmia | 26 | 4 | 10 | 6 |
| Bradycardia | 26 | 49 | 49 | 0 |
| Erectile dysfunction | 26 | 49 | 49 | 0 |
| Tachycardia | 26 | 49 | 49 | 0 |
| Chest pain | 25 | 47 | 49 | 2 |
| Halitosis | 25 | 44 | 44 | 0 |
| Headache | 25 | 13 | 15 | 2 |
| Middle back pain | 25 | 49 | 49 | 0 |
| Pruritus ani | 25 | 47 | 49 | 2 |
| Dysgeusia | 24 | 10 | 11 | 1 |
| Hair loss | 24 | 48 | 49 | 1 |
| Hypoxia | 24 | 49 | 49 | 0 |

**Table 5: Lead Time of Georgia's symptom search trends on weekly covid cases in Alabama.**

| Top k symptoms | Match | $t^*$ | $t^*$ wrt Georgia | Gain in $t^*$ |
|---|---|---|---|---|
| Chest pain | 25 | 49 | 49 | 0 |
| Shortness of breath | 25 | 12 | 47 | 35 |
| Chills | 24 | 0 | 0 | 0 |
| Dysgeusia | 24 | 9 | 13 | 4 |
| Hypoxemia | 24 | 7 | 9 | 2 |
| Ageusia | 23 | 10 | 15 | 5 |
| Bradycardia | 23 | 47 | 49 | 2 |
| Erectile dysfunction | 23 | 49 | 49 | 0 |
| Eye pain | 23 | 48 | 11 | -37 |
| Low-grade fever | 23 | 0 | 12 | 12 |
| Night sweats | 23 | 47 | 47 | 0 |
| Tachycardia | 23 | 49 | 49 | 0 |
| Anosmia | 22 | 5 | 10 | 5 |
| Fever | 22 | 5 | 13 | 8 |
| Hemoptysis | 22 | 48 | 1 | -47 |

**Table 6: Lead Time of Georgia's symptom search trends on weekly covid cases in South Carolina.**

- Data collection : The data sources suitable for our work was initially difficult to find (Some had permission issues, e.g. AWS data)
- Anomaly prediction not easily quantifiable(as time-series is limited) : In our earlier approach, we found that the dataset for Covid is still not having sufficient amount of seasons to detect anomoly. But now we are using anomaly prediction in a more general sense wrt. to neighbouring states symptoms.

## 11 FUTURE WORK

- We have explored whether google search trends are correlated with covid cases on different spatial levels, and we now aim at including the other trends data sets in the upcoming weeks.
- We have found several correlated symptoms which also match with the covid-19 symptoms listed by the CDC. We have utilized these symptoms to study lead time at a national level and between neighboring states.
- Moving forward, we want to consolidate our top-k correlated symptoms and perform our subsequent experiments using them. We will further expand our spatio-temporal study of the impact of neighboring states google trends to study lead time.
- We have to also begin to work on forecasting models.
- Finally we also have to address model improvement (as described in the section 'Introduction')

More detail can be found on the Introduction section.

## REFERENCES

[1] Shailesh Bavadekar, Andrew Dai, and Davis et al. 2020. Google COVID-19 Search Trends Symptoms Dataset: Anonymization Process Description (version 1.0). https://doi.org/10.48550/ARXIV.2009.01265

[2] Steve Chien and Nicole Immorlica. 2005. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th international conference on World Wide Web*. 2–11.

[3] Vinay Kumar Reddy Chimmula and Lei Zhang. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals* 135 (2020), 109864.

[4] Ramon Gomes Da Silva, Matheus Henrique Dal Molin Ribeiro, Viviana Cocco Mariani, and Leandro dos Santos Coelho. 2020. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. *Chaos, Solitons & Fractals* 139 (2020), 110027.

[5] Jacques Demongeot, Yannis Flet-Berliac, and Hervé Seligmann. 2020. Temperature decreases spread parameters of the new Covid-19 case dynamics. *Biology* 9, 5 (2020), 94.

[6] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[7] Najmul Hasan. 2020. A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model. *Internet of Things* 11 (2020), 100228.

[8] Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. 604–613.

[9] Amaryllis Mavragani et al. 2020. Tracking COVID-19 in Europe: infodemiology approach. *JMIR public health and surveillance* 6, 2 (2020), e18941.

[10] William JM Probert, Chris P Jewell, Marleen Werkman, Christopher J Fonnesbeck, Yoshitaka Goto, Michael C Runge, Satoshi Sekiguchi, Katriona Shea, Matt J Keeling, Matthew J Ferrari, et al. 2018. Real-time decision-making during emergency disease outbreaks. *PLoS computational biology* 14, 7 (2018), e1006202.

[11] Siva R Venna, Amirhossein Tavanaei, Raju N Gottumukkala, Vijay V Raghavan, Anthony S Maida, and Stephen Nichols. 2018. A novel data-driven model for real-time influenza forecasting. *Ieee Access* 7 (2018), 7691–7701.

[12] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D Corley. 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one* 12, 12 (2017), e0188941.

[13] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1

[14] Albert C Yang, Jong-Ling Fuh, Norden E Huang, Ben-Chang Shia, Chung-Kang Peng, and Shuu-Jiun Wang. 2011. Temporal associations between weather and headache: analysis by empirical mode decomposition. *PloS one* 6, 1 (2011), e14612.