Expanding eVision's Granularity of Influenza Forecasting

Navid Shaghaghi(✉), Andres Calle, George Kouretas, Supriya Karishetti, and Tanmay Wagh

BioInnovation and Design Laboratory, Santa Clara University, Santa Clara, CA 95053, USA
{nshaghaghi,acalle,gkouretas,skarishetti,twagh}@scu.edu

Abstract. According to the United States' Center for Disease Control and Prevention (CDC) between 39 and 56 million people in the US suffered from Influenza Like Illnesses (ILI) in the 2019-20 flue season. From which, 410 to 740 thousand were hospitalized and 24 to 62 thousand succumbed to the disease. Therefore, the existence of an early warning mechanism that can alert pharmaceuticals, healthcare providers, and governments to the trends of the influenza season well in advance, would serve as a significant step in helping combat this communicable disease and reduce mortality from it.

As reported in the [ACM Special Interest Group in Computers and Society (SIGCAS) 2020 Computers and Sustainable Societies (COMPASS)], [IEEE Technology and Engineering Management Society (TEMS) 2020 International Conference on Artificial Intelligence for Good (AI4G)], and [IEEE Global Humanitarian Technology Conference (GHTC) 2020] Long Short-Term Memory (LSTM) neural networks are utilized by Santa Clara University's EPIC (Ethical, Pragmatic, and Intelligent Computing) and BioInnovation & Design laboratories for continued research and development of an eVision (Epidemic Vision) machine learning tool to predict the trend of influenza cases throughout the flu season.

There we reported eVision's success in making 3, 7, and 14 weeks in advance predictions for the 2018–2019 United States flu season with 88.11%, 88%, and 74.18% accuracy respectively and delineated future steps of expanding eVision's granularity by 1) adding state level predictions in order to enhance national predictions and 2) utilizing metropolitan area keyword trends to improve both state level and national predictions. This resulted in the improvement of the model's accuracy to 90.38%, 91.43%, and 81.74% for 3, 7, and 14 weeks in advance predictions respectively. This paper is to report on the methodology of obtaining these improved results.

Keywords: Flu trend prediction · Google Trends · Health care technology · Influenza incident rate forecasting · Long Short-Term Memory (LSTM) neural networks · Medical machine learning

J. Ye et al. (Eds.): MobiHealth 2020, LNICST 362, pp. 227–243, 2021.
https://doi.org/10.1007/978-3-030-70569-5_14

1 Introduction

Influenza (a.k.a. the flu) is a pervasive respiratory infection caused by Influenza viruses with an estimated 3 to 5 million severe cases annually, which lead to between 290 to 650 thousand respiratory deaths world wide [12]. For the 2019–2020 US flu season, which started October 1, 2019, and ended April 4, 2020, the United States' Center for Disease Control and Prevention (CDC) estimates between 39 and 56 million cases of flu illness, which led to between 410 and 740 thousand hospitalizations and between 24 and 62 thousand deaths [2].

During the 2018–2019 flu season, influenza vaccines prevented between 3.4 and 7.1 million flu cases and, thus, prevented 30 to 156 thousand hospitalizations as well as 1 to 13 thousand deaths [3]. At the time of this writing, the 2019–2020 flu season's flu vaccine effectiveness statistics were not yet finalized and released by the the CDC.

Since Influenza vaccination is the primary strategy to prevent influenza [16], an accurate prediction model is essential for pharmaceutical companies and healthcare providers to be able to properly prepare for an upcoming flu season. For instance, vaccine manufacturers in the US rely heavily on seasonal influenza data provided by the CDC [1] which, due to the two-week reporting lag of the CDC, leaves the vaccine manufacturers insufficient time to produce enough flu vaccines for the appropriate flu strains that can be distributed through the health care network in time.

However, the CDC only collects US data and thus for the rest of the countries the World Health Organization (WHO)'s global estimates must be used as a basis for a prediction model. Though, improvements are required to gain more accurate results, as the WHO only extrapolates based off of the limited data it receives from the countries [12].

2 Related Work

Between 2008 and 2015, the Google Flu Trends project provided an influenza activities forecaster with a linear model [9]. The idea being that since many potential patients or relatives and friends of potential patients will use Google Searches as a first attempt at diagnosis, by monitoring a region's population's Google search queries into influenza related terminology and symptoms, the presence of ILI in the population of that region may be predicted. However, no actual flue statistics from the CDC or WHO were used to validate or enhance the predictions.

Ginsberg et al. estimated weekly influenza activities by finding and monitoring Google search queries that are highly correlated with CDC data, achieving an accurate estimate with a one-day reporting lag [8]. However, their aim was only to overcome the two week reporting lag of the CDC. No attempt was made to help predict future numbers of ILI cases.

Dugas et al. applied a generalized linear model to Google Trends data [6] on a city level. Similarly to Ginsberg et al. predictions of future influenza trends was not within the scope of the research.

Paul et al. used both Google Trends and Twitter data to forecast influenza outbreak, but because people usually only tweet about influenza after the outbreak has happened, their research can only be used for post-verification [17].

Xie used a vector auto-regression model which factors state population density, weekly temperature, and precipitation as predictors to forecast ILI incidence rate based on the Google Flu Trends and the CDC ILI incidence [23]. However, the goal of her project was not as narrowly focused as eVision. It does not aim to provide companies and health providers with an easily understood forecast of an upcoming influenza season, and as such it cannot provide a long term forecast.

3 Vector Autoregression (VAR) Model

Regression modeling is a technique which provides a relationship between dependent and independent variables. One such model is the Vector Auto Regression (VAR) model, which generalizes the uni-variate auto regression model by allowing it to include more than one predictor variable. The VAR model is thus an extension of the Autoregression model that is used to predict multiple time series variables using a single core model. Therefore, VAR helps in performing multivariate time series forecasting between multiple predictors and a response variable. This model works on the concept of lags, which means that each variable is a linear combination of past lags of itself and past lags of the other variables [18].

For example to measure three different time series variables, denoted by $x_{t,1}, x_{t,2}, x_{t,3}$. the Vector Autoregression model of order 1, denoted as VAR(1), is as follows:

$$\begin{aligned}
x_{t,1} &= \alpha_1 + \phi_{11} x_{t-1,1} + \phi_{12} x_{t-1,2} + \phi_{13} x_{t-1,3} + w_{t,1} \\
x_{t,2} &= \alpha_2 + \phi_{21} x_{t-1,1} + \phi_{22} x_{t-1,2} + \phi_{23} x_{t-1,3} + w_{t,2} \\
x_{t,3} &= \alpha_3 + \phi_{31} x_{t-1,1} + \phi_{32} x_{t-1,2} + \phi_{33} x_{t-1,3} + w_{t,3}
\end{aligned}$$

3.1 Utilization of the VAR Model for Flu Prediction

The VAR model was built in MATLAB. It was entirely constructed with the functions provided by MATLAB's Econometric Toolbox. Initial pre-processing of the data was carried out and then the VAR(4) model was created. The model was constructed using a function called varm() provided in the aforementioned toolbox, which returns a varm object, which in turn characterizes the model [14].

The VAR model was constructed to take in the same data as the eVision model to predict across the same distances. However, as the results (depicted in Sect. 6.1) show, this model does not perform as accurately as eVision's LSTM model described below (with results in Sect. 6.2).

4 Modifications to eVision

Prior work on eVision has established the base LSTM model which takes in the number of ILI reported cases along with Google Trends data to make long-term forecasts on the number of cases [20–22]. While work had previously been done on making national level predictions using state-level data, the optimal selection of states was not yet found and the effects of lower level division data were not explored.

4.1 Selecting States

Adding states was the first step in augmenting national level forecasts with more granular data.

The CDC, in addition to national level influenza statistics, provides statewide statistics for influenza. Similarly, Google Trends provides popularity of keywords by state. Thus eVision is capable of incorporating these state level indicators as features to augment its predictions.

4.2 Adding Metropolitan Data

Adding metropolitan data was done as an attempt to see if the granulation of Google Trends data would correlate to a higher level of prediction accuracy.

The municipalities on Google Trends are broken up into what were known as Designated Market Areas (DMA). DMA are 210 regions in the Unites States which receive the same radio and television options created by the Neilsen Media Research firm [11]. Having the option of a metropolitan level, it allows further testing to see if the granulated Google Trends data leads to more accurate predictions. Each of these DMA has a distinct three digit code, which Google Trends used to differentiate the different metropolitan areas from one another.

5 Data Acquisition

5.1 Google Trends

Google Trends data was used as the basis for the LSTM and VAR models. It provided a great level of flexibility because of the volume and scope of Google searches that people frequently make. Google Trends data is presented in time intervals that can range from the last 24 h to the last 10 years. The data in Google Trends is normalized from data points that correspond to searches at a given time and place. The data is normalized on a scale of 0–100 with respect to the time interval allocated, with time periods of higher search frequency corresponding to a higher number [10].

Google Trends provides data on three levels: region-wide, state-wide, and metropolitan. The region-wide levels consist of countries across the world and the state-wide areas consist of the 50 states plus the municipality of the District of Columbia.

5.2 Google Search Keywords

eVision uses four key terms: cough, flu, sore throat, and tamiflu. The search frequency for these influenza related terms strongly correlate to the frequency of influenza cases, making them an excellent source of information for training the model.

5.3 Data Acquisition Accommodations Due to COVID-19

The current COVID-19 pandemic has severely skewed data for many of our relevant search terms. As mentioned earlier, Google Trends comparatively ranks search frequency on a normalized scale from 0–100. When a significant and irregular event occurs, such as a pandemic, there is usually a corresponding alteration in search frequency for relevant terminology. This has caused an intense spike in the number of searches for virus related keywords (cough, flu, etc.), which due to Google's data post processing, eliminates the variance in weekly data.

A prime example of this is the search term "fever", which is a common symptom for both COVID-19 and influenza. Figure 1 illustrates this discrepancy by showing search frequency for a given time before and after COVID-19. When using a custom time range that does not encroach upon the hysteria of COVID-19 related Google searches, its magnitude becomes comparatively much smaller than the time range which includes COVID-19 related searches.

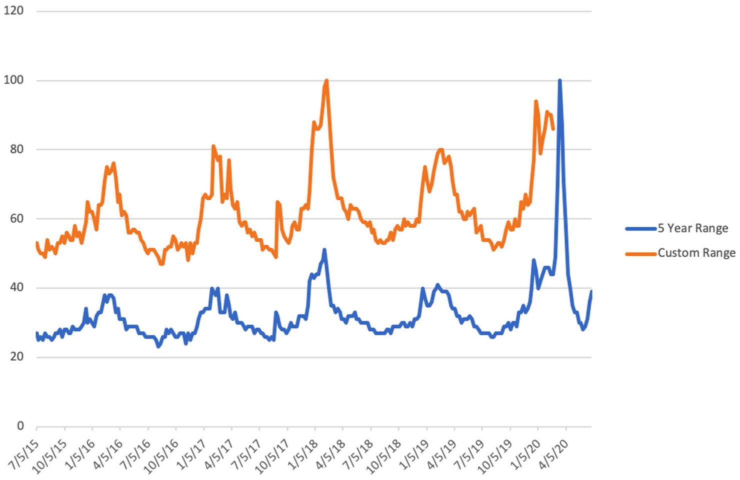


Fig. 1. Past 5 Years (blue) vs. Custom Date Range (orange) Search Frequency for Search Term "fever" (Color figure online)

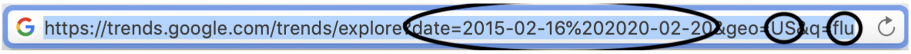
The date range used for testing was a five year interval between February 16, 2015 to February 16, 2020, who's difference with modern data can be seen in Fig. 1.

5.4 Python Scraper

To obtain data from Google Trends, the publicly available Google Trends API was utilized. Using this API, a scraper was created that was able to extract selected data from Google Trends.

An older scraper created for the previous version of eVision [21], did not meet all requirements for the new additions to eVision. The main features implemented for the scraper were to allow for settings that can be toggled in order to extract results by the regional levels that Google supported: state, metropolitan, and country. It allows for the mass extraction of search data from any region in a matter of seconds.

Geographical codes were needed in order to successfully distinguish between the different regions being scraped. Since there are different scopes of regions that can be searched, Google Trends differentiates the geographical scopes in distinct ways. The way that the API accepts inputs for search terms such as geographical region, date range, etc., is by the URL of a search term on the Google Trends site as depicted in Fig. 2.



https://trends.google.com/trends/explore?date=2015-02-16%202020-02-20&geo=US&q=flu

Fig. 2. Standard URL for Google Trends search

In order for the scraper to be able to yield data from different regions, a database of country, state, and metropolitan codes was needed. This is because Google differentiates locations by ISO 3166-2 codes for countries and states, while using DMA codes for metropolitan areas. These codes would be needed as an essential parameter within the scraper that would allow it to scrape data from any area in the world.

A list of all the existing DMA codes in the United States was found online [15], along with a comprehensive database of country and state codes from a public GitHub repository [7]. This proved to be sufficient to allow the scraper to swiftly and efficiently extract data from any region in the US recorded by Google Trends.

5.5 Data Selection

In addition to including a national forecast without any states and a forecast with all states to serve as comparative baselines, two main approaches for selecting states with which to make the prediction were undertaken: selecting states with the highest population and selecting those which are the largest transit hubs with the highest level of traffic.

For selecting the states that have the highest level of international transportation, it was decided that the largest ports of entry and the states with the busiest airports would be used. To find the busiest airports, the total number of passengers each airport reported to have serviced in the year 2019 was studied, and the six airports that serviced over 60 million passengers that year were selected. These airports were located in the states of Georgia, California, Illinois, Texas, Colorado, and New York.

Airports were selected as they represent the most rapid and commonly used method of transportation into the United States, especially from countries that do not directly border it. As such, it represents a major vector of disease transmission from abroad, and it could be argued that these busiest transit states would exhibit a growth in infections before the national average begins to in any significant way. Thus, under this hypothesis, data of infections in these states would be useful for predicting the total amount of infections in the future.

Data on the busiest ports of entry into the United States from Mexico and Canada was gathered from the Department of Transportation's Bureau of Transportation Statistics. Data from 2019 shows that San Ysidro, California and El Paso, Texas were by far the largest ports of entry with Buffalo, New York coming in as a distant third. The logic behind these three states serving as useful precursors to a national epidemic is the same as with airports.

The six highest population states were chosen to contrast with the airport selection, with an anomaly in the Floridian data resulting in two versions being created with and without the state. The anomaly in question is that in the CDC FluView state by state records of influenza like illnesses, data from Florida is not included resulting in it appearing as if it has always had no cases. While this does not prevent data gathered on the google keyword trends in Florida, it was determined that this could be harmful to the model and a version of the data without Florida was generated to determine if this was the case.

As previous research has determined that national level predictions can be enhanced with state level data, it raised the question of whether or not metropolitan level data could enhance these predictions further.

In order to explore this possibility, four data sets were made consisting of the top five, ten, fifteen, and twenty most populated metropolitan areas in the United States and their Google keyword search results. National level predictions were made using only national data and the metropolitan data sets. Predictions were run with and without state data as well to observe the effect of including all three levels.

For the purposes of investigating the ability for metropolitan level data to boost state level predictions, a simple set of predictions were made for California, Texas, and New York, using the state data sets alone for each of them, followed by collecting metropolitan data sets for every metropolitan area that Google Trends collected data for in each state.

6 Results

The calculation error is measured using the same metrics established in the previous paper on eVision [21]. Originally Mean Absolute Percentage Error (MAPE) [4] was used to determine error, but after review it was determined that Symmetric Mean Absolute Percentage Error (SMAPE) [13] would be a more effective metric to make use of. The methodology behind the construction of the confidence intervals in use for the LSTM results were also not changed from the aforementioned paper.

6.1 VAR Results

Various forecasts were conducted with the VAR model in order to compare its results with the ones of the LSTM model. Table 1 contains the series of national VAR forecasts, including SMAPE scores for 3, 7, and 14 weeks ahead predictions.

Table 1. VAR national forecast results

Forecast	States	3 week SMAPE	7 week SMAPE	14 week SMAPE
National all	All states	28.31	36.98	41.21
National ports of entry	CA, TX, NY	30.12	43.98	51.15
National population	CA, TX, FL, NY, PA, IL	32.96	48.10	67.80
National only	N/A	32.10	40.79	72.22

The best results were obtained for the national level prediction when all the states were included. Across every forecast, the level of error increased the further out the prediction was made. The best national results given by VAR was an error rate of 28.31%, 36.98%, and 41.21%, for 3, 7, and 14 weeks respectively. In the case of a curated selection of states, the Top 3 largest Ports of Entry proved to be a better selection of states than using the 6 largest population states. The results for 3 weeks ahead prediction between both selections was 2.84%, and the difference only rose to 4.12% for 7 weeks, but for the 14 week forecast the difference became a significant 16.65%.

Finally, the model provided with national data only had the most varied performance. With SMAPE errors of 32.1%, 40.79%, and 72.22% for 3, 7, and 14 weeks, its placement varies from third to second to last place respectively. The resulting graphs produced by these predictions can be seen in Fig. 3 For all the results obtained, it can be seen that the forecast for initial weeks matched the number of cases but completely missed the peak period, leading to high SMAPE as compared to the LSTM model.

The results in Fig. 3 were promising for 3 weeks ahead predictions but failed as the length of the prediction was increased. Considering different combinations of states for forecasting on the National level, the results were almost the same. For all the results, the model is unable to predict the peaks at the expected time interval.

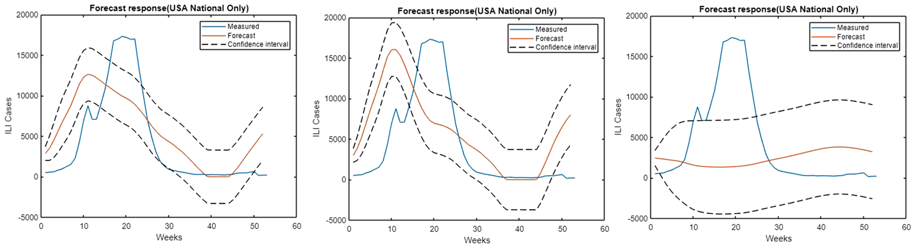


Fig. 3. National only (3, 7, 14 weeks)

State Level Forecast. Forecasts were also created to for state level predictions using no further outside data to augment them. The states selected for examination were California, Texas, and New York as they are held in common between the population selection and the ports of entry selection.

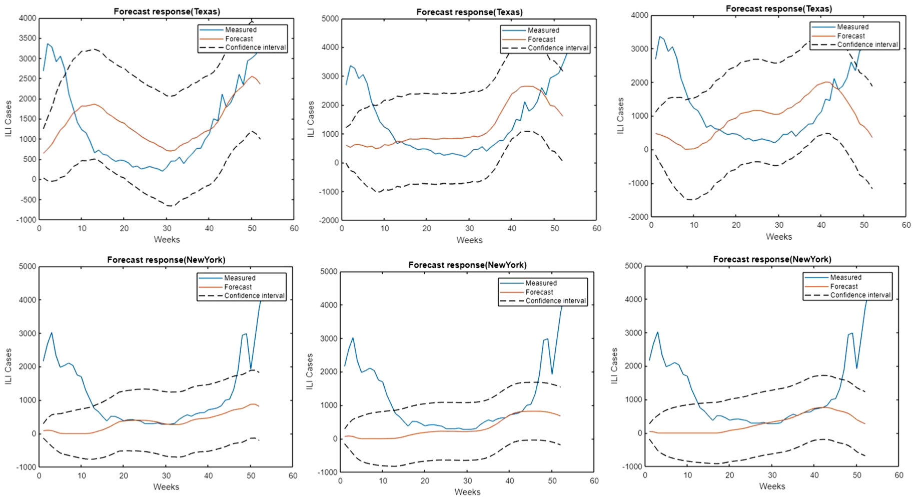


Fig. 4. Top: Texas only (3, 7, 14 weeks) Bottom: New York only (3, 7, 14 weeks)

The following Table 2 contains the SMAPE scores for the state forecasts at 3, 7, and 14 weeks ahead predictions.

Figure 4 demonstrates the results for the Texas and New York predictions. It should be noted that although the Texas forecasts have significantly higher SMAPE scores than the New York forecasts, the utility of the predictions generated are both abysmal as can be seen in the figure. Although both California and Texas manage to obtain error rates of 30% in their 3 week forecasts, and 7 week forecast, in the case of California these forecasts fail to consistently provide accurate information on the start, peak, and magnitude of an influenza outbreak.

Table 2. VAR state forecast results

Forecast VAR	SMAPE 3 week	SMAPE 7 week	SMAPE 14 week
California only	23.52	27.00	41.62
New York only	56.16	57.17	61.76
Texas only	27.43	31.69	47.64

Overall Model Analysis. The regression model used here generates hypothesis functions which produce a nonlinear curve. From the results obtained, it can be seen that the model was unable to predict the peak week of an outbreak, missing its mark by 10 weeks when it predicts an outbreak at all.

Therefore, it can be inferred from these results that this model under-fits on the data, and as such fails to extrapolate useful patterns with which it can create accurate predictions. It even fails to capture patters as basic as continuing a steady rise in cases until a shift downwards is noticed. All results generated in the VAR model would be greatly improved upon with the LSTM model, which makes use of recursive neural networks to ensure that the problem of under-fitting would be avoided and that long term patterns could be noticed in order to provide accurate, and long term forecasts.

6.2 LSTM Results

Numerous trails were conducted with the LSTM in order to determine the effects of various combinations of states, as well as the inclusion of Google keyword popularity in metropolitan areas on the accuracy of national and state forecasts. The results of these trails are included in the tables below, with SMAPE scores for three different extents of prediction, 3 weeks, 7 weeks, and 14 weeks ahead of the present week.

Two other important measures consist of the ability for a model to predict the peak week of a influenza outbreak, and its ability to predict the number of reported cases. While these two are related to the SMAPE score, severe failures on either measure would cause significant damage to the score as they are not directly related and it is possible for one model to have a higher SMAPE score than another yet fall behind on other metrics.

Most Effective State Selection for National Forecast. While there is no one selection of states that performed the best across all levels of forecasts, in fact each level performs best with a different selection, there are some important patterns that can be gleaned from the data.

The first point that stands out is the clustering that occurs in the accuracy between the levels of forecast. Across every national forecast, the difference between the SMAPE score for the 3 week and 7 week forecasts are less than the difference between either level of forecast and the 14 week forecasts.

Table 3. State selection for national forecast

Forecast	States	3 weeks SMAPE	7 weeks SMAPE	14 weeks SMAPE
National airports	GA, CA, IL, TX, CO, NY	10.85	10.43	18.26
National all	All states	19.69	16.08	23.22
National ports of entry	CA, TX, NY	09.87	08.57	22.56
National population	CA, TX, FL, NY, PA, IL	11.85	09.10	19.80
National population Sans Florida	CA, TX, NY, PA, IL	09.62	08.96	20.68
National only	N/A	11.89	12.00	25.82

Table 4. Effect of metropolitan data on national forecast

Forecast	3 weeks SMAPE	7 weeks SMAPE	14 weeks SMAPE
National top 5 metros	11.02	12.72	26.39
National top 10 metros	10.87	10.33	19.47
National top 15 metros	12.67	09.89	23.77
National top 20 metros	12.27	11.06	23.73
National top 10 with states	11.56	10.23	21.47
National top 10 states only	10.31	10.08	21.33
National top 20 with states	15.07	10.46	25.62
National top 20 states only	10.60	12.01	20.22

Table 5. Effect of metropolitan data on state forecast

Forecast	SMAPE 3 week	SMAPE 7 week	SMAPE 14 week
California metro	14.34	16.07	20.84
California only	37.01	23.28	18.54
New York metro	20.43	22.79	33.35
New York only	38.42	13.03	29.32
Texas metro	19.85	19.40	41.37
Texas only	20.20	25.13	35.77

The 14 week forecasts also notably always show a higher level of SMAPE error than any of the earlier weeks. However, as can be seen in Fig. 5, a model's predictions can still be useful even when they do not follow the results of the outbreak perfectly. For the first outbreak in the testing data, the model is able to determine the peak week of the outbreak within one week of error, while keeping the number of cases comfortably within the confidence intervals. Although the

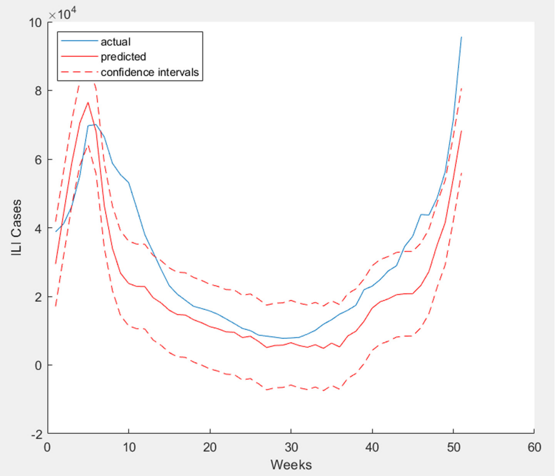


Fig. 5. National airports, 14 weeks

model performs more poorly after that point, it maps a general path of the virus in the off season, and more importantly, manages to keep the second outbreak at the end of the testing data close to its maximum confidence interval. Another consistent pattern is that the magnitude of the sharp, second outbreak is best captured by the 14 week forecasts.

The second point of note is that the no states added and all states added categories both performed worse than any of the curated state selections. As can be seen in Table 3 the all state model demonstrates the worst performance in the 3 and 7 week levels achieving 19.69% and 16.89% error rates respectively, far worse than any other model. While it does perform better in the 14 week forecast, it only does so by 2.6%. Furthermore, the accuracy with which the all state model predicts the magnitude and location of the peak week is worse than the no state model, which are the main benefits of the 14 week forecast to begin with.

The models based on largest ports of entry and highest population states, excluding Florida, are the only models that manage to break below 10% error in the 3 week forecast, and 9% error in the seven week forecast. Of the two, the model based on population performs best in the 3 week and 14 week forecast, but the ports of entry model achieves the lowest SMAPE score of only 8.57% error in the 7 week forecast. As can be seen in Fig. 6 the overall results are qualitatively similar, and it should be noted that the major difference between the two data sets is the inclusion of the states of Pennsylvania and Illinois in the population model.

Finally, the last major point of note can be seen in the effect that the inclusion of the state of Florida has in the population model compared to the one that excludes it. Similar to the no states/all states comparison, excluding Florida allowed the population model to perform better in the 3 and 7 week levels, but

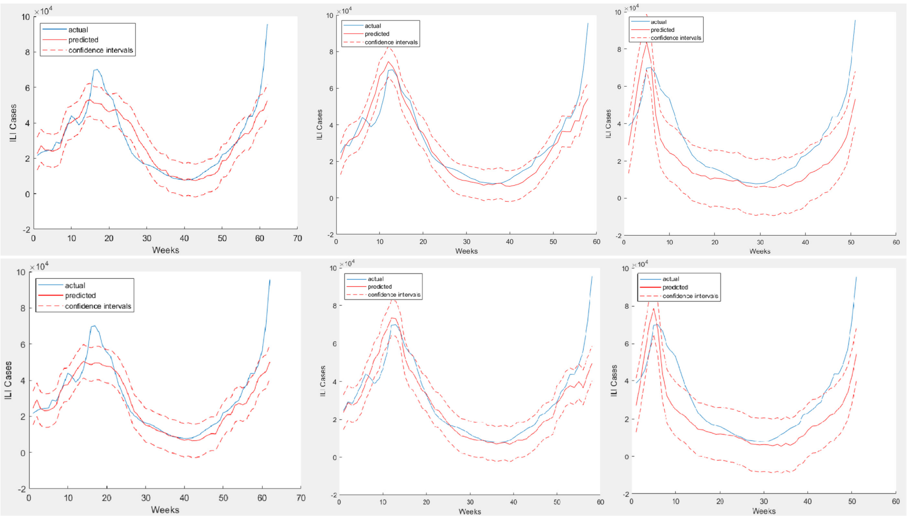


Fig. 6. Top: Ports of Entry (3, 7, 14 weeks) Bottom: Population sans FL (3, 7, 14 weeks)

in the 14 week level it performed 0.88% worse than the model containing Florida's ILI cases and keyword trends. In terms of the magnitude and peak week location measures, the two models also perform similarly, with only minor differences in the 3 week and 7 week forecasts where the Florida excluding model provides a better measure of the number of cases in the first outbreak in the testing data.

Effect of Metropolitan Data on National Forecast. For the models consisting of the top metropolitan areas, the results follow a pattern similar to state selection. The model performs worse both in the case of being provided with too little supporting data, and when provided with too many features. As each metropolitan data has four keywords to keep track of each as their own independent feature, the total number of metropolitan features can reach as many as 80 in the case of the Top 20 model.

Overall, the best performing model of the four was the Top 10 model as can be seen in Table 4. Achieving the best SMAPE scores for the 3 week and 14 week levels, and coming just 0.44% short of the best 7 week result, the model provides the most consistently accurate results across the three levels of forecast. With a total of 40 features added from the keyword trends, it has a higher feature count than most state selections, though only by ten.

Two additional models were created for the Top 10 and Top 20 data sets, adding the data for every state of the metropolitan areas as well as including a data set with only the state data. In all but a single case, the 7 week forecast for the Top 20 model, the models with both state and metropolitan data performed worse than either of the two data types alone. This was expected as adding state data would increase the number of features and harm the LSTM's ability to detect patterns.

When placing the two data types head to head against each other, it can generally be said that state data performs better at the 3 and 7 week levels, though the difference is more pronounced in the case of the Top 20 model. In the case of the 14 week forecast, the metropolitan data does manage to outperform the state data by 1.86% in the Top 10 model, where as the Top 20 shows the state data 3.51% more accurately. The most likely source of the superiority of the state level data would be their ILI data. Though as can be seen in the results, the benefits of the ILI data prove to be mostly marginal compared to the keyword trends data.

Effect of Metropolitan Data on State Forecast. The major outstanding pattern with regards to applying metropolitan data to state level forecasts is that there is no major pattern.

Smaller patterns do exist, such as metropolitan augmented forecasts performing worse at the 14 week forecast across all three states as can be seen in Table 5. But beyond that the results become more varied, such as the metropolitan data increasing accuracy in the 3 week forecasts, but with its improvements varying from highly significant in California (22.67%), to almost negligible in Texas (0.35%). In the case of the 7 week forecast, metropolitan data aids in the case of California and Texas, but adds even higher error in the case of New York. Furthermore, it should also be noted that this was not an exhaustive study of the effects of metropolitan data on state level forecasting. The states of California, New York, and Texas were examined as they were the states that appeared in every stat selection for national level predictions. The inconsistencies in the results here may suggest that the utility of metropolitan level forecasts may vary depending on the state in question. Further study will be required to draw serious conclusions, particularly in the case of low population states.

7 Future Work

7.1 Google Trends Data Ranges and Adjustments

As a result of the COVID-19 outbreak, a lot of the data has been skewed. Because of this, the model is currently trained on data predating the outbreak so that it would not be affected by this anomaly. However, the end goal of this software is for it to be practical for commercial use by pharmaceutical companies, which necessitates the creation of a solution to the current skewing of Google Trends data. This is because, in the future, there will likely still be a level of corruption in Google Trends data from COVID-19.

It may be possible to simply omit that data and work around it, but it is unknown how a missing chunk of data will affect the model's ability to make accurate predictions.

7.2 Influenza Strain-Level Predictions

The model is also expected to be able to predict the trends of influenza strains. There are four distinct strains of influenza: A, B, C, and D. Of these strains, influenza types A and B lead to the majority of influenza cases [5].

Ideally, predictions for different influenza strains would yield similar results as the influenza forecaster. However, this may need to be achieved through a different means than what is currently done. Types A and B do not have distinguishing symptoms [19], therefore symptoms of the strains cannot be used to predict the trends.

However, there are some general trends of the timing of the dominant strain, with type A being most prevalent at the start of the flu season and type B becoming more frequent in the latter half of the season [19]. Common trends of timing like this will be the starting point in helping the model determine a dominant strain during a given period of time.

7.3 Ease of Use

For future versions of eVision, there are hopes for a more uniform prediction process. Currently, there exists a multi-step prediction process involving running the Python scraper, acquiring the data, and running the MATLAB script that makes the prediction. There is the end goal of making the entirety of the model mostly autonomous by having the model run continually on a server. This will allow the model to make predictions more frequently and no longer require trained programmers to make edits to allow for said predictions.

There is also hope to incorporate a user-friendly and simplistic UI. The end goal for eVision has always been for it to be a tool used by pharmaceutical companies and healthcare providers to gauge the quantity of tester kits, vaccines, and medication they need to manufacture or resources they need to allocate in order to prevent and treat Influenza. Having a quality GUI for instance, will allow for its ease of use by even nontechnical staff at said organizations.

8 Conclusion

Through adding state level predictions in order to enhance national predictions and utilizing metropolitan area keyword trends to improve both state level and national predictions, eVision's success in making 3, 7, and 14 weeks in advance predictions were improved from 88.11%, 88%, and 74.18% accuracy to 90.38%, 91.43%, and 81.74% respectively. Furthermore, it was determined that the LSTM model is superior to the VAR model on all counts, and that generally speaking for national level forecasts state level data is superior to metropolitan data or a mixture of the two. Meaning, granularity in prediction is helpful in improving overall prediction as long as the grains are not selected to be too small, and not too many of them are selected such that the model is overwhelmed with features.

Acknowledgment. Many thanks are due to Ben Dorty, the New-Technology Senior Director at Cepheid Inc. for inspiring the project, and supporting it throughout development. Also to Prashanth Asuri, director of SCU's BioInnovation and Design Lab for obtaining financial support from Cepheid Inc. And to the School of Engineering's Frugal Innovation Hub as well as the Departments of Bio Engineering, Computer Science & Engineering, and Mathematics & Computer Science for their continued support of the eVision project. And lastly to other eVision team members: Yash Kamdar and Ron Huang, as well as past eVision team members: Yuhang Qian, Anika Shahi, Liying Liang, and Meghan McGinnis for their hard work on research, data collection, and software development in the earlier stages of the project.

References

1. Centers for Disease Control and Prevention (CDC): Preliminary in-season 2018–2019 burden estimates (2019). https://www.cdc.gov/flu/about/burden/preliminary-in-season-estimates.htm
2. Centers for Disease Control and Prevention (CDC): 2019–2020 U.S. flu season: Preliminary burden estimates (2020). https://www.cdc.gov/flu/about/burden/preliminary-in-season-estimates.htm
3. Chung, J.R., et al.: Effects of influenza vaccination in the United States during the 2018–2019 influenza season. Clin. Infect. Dis. **71**, e368–e376 (2020)
4. De Myttenaere, A., Golden, B., Le Grand, B., Rossi, F.: Mean absolute percentage error for regression models. Neurocomputing **192**, 38–48 (2016)
5. Centers for Disease Control and Prevention: Types of influenza viruses, 18 November 2019. https://www.cdc.gov/flu/about/viruses/types.htm
6. Dugas, A.F., et al.: Google flu trends: correlation with emergency department influenza rates and crowding metrics. Clin. Inf. Dis. **54**(4), 463–469 (2012)
7. Gada, D.: Country state city DB demo (2020). https://dr5hn.github.io/countries-states-cities-database
8. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**(7232), 1012–1014 (2009)
9. Google: Google flu trends (2019). https://www.google.org/flutrends/about
10. Google Trends Help: FAQ about google trends data (2020). https://support.google.com/trends/answer/4365533
11. Halbrooks, G.: What is a designated market area (DMA)?, 25 June 2019. https://www.thebalancecareers.com/what-is-a-designated-market-area-dma-2315180
12. Lee, V.J., et al.: Advances in measuring influenza burden of disease. Influenza Other Respir. Viruses **12**(1), 3–9 (2018)
13. Makridakis, S.: Accuracy measures: theoretical and practical concerns. Int. J. Forecast. **9**(4), 527–529 (1993)
14. MATLAB Help Center: Create vector autoregression (VAR) model (2020). https://www.mathworks.com/help/econ/varm.html
15. MDR ADC: 2014–2015 DMA's. https://www.mdreducation.com/pdfs/dma.pdf
16. Rolfes, M.A., et al.: Effects of influenza vaccination in the united states during the 2017–2018 influenza season. Clin. Inf. Dis. **69**(11), 1845–1853 (2019)
17. Santillana, M., Nguyen, A.T., Dredze, M., Paul, M.J., Nsoesie, E.O., Brownstein, J.S.: Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput. Biol. **11**(10), e1004513 (2015)

18. PennState Eberly College of Science: Vector autoregressive models VAR(p) models. https://online.stat.psu.edu/stat510/lesson/11/11.2
19. Seladi-Schulman, J.: How are influenza a and b different? 28 March 2020. https://www.healthline.com/health/cold-flu/influenza-a-vs-b#types
20. Shaghaghi, N., Calle, A., Kouretas, G.: eVision: influenza forecasting using CDC, who, and google trends data. In: 2020 IEEE Technology and Engineering Management Society (TEMS) 2020 International Conference on Artificial Intelligence for Good (AI4G). IEEE (2020)
21. Shaghaghi, N., Calle, A., Kouretas, G.: Expanding eVision's scope of influenza forecasting. In: 2020 IEEE Global Humanitarian Technology Conference (GHTC). IEEE (2020)
22. Shaghaghi, N., Calle, A., Kouretas, G.: Influenza forecasting. In: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2020, pp. 339–341. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3378393.3402286
23. Xie, W.: Spatial panel VAR and application to forecast influenza incidence rates of us states. Available at SSRN 2646870 (2015)