Depression Surveillance based on Google Trends

Wenqing Qian, Xin Li, Yuting Duan

2022-12-17

Introduction

Depression is a common mental disorder in the United States, with over 21 million adults reporting at least one major depressive episode in 2020[1]. However, there are some disadvantages of traditional survey-based methods, such as low response rates, the stigma of depression, concealment, report bias, and high costs[2]. Major organizations such as the National Institute of Mental Health (NIMH), Anxiety & Depression Association of America (ADAA), and CDC provide only limited data specific to the time and population being studied from their surveys[3].

Google Trends is a free online tool that allows users to obtain big data on the search volume of different queries across various regions over time. The usefulness of internet search trends in the surveillance of mental health has been demonstrated. In 2019, Barros et al. utilized Google Trends search volumes for the prediction of national suicide rates in Ireland[4]; in 2021, Zhang et al. developed an effective method to monitor depression trends on Twitter during the COVID-19 pandemic[5].

This project explores the factors that may impact the search volume of depression and build an R Shiny Application to visualize the search volumes of depression-related words by state compared with different factors. The code for the project is on https://github.com/BIOSTAT625-Project/Final-project. A fantastic R Shiny Application can be accessed via https://conchaespina.shinyapps.io/FinalProject625/ .

Data collection and Processing

Since there is no established database including data we need, we had to collect data from different sources and build up our own database first.

Search volumes of depression-related words

Weekly data on the mental health session were extracted from Google Trends for an 18-year period (2004/01/01-2022/11/01) and separated by US state for the following terms: “feeling sad,” “depressed,” “depression,” “empty,” “insomnia,” “fatigue,” “guilty,” and “suicide.” Since Google Trends only provide search volume relative to the highest point for the given region and time, we have to download the absolute search volumes using GLIMPSE in order to compare search volumes between different time periods and states. Due to the limitation of GLIMPSE, we only get absolute and relative search volumes for the recent five years (2017/11/01-2022/11/01). Then, data of relative search volumes for the rest of the years (2004-2017) was downloaded by using the “gtrendsR” package. In order to get weekly data, we had to download the data every five years; and then, to make sure that the relative search volumes in different periods of time are comparable, we make sure that every two subsets of five years have at least one year of overlap. Linear regression models through the origin were adopted to: first, normalize the relative search volumes in different periods based on the overlapping year; and then, transform the relative search volumes to absolute search volumes.

Environmental data

- Data of local weather was downloaded from VisualCrossing. The dataset includes daily data of the average temperature, feels like temperature, precipitation volume, precipitation type and solar radiation of each state.
- Data of local air quality was downloaded by using RAQSAPI package from EPA. The dataset includes daily air quality index (AQI) of each state.

Socioeconomic data

- Monthly data of unemployment rates (2009-2022) by state was obtained from the website of National Conference of State Legislatures NCSL.
- Annually data of per capita personal income (2004-2021) by state was obtained from the website of Federal Reserve Economic Data FRED.
- Median age by state was obtained from the website of World Population Review.
- Data of education level by state was obtained from the website of U.S. Department of Agriculture USDA. The dataset includes the percentage of the population in each state that fit in the following three education levels: not completing high school, completing high school only, and completing college.

Prediction

We explore the model of population depression by Google trends and its relationship with natural and social factors and the prediction of population depression is given.

First, we focus on overall Depression in US. We split the data into training data and test data. The first 80% is training data and the last 20% is test data. The first model we consider is the autoregressive integrated moving average (ARIMA) time series model. Without any other information, we use the training data to fit the model and then forecast the test data. The test mean square error is 36.77.

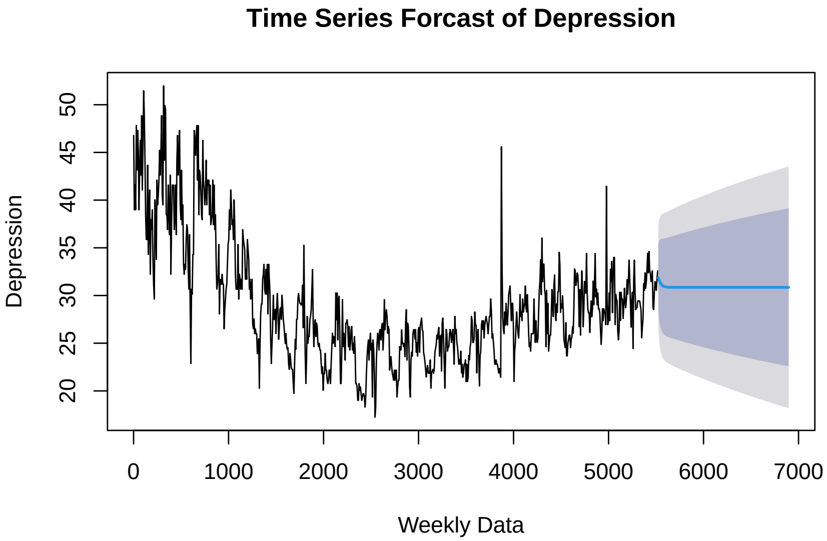


Figure 1: ARIMA Forecast

Then we add predictors to forecast the depression in US. Natural and social-economic factors are applied here, including weather and employment. After merging the depression data and predictors by date, time series regression model and randomForest model are buit for prediction. The p-value of the coefficient test of the model is small enough, which verifies the significance of the selected predictors. Higher employment rate and higher temperature results in lower population depression. The test mean square error for time series regression and randomForest are 24.52 and 24.43, respectively, which are lower than ARIMA model. The following graph is the prediction(blue) and true(green) depression of the test data.

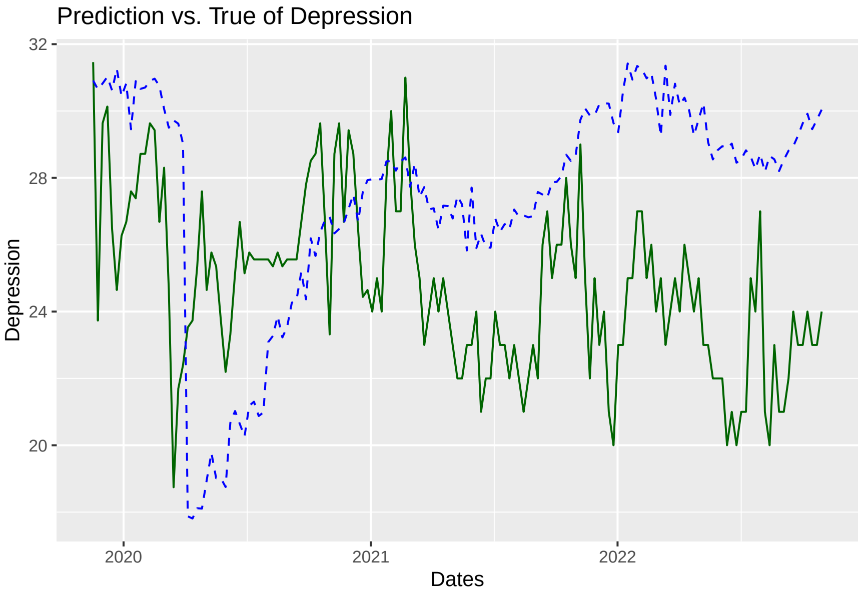


Figure 2: Prediction vs. True of Depression

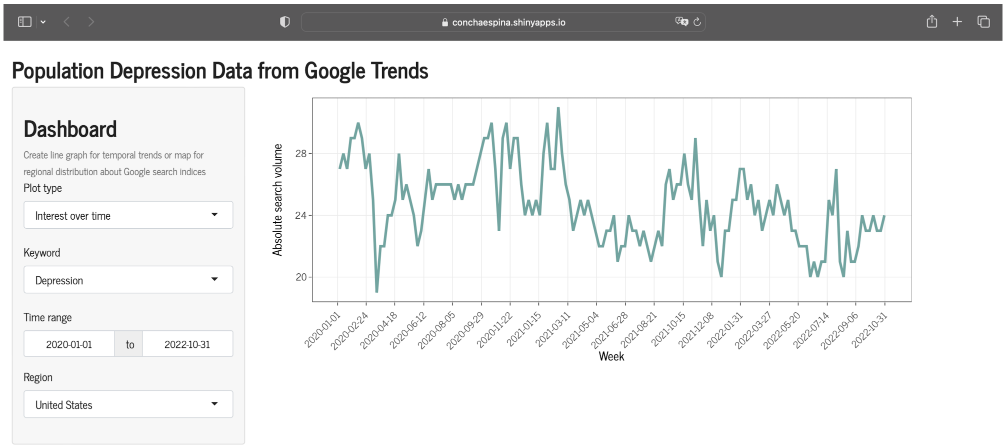
Second, depression by state of US is studied with more factors. Depression data of 50 states and Washington, D.C. are available and more demographics factors for each state are considered for modelling, such as education and median age. ARIMA, time series regression and randomForest by group of keyword and region are used for prediction. Forecasting of future depression-related keywords search volume by ARIMA is passed for visualization by R Shiny Application.

Generally, mean square error of time series regression and randomForest is lower than that of ARIMA. This is not surprising because for the regression model, more factors of depression are considered for prediction.

Shiny app: Population Depression Data from Google Trends

We built a Shiny app named *Population Depression Data from Google Trends* to visualize our data and prediction results. The app is composed of a side bar serving as a dashboard and a main panel to display plots. Its default interface is shown in Figure 3.

There are two plot types available to choose. If selecting “Interest over time”, a line graph tracing temporal change of absolute Google Trends search volume of given keyword, time range, and region will appear in the main panel. When time range includes a future date, predicted search volumes will be plotted in a dotted line following historical trend drawn in a solid line (Figure 4). If switching to “Interest by region”, a US map will be displayed whose filled color indicates the total search volume within specified time period across all the states. The menu in section “Other indicators” allows users to choose an external indicator like unemployment rate or sunlight index to compare it with search volume (Figure 5). Values of this indicator will be printed as a text label at each state. If there is no label after selecting a certain indicator, it means data are not available for the current time range. Besides, both line graph and map are rendered by `plotly` and thus are interactive. Users can zoom in or check hover texts by moving cursor.



conchaespina.shinyapps.io

Population Depression Data from Google Trends

Dashboard

Create line graph for temporal trends or map for regional distribution about Google search indices

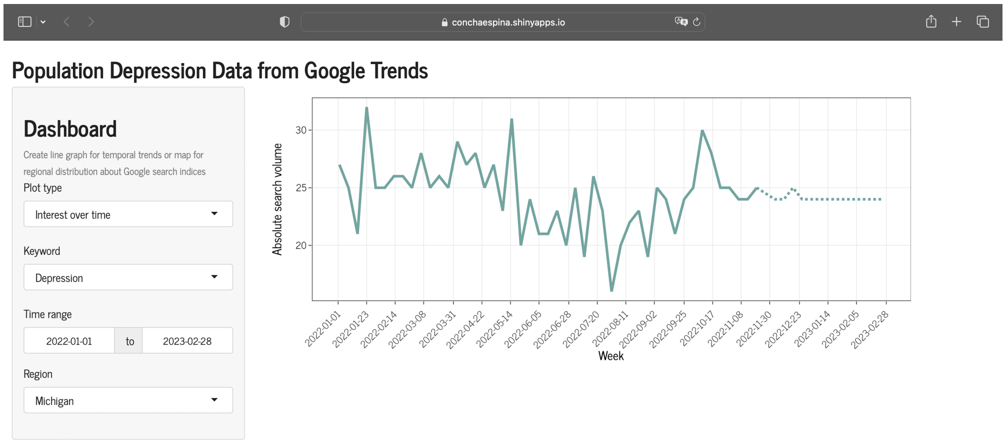
Plot type: Interest over time

Keyword: Depression

Time range: 2020-01-01 to 2022-10-31

Region: United States

Figure 3: Default interface



conchaespina.shinyapps.io

Population Depression Data from Google Trends

Dashboard

Create line graph for temporal trends or map for regional distribution about Google search indices

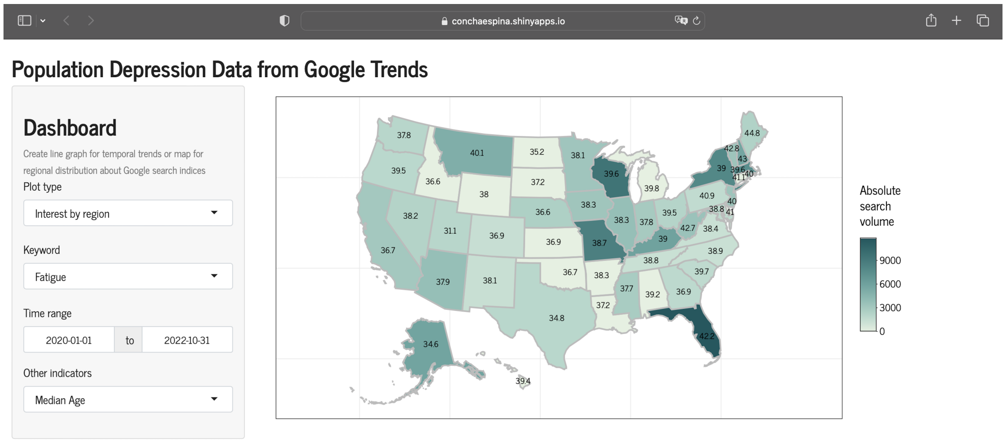
Plot type: Interest over time

Keyword: Depression

Time range: 2022-01-01 to 2023-02-28

Region: Michigan

Figure 4: Line graph: Absolute search volume of keyword "Depression" between 1/1/22 and 2/28/23 in Michigan



conchaespina.shinyapps.io

Population Depression Data from Google Trends

Dashboard

Create line graph for temporal trends or map for regional distribution about Google search indices

Plot type: Interest by region

Keyword: Fatigue

Time range: 2020-01-01 to 2022-10-31

Other indicators: Median Age

Figure 5: Map: Absolute search volume of keyword "Fatigue" between 1/1/20 and 10/31/22 with median age of each state

Conclusion

This project starts with no existing dataset. We collect great amount of raw data from Google trends and other websites and then process these data with some algorithm. Relationship between search volume of depression and other factors is studied and predictions are made. It shows that higher employment and higher temperature are related to lower population depression. We further build an R shiny Application to visualize the search volumes of depression-related words by state compared with different factors. Both line graph and map can be shown interactively.

References

1. 2020 National Survey of Drug Use and Health (NSDUH) Releases | CBHSQ Data. Accessed February 27, 2022. https://www.samhsa.gov/data/release/2020-national-survey-drug-use-and-health-nsduh-releases
2. Boogaerts T, Degreef M, Covaci A, van Nuijs ALN. Development and validation of an analytical procedure to detect spatio-temporal differences in antidepressant use through a wastewater-based approach. Talanta. 2019;200:340-349. doi:10.1016/j.talanta.2019.03.052
3. Wang, A., McCarron, R., Azzam, D., Stehli, A., Xiong, G., & DeMartini, J. (2022). Utilizing Big Data From Google Trends to Map Population Depression in the United States: Exploratory Infodemiology Study. JMIR Mental Health, 9(3), e35253. https://doi.org/10.2196/35253
4. Barros, J. M., Melia, R., Francis, K., Bogue, J., O'Sullivan, M., Young, K., Bernert, R. A., Rebholz-Schuhmann, D., & Duggan, J. (2019). The Validity of Google Trends Search Volumes for Behavioral Forecasting of National Suicide Rates in Ireland. International Journal of Environmental Research and Public Health, 16(17), Article 17. https://doi.org/10.3390/ijerph16173201
5. Zhang, Y., Lyu, H., Liu, Y., Zhang, X., Wang, Y., & Luo, J. (2021). Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study. JMIR Infodemiology, 1(1), e26769. https://doi.org/10.2196/26769

Contribution

Xin Li: Data collection and processing

Yuting Duan: Prediction

Wenqing Qian: R shiny Application