Covid 19 regional analysis and statistics using Data Science

Harpal Kaur Dhindsa, Ankita Sudhakar Chaudhari

Department of CSE with spl in big data analytics
VIT Vellore
harpalkaur.dhindsa2020@vitstudent.ac.in

Department of CSE with spl in big data analytics
VIT Vellore
ankita.sudhakar2020@vitstudent.ac.in

***Abstract* – The job growth decelerated in July as the pandemic took toll on the economy all around the world. The unemployment rate from a peak of 14.7% seen in April fell to 10.2%. Unemployment rate is a precarious factor concerning the development of a country. As of February, the U.S.A has lost more than about 12 million positions and seen joblessness take an ascent from an around long-term low of 3.5%, aggravating it than the money related emergency. In the three months before the finish of June, the U. S's economy endured its most keen quarterly constriction in over 70 years of record continuing, diminishing to a yearly pace of 33 percent or almost 10 percent year-on-year. In order to get a proper view about the local industries which suffered loss during the pandemic, we propose to make use of BLS data along with IPUMS data, and making use of tracts for geographic visualization. To be able to predict the accurate unemployment rate in this paper we propose to use R language statistical methods in order to analyze the data.**

Index Terms - ACS, COVID, IPUMS, Tigris, BLS, NAICS.

I. Introduction

These days unemployment is one of the greatest social early stage struggles on the planet. The joblessness issue exists in pretty much every nation, regardless of how ground-breaking the nations are. The thing that matters is the difference in severity.

International Labor Organization (ILO) extended that overall, in excess of 25 million businesses would be powerless because of the spread of Covid. It is extended that four out of five individuals (81%) in the worldwide labor force out of 3.3 billion are at present influenced by complete or halfway working environment conclusion.

The US, UK, Canada, and larger part of the European and Asian nations have emerged to record monstrous employment misfortunes prompting an important ascent in the joblessness rate. As per the record of the Center for Monitoring Indian Economy (CMIE), just somewhat over one-fourth (27.7%) of the absolute working-age people, that is, individuals with ages between 15 to 59 years, out of 1003 million, for example 285 million individuals were working in the week ensuing the lockdown. This exhibits that within the fourteen-day time of lockdown, 119 million working individuals had lost their positions. Naturally, this assigns that the current cross-country lockdown has been the greatest employment destroyer ever. Notwithstanding, these assessments just unveil the effect on occupations for the term of the lockdown, and ought not be reflected as a perpetual loss of wellspring of income. A large number of them may perhaps have the option to return to support after the lockdown would be finished. The easygoing laborers are the most helpless because of the instable idea of their work and every day wage installment, which are notable in the development area. The areas hit hardest by COVID-19 employment misfortunes are home to labor forces in organizations like the travel industry and transportation, which are bearing the effect of the monetary closure. To pinpoint which areas are most in risk, we assess the number of low-pay jobs have been gotten sidetracked, by utilizing occupant's information in each statistics lot or are in jeopardy while stay-at-home requests are set up. We base our appraisals on the US Bureau of Labor Statistics (BLS) Current Employment Statistics data for a month to month working representatives by industry alongside 2014–18 five-year American Community Survey (ACS) IPUMS USA micro data.

To appraise the quantity of positions lost by neighborhoods cross-country, we apply the BLS data on work change at all compensation levels based on industries, accustomed to conform to state-level BLS data on business change and to ACS micro data at the public utilize micro data area level to survey low-pay work adversity by the business locally. We by then apply these appraisals to insights plot-level information enumerating the quantity of low-pay occupations based on industry.

II. Literature survey

A method of data mining using neural networks that has been previously anticipated to predict the unemployment rate [1]. Various feature selection algorithms are used to pick the appropriate feature for the unemployment rate series. After this, the proper neural network is used to choose a subset of features and an effective training function. The grid search algorithm, correlation coefficient, and genetic algorithm are used as a feature selection algorithm. Mainly they proposed a new unemployment forecast by means of search engine inquiry data. In particular, the estimation of the unemployment rate is efficient due to the search actions. Another method that has been used focuses on forecasting the unemployment rate in an effort to show the implementation of an ARIMA model in time series forecasting [2]. ARIMA States ARIMA (Autoregressive Integrated Moving Average). It permits disintegrating the time arrangement into patterns and commotion to be communicated. There are two orders of polynomial graphs used to show accuracy of an effective prediction. One is the high order graph and another is a low order graph. To keep any commotion from upsetting the financial information, straight and nonlinear, smoothing can be applied making the time arrangement examination more productive and precise. Gumbel distribution is utilized for time series examination and forecasting. This Gumbel Distribution is a return that how every now and again or rarely certain information esteems show up.

In this digital world, Social media has proved to be a very big platform and used by millions of web users for showing their current status. Sentiment analysis can be used on the tweets which are expressions of the public regarding their current status of the job [3]. They have used the R language for efficient and accurate output. With the support of R language inbuilt packages, they have demonstrated the correlation between public expression with their status and show the unemployment rate.

Selection of indicators proposed by the Commission Experts, the unemployment projection model being constructed is another approach [4]. Until the information is gone into the neural back-propagation Network (BPNN) by the utilization of Principal Component Analysis (PCA) the methodology used to limit the size of the information can be utilized. Decline the size of the BP neural network and afterward increment the intensity of the network to sum up. Finally, the analysis of the correlation is adopted to verify the consistency of the BP neural training result Network. There are a number of linear as well as nonlinear time series simulations which can be opted for forecasting via the U.S. unemployment rate [5]. Our primary focus is on calculating forecast performance throughout economic contraction and expansion by leveraging the asymmetric cyclical behavior of unemployment figures, constructing vector models as a leading indicator that integrate initial unemployment claims, and using additional details given by the monthly rate for quarterly rate forecasting. Comparisons with the consensus projections from the Experienced Forecasters Survey can also be made. The findings indicate that it is possible to achieve substantial improvements in forecast accuracy over existing methods.

III. Data and Methodology

A. Data description

1) Census Tract: A census tract or mesh block is a geographic district characterized for the resolve of acquiring a statistic. Once in a while these cover with the constraints of urban areas, towns or other authoritative locales and a few plots normally exist inside a region. In self-governing areas of the United States these are often subjective, except for concurring with politically aware lines.

Census tracts denote the smallest provincial entity for which populace data is accessible in many countries. In the United States, census tracts are sectioned into block groups and census blocks.

2) PUMA: Public Use Micro Data Areas (PUMAs) are statistical geographical regions defined for the distribution of Public Use Micro Data Sample (PUMS) data. They are also used for publicizing American Community Survey (ACS).

3) ACS: The American Community Survey (ACS) helps local executives, community front-runners, and businesses apprehend the fluctuations taking place in their communities. It is the leading source for comprehensive population and housing information about U.S.

4) CBSA: A core based statistical area (CBSA) is a U.S. topographical locale which comprises of at least one county (or reciprocals) moored by a metropolitan focus of in any event 10,000 individuals in addition to contiguous areas that are financially attached to the metropolitan place by commuting. The annotation "CBSA" refers to the communal cooperation of the metropolitan statistical areas and micropolitan areas.

5) NAICS: The North American Industry Classification System (NAICS) is the specification used by Federal statistical agencies in categorizing business establishments for the resolve of gathering, evaluating, and broad casting statistical data related to the U.S. commercial economy. NAICS is designed as follows: Subdivision: 2-digit code, the NAICS practices a six-digit coding structure to catalogue and pin point discrete sectors in this tiered classification scheme arrangement. The initial two digits of the code perceives the area, the third as its subsector, the fourth as the business network, the fifth as the NAICS business, and the 6th as the public business.

B. Methodology

1) The first step would consist of defining the static data. That is collecting all the necessary data which is the base data using which the calculation is going to be done. Hence, it will remain same during the analysis, such as the Census tract, states, PUMA, and CBSA data. The Census LODES data are aggregated to the Census tract which are available on the Urban Data Catalog.

2) To be able to define the geographic area of the counties, we need to produce few intermediary geographic files including geojsons of all CBSA's tract, and counties in the US, a tract<>CBSA crosswalk, and a tract<>PUMA crosswalk. Crosswalk means what percent the data needs to be a part of another dataset.

The crosswalk is generated by spatially joining tracts to CBSA’s and using 99.5% area cutoff. The CBSA’s are made up tracts, hence all tracts should be 1 in CBSA max.

We will change the projections of the intersection of tracts and CBSAS, to Albers equal area as we want a projected CRS when doing area calculations.

Coordinates can be transformed from a sphere-shaped datum into Albers equal-area conic projection coordinates by making use of the following formulas, where R stands for the radius, λ stands for the longitude, λ_0 stands for the reference longitude, φ stands for the latitude, φ_0 stands for the reference latitude, and φ_1 and φ_2 stand for the standard parallels:

$$x = \rho\, sin\theta$$
$$y = \rho_0 - \rho\, cos\theta \qquad (1)$$

Where,

$$n = \frac{1}{2}\,(\sin\varphi_1 + \sin\varphi_2)$$
$$\theta = n\,(\lambda - \lambda_0)$$
$$c = cos^2\varphi_1 + 2n\, sin\varphi_1$$
$$\rho = \frac{R}{n}\,\sqrt{c - 2n\sin\varphi}$$
$$\rho_0 = \frac{R}{n}\,\sqrt{c - 2n\sin\varphi_0} \qquad (2)$$

We keep track of tracts not in any CBSA's, which can be used as a masking polygon for the data visualization.

To create the tract<>Puma crosswalk we will generate tract population centroids, used to set the center of tracts. 25 tracts don't exist in the 2010 Census provided population centroids, so we will just calculate the area centroids and append.

3) The next step consists of getting the Unemployment data from BLS and WA. Also getting information about all the industries according to the state.

To be able to display the data in form of time series we use the week number to get unemployment data.

4) To calculate the job loss by LODES Industry Sector using the BLS data, we need to generate BLS percent change of job loss by industry for detailed industries, projecting forward those a month old to the current month. We take the following parameters into consideration, the BLS month to utilize a benchmark to quantify work misfortune rate change, the BLS year to utilize a measure to quantify work misfortune rate change, the CES data, and the crosswalk of CES industry codes to ACS codes. The output generated will be a dataframe with all detailed industries by projected job loss from the base line month and base line year to the most recent month.

5) To calculate the job loss by CES detailed Industry Sector using the BLS data the methodology of combining CES, SAE, and IPUMS data will be used, that is, creating a PUMA-level estimate of job loss. To generate BLS percent change job loss by industry the following parameters will be taken into consideration, the BLS month to utilize a benchmark to quantify work misfortune rate change, the BLS year to utilize a measure to quantify job loss percent change, CES data, SAE data, CES-SAE crosswalk, and CES estimates previously calculated. To get the output as a dataframe having every row as a unique state and CES industry. The dataframe shows the measure of percent change in net employment for each state-based industry in relation with the base line month and year.

6) Now we need to calculate the Job loss by ACS IND sector using the BLS data, by combining CES, SAE, and IPUMS data. To generate ACS job, change in industry the following parameters are taken into consideration, the BLS month to use as a base line to measure job loss percent change, the BLS year to use as a base line to measure job loss percent change, SAE estimates, CES-ACS crosswalk, and the most recent month and year. Generating the output as a dataframe where every row is a unique state with ACS industry. The dataframe being the measure of percent change in the net employment for each state-based industry in relation with the base line month and year.

7) Now we will use the data previously generated to calculate the ACS job change by industry. Using the following parameters, the BLS month to utilize a benchmark to quantify work misfortune rate change, the BLS year to utilize a measure to quantify job loss percent change, the IPUMS data, the ACS estimates, the most recent month and year, and the IPUMS vintage year's data for analysis. The output generated is a dataframe, where every row is a distinctive PUMA and a 2-digit NAICS industry. This dataframe is the measure of percent change in net employment for each PUMA-NAICS in relation to the base line month and year for the people earning less than $40,000 per year in wages, in 2018 inflation adjusted dollars.

8) Next, we need to generate job loss estimates by tract by using the Job loss by industry IPUMS file generated earlier in step 6. We need to generate job loss estimate for each industry across all tracts, by making use of the PUMA-tract crosswalk.

9) Generating the summary statistics, used to make decisions about data visualization breakpoints.

10) The final step would be to be able to get the maximum value of the total job loss in any industry for the data aggregated by the geographic points. Also, to be able to get the maximum value of any one tracts job loss in any industry for the data aggregated by the geography. And finally, the job loss estimates for whole US.

C. Packages

Packages which are used to download the data:

1) Tigris: - Tigris is an R package that permits users to download and use TIGER/Line shape files (https://www.census.gov/geo/maps-data/data/tiger-line.html) directly from the US Census Bureau. Tigris only yields geometric features for US Census data

which are defaulted to the coordinate reference system NAD. For US Census demographic data, the tidycensus package can also be used. The crsuggest usually helps deciding on an appropriate coordinate reference system.

2) *ipumsr:* - The ipumsr package permits users to read data via the IPUMS extract into R alongside the related metadata like variable labels, value labels and more. IPUMS is a wonderful source of international census and survey data. IPUMS makes available census and survey data from all around the globe fused across existence.

3) *urbnthemes: - The* urbnthemes package provides few tools for creating Urban Institute-themed plots and maps in R. The package extends ggplot2 with print and map themes as well as tools that make plotting easier for the Urban Institute.

IV. Output

The Fig. 1 represents the histogram that shows the maximum number of job loss in any tract-industry across the counties.

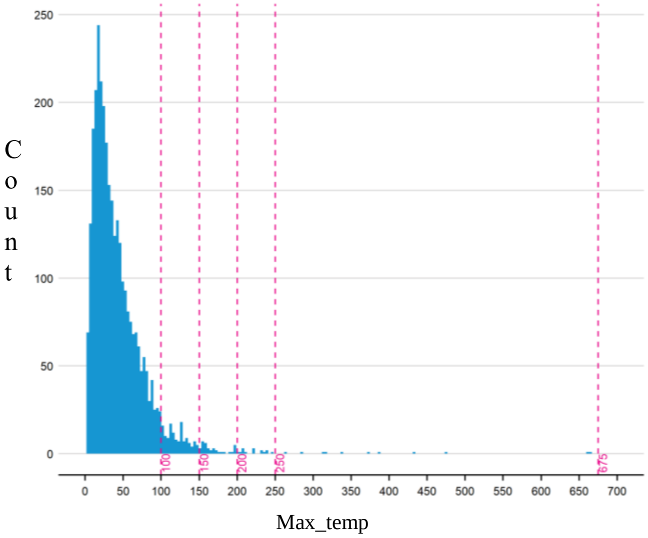

Fig. 1 Maximum of tract-level job loss at county level with 665 as maximum

The figure 2 represents the histogram that shows the maximum number of job loss in any tract-industry across CBSAs.

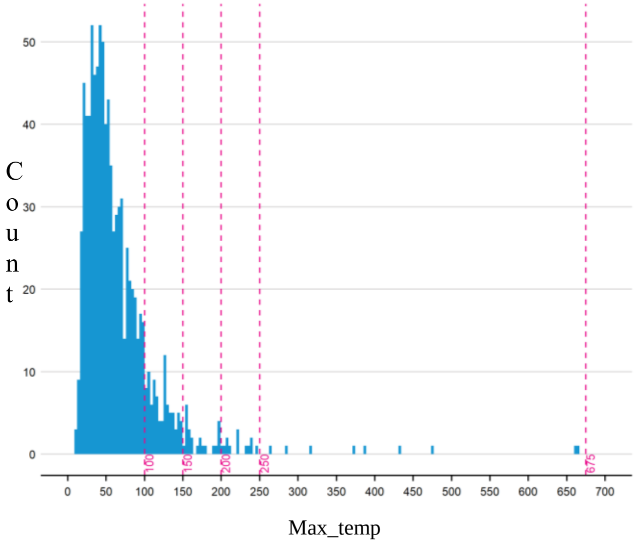

Fig. 2 Maximum of tract-level job loss at cbsa level with 665 as maximum

The figure 3 represents a histogram that shows the maximum number of job loss in the industries across the Counties.

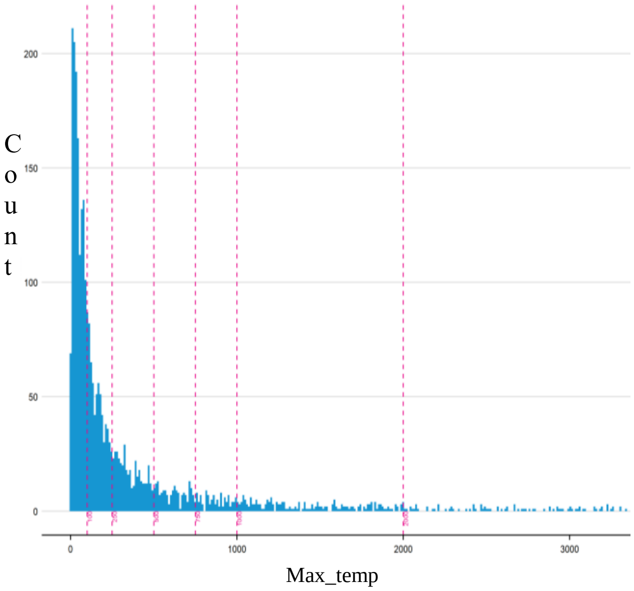

Fig. 3 Maximum job loss at county level for all industries

The figure 4 represents the histogram that shows the maximum number of job loss in the industries across the CBSAs.

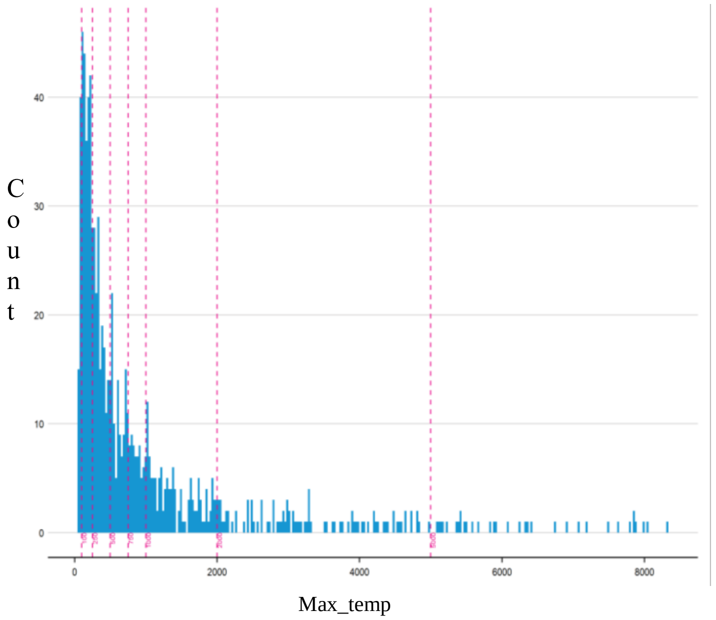

Fig. 4 Maximum job loss at cbsa level for all industries

Hence, considering the sum of job losses over the CBSA, sum of job losses over the County, and the sum of job losses overall in the United States of America we are able to plot a map. The map shows the variations among the various industries for the jobs lost in the United States of America.

REFERENCES

[1] W. Xu, Z. Li and Q. Chen, "Forecasting the Unemployment Rate by Neural Networks Using Search Engine Query Data," *2012 45th Hawaii International Conference on System Sciences,* Maui, HI, 2012, pp. 3591-3599, doi: 10.1109/HICSS.2012.284.

[2] A. Kyung and S. Nam, "Study on Unemployment Rate in USA Using Computational and Statistical Methods," *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, USA, 2019, pp. 0987-0993*, doi: 10.1109/UEMCON47517.2019.8993016.

[3] C. R. Nirmala, G. M. Roopa and K. R. Naveen Kumar, "Twitter data analysis for unemployment crisis," *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere*, 2015, pp. 420-423, doi: 10.1109/ICATCCT.2015.7456920.

[4] G. Wang and X. Zheng, "The Unemployment Rate Forecast Model Basing on BP Neural Network," *2009 International Conference on Electronic Computer Technology, Macau*, 2009, pp. 475-478, doi: 10.1109/ICECT.2009.58.

[5] Clements, Michael & Franses, Philip & Swanson, Norman. (2004), “Forecasting economic and financial time-series with non-linear models”, *International Journal of Forecasting*, 20. 169-183. 10.1016/j.ijforecast.2003.10.004.