Exploring Human Mobility Patterns Using Twitter Data

Alden Felix, Kyndall Brown, and Zeyu Sun

EPPS 6302 Methods of Data Collection and Production

December 9, 2022

Abstract

This project focuses on the differences between DFW residents' mobility patterns before, during, and after COVID-19 social distancing restrictions. Mobility data in the form of geotagged tweets was scraped from twitter for these three periods using API and non-API methods, and median distance traveled for the user sample in all three periods was computed. User movement was also animated on a map at a national and local level to showcase the effectiveness of studying mobility data through animation. This study finds that for the users in the sample human mobility during social distancing restrictions suddenly and drastically decreased as expected, but in the period after restrictions ended mobility increased to a level greater than what was observed in the period before social distancing restrictions were put in place.

Introduction

Human mobility, the study of human movement across space and time, is an important topic with applications in many fields. Seeing as transportation makes up 15 to 25 percent of the average household income in the United States and Europe while also accounting for the second greatest source of greenhouse gas emissions globally, the significance of quantifying human movement is made apparent (Barbosa et al., 2018). Studies on why and how humans move on an individual and group level reveal patterns that are used in urban planning, transportation networks, tracing the impact of socio-economic and natural factors, and numerous other applications. One such application is modeling epidemic spread, giving policy makers and institutions data driven information that has the potential to provide valuable insight and save countless lives (Colizza et al., 2007; Tizzoni et al, 2014). This study focuses on how the COVID-19 pandemic has impacted human mobility during and after social distancing restrictions. A result other than a return of human mobility to its state before the pandemic may indicate the presence of lasting factors that influence human movement, although this study will not explore those factors. Many sources of human mobility data are readily available with current technology, including census data, tax revenue data, travel surveys, bank notes, mobile phone records, GPS data, and online data. Although this list is not comprehensive these are the forms most found in human mobility research (Barbosa et al., 2018). Most traditional works on the topic have been conducted on census or survey data (Nestorowicz 2019; Chi et al. 2020). The study of human mobility has been a core subject for both scholars and policymakers for a long time. Both government agencies and international organizations have put in substantial effort in tracking and recording migration activities through the years.

Literature Review

The study of human mobility is based on two dimensions: temporal and spatial. For a mobility activity to occur, one needs to move from the "home" location where they stay over a period of time to a new geographic space (Willekens 2008; Rogers 2003). It is challenging to draw a complete picture of one's migration pattern based on sporadic data. With the development of technology, more and more scholars employ digital trace data, including mobile service data, social media, and other "big data" sources, to study mobility and migration patterns to better reflect the temporal and spatial nature of the phenomena (Armstrong et al. 2021; Yin et al. 2022).

Despite its limitation on representation, Twitter data is still the most commonly used for this type of digital trace data analysis due to its easy accessibility and mass quantity. Many scholars recognize this limitation and conduct research on Twitter data compared to data collected from other traditional sources to validate the findings. Chi et al. (2020)'s analysis predicts mobility events using Twitter data based on mobile services data collected over four years and finds similar results for both. Yin et al. (2021) evaluate geo-located Twitter data with county-based migration flows with records from the U.S. IRS and conclude that digital trace data tends to be over-represented in metropolitan areas with many tourists and high mobility.

Besides within-country analysis, scholars also use Twitter data to analyze international migration for refugees (Zagheni et al. 2014; Hubl et al. 2017). Despite its limitation on the power of inference, the use of social media data, particularly Twitter data, provides new insight into the spatial and temporal activities of human mobility. With the rapid development of technology and the constraint posed during the pandemic, the use of social media has become more and more proliferated. A study conducted on human mobility dynamics during COVID-19 using Twitter data finds that the mobility activities coincide with each country's protection measures (Huang et al. 2020).

As most countries reduce protection measures to return to normal activities, this study will compare mobility activities before and after COVID using Twitter data to understand the impact of COVID on human mobility.

Data Collection and Visualization Methods

During the data collection phase, we initially intended to only utilize R and the "rtweet" package to scrape and analyze Twitter data. However, because the Twitter API restricts access to the "full archive" to academic level developer accounts, and we were not able to get this access level, we were only able to access tweets made within the last 7 days. Therefore, we employed a non-API method in Python in combination with an API method in R to scrape twitter data from the 3 time periods we had selected: March 2018 to October 2018, September 2019 to May 2020, and October 2021 to July 2022. To ensure randomization, we utilized the stream function in the "rtweet" package to collect live tweets in the Dallas Fort Worth area over a 2-hour period. Through this function we were able to obtain users IDs and names for scraping historical tweets utilizing the non-API method. We were able to obtain 1000 usernames at this stage. We then utilized the Python package "snscrape" to scrape tweets posted by individual users in our sample in the three time periods. The data collected contained ten variables: date and time posted, tweet ID, user ID, coordinates of where the tweets were posted from, place of where the tweets were posted from, the retweet count, the reply count, the like count, the count of the tweets, and the Username. Due to the size of the data and processability concerns, we reduced the sample data down to 350 Users. The total data collected contained 1,224,565 observations, where 741,249 contain coordinates. The coordinates contain longitude and latitude information for our further analysis.

All of our visualization was performed in R using multiple methods. We identified and ranked the users in our sample based on the number of unique coordinates they had posted throughout the

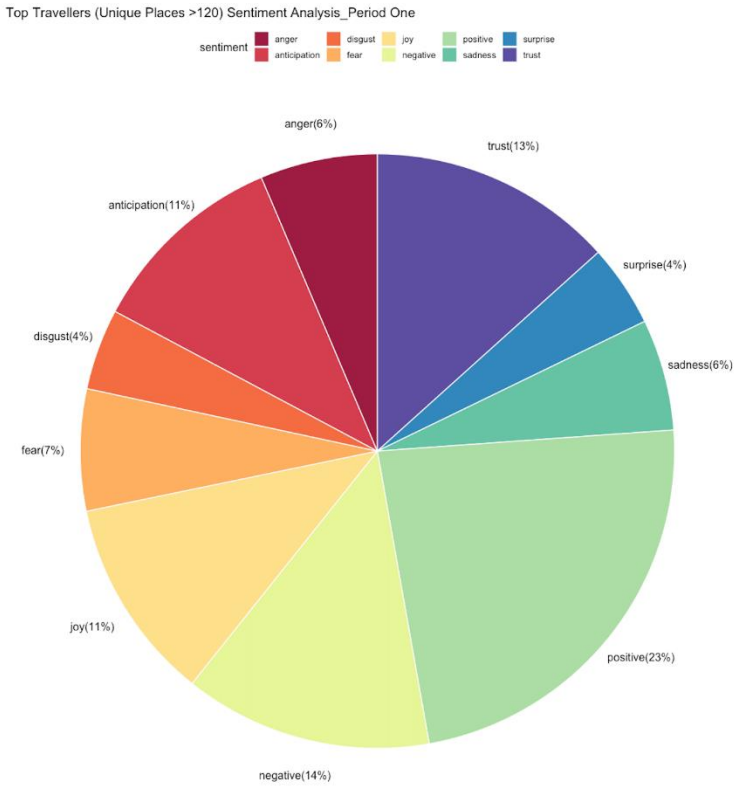
time periods. The top traveler has 192 unique coordinates which indicates frequent travel. Based on the number of unique coordinates, we classified users with more than 92 unique coordinates as the top travelers. Thirteen people among the 350 randomly selected Users are considered as top travelers. We classified users with less than 6 unique coordinates as the least travelers. Among the 350 Users, 44 people posted tweets with less than 6 unique coordinates in a 5-year period. To better understand the tweets, we performed a sentiment analysis (Hack Your Data Beautiful; Saif 2013) on both top travelers and least travelers among the three time periods. For this analysis, we used the R package "tidytext" and "tidyverse."

We also calculated the distance traveled between consecutive tweets and cumulative distance traveled for a subgroup of users, then created scatterplots and line graphs showing the distances traveled by the users. We identified a list of 112 unique usernames that tweeted in all data collection periods. The distance traveled between consecutive tweets for each unique user was calculated using the Haversine formula for finding the shortest great circle distance between two points on the surface of the Earth (Pineda-Krch 2011). Then those distances were summed as cumulative total distances traveled for each tweet, meaning total distance traveled at the time of each tweet as a sum of all previous between-tweet distances. The calculations were validated by comparing the results for a small number of points with an online distance calculator to ensure the code had no errors. Although the calculation, being based on the Law of Cosines, contains a small amount of error due to the Earth being an ellipsoid instead of a perfect sphere, the approximation is acceptable for our purposes as we wanted to map human movement over larger distances. Median cumulative distance traveled for the 112 unique users was calculated for each of the three periods. The "ggplot2" package was used to two types of graph, a scatterplot showing the distance between tweets across time for a selected user and a line graph showing cumulative distance traveled across time for a select group of users.

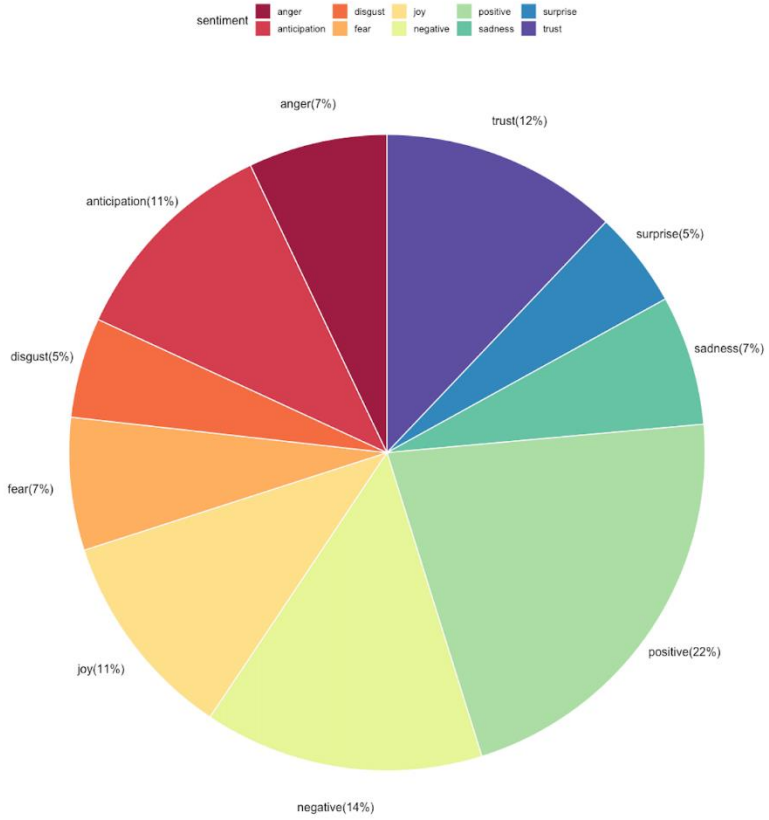
The main packages used to animate our movement data were "sf" and "moveVis". "Sf" was used for general geographic data manipulation while "moveVis" provided the means of animating our movement data onto a basemap. The main difficulty we faced in creating these visualizations was the time that the functions in the "moveVis" package took to execute with a dataset of about 400,000 observations spread between three separately animated objects. Creating the frames and stitching them together into a video, even on a mid-range desktop computer, led to significant downtime spent waiting for the functions' outputs.

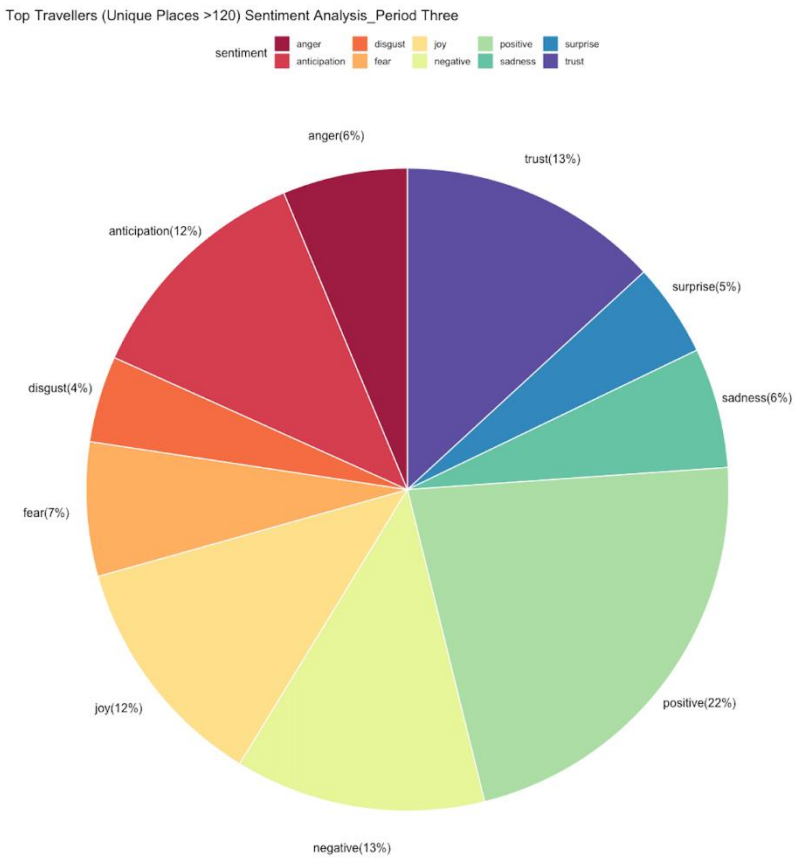
Data Analysis

Sentiment Analysis

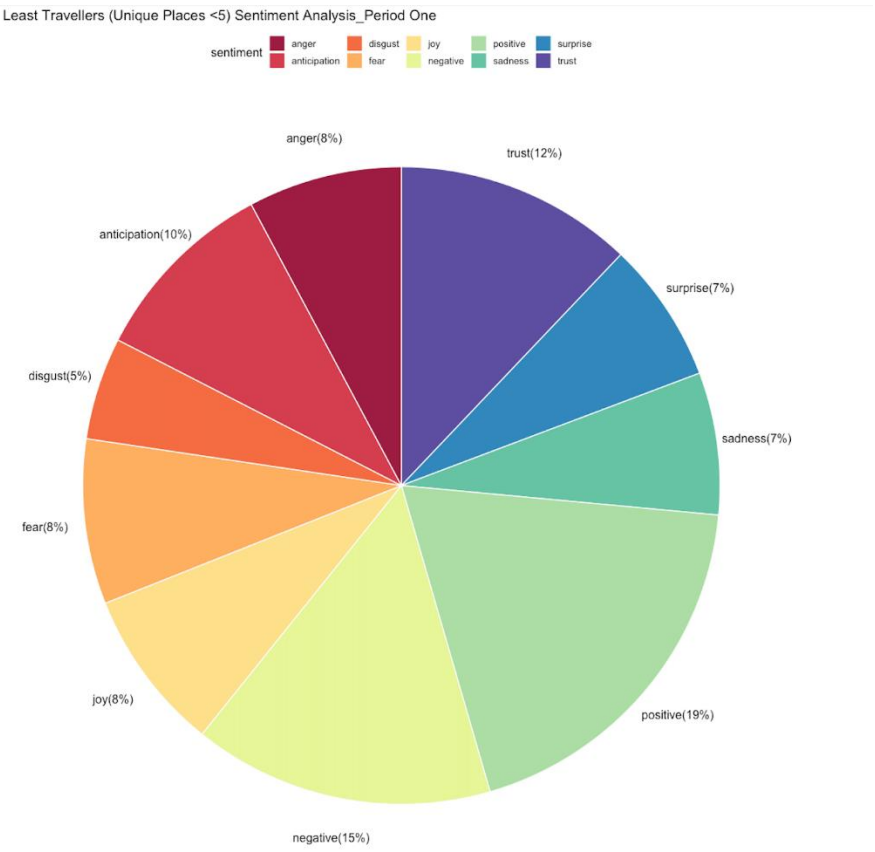


Top Travellers (Unique Places >120) Sentiment Analysis_Period Two

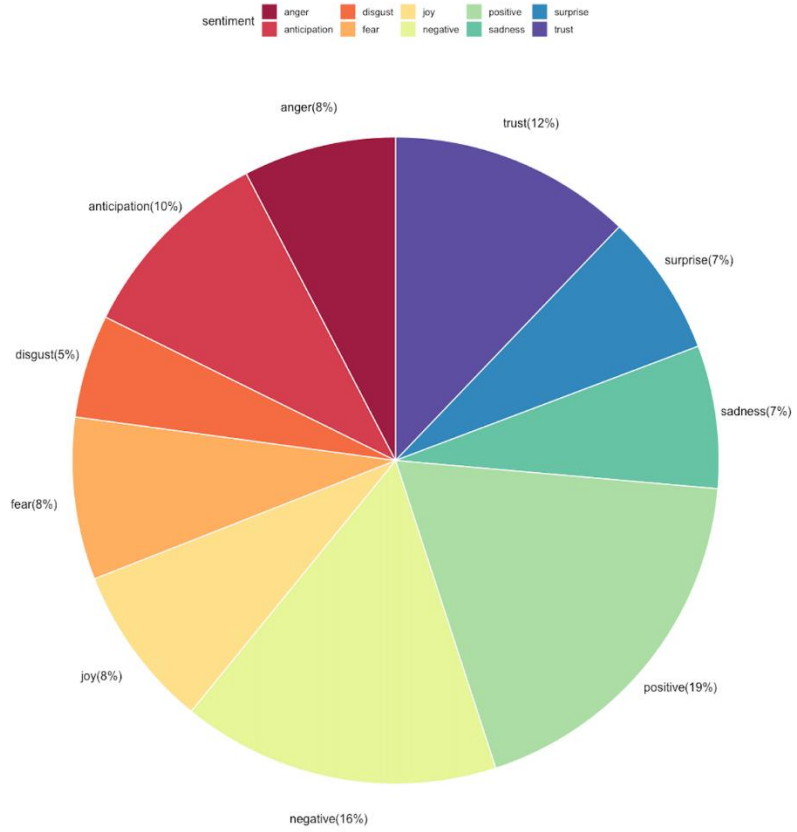


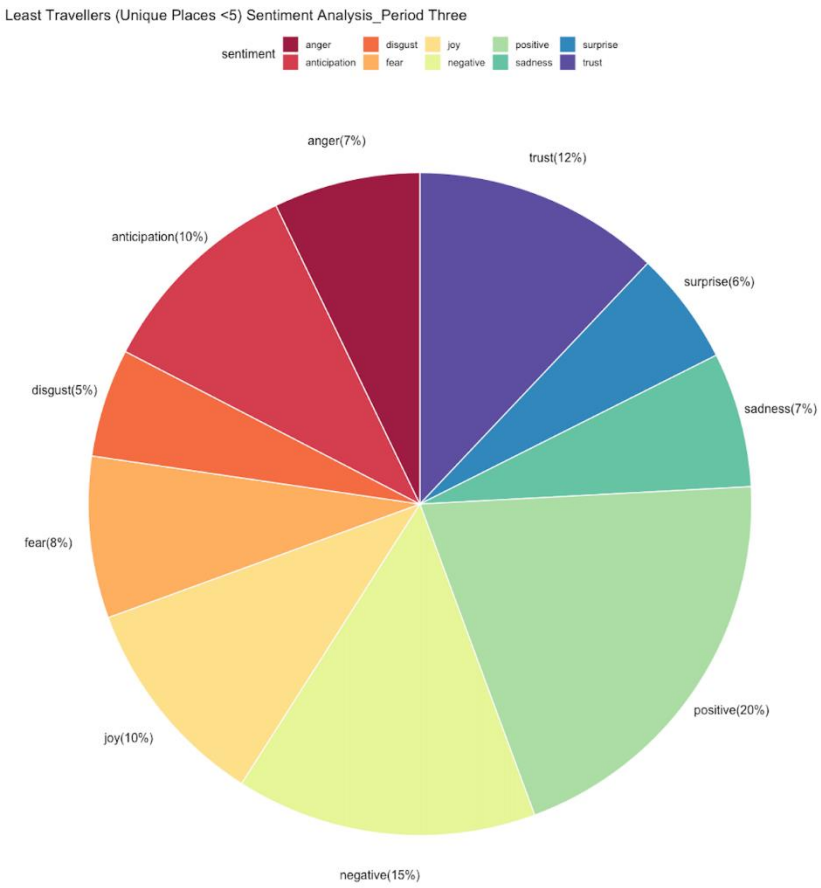


The above graphs show the sentiment analysis of the top travelers among the three time periods: March 2018 to October 2018, September 2019 to May 2020, and October 2021 to July 2022. The sentiment analysis among the three time periods does not show significant emotional change as we expected. The top travelers posted slightly more negative content during the second time period at the beginning of the pandemic and showed more negative emotions. However, the change is not as significant with only one percent difference on the emotion of "negative," "sadness," "trust," and "anger." People's sentiment becomes more positive from October 2021 to July 2022 with one percentage increase on positive emotions, such as "anticipation," and "joy." Among the three periods, people's positive and negative emotions stayed around 22% and 14%, respectively. Compared to the top travelers, we performed the same sentiment analysis on the least travelers below.


Least Travellers (Unique Places <5) Sentiment Analysis_Period One
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
anger(8%)
trust(12%)
anticipation(10%)
surprise(7%)
disgust(5%)
sadness(7%)
fear(8%)
joy(8%)
positive(19%)
negative(15%)

Least Travellers (Unique Places <5) Sentiment Analysis_Period Two





The analysis on least travelers showed more negative emotions compared to the top travelers. Among the three periods, least travelers' positive emotions range around 19% compared to the 22% percent for top travelers. Least travelers also showed more emotions on "anger," "fear," and "sadness." Like the top travelers, sentiment analysis shows less impact of the pandemic to people's emotions for least travelers than we expected. To get a closer look at the content of top and least travelers' tweets, we performed a wordcloud using the R package "wordcloud2" and "dplyr." (Kolapo Obajuluwa)

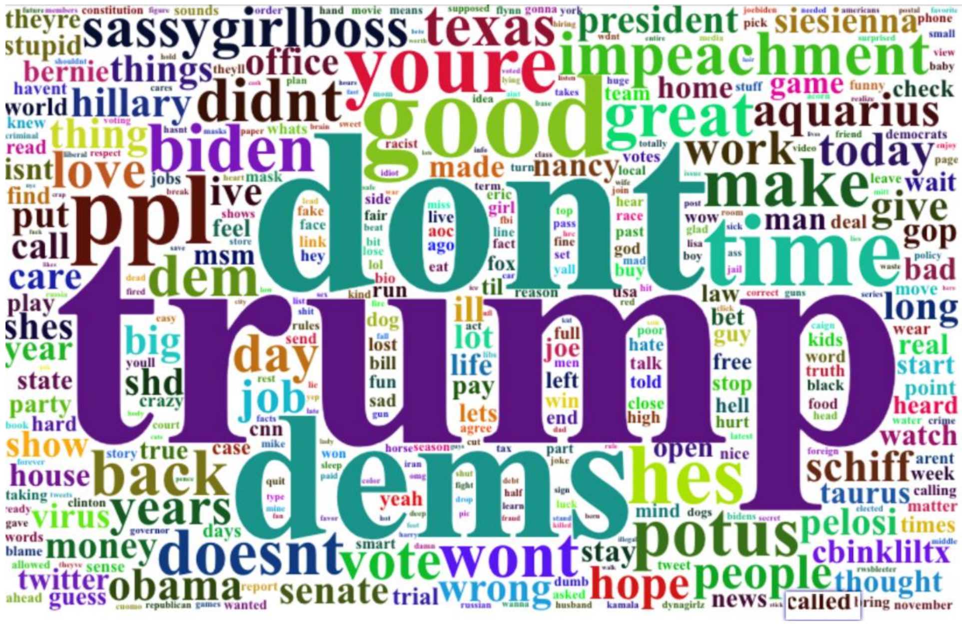
Wordcloud for Top Travelers Period One:

Wordcloud for Top Travelers Period Two:

Wordcloud for Top Travelers Period Three:

Wordcloud for Least Travelers Period One:

Wordcloud for Least Travelers Period Two:

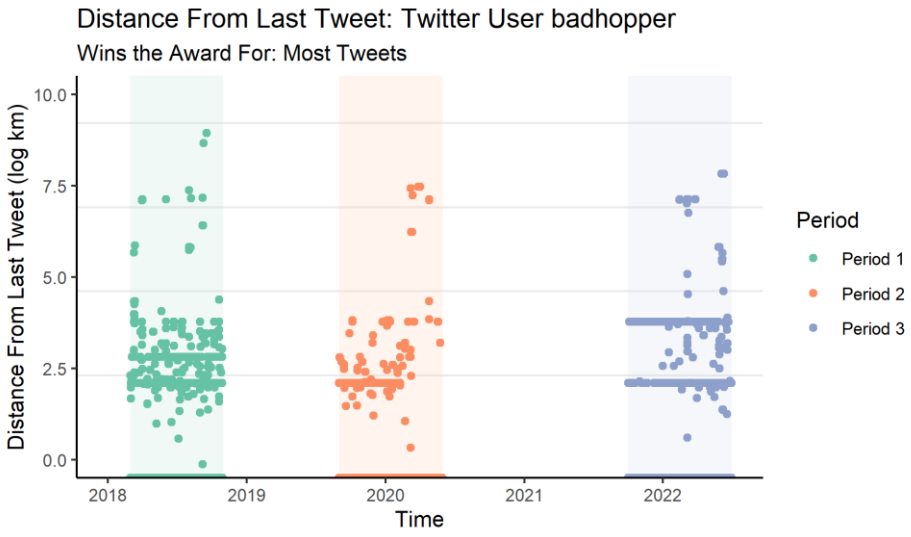


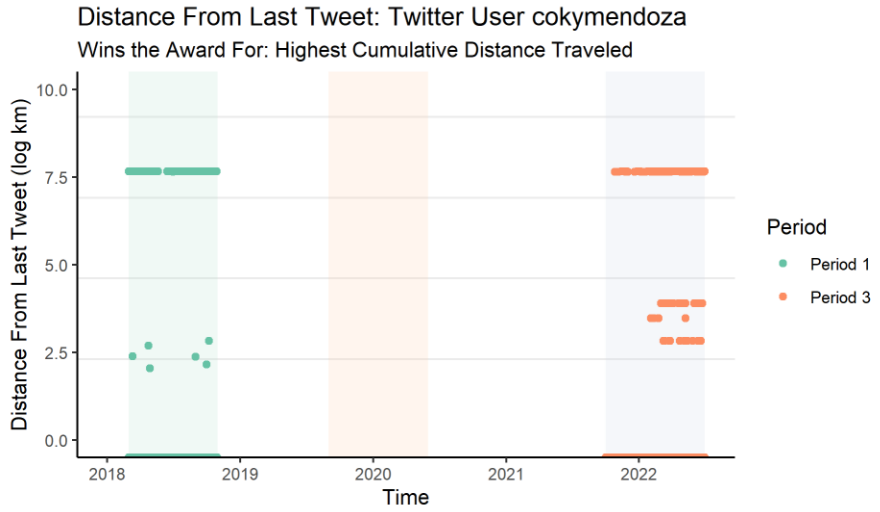
Wordcloud for Least Travelers Period Three:

Among the words most frequently appeared in the top travelers' tweets, we can see more positive words, such as "Cheers" throughout the three time periods. This coincides with the sentiment analysis as top travelers generally have more positive emotions. Among the least travelers, negative words tend

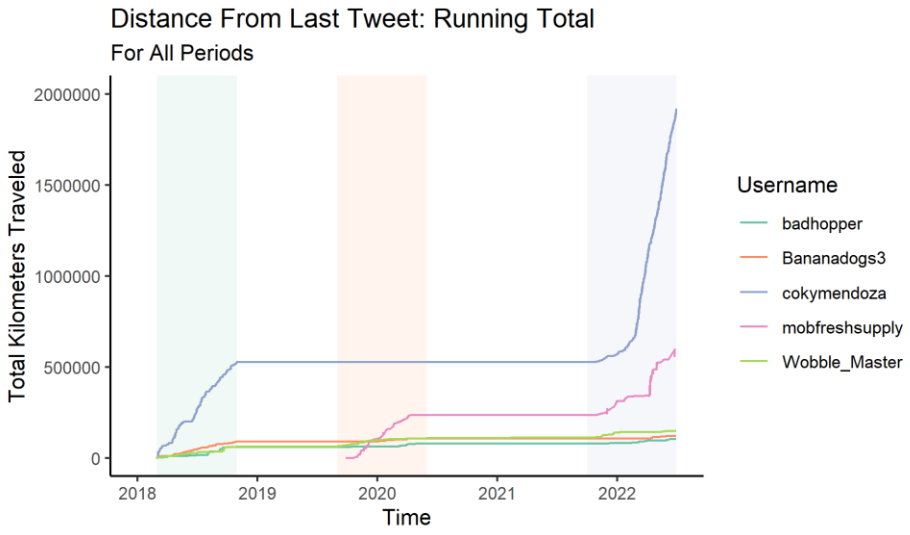
to appear more in the tweets. Specifically, topics mentioning politics and the president tend to appear more. Topics mentioning the pandemics did not seem to occupy both travelers' tweets as we expected. Based on the analysis, least travelers tend to pay more attention to politics and express negative emotions towards political events where top travelers tend to focus less on politics and express more positive emotions.

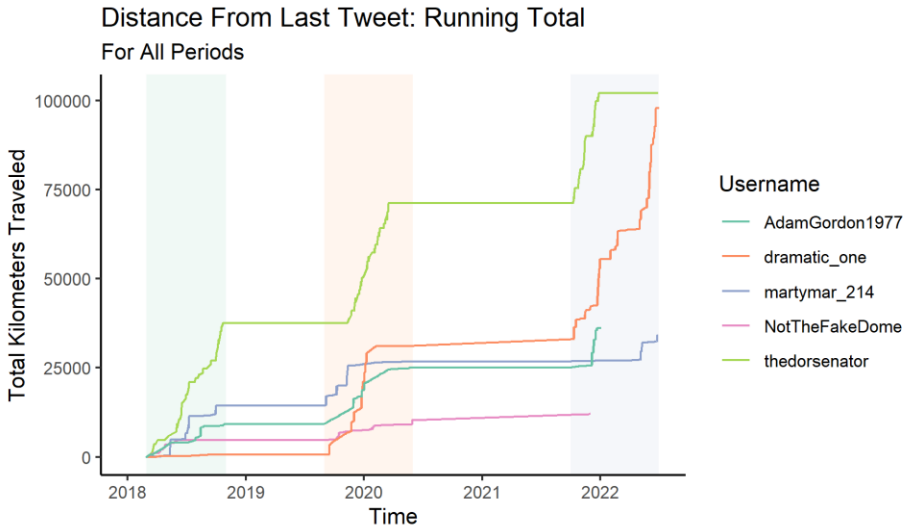
Travel Distance Visualizations





Above are the scatterplots for two selected usernames showing distance from the last tweet. The y-axis shows the distance in log kilometers, so to assist with interpretation, we added grey lines at the values of 10 kilometers, 100 kilometers, 1,000 kilometers and 10,000 kilometers. An interesting pattern of straight lines emerged from the points, indicating the user traveling back and forth from the same location regularly, most likely school or work. This means the more scattered the points appear, the more variation in locations the user traveled to. For some users, the points showed more variation early in the period before the pandemic before showing less variation after 2020.



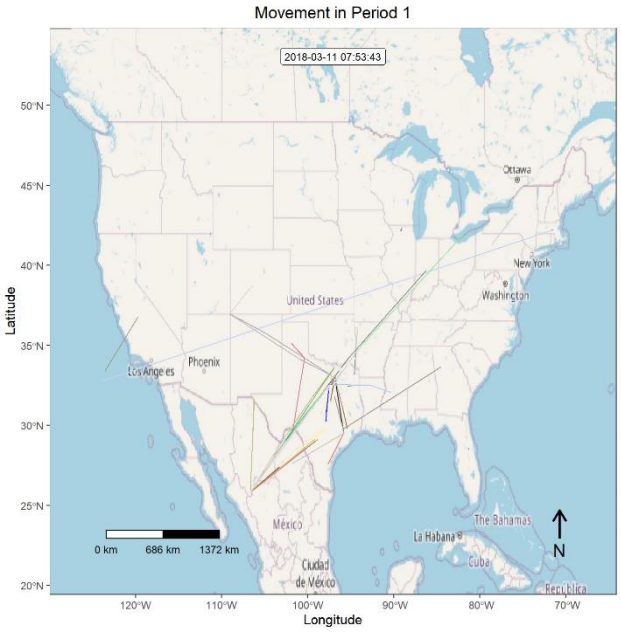


The above graphs show the cumulative distance traveled for two groups of five users in kilometers. The focus of the analysis is on comparing the slope of a single user's line in each of the highlighted periods. For the selected users, there was large variation in the rate of change of total kilometers traveled for each user. Some users showed no visible change of movement pattern in any period, while others showed an acceleration of travel in the last period and relatively little travel in the previous periods.

We found the median total distance traveled for the 112 unique users who tweeted in all periods to be 13,755 km for the first period, 13,798 km for the second period, and 16,425 km for the third period. This indicates that these users traveled approximately the same amount between the first and second periods, but then movement increased dramatically in the post-lockdown period. This could be due more to the cutoffs for the data collection periods, because even though we expected to see a decrease in movement during the lockdowns, as the second period contains data points from September 2019 through May 2020, it contains mostly data from before the lockdowns that began in April 2020.

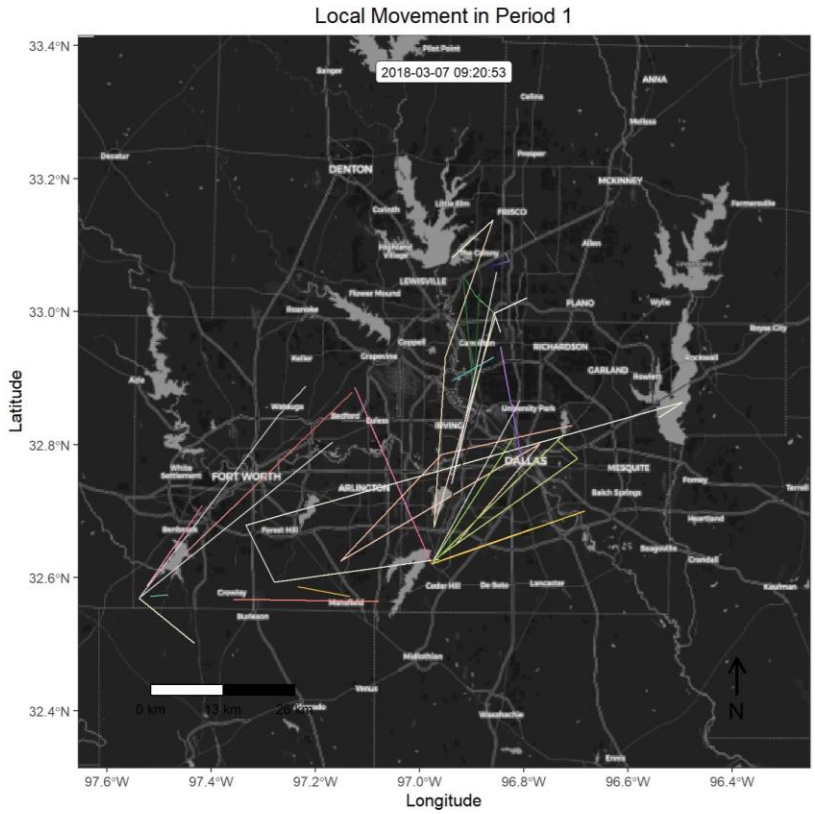
Movement Data Animated

Period 1 of our data collected from Twitter ranges from March 1, 2018 to October 21, 2018, and serves as a benchmark of normal human mobility for the users in our sample. Sampled user movement during this period can be viewed as a video by clicking the link embedded in the map to the right. As this is a benchmark, comparisons



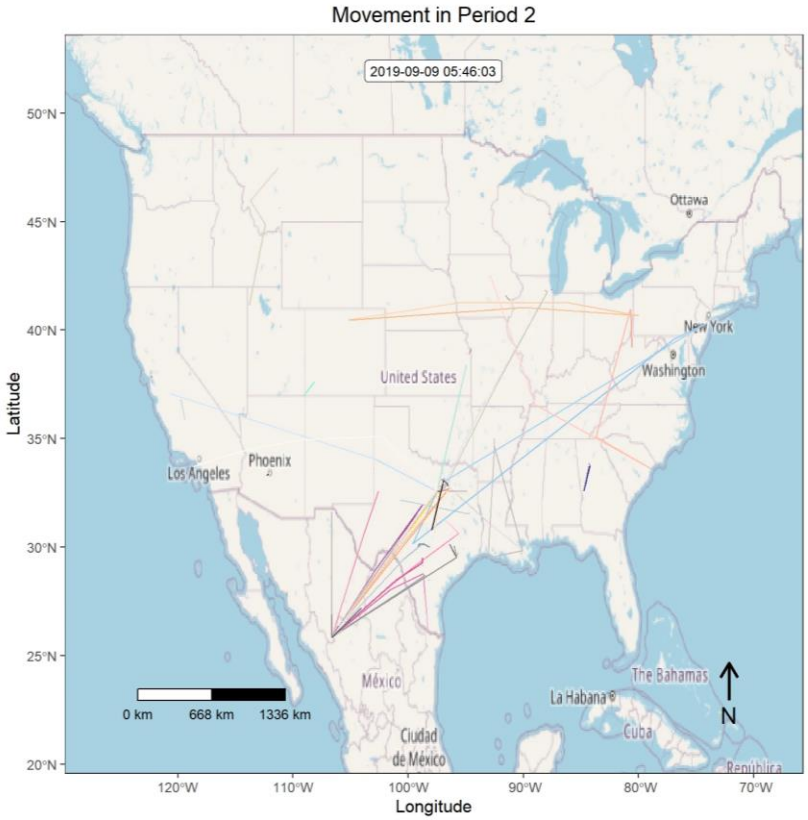
can only be made after also viewing the other two periods, however, observations can be made on patterns in the data. For example, there is significant travel to Mexico, as well as to hotspots on the East Coast of the United States such as Washington D.C. and New York. Travel within Texas to Austin, San Antonio, and Houston is also apparent.

Data from period 1 was also visualized on a more local level, focusing on movement in only the DFW area. This visualization can be viewed as a video by clicking the link embedded in the map below. Although patterns become harder to visually discern with the data restricted to this smaller area, the application of human mobility models or machine learning methods could reveal more information. However, this is outisde the scope of our project.

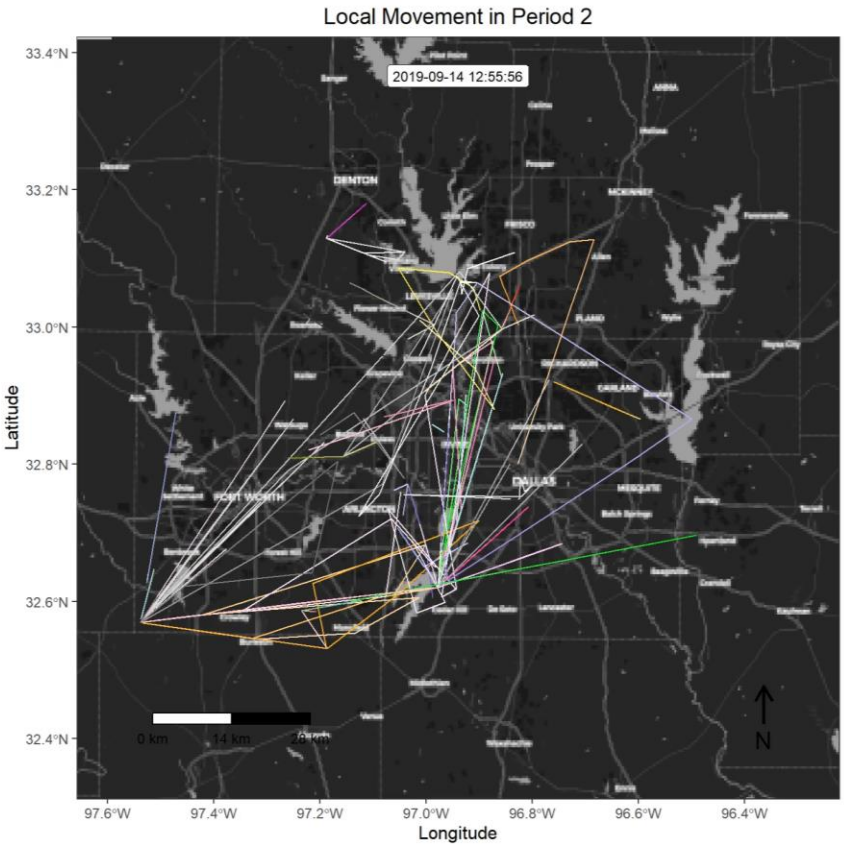


Period 2 of our data ranges from September 1, 2019 to May 31, 2020, and is significant in showcasing how effectively shocks to human mobility can be visualized on a map. Near the end of

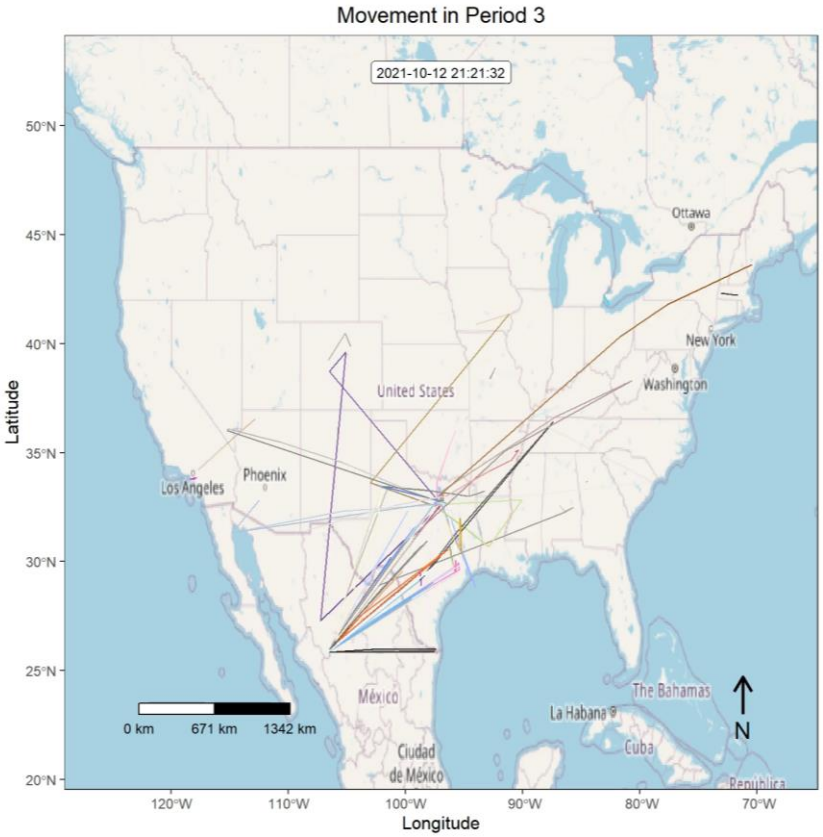
this period marks the start of social distancing restrictions brought on by the COVID-19 Pandemic, and as can be seen in the visualization user movement suddenly decreases to a significantly lesser amount than is observed throughout period 1 or for the beginning of period 2. Many of the same movement patterns observed in period 1 are also present before the shock in period 2. However, it is important to note that the time of year is flipped between period 1 and 2, as period 1 starts in the spring and ends in the fall, whereas period 2 starts in the fall and ends in spring. Period 2 and period 3 are only offset by 1 month. Therefore, comparisons should be made with this in mind. This visualization can be viewed as a video by clicking the link embedded in the map below.



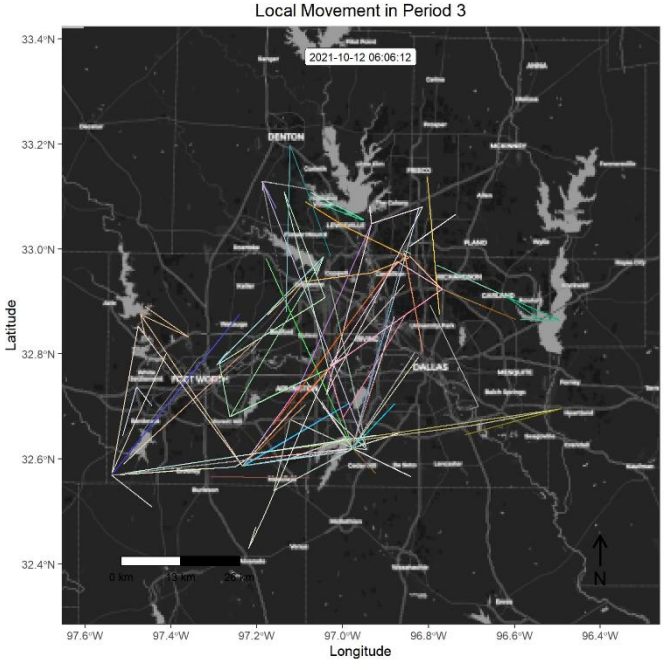
The local level visualization for period 2 shares the same shock as its national counterpart and gives a good idea of the extent to which the pandemic limited movement. This visualization can be viewed as a video by clicking the link embedded in the map below.



Period 3 of our data ranges from October 1, 2021 to June 31, 2022, and can be viewed as a video by clicking the link embedded in the map below. Comparing this period to period 1 will help us to answer our research question on how DFW residents' mobility patterns may have changed from before the COVID-19 Pandemic to after social distancing restrictions had mostly ended. The decrease to residents' mobility during the pandemic is made apparent by the visualization of period 2, and as covered in the literature review is a topic that has already been explored. Using only the videos to compare the two periods reveals some of the flaws in a visual analysis. Although it may seem like user mobility increased from period 1 to period 3, a quantitative comparison is necessary to reach accurate conclusions.



The local level visualization for period 3 can be viewed as a video by clicking the link embedded in the map to the right. Comparing local level visualizations between periods 1 and 3 leads to the same problem of being unable to reach definitive conclusions without a quantitative analysis to support them.



Conclusion

This analysis contains certain limitations. Due to the nature of Twitter data, the analysis lacks generalization due to the targeted population. As an online social media platform, the user demographic is limited. In addition, due to Twitter's recent change on Twitter's geographic metadata, only a limited number of tweets contain geographic information. For our analysis, slightly over half of the tweets generated contained coordinate information for us to analyze. These limitations are recognized by many scholars using tweets for analysis as well. For our analysis, due to time constraints and concerns on the processability of the data, we were only able to analyze tweets from 350 users. To decrease selection bias in a future analysis we would like to expand the sample to more users. Furthermore, the time periods selected for this analysis may not cover the full impact of the pandemic on the population. To better analyze the impact of the pandemic, we would like to expand the time period to the end of 2020. Lastly, we would like to expand the research on wider geographic areas, such as rural areas, to cover more demographics.

Our project yields important conclusions for the study of human mobility, especially human mobility in relation to the COVID-19 pandemic. As expected, and as studied in existing literature, for the users in the sample human mobility during social distancing restrictions suddenly and drastically decreased. The more significant finding is that in the period after social distancing restrictions ended human mobility in our sample increased to a level greater than what was observed in the period before social distancing restrictions were put in place. Validating this study and generalizing it to a broader population would have implications for if an event with an impact on human mobility similar to COVID-19 were to occur in the future.

In relation to data analysis methods, we found that animation can be an effective visualization for mobility data, especially for shocks in the data. As seen in our visualization the sudden decrease in human mobility at the start of the pandemic was easily observable. However, in order to reach more definitive conclusions for not as easily observable events such as the increase in mobility we saw from users after the pandemic, a quantitative analysis was necessary.

In relation to data collection, Twitter remains an easily accessible source for large amounts of movement data, even after changes by Twitter that decreased the number of geotagged tweets available. Whether this will remain true in the future as data privacy becomes a larger concern is questionable. Also, as explored in other works, depending on the target population of a study, Twitter data might not yield a representative sample. In the process of generalizing the study, it is possible that specific demographics will not be captured.

Synergy Report

Alden Felix: Collected Twitter user handles in the DFW area that were passed on to team for collecting tweets. Visualized national and local level movement animations for three analysis periods using sample of tweets received from team. Worked on proposal, proposal presentation, final presentation, and final project paper.

Zeyu Sun: Presented idea for the project and performed initial literature review. Used Twitter user handles to collect tweets made during the 3 periods chosen for analysis. Performed and visualized sentiment analysis on collected tweets for top and least travelers, and visualized word clouds for these groups' tweets. Worked on proposal, proposal presentation, final presentation, and final project paper.

Kyndall Brown: Performed quantitative analysis on collected movement data for all three periods, calculating and visualizing distance between tweets for selected users and cumulative distance traveled. Compiled statistics on data and calculated median total distance between tweets to answer the research question. Worked on proposal, proposal presentation, final presentation, and final project paper.

Appendix

Link to project repository: https://github.com/aldenfelix/EPPS_6302

References

Armstrong, Caitrin, et al. "Challenges when identifying migration from geo-located Twitter data." *EPJ Data Science* 10.1 (2021): 1.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C., Lenormand, M., Louail, T., Menezes, R., Ramasco, J., Simini, F. and Tomasini, M., 2018. Human mobility: Models and applications. *Physics Reports*, 734, pp.1-74.

Chi, Guanghua, et al. "A general approach to detecting migration events in digital trace data." *PloS one* 15.10 (2020): e0239408.

Colizza V, Barrat A, Barthelemy M, Valleron A-J, Vespignani A (2007) Modeling the Worldwide

Hack Your Data Beautiful, "Scraping and Visualising Twitter Data", https://psyteachr.github.io/hack-your-data/scrape-twitter.html

Hübl, Franziska, et al. "Analyzing refugee migration patterns using geo-tagged tweets." *ISPRS International Journal of Geo-Information* 6.10 (2017): 302.

Kolapo Obajuluwa, "Twitter Word Clouds with R," https://rpubs.com/kolaoba/twitterwordclouds

Nestorowicz J, Anacka M. Mind the Gap? Quantifying Interlinkages between Two Traditions in MigrationLiterature. International Migration Review. 2019; 53(1):283–307. https://doi.org/10.1177/0197918318768557

Pineda-Krch, M. (2011, May 12). Great-circle distance calculations in R. R-bloggers. https://www.r-bloggers.com/2010/11/great-circle-distance-calculations-in-r/

Rogers A, Raymer J, Newbold KB. Reconciling and translating migration data collected over time intervals of differing widths. The Annals of Regional Science. 2003; 37(4):581–601. https://doi.org/10.1007/s00168-003-0128-y

Rudis, Bob. 2018. 21 Recipes for Mining Twitter Data with rtweet(https://rud.is/books/21-recipes/)

Saif M. Mohammad and Peter Turney. (2013), ``Crowdsourcing a Word-Emotion Association Lexicon." Computational Intelligence, 29(3): 436-465.

Spread of Pandemic Influenza: Baseline Case and Containment Interventions. PLoS Med 4(1): e13. https://doi.org/10.1371/journal.pmed.0040013

Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, et al. (2014) On the Use of Human Mobility Proxies for Modeling Epidemics. PLoS Comput Biol 10(7): e1003716. https://doi.org/10.1371/journal.pcbi.1003716

Willekens F. Models of migration: Observations and judgement. International migration in Europe: Data, models and estimates. 2008; p. 117–147.

Yin, Junjun, Yizhao Gao, and Guangqing Chi. "An evaluation of geo-located Twitter data for measuring human migration." *International Journal of Geographical Information Science* 36.9 (2022): 1830-1852.

Zagheni, Emilio, et al. "Inferring international and internal migration patterns from Twitter data." *Proceedings of the 23rd international conference on world wide web*. 2014.