“Using Twitter data to determine the impact of poverty on political engagement. Evidence from the US Capitol Hill insurrection.”

Author:

Omiros Petrou

Supervisor:

Andros Kourtellos

Submitted in partial fulfillment of the requirements for the MSc degree in Economic Analysis at the University of Cyprus.

May 2022

Abstract

Social media websites have established themselves as an essential part of our daily lives, where people can share their opinions on any topic of their interest, and have been consistently growing over the last decade. It is estimated that in 2017 the number of social media users worldwide was 2.86 billion and it is expected to reach 4.41 billion by 2025[1]. Despite their exponential growth and the large volumes of data that are generated, big data, that is, data generated from social media networks, have not yet established its significance and are not being utilized to the degree one could think, especially in academic literature. In recent years, however, this has changed, as more and more academic papers using data from the social media website Twitter have made their debut.

In this thesis, we attempt to investigate the role of poverty on the decision of Twitter users to politically engage/express themselves once an unprecedented event is realized. The objective is to classify geographical areas in the United States as high or low poverty areas on the county level, and then using the global position system (GPS) we identify the location of users at the time they sent a tweet. We then proceed by aggregating the total volume of tweets sent across all areas of interest for the period we investigate. By employing a dynamic panel model to control for intrinsic characteristics across counties and time effects to control for changes, we identify that poverty on the county level is a contributing factor to the decision of Twitter users to politically engage/express themselves during the US Capitol Hill insurrection.

Keywords:

Social media data, Twitter, dynamic panels, GIS, Insurrection.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Andros Kourtellos for the continuous support and guidance throughout my MSc Thesis. His overall contribution and academic expertise helped me overcome challenges throughout this journey and set the ground for new learning experiences that were within the scope of my thesis and beyond.

Contents

List of Tables

List of Figures

1. Introduction

1.1 Context and motivation

A large and diverse literature is based on longitudinal surveys, millions of administrative tax records, and randomized control trials that provide convincing evidence that growing in poor areas undermines life opportunities. A major factor that contributes to this effect is that poor neighborhoods are geographically isolated from middle-class environments of opportunities, and as a result, there is limited access to middle-class role models, safe environments, and institutional resources, as well as information about vacancies and general information about the labor market. Nevertheless, the academic literature that studies neighborhood isolation so far assumes that social interactions are restricted in one's neighborhood of residence. This is assumption is highly questionable in interconnected economies where i) individuals have access to public transportation; ii) the cost of owning a private vehicle is moderate and as time passes it becomes more affordable, and iii) the internet is widely accessible in most developed countries is a place with abundant resources of information that can be accessed for free without discrimination. Qi Wang et. Al. (2018) attempt to address this issue, that is, the assumption that social interactions are limited to one's neighborhood, by leveraging geotagged tweets where they develop a test for neighborhood isolation and their analysis suggests that minority neighborhoods in poor areas do not limit themselves in their neighborhood and have travel patterns similar to other groups.

So far, the scholarly community has shown great interest in isolation that stemmed from physical barriers or isolated geographic areas. Since the internet enables us to socialize

ourselves in a spectrum where income, physical barriers, and distance are insignificant, a worthwhile question would be: If physical barriers are no longer in place, are people living in disadvantaged areas isolated in one way or another?

In this paper, we attempt to fill this gap of knowledge by using unsolicited messages posted by users living in the United States on the micro-blogging website Twitter. With the exponential growth of social media websites and the internet, given the availability of devices that provide access to social media websites via a browser or an application, offers a great opportunity to collect data that can be used for research by statistically analyzing them. Unlike traditional methods, using data from social media websites can provide real-time insights and measure fluctuations from occurrences of extemporaneous events.

In this paper, we argue that social media posts can be utilized in the following applications:

1. Establish the significance of different events (e.g., political) and compare their impact by measuring changes in the volume of messages.
2. Test whether demographic characteristics from different areas play a role in the decision of users to engage themselves.

For this exercise, we collect all geotagged tweets sent from December 17, 2020, to January 25, 2021, in states with metropolitan areas. For the collection of data, we used the snscrape[1] (development version) which is a python library that allows users to scrape things like user profiles, hashtags, or searches and returns the relevant posts. Unlike Twitter API, which is a service offered by Twitter, snscrape is free of charge but limits the posts users can download, to approximately 20 posts per second. Because of the limitation imposed by snscrape, we focus on states with metropolitan areas for our research.

[1] https://github.com/JustAnotherArchivist/snscrape

The dataset consists of roughly 8.27 million tweets, with their corresponding, time of posting, text, precise longitude and latitude coordinates, a unique tweet identifier, and a unique user identifier. By locating and analyzing the daily volume of millions of social media posts, we create dynamic panels and find that i) the US Capitol Hill insurrection was a significant event that caused an abrupt change to the volume of tweets sent by users; and ii) counties classified as high poverty areas had a relatively lower response to the event, compared to the response of low poverty areas.

1.2 Thesis outline

This paper is organized as follows:

- Section 2 is the literature review that summarizes the existing literature on how poverty relates to the isolation and how data have been used so far.
- Section 3 summarizes our sample, and we show choropleth maps with patterns of user activity forming.
- Section 4 summarizes the steps and rationale of how we build our model.
- Section 5 presents our findings once we have fitter our model specified in section 4.
- Section 6 summarizes our result and addresses limitations.
- Section 7 discusses future work and improvements.

2. Literature Review

Increasingly, the research community is turning to big data, and more precisely to Twitter as it enables researchers identify the location of users. A prime example of how the location of users can be used is from Agustín Indaco (2020) who uses geotagged tweets with images shared in 2012-2013 and finds that the volume of tweets aggregated at the country level can be used as a proxy to estimate current GPD in USD. He argues that traditional methods of measuring GPD are often expensive, complicated, and might result in measurement error, especially in developing countries. Another concern that is stated is the incentive of manipulating official GPD estimates in terms of both market fluctuations as well as favorable shifts in the public's opinions on political figures. Additionally, they use night-light data to detect economic activity and find that the goodness-of-fit of Twitter data is comparable to that of the night-light data. Because of the geographic granularity of tweeter data as they provide the location of a user with high precision, they exploit them to estimate GPD at the sub-national level and conclude that Twitter data can be used to measure economic activity in a timely and spatially disaggregate manner relatively to conventional data.

The content of tweets can also be used for research as Curini et. Al (2014) examined Twitter posts sent by Italian users to investigate which idiosyncratic shocks affect happiness. They argue that contrary to traditional questionnaires, using tweeter data can provide real-time insights about happiness and can measure the impact of extemporaneous facts based on the fluctuations of their original happiness index (iHappy). To create iHappy they used machine learning models to classify their data as either positively or negatively at the provincial level and find that the static variables such as overall quality institutions have a marginal effect on the overall level of happiness. On the contrary, dynamic variables that might not necessarily be extemporaneous such as

the spread of Italian and German bonds or the day of the month where salaries are deposited have the greatest impact.

Twitter data can be used on the macro level as mentioned with night lights, but it can also be used on the micro level as Q. Want et. al. (2018) attempt to shed light on urban mobility flows in America's 50 largest cities and examine the impact of poverty on mobility. Contrary to previous works, they do not rely on the implicit assumption that residents limit themselves to their neighborhoods' boundaries. Using geotagged Twitter data, they were able to measure the radius and spread of travel for each user and find clear discrepancies in users' exposure to "mainstream" areas. Their analysis shows that even though residents in disadvantaged neighborhoods travel as far and wide as their counterparts, their relative segregation and isolation have some persistence.

Despite their macro and micro level potential, Twitter data can not be used to make inference about the greater population. The research community seeks to overcome this challenge as Yildiz et. al attempt to identify methods that can be used to do demographic research using Twitter data. They provide alternative ways that can be used to determine the sex and age of users and propose best practices for estimating Twitter user's demographic characteristics and calibration methods to address selection bias in the Twitter population, enabling researchers to generalize findings and use Twitter data to make inference about the general population.

3. Methodology

Twitter is an online social media website that allows its users to post short messages of any subjet of their choosing. These messages are referred to as tweets. Twitter was founded in 2006 and by 2020 Twitter reached 3.6 billion users with an average of 500 million tweets per day. Historically, Twitter was designed to be accommodated as an SMS mobile platform and the number of characters per tweet was limited to 140. As Twitter started growing, the platform switched from a phone-based platform to a web-based platform and eventually allowed users to add images to their tweets, however, the character limit remained as a type of branding. Tweets can be accessed publicly and can be read from either using Twitter's application or any browser as long as users have an internet connection and a registered Twitter account. Users have the option to restrict their tweets and keep a private profile that is inaccessible to other users and scrapers[2].

Tweets and their metadata can be purchased directly from Twitter if the tweets were sent by users who have their profile public, meaning anyone can read their tweets from their browser. Twitter allows users to download tweets free of charge via the Twitter API with a limit of 15,000 tweets per day, preventing users to download large-scale datasets.

To overcome this obstacle, we use snscrape (development version) which is a scraper for social networking services (SNS) developed in python. This package is free of charge and can download up to 1.5 million tweets per day for each instance. Snscrape requires no API key so multiple queries can be instantiated.

Because of the limited download rate offered by snscrape, we focus our analysis on states with metropolitan areas. Out of the 50 states in the United States, we select to query for

[2] Software that collects information.

all tweets sent from Washington, Florida (Miami), California (Los Angeles), Texas (Houston, Dallas), Pennsylvania (Philadelphia), New York, and Illinois (Chicago).

Our dataset contains all geotagged tweets posted from December 17, 2020, to January 25, 2021, across the areas of interest. The dataset contains 8.27 million tweets including i) a user identifier; ii) a tweet identifier; iii) content of tweet; iv) time and date of posting; v) longitude and latitude coordinates.

Table 3-1 provides the distribution of tweets across all areas of in (AoI) including the area, population, and the percentage of the population that has been sampled. We use the unique user identifier to identify how many users we have in our sample and normalize it with the total population of their corresponding state as per the ACS 2020 5-year estimates. We observe that Florida has the highest number of engaging users on Twitter with 0.41% and Pennsylvania has the lowest number of users with roughly 0.16%. On average, 0.2984% of the US population living in the areas of our study is a tweeter user with the geotagging option enabled.

Table 3-1 Population and sample for AOI

State	Area (km²)	Population (millions)*	No. Tweets	Users	Sample
California	423,970	39.346	1,854,454	103,624	0.2634%
Florida	170,312	21.217	1,557,907	88,634	0.4178%
Illinois	149,998	12.716	734,241	38,536	0.3030%
New York	784	19.515	1,413,909	64,037	0.3281%
Pennsylvania	119,282	12.795	352,101	20,954	0.1638%
Texas	695,662	28.635	1,979,402	104,654	0.3655%
Washington	184,666	7.512	403,976	18,583	0.2474%

*ACS 2020 5-year estimate

Table 3-2 summarizes the mean and standard deviation of tweets for each group of our study. We define low poverty areas, counties where the poverty level is below the 5^{th} percentile, and high poverty areas where the poverty level is above the 95^{th} percentile of their corresponding state. On average, the 5th percentile represents areas where the poverty level is below 6% and the 95th percentile represents areas where the poverty level higher than 15.4%.

By classifying counties as low, high, and normal poverty areas, we observe that areas with high poverty on average have the lowest user activity, whereas low poverty areas have on average a higher user activity. It is worth mentioning that tweets sent from areas classified as normal poverty have the highest number of tweets. Our results from table 3-2 could indicate the following: i) tmoreterest people are living in low poverty areas than high poverty areas and the majority of the population lives in areas classified as normal poverty; ii) after a certain level of income, users reduce their activity in social media (Twitter). Further analysis is required to make more concrete conclusions.

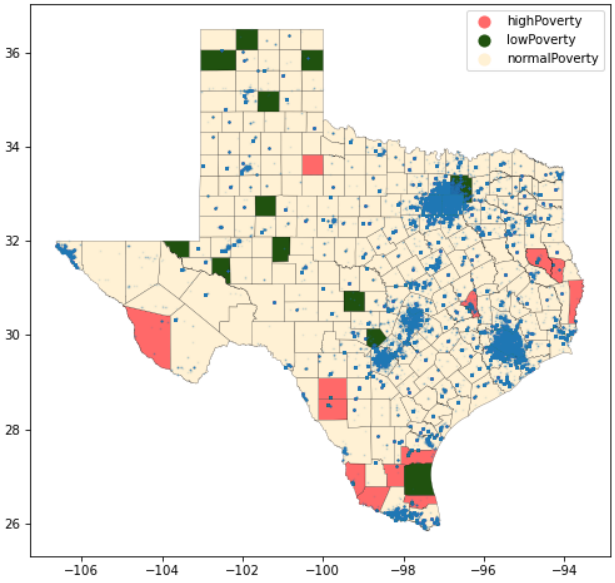
Table 3-2 Summary statistics: Mean and S.d of the volume of tweets

Poverty Level	Mean	Counties
High Poverty	259.91	19
	(435.90)	
Low Poverty	500.14	338
	(577.85)	
Normal Poverty	563.03	23
	(1636.99)	

Below we present choropleth maps for the states our study focuses on, embedded with the tweets we have sampled. Choropleth maps are maps that use different shading, and coloring on predefined areas to indicate the average values of a particular quantity in those areas. To better illustrate which areas we consider as high, low, and normal poverty areas we color our maps based on our county classification and not on the poverty level. We color high poverty areas with the red color, low poverty areas with the green color, and normal poverty areas with beige. Counties can be distinguished from one another

using a light gray color that highlights their border and tweets are represented by the blue dots.

Figure 3-1 State of Texas embedded with tweets

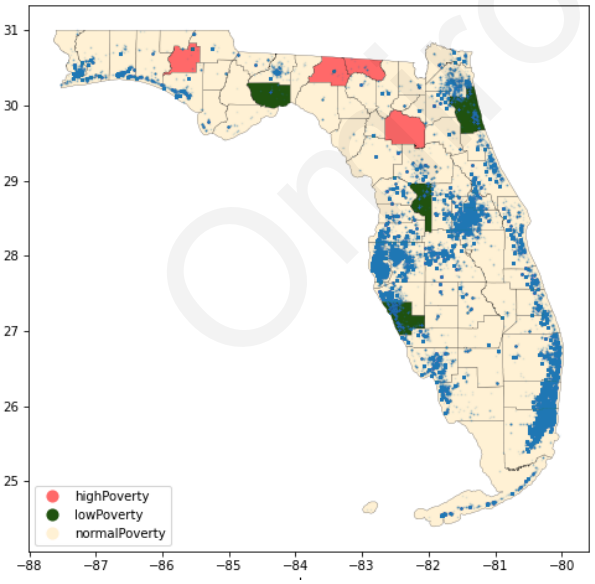


Note: Tweets are represented by a small blue dot using the precise location of the latitude and longitude metadata from

and tweets are represented by the blue dots.

Figure 3-1 illustrates the map of Texas and shows four major clusters of tweets that represent four metropolitan areas with the highest population density. Dallas with an approximate 7.58 million residents has the highest concentration of user activity; Houston is the second most populated area with 7.066 million residents has the second highest user activity and it is located on eastern part of Texas; and Austin with san Antonio have 2.55 million and 2.23 million residents respectively and are in proximity from the center.

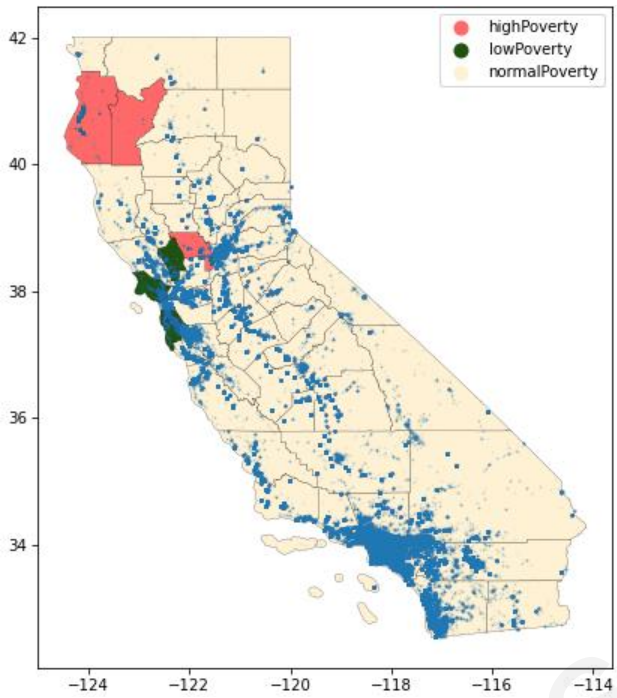
Figure 3-2 State of Florida embedded with tweets



Note: Tweets are represented by a small blue dot using the precise location of the latitude and longitude metadata from Twitter.

Figure 3-2 shows a detailed map of Florida that depicts clear visual patterns of tweets sent from the eastern and western parts of Florida along the coastline. This figure contains four clusters: the Miami metropolitan area, which is primarily consists of the west coastline, Jacksonville which is at the north-western coastline; Orlando which is in central Florida; and Tampa which is along Florida's Gulf Coast which is a major business center. We identify high user activity across Florida's shorelines and areas where high economic activity takes place.

Figure 3-3 State of California embedded with tweets

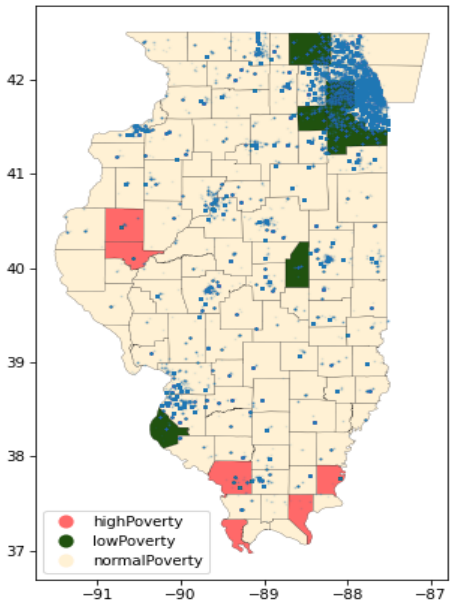


Note: Tweets are represented by a small blue dot using the precise location of the latitude and longitude metadata from Twitter.

Moving to California, Figure 3-3 shows that there exist two distinct clusters located at the north-western and south-eastern parts of the figure. The northern cluster constitutes of the San Fransico and San Jose, and the southern cluster represents Los Angeles and San Diego. We identify similar patterns from figure 3.1 and figure 3.2 where user activity can be found near shorelines.

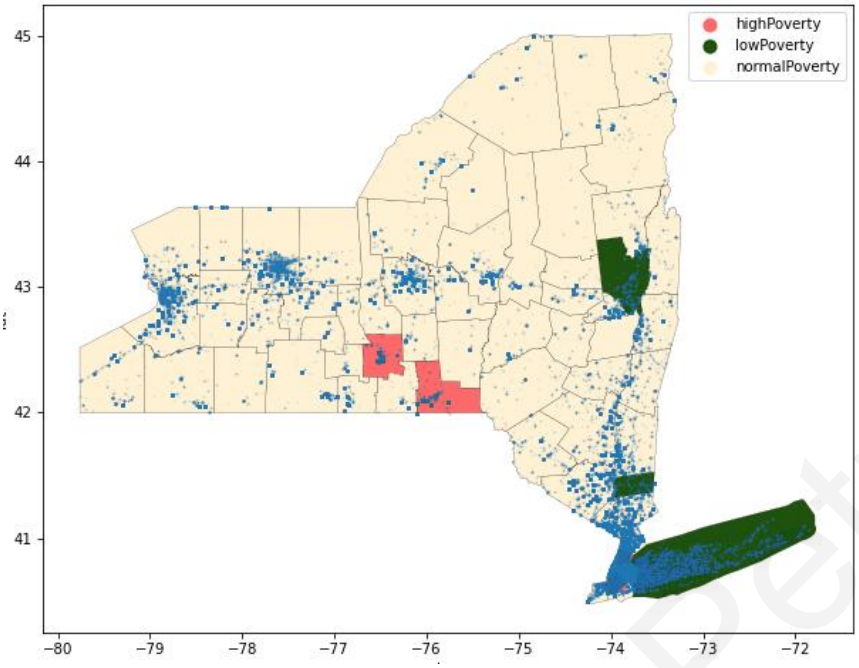
Figure 3-4 State of Illinois embedded with tweets

Figure 3-4 shows the map of Illinois with a distinct cluster located in the north-eastern part. High user activity is located in the Chicago area which is located on Lake Michigan and it is among the largest cities in the United States.



Note: Tweets are represented by a small blue dot using the precise location of the latitude and longitude metadata from Twitter.

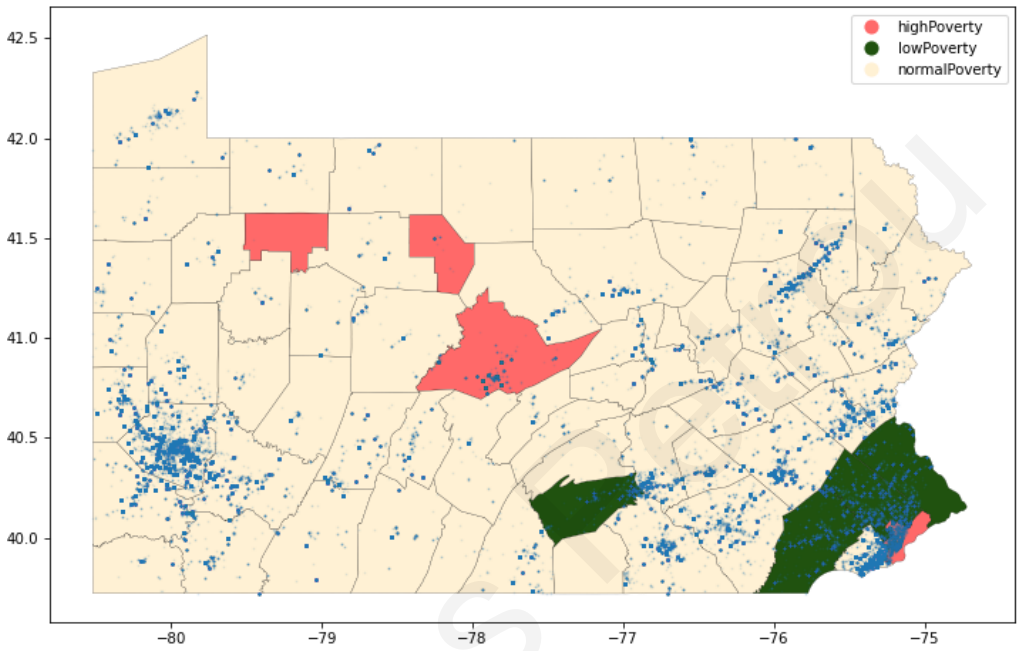
Figure 3-5 State of New York embedded with tweets



Note: Tweets are represented by a small blue dot using the precise location of the latitude and longitude metadata from Twitter.

Consistently with the previous observations, we see that in New York's State map in Figure 3-5 the highest volume of tweets can be found in areas with shorelines. Most user activity can be found in areas affected by the North Atlantic Ocean which also includes a port there. The other four clusters that can be found in the center starting from left to right are Buffalo; Rochester; Syracuse and Albany. All four clusters are linked with interstate roads that are depicted by the scattered tweets that can be found between these Areas.

Figure 3-6 State of Pennsylvania embedded with tweets



Note: Tweets are represented by a small blue dot using the precise location of the latitude and longitude metadata from Twitter.

In Pennsylvania's map in Figure 3-6 we identify two distinct clusters. The western cluster is the city of Pittsburgh which is located at the junction of three rivers and the eastern cluster is Philadelphia.

4. Empirical Estimates and methods

The main goal of this paper is to explore whether Twitter data can be used as a proxy to test whether poverty on the county level plays a role in the decision of a user to express/engage themselves, in a context where geographic isolation are less significant. For this exercise, we collect data around the date of Jan 6, 2021, when the US Capitol Hill Insurrection took place, which was undoubtedly an unprecedented event in modern American history. By focusing on dates around January 6, we expect an unusual increase in the volume of tweets for both high poverty and low poverty areas. We believe that by comparing the change in user activity for both areas with their typical user activity, we will be able to assess whether income plays a role in this type of engagement.

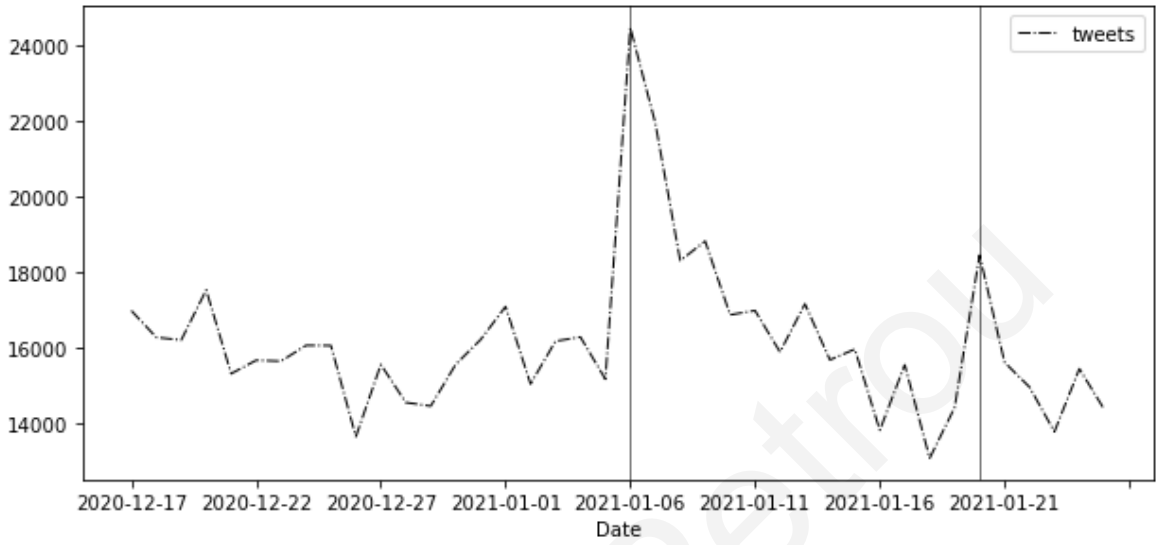
First, we aggregate all tweets on the county level by using the precise location of longitude and latitude coordinates from each tweet metadata and identify the county of origin each tweet was sent. We then proceed with aggregating the volume of tweets by county for each date from 14 December 2020 to 21 Jan 2021 and create panels for a total of 42 counties, of which 19 are considered to be high poverty and 23 are considered to be low poverty for a total of 40 days.

Before we run our econometric model, we attempt to quantify whether the US Capitol Hill Insurecttion is indeed an appropriate event for our analysis and whether there was indeed an increase of tweets during and after that event had taken place. For this reason, our first task is to measure whether there is a structural break on January 6, 2021. A structural break is defined as when a time series abruptly changes at a point in time. In our study, a structural break will involve a change in the mean and that will help us determine whether the event of choice is appropriate for our analysis.

Figure 4-1 depicts the total volume of tweets sent from December 17, 2020, to January 26, 2021, across all states in areas that are classified as either low poverty or high poverty. We observe that there are two structural breaks in figure 4-1, which take place on January 6 which was the insurrection date on the US Capitol Hill, and January 20,

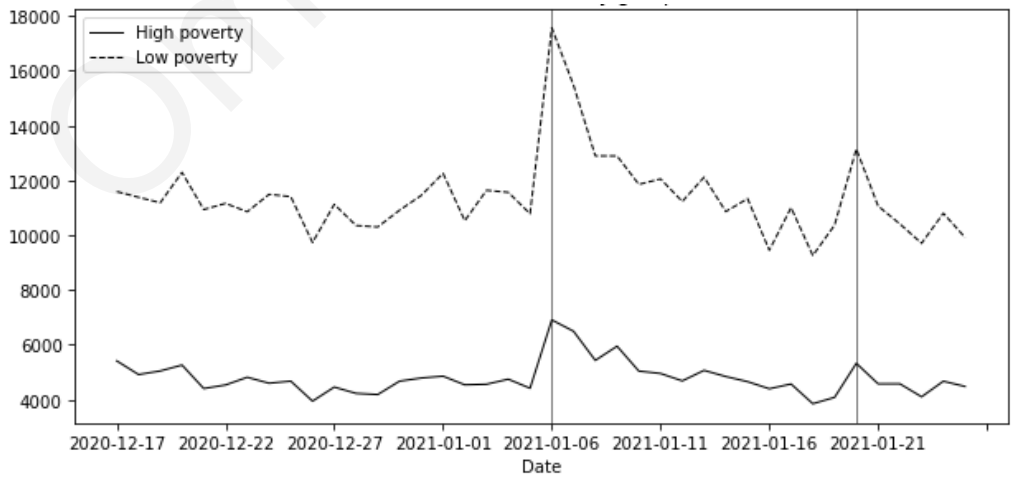
which is the Innaugeration date for every new president. This date marks the commencement of a new four-year term for the president of the United States.

Figure 4-1 Volume of tweets from AOI



In Figure 4-2, break our series from figure 4.1 to low poverty and high poverty areas, and our new time series follows similar patterns to figure 4-1, where we found structural breaks on January 6 and January 20. We also observe that during the US Capitol Hill riots low poverty and high poverty areas peak at their highest level at, 17,500 and 5,500 respectively, gradually decaying, showing a persistent effect of that event.

Figure 4-2 Volume of tweets from AOI by poverty status



More concretely, we proceed to quantify structural breaks more formally, by estimating equations (1) and (2) below:

$$y_{it} = \beta_0 + \beta_1 X_{it} + a_i + e_{it}\,, \qquad X = \begin{cases} 1\ for\ t = t_0 \\ 0\ otherwise \end{cases} \tag{1}$$

$$y_{it} = \beta_0 + \beta_1 X_{it} + a_i + e_{it}\,, \qquad X = \begin{cases} 1\ for\ \ t_0 \le t\ \le t_{0+\Delta_t} \\ 0 \qquad\qquad otherwise \end{cases} \tag{2}$$

where, y_{it} is the level of tweets, β_0 is the intercept, X_{it} is a time dummy variable, a_i are the time-invariant characteristics on the county level and e_{it} is the error term. To estimate equation (1), we first create a time dummy, where we set its value equal to 1 if the date is January 6, 2021, and 0 otherwise. Respectively, we proceed by doing the same exercise for equation (2), however, our time dummy variable is set to 1 for dates starting from January 6 to January 9 (included). Both equations (1) and (2) are estimated using fixed time effects to eliminate unobserved time invariant heterogeneity.

We present our results when we estimate equations (1) and (2) in Table 4-3. In our sample, there are 42 counties, of which 19 are classified as high poverty areas and 23 are classified as low poverty for all 40 days this paper covers. Columns (1) and (2) represent the estimates for equations (1) and (2) respectively, and for each panel group (i.e., high, or low poverty). From Table 4-3, we find that the volume of tweets is greater in low poverty areas relative to high poverty areas and that there exists a structural break in both panels on January 6. We denote the time dummy variable as a break in the table, and it is statistically significant across both models for both panels at the 95% confidence level. Our estimates for structural breaks for low and high poverty areas are 562 and 394 respectively with a positive coefficient, indicating that the Capitol Hill Insurrection in 2021 creates a positive influx of tweets and therefore is a suitable reference date for this study. Additionally, we see that the level of tweets is lower in high poverty areas and this can be explained by the following two facts: (i) From our figures we observe that high poverty areas are mostly located outside city centers where the population density is significantly

lower, and naturally fewer tweets will be sent from these areas; (ii) Twitter is a free of charge social media website with no additional fees for usage, however, there is a cost for purchasing and owning a smartphone and additionally users are required to have an internet subscription. We believe that both are valid reasons to explain this phenomenon, but we believe the former seems a more plausible explanation than the latter, as free internet is widely available, and the prices of smartphones have dropped significantly the recent years.

Figure 4-3 Structural breaks for high poverty and low poverty panels

Dept. var.: Tweets	(1)		(2)	
Area type	Low pov.	High pov.	Low pov.	High pov.
Constant	562.2103***	394.173***	551.7819***	385.701***
	(1.909)	(1.951)	(4.104)	(5.645)
Break	315.989***	180.826**	183.280***	129.923**
	(76.386)	(78.077)	(41.045)	(56.455)
R^2	0.1547	0.0714	0.252	0.241
Num. obs.	800	480	800	480
Vce (robust)	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓

* p<0.10, ** p<0.05, *** p<0.010

4.1Determining the optimal level of AR term

When building a time series model often several models can be considered, for instance, an autoregressive (AR) model of a certain order or an AR model of a different order and there is a level of uncertainty about which model is more appropriate. In the context of linear models, a common way to compare models is using the Akaike Information Criterion (AIC) which follows similar principles with the "adjusted R^2", but AIC uses the estimated log-likelihood to compare models. The AIC was proposed by Akaike (1974) and aims to balance the goodness of fit of a model with the number of features by

rewarding a high goodness-of-fit score and penalizing models that become overly complex.

$$AIC = 2k - 2\ln(L), where\ L = L(\hat{\theta})\ (3)$$

Equation (3) shows the AIC function, where k is the number of features and is the penalty parameter, similarly to the adjusted R^2, and $\ln(L)$ is the maximum value of the log-likelihood function of the model. As shown, there is an inverse relationship between the log-likelihood value and AIC, meaning that a lower value of AIC is desired. Models with larger sample sizes and therefore lower unexplained variance will be preferred as they use fewer parameters over models with a higher number of parameters and a lower value of n. We compare different values of lags, and we select the most appropriate number of lags based on the value of AIC. Table 4-1 and Table 4-2 show our results for the low poverty and high poverty groups respectively.

Table 4-1 AIC Test for low poverty panels

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$Constant$	579.250***	366.126***	317.017***	392.212***	332.281***	341.250***	304.018***
	(24.5)	(44.85)	(34.48)	(67.42)	(62.31)	(73.57)	(65.05)
$Tweets_{t-1}$		0.350***	0.322***	0.332***	0.344***	0.364***	0.363***
		(0.06)	(0.05)	(0.05)	(0.04)	(0.03)	(0.03)
$Tweets_{t-2}$			0.101***	0.116**	0.106**	0.097	0.098
			(0.03)	(0.04)	(0.05)	(0.06)	(0.06)
$Tweets_{t-3}$				-0.052	-0.017	-0.012	-0.01
				(0.04)	(0.04)	(0.04)	(0.04)
$Tweets_{t-4}$					-0.08	-0.103	-0.111*
					(0.07)	(0.06)	(0.06)
$Tweets_{t-5}$						0.040**	0.028
						(0.02)	(0.02)
$Tweets_{t-6}$							0.059**

							(0.03)
R^2	0.352	0.438	0.448	0.453	0.468	0.476	0.48
$Adj.\ R^2$	0.319	0.408	0.418	0.423	0.438	0.445	0.449
AIC	9692.418	9345.457	9109.17	8878.869	8619.184	8384.753	8154.485
BIC	9781.426	9433.983	9197.203	8966.395	8706.19	8471.224	8240.405
$Obs.$	800	780	760	740	720	700	680

* p<0.10, ** p<0.05, *** p<0.010

Table 4-2 AIC test for high panels

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$Constant$	579.250***	366.126***	317.017***	392.212***	332.281***	341.250***	304.018***
	(24.5)	(44.85)	(34.48)	(67.42)	(62.31)	(73.57)	(65.05)
$Tweets_{t-1}$		0.350***	0.322***	0.332***	0.344***	0.364***	0.363***
		(0.06)	(0.05)	(0.05)	(0.04)	(0.03)	(0.03)
$Tweets_{t-2}$			0.101***	0.116**	0.106**	0.097	0.098
			(0.03)	(0.04)	(0.05)	(0.06)	(0.06)
$Tweets_{t-3}$				-0.052	-0.017	-0.012	-0.01
				(0.04)	(0.04)	(0.04)	(0.04)
$Tweets_{t-4}$					-0.08	-0.103	-0.111*
					(0.07)	(0.06)	(0.06)
$Tweets_{t-5}$						0.040**	0.028
						(0.02)	(0.02)
$Tweets_{t-6}$							0.059**
							(0.03)
R^2	0.352	0.438	0.448	0.453	0.468	0.476	0.48
$Adj.\ R^2$	0.319	0.408	0.418	0.423	0.438	0.445	0.449
AIC	9692.418	9345.457	9109.17	8878.869	8619.184	8384.753	8154.485
BIC							
$Obs.$							

	9781.426	9433.983	9197.203	8966.395	8706.19	8471.224	8240.405
	800	780	760	740	720	700	680

* p<0.10, ** p<0.05, *** p<0.010

We test from one up to six lags, and we report four metrics for benchmarking all candidate models. We report the R^2, $Adj.R^2$, AIC and BIC. Comparing all models across all four metrics, we find consistently find across all tests for both groups that an autoregressive model of order 6 is the optimal choice of model as adding more lags improves our models. Nonetheless ,an AR(6) model is highly unlikely that is a parsimonious model in the context of our study, and we proceed by comparing our AR(0), AR(2), and AR(6). By comparing the $Adj.R^2$ of the AR(0) and the AR(2) model, the $Adj.R^2$ increases from 0.319 to 0.418 an improvement of 0.109 ,and comparing the AR(2) with the AR(6) model we find that by adding four more lags the adjusted R^2 increased from 0.418 to 0.449 with a slight improvement of 0.031. Additionally, we also observe that adding tweets from $t-1$ period is consistently significant across all models, and tweets from $t-2$ period are also statistically significant for our models from columns (2) to (4). We proceed by selecting an AR(2) model as it's the model that balances out complexity and statistical significance.

4.2 Random Effects vs Fixed effects estimator

In this section, we briefly discuss the use of a fixed effects or a random effects estimator.Consider the model below:

$$y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_{it} + a_i + u_{it} \text{ (4)}$$

The random-effects estimator assumes that :

$$Cov(a_i, X_{it}) = 0$$

$$Cov(a_i, Z_{it}) = 0$$

These two assumptions ensure that $\hat{\beta}_1^{RE} \xrightarrow{p} \beta_1$ and $\hat{\beta}_2^{RE} \xrightarrow{p} \beta_2$. The benefit of this approach is that $SE(\widehat{\beta_{RE}}) < SE(\widehat{\beta_{FE}})$, that is under the assumptions (1) and (2) the random effects estimator is more efficient than then fixed-effects estimator. Another benefit of the random effects estimator, is that it allows us to estimate the effect of time constant variables on the dependent variable. Nevertheless, random effects estimator is no perfect. Probably the most obvious reason and almost certainly not going to be the case that assumptions (1) and (2) hold true, that is, the covariance of a_i with independent variables are equal to zero. That is a strange set of circumstances that must take place to use random effects. If assumptions (1) and (2) are not met then $\hat{\beta}_{RE}$ is inconsistent, whereas the fixed-effects estimator is always consistent independent of whether this covariance is equal to zero or not. The random-effects estimator hinges on these assumptions being true and makes it impossible to estimate the parameter α_i whereas we can by using fixed-effects or least-squares dummy variables (LSDV) estimation.

4.2.1 The Hausman test

$$W = \frac{\left(\widehat{\beta_{FE}} - \widehat{\beta_{Re}}\right)^2}{Var\left(\widehat{\beta_{FE}}\right) - Var\left(\widehat{\beta_{RE}}\right)} \sim x_1^2 \ (5)$$

With $H_0: cov(a_i, x_{it}) = 0$, we can use random effects and $H_1: cov(a_i, x_{it}) \neq 0$ not being true. The intuition behind this statistic is that if the null hypothesis is true then we know that: i) the random effects and fixed effects estimators are consistent; ii) the random effects estimator is more efficient than the fixed effects estimator and therefore if both are estimators are consistent, the difference of these two estimates which is the nominator of the equation should be very small. Secondly, if the null hypothesis is true then we know that $Var\left(\widehat{\beta_{FE}}\right) > Var\left(\widehat{\beta_{RE}}\right)$ and consequently the denominator will be quite large, which means the value of our statistic will be small. Figure 4-4 **Chi-square distribution, k=1**shows the chi-square distribution with 1 degree of freedom. This figure shows that for small values of W we accept the null hypothesis and consequently random effects are more appropriate.

Figure 4-4 Chi-square distribution, k=1

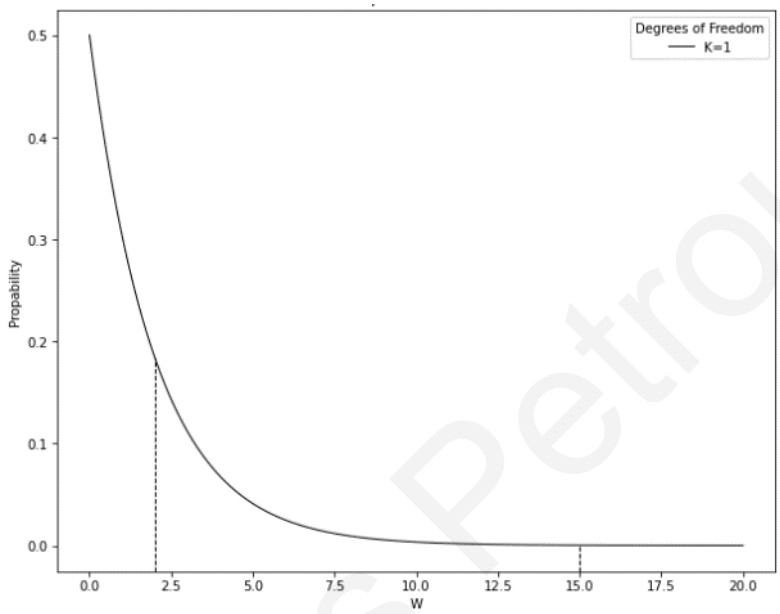


Table 4-3 summarizes the Hausman test performed for each individual panel. The results in Table 4-3 indicate that we should reject the null hypothesis at the 99% confidence level, and we proceed with our analysis using fixed-effects estimation.

Table 4-3 Hausman test for random and fixed effects

	Critical Value	*P-value*	*Result*
High Poverty			
Houseman Test	*81.49*	*0.00*	*Reject Ho*
Low Poverty			
Houseman Test	*144.97*	*0.00*	*Reject Ho*

4.3Unit roots in panel data

4.3.1 What are unit roots

Unit roots are important as they get in, they end up something that gets in our way when attempting to properly model a time series. The reason for that is that when we have a time series with a unit root then it's not stationary and we cannot proceed by using conventional ar models and require transformations to eliminate or at the minimum we should be aware that our time series suffer from this issue and maybe we can try some other methods of analysis. Typically, a visual check can provide great insights into whether a time series is stationary or not and can easily be formalized using a Dickey-Fuller dest.

Below we present an example of a simple AR(1) model.

$$a_t = \varphi a_{t-1} + \varepsilon_t$$

$$= \varphi^t a_0 + \sum_{k=0}^{t-1} \varphi^k \varepsilon_{t-k}$$

$$Var(a_t) = \sigma^2\left[\varphi^0 + \varphi^2 + \varphi^4 + \cdots + \varphi^{2(t-1)}\right]$$

$$\mathbb{E}(a_t) = \sigma^2\mathbb{E}(a_{t-1}) = \varphi^2\mathbb{E}(\alpha_{t-2}) = \cdots = \varphi^t a_0$$

There are three distinct cases of what the value of φ could be.

When $|\varphi| < 1$

$$\lim_{t\to\infty} \varphi^t a_0 = 0$$

And therefore, $\mathbb{E}(a_t) \to 0 \; and \; Var(a_t) \to \frac{\sigma^2}{1-\varphi^2}$ making our time series stationary with a constant mean and variance.

When $|\varphi| > 1$

The time series will be non-stationary as $\lim_{t\to\infty} \varphi^t a_0 = \pm\infty$.

When $|\varphi| = 1$ (the unit root case)

When φ is equal to 1, then the mean of the time series will be stationary and will be equal to its initial value.

$$\mathbb{E}(a_t) = a_0$$

Nonetheles, the variance of the time series becomes

$$Var(a_t) = t\sigma^2$$

Making our time series non-stationary as it violates the constant variance assumption.

4.3.2 Unit root test in a simple AR(1) model

Unit roots cannot be identified with visual inspections, and researchers rely on the Dickey-Fuller test to identify whether a time series has a unit root or not.

$$Y_t = \rho Y_{t-1} + \varepsilon_t$$

With $H_0: \rho = 1 \; and \; H_1: \rho < 1$.

Testing the ρ term cannot be done as under the null hypothesis both $Y_t \; and \; Y_{t-1}$ are non-stationary, and when a time series is non-stationary the central limit theorem can not be applied, therefore we cannot simply estimate ρ using a t-test. An alternative approach that eliminates unit roots is to take the first difference so that our problem transforms to:

$$x_t - x_{t-1} = (\rho - 1)x_{t-1} + \varepsilon_t$$

$$\Delta x_t = \delta x_{t-1} + \varepsilon_t$$

Given that under the null hypothesis ρ = 1, the δ term would be in fact eliminated and x_{t-1} wouldn't be on the right-hand side of the equation, so we have eliminated any non-starionary components, and as a result we are better off from where we started. If ρ<1, then x_{t-1} will exist in the right-hand side. Using a t-statistic seems the appropriate way of esimating ρ on the δ term and then we could compare our t-statistic with the t distribution. However, under the null hypothesis x_{t-1} is itself non-stationary, so the ordinary central limit theorems do not apply when we think about the ordinary least squares estimator for $\hat{\delta}$, so its not the case that under a large sample size that delta has a given t-distribution or normal distribution. Dickey and Fuller (1979) tabulated the asymptotic distribution of least squares estimators for $\hat{\delta}$ under the null hypothesis of it being a unit root. We can actually just compare our ordinary T statistic with the values of this Dickey-Fuller distribution and if $t < DF_{critical}$ then we reject the null hypothesis.

4.3.3 Unit root test in dynamic panel models

In sections 4.3.1 we presented the problems caused by unit roots and in section 4.3.2 we presented the Dickey-Fuller test to identify whether there is a unit root in a simple (1) model. In section 4.3.3 we present the Levin-Lin-Chu (LLC) test which identifies unit-root in panel data models. There are different assorted tests for identifying unit roots such as the Harris-Tzavalis (1999), Breitung (2000), and the Lagrange Multiplier (LM). All tests make different asymptotic assumptions regarding the number of panels in the data and periods for each panel. The Levin-Lin-Chu test requires panel-specific means and no time trends, as well as that the number of periods grows faster than the number of panels so that the ratio of panels to periods tends toward zero. Intuitively, the LLC test fits an augmented Dickey-Fuller regression for each panel in our data. As mentioned earlier, our dataset satisfies the LLC requirement that the number of panels is lower than the number of periods, on the contrary, the Harris-Tzavalis test is suited to datasets with a large number of panels and relatively few periods. We proceed by estimating the LLC test with

$H_0 = Panel\ contains\ unit\ roots$ and $H_1 = Panel\ are\ stationary$

Table 4-4 summarizes the results of the LLC test. With critical values of -9.7606 and -8.8676 for high and low poverty panels respectively and both $p_{val} < 0$, we reject the null hypothesis and conclude that our data are stationary.

Table 4-4 Levein-Lin-Chu test for stationary

Panel Type	Unadjusted t	Adjusted t*	P-value	Outcome
High Poverty	-14.424	-9.7606	0	Reject H0
Low Poverty	-15.0361	-8.8676	0	Reject H0

4.3.4 Dynamic Panels Induced Endogeneity

Nickel (Econometrica,1981) shows that in the context of dynamic panel data models when demeaning the value of Y and X from the respective variable, which is how the fixed effects estimator works, it creates a correlation between the regressor and the error term. Equation (4) shows a simple linear model where Y_i is the dependent variable, X_{it-1} is the independent variable, Y_{it-1} is the lagged value of the dependent variable, a_i are time-invariant characteristics and uit is the error term. The ai term must be eliminated as it creates endogeity in our model. We can eliminate the ai term by either using first differences if the second dimension of the panel is a time series or by using a within transformation (demean), similarly to one-way fixed effects model.

$$y_{it} = \beta_0 + \beta_1 x_{it-1} + \beta_2 y_{it-1} + a_i + u_{it} \ (4)$$

Equation (5) applies the first-difference transformation in our model from equation (4). By applying the first differencing we have successfully dealt with the a_i term. Nonetheless, by using first differencing and having a lagged version of the dependent variable on the right-hand side of the equation, we have introduced the Nickel bias.

$$y_{it} - y_{it-1} = \beta_0 + \beta_1(x_{it-1} - x_{it-2}) + \beta_2(y_{it-1} - y_{it-2}) + (u_{it} - u_{it-1}) \ (5)$$

Endogeneity

Endogeneity issues in economic literature are dealt with instrumental variables and either IV estimation or the 2sls method. Typically finding appropriate instruments is a challenging task by itself and often times researchers might fail to do so. However, in the context of this study and in the case of this simple example AR(1) model we identify that we can use it as an appropriate instrument. satisfies both the relevance and exclusion restriction as of the results of the auto-regressive paths and the sequential exogeneity assumption respectively. Equation (6) shows that can be used as an instrument as it can be used to explain (auto-regressive path) and that it is independent of (sequential exogeneity assumption).

$$y_{it-1} = \beta_0 + \beta_1 x_{it-1} + \beta_2 y_{it-2} + a_i + u_{it-1} \ (6)$$

Therefore, in our real-life model we can use Y_{it-3} as an instrumental variable for Y_{it-1} and Y_{it-4} as an instrument for Y_{it-2}. Now that we have identified our instruments, we proceed by expanding the discussion on which modeling approach deemed to be more appropriate.

4.3.5 Arellano-Bond and IV estimation

In section 4.3.4 we showed what happens once we sweep out time-invariant characteristics using a within transformation from a dynamic panel data model. A trivial solution to eliminate the bias we introduced to our model is to use instrumental variables and proceed with IV estimation. Additionally, we also discussed how we can construct instruments from lagged versions of the dependent variable from the second, third lag, and so on. As long as the error term u_{it} satisfied the independent and identically distributed assumption (i.i.d), the lags of the dependent variable y_{it} will be correlated with the dependent variable and consequently with its difference but uncorrelated with the composite error process. The Anderson-Hsiao (AH) estimator follows a "backing off" strategy where we use past values of the dependent variable as instruments.

Arellano and Bond (1991) show that the Anderson-Hsiao estimator fails to take into consideration potential orthogonality conditions despite being a consistent estimator. What makes the AB estimator different, is that AB assumes that the required instruments

are "internal": that is, based on lagged values of the instrumented variable(s). The Arellano-Bond estimator uses a generalized method of moments (GMM) problem where the model specifies a system of equations for each period in the panel, where a different number of instruments for each period can be applied. The Arellano-Bond estimator and its extension System GMM is suited for situations: i) a large number of panels i and a small number of periods t; ii) linear functional relationship; iii) lagged values of the dependent variable are used; iv) the independent variables are not strictly exogenous; v) there exists unobserved time-invariant heterogeneity, and vi) heteroskedasticity and autocorrelation may exist within each entity but not across them.

Other than the more technical benefits of using the AB estimatore over the AH estimator there are two more practical benefits. The advantage of AB over the AH estiamtor is that using the AB estimator we are not required to lose observations in order to construct our instruments, whereas in the case of AH estimator we need to drop observations equal to the number of instruments time the number of entities in our model. The second benefit of the AB estimator is that it enables us to easier select distant values of lags as instruments and not limit ourselves to closer values of lags.

Before we provide the results of our model, it is worthwhile mentioning some concerns and critics of the Arellano-Bond estimator. Allison et. Al (2017) published a paper with their criticism over the AB estimator and argued that the AB estimator suffers from: i) small sample bias; ii) Inefficient as they do not use all moment conditions imposed by the model, and iii) Lack of certainty about which instruments are more appropriate.

5. Results

Table 5-1 summarizes our results based on the Arellano-Bond and Blundell bond estimators.

Table 5-1 Arellano-Bond and Blundell bond estimators.

	Arellano-Bond		Blundell-Bond	
Panel type	High Poverty	Low Poverty	High Poverty	Low Poverty
	(1)	(2)	(3)	(4)
$Constant$	278.444**	413.358***	51.937	-39.38
	(131.74)	(107.55)	(58.92)	(28.36)
$Tweets_{t-1}$	0.301***	0.281***	0.548***	0.570***
	(0.04)	(0.04)	(0.03)	(0.02)
$Tweets_{t-2}$	0.149**	0.064**	0.401***	0.394***
	(0.07)	(0.03)	(0.04)	(0.03)
d3	-48.95	-51.669*	-63.836	38.972**
	(30.89)	(27.41)	(48.05)	(17.53)
d4	-28.016	7.158	-20.902	115.532**
	(65.63)	(23.75)	(82.94)	(52.22)
d5	-106.465*	-75.506**	-103.884	-1.526
	(63.93)	(31.95)	(71.07)	(38.37)
d6	-76.832*	-48.876**	-56.431	34.794
	(40.75)	(24.78)	(39.10)	(25.00)
d7	-46.015	-62.798**	6.343	41.326
	(44.08)	(27.74)	(60.80)	(28.07)
d8	-72.76	-27.657	-47.284	78.163***
	(49.21)	(20.51)	(53.06)	(26.94)
d9	-64.481	-39.565*	-39.717	82.957**
	(50.71)	(20.59)	(57.46)	(32.74)
d10	-124.604*	-124.225***	-90.982	-29.426
	(67.33)	(35.14)	(71.25)	(20.08)

d11	-64.389*	-30.826	6.974	95.273***
	(35.48)	(29.84)	(33.25)	(29.13)
d12	-87.35	-83.647***	-54.804	50.725
	(72.74)	(27.81)	(85.31)	(32.54)
d13	-91.477	-79.914***	-47.855	30.685
	(56.71)	(22.40)	(59.92)	(26.15)
d14	-46.896	-46.081*	-4.007	80.430**
	(35.29)	(27.54)	(47.18)	(33.56)
d15	-49.026	-27.462	-2.63	111.458***
	(46.30)	(26.21)	(36.93)	(43.15)
d16	-52.395	3.264	-23.693	118.605***
	(59.08)	(24.17)	(71.80)	(35.32)
d17	-81.839**	-96.305***	-63.056	-26.905*
	(37.46)	(34.04)	(43.50)	(15.84)
d18	-72.916	-19.086	-42.05	87.701***
	(48.34)	(25.95)	(76.27)	(32.94)
d19	-54.204	-32.867	-18.222	93.619
	(58.40)	(29.27)	(78.27)	(57.54)
d20	-86.674*	-74.436**	-52.677	14.492
	(51.22)	(33.14)	(59.51)	(24.97)
$Insurrection$	127.169*	276.182***	219.546*	447.400***
	(66.00)	(69.10)	(123.03)	(147.64)
d22	34.257	78.393**	56.843	121.358**
	(46.35)	(32.46)	(74.66)	(52.54)
d23	-74.967	-42.823*	-151.839	-169.002**
	(50.51)	(24.13)	(102.59)	(72.08)
d25	-75.432*	-43.562*	-85.921	4.157
	(40.04)	(26.45)	(64.27)	(21.51)
d26	-66.443	-19.013	-81.613	55.736
	(50.97)	(20.79)	(89.97)	(39.32)
d27	-75.443	-59.668**	-58.066	18.159
	(50.41)	(27.07)	(62.04)	(26.87)
d28	-35.579	-4.378	5.38	98.120**
	(47.25)	(17.98)	(63.68)	(41.76)
d29	-60.598	-76.892***	-30.295	16.346
	(57.15)	(25.31)	(75.64)	(26.16)
d30	-75.317	-38.573	-51.396	69.652
	(47.86)	(29.12)	(51.68)	(42.83)
d31	-89.496*	-135.305***	-69.416	-45.861**
	(46.51)	(37.63)	(55.91)	(20.55)
d32	-66.068	-32.657	-32.237	106.582***
	(42.03)	(25.62)	(50.67)	(31.79)

d33	-127.305*	-135.736***	-96.062	-7.467
	(76.56)	(33.05)	(92.74)	(27.59)
d34	-92.072	-60.849**	-45.461	81.962**
	(62.81)	(25.35)	(57.77)	(34.09)
$Inaugeration$	14.663	68.213**	80.868*	262.045***
	(46.38)	(30.55)	(45.23)	(93.00)
d36	-81.72	-78.423***	-51.87	6.378
	(57.35)	(20.42)	(67.45)	(22.98)
d37	-78.344*	-90.888***	-60.461	-38.371
	(45.73)	(22.78)	(45.99)	(31.01)
d38	-108.649*	-110.286***	-74.808	-0.082
	(60.73)	(27.44)	(59.72)	(19.13)
d39	-49.335	-43.505*	19.299	99.893**
	(45.18)	(23.69)	(49.90)	(38.85)
d40	-73.693	-100.872***	-35.638	39.633
	(60.32)	(29.72)	(65.50)	(41.33)
$Obs.$	444	740	456	760

* p<0.10, ** p<0.05, *** p<0.010

Columns (1) and (2) summarize our result for the Arellano-Bond estimator for high poverty and low poverty areas respectively. Looking at the constant term we find that the volume of tweets sent from low poverty areas is 413 tweets whereas in high poverty areas are 278. This can be explained by two reasons: i) people in high poverty areas potentially have limited access to Twitter or the cost of owning a device with an internet connection is relatively high to their income, and ii) as we observed from our analysis high poverty areas are located in the suburbs where population density is lower than areas in the city centers. High poverty areas account for 67% of the total tweets sent from low poverty areas. Additionally, we observe that the first lag of the dependent variable is statistically significant for both models with a p-value <0.01 and the second lag of the dependent variable is also statistically significant with a pvalue<0.05. The variables d3 to d40 are time dummy variables where they take the value 1 if t=to and 0 otherwise. Our time dummies start from December 19, 2020, until January 25, 2021, which is indicated by our d40 variable. We change the values of d21 and d35 which represents the dates of insurrection and inauguration dates respectively. Our time dummy variable that is

responsible for capturing the effect of the insurrection date on the volume of tweets shows that the volume of tweets changed by 127 in high poverty areas and 276 for low poverty areas and are both statistically significant. We estimate the effect of the inauguration date to be 14 tweets for high poverty areas which are statistically insignificant and 68 for low poverty areas which are significant at a p-value<0.05.

Columns (3) and (4) show our results for the high poverty and low poverty areas respectively for the Blundell-Bond estimators. For the insurrection date, we find that for high poverty areas the contribution is 219 and 447 for low poverty areas and are both statistically significant and for the inauguration date we find that for high poverty areas we estimate an effect of 80 tweets and 262 tweets respectively and are both statistically significant. As we found earlier in previous sections, we found that the volume of tweets sent from low poverty areas is higher than the high poverty areas. To account for their differences, we proceed to normalize our estimations with the mean of our panels for each corresponding group. In section 4 we also found structural breaks that changed the mean of the time series after the events of January 6, 2021. For that reason, we proceed by normalizing our results both with the mean of the entire time series for the time of interest and also the mean of the time series before the events took place.

Table 5-2 shows are post-estimation results for both high and low poverty groups with their mean before and after the events of the US Capitol Hill.

Table 5-2 post-estimation of results

	(1)	(2)	(3)	(4)	(5)
	Estimate	Mean	Ratio	Mean (BI)	Ratio(BI)
Arellano/Bond					
Low poverty	276.182	570	0.485	557	0.496
High poverty	127.169	398	0.320	387	0.329
Blundell/Bond					
Low poverty	447	570	0.784	570	0.803
High poverty	219.546	398	0.552	398	0.567

Column (1) summarizes our estimates of the effect of the US Capitol Hill insurrection on January 06, 2021. Column (2) is the mean of the time series and column (3) is the ratio of the estimate over the mean. Columns (4) is the mean of our panels before the insurrection (BI) took place and column (5) is the ratio of our estimate from column (1) over the mean BI from column (4). From our results above we find that the effect of the Capitol Hill insurrection is greater across both models and after normalizing the mean and the mean before the insurrection for low poverty areas compared to high poverty areas. Our results suggest segregation and isolation may exist even in a context where physical barriers do not take place.

6. Conclusion

The main goal and purpose of this paper is to examine whether big data and more precisely big data generated from social media websites such as Twitter, can be used to determine the role of poverty on the decision of an individual to politically engage/express themselves. Our analysis shows that using fluctuations in the volume of tweets can be used to identify important events. With a sample size of 8.3 million geotagged tweets from seven states with metropolitan areas, we find that areas classified as high poverty areas which on average are areas with 15.4% poverty and above, find that the impact of an unprecedented event on those communities is lower than the impact of that even in areas classified as low poverty areas which on average is a poverty level of 6% and less. After normalizing our estimates, we find that the typical behavior of low poverty areas changed by 48.5% whereas for high poverty areas their typical behavior changed by 32% on the day of insurrection.

It is important to also note that our sample is not representative and cannot be used to make inferences for the population of the United States, however, at a minimum we can make inferences about sample which is Twitter users with the geo-tagged option enabled.

7. Future work

Our analysis was limited to seven states with metropolitan areas and was centered around the events of the US Capitol Hill insurrection for approximately 42 days. The number of states and number of days were selected as a result of limited resources and processing power. This analysis can be extended in the future by including all 50 states for multiple years and we could focus not only on the daily change of volume of tweets, but we could also measure fluctuations in 15-minute intervals and see whether different information diffusion patterns exist in different areas based on income. Additionally, we could use machine learning models to classify which tweets were related to the particular event, not just the volume, and also use sentiment analysis to get greater insights into the impact of the events in the communities.

Twitter is probably the most popular medium when using big data in economics, nonetheless, we could extend our analysis to other platforms where we could compare our results among them or simply add them to our sample. Further, we could also run our model in a single group and use all three types: high, low, and normal poverty with a spatial model to capture spatial dependencies in our panels. Finally, we could attempt to classify the gender and ethnicity of our users so that we could re-weight our sample in a way that gives us greater insights about the general population and not limit ourselves to Twitter users with the geotagged option enabled.

8. Bibliography

[1] Akaike, H., 1974. A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), pp.716-723.

[2] Raftery, A. E. 1995. Bayesian model selection in social research. In Vol. 25 of Sociological Methodology, ed. P. V. Marsden, 111–163. Oxford: Blackwell.

[3] Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike Information Criterion Statistics. Dordrecht, The Netherlands: Reidel.

[4] Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6: 461–464. https://doi.org/10.1214/aos/1176344136.

[5] Hausman, J. and McFadden, D., 1984. Specification tests for the multinomial logit model. *Econometrica: Journal of the econometric society*, pp.1219-1240.

[6] Hausman, J.A., 1978. Specification tests in econometrics. *Econometrica: Journal of the econometric society*, pp.1251-1271.

[7] Baltagi, B. H. 2011. Econometrics. 5th ed. Berlin: Springer.

[8] Gourieroux, C. S., and A. Monfort. 1995. Statistics and Econometric Models, Vol 2: Testing, Confidence Regions, Model Selection, and

Asymptotic Theory. Trans. Q. Vuong. Cambridge: Cambridge University Press.

[9] Levin, A., C.-F. Lin, and C.-S. J. Chu. 2002.Unit root tests in panel data: Asymptotic and finite-sample properties. Journal of Econometrics 108: 1–24.

[10] Harris, R. D. F., and E. Tzavalis. 1999.Inference for unit roots in dynamic panels where the time dimension is fixed. Journal of Econometrics 91: 201–226.

[11] Indaco, A., 2020. From twitter to GDP: Estimating economic activity from social media. *Regional Science and Urban Economics*, *85*, p.103591.

[12] Yildiz, D., Munson, J., Vitali, A., Tinati, R. and Holland, J.A., 2017. Using Twitter data for demographic research. Demographic Research, 37, pp.1477-1514.

[13] Curini, L., Iacus, S. and Canova, L., 2015. Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case. Social Indicators Research, 121(2), pp.525-542.

[14] Wang, Q., Phillips, N.E., Small, M.L. and Sampson, R.J., 2018. Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences*, *115*(30), pp.7735-7740.

[15] *https://github.com/JustAnotherArchivist/snscrape*

[16] *Dickey, D.A. and Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American statistical association, 74(366a), pp.427-431.*