# From Twitter to GDP: Estimating Economic Activity From Social Media

Indaco, Agustín

The Graduate Center, CUNY

19 March 2019

# From Twitter to GDP: Estimating Economic Activity From Social Media

Agustín Indaco
The Graduate Center, CUNY

March 19, 2019

**Abstract**

Using all geo-located image tweets shared on Twitter in 2012-2013, I find that the volume of tweets is a valid proxy for estimating current GDP in USD at the country level. Residuals from my preferred model are negatively correlated to a data quality index, indicating that my estimates of GDP are more accurate for countries with more reliable GDP data. Comparing Twitter with more commonly-used proxy of night-light data, I find that variation in Twitter activity explains slightly more of the cross-country variance in GDP. I also exploit the continuous time and geographic granularity of social media posts to create monthly and weekly estimates of GDP for the US, as well as sub-national estimates, including those economic areas that span national borders. My findings suggest that Twitter can be used to measure economic activity in a more timely and more spatially disaggregate way than conventional data and that governments' statistical agencies could incorporate social media data to complement and further reduce measurement error in their official GDP estimates.

*JEL Classifications*: JEL codes C53, C55, E01, O11

*Keywords*: National Accounts, Big Data

*The Graduate Center at the City University of New York, Ph.D. Program in Economics, 365 Fifth Avenue, New York, NY 10016, (email: aindaco@gradcenter.cuny.edu). I am grateful to my advisors David Jaeger, Francesc Ortega, and Lev Manovich for their advice, guidance, and support. I also thank Tahir Butt and Agustín Mario for immensely helpful conversations, and Damon Crockett for putting together the maps that are shown in Figures 1-5.

# 1 Introduction

Despite incessant debate about its ability to accurately measure the state of the economy, the gross domestic product (GDP) is still the most widely used indicator to gauge countries' economic performance (Masood (2014)). One of the many difficulties with estimating GDP is that its measurement is often complicated and expensive. This could lead to measurement error, particularly in developing countries, which in turn mislead policy makers and businesses. Another concern is that given the importance surrounding official GDP estimates both in terms of market fluctuations as well as public perception of politicians' performances, governments can find short-term benefits in manipulating these estimates. Additional shortfalls of GDP estimates are the delay with which they are released and their geographic rigidity. In light of this, much research has been focused on alternative ways of measuring GDP other than the traditional sample survey method, both to corroborate as well as a control mechanism.

In this paper I argue for the use of data from social media posts as a proxy for measuring GDP. By locating and analyzing the volume of hundreds of million social media posts, I show that one can accurately estimate GDP at the country level and that it compares well with satellite night-light images that are widely used as a proxy for economic activity. For this exercise, I collect all geo-located image tweets shared on Twitter for the years 2012 and 2013. In total the dataset includes roughly 140 million tweets, with their corresponding text, time of posting, precise latitude and longitude and unique user identifier. I then aggregate the number of tweets sent from each country in each year to estimate annual GDP at the country level.

This paper has five main findings. First, I show that the volume of tweets can be used as a proxy for estimating current GDP at the country level; my preferred model can explain 87 percent of the cross-country variation in GDP. Second, I compare the strength of social media as a proxy for estimating GDP relative to the more commonly used satellite night-lights data. I find that Twitter data explains slightly more of the variance in cross-countries' GDP estimates and that the two proxies could be used together for more accurate predictions. Third, I find that social media data can be used to estimate annual variation in GDP at the country level. Despite not being very precise, the relationship between annual changes in social media posts and changes in economic activity are informative. Fourth, I present and study a hypothesis of the underlying relationship between image tweets and economic activity in which image tweets are a byproduct of consumption. It seems that social media is used as a medium to share conspicuous consumption among its users. Finally, I exploit the continuous time and geographic granularity aspect of social media posts to compose a monthly and weekly estimate of GDP for the US. Along the same line, I study the local economic effect of a large oil reserve finding in an underdeveloped region in Argentina. Despite widespread enthusiasm regarding the local economic effect the discovery would have, no official estimates had been reported: for the first time, I estimate a 38% annual increase in economic activity in the region.

Social media can contribute greatly to economic research. Measures taken from social media, as well as from other alternative data sources, can serve both as substitutes as well as complements to traditional survey data. In particular, social media data has several properties that are beneficial when estimating economic measures. First, social media data is available in real time, which allows to nowcast economic activity and thus provide companies and individuals updated information when making economic decisions[1]. Furthermore, the continuous time flow aspect of social media posts

[1]A few papers have looked at proxies to nowcast economic activity at the local level. Askitas and Zimmermann (2013) estimate the German business cycle at a monthly level by measuring toll activity on important highways by heavy transport vehicles. Whereas Glaeser et al. (2017) use data from Yelp to estimate business establishment entry and exit across zip codes in the US

enables weekly or monthly estimates of economic data which through official agencies are only available every quarter or year. Second, given that geo-tagged social media posts can be geographically assigned to a precise location within approximately a 10 meter radius, one can aggregate social media posts at any sub-national geographical level one deems interesting. This includes aggregating data between areas that are not bound together inside political borders and thus fabricate meaningful areas of study that are not possible with official datasets. Third, unlike survey data that are costly to recollect, social media data is organically being generated by users from all over the world and available to data agencies, international organizations and non-governmental organizations at a relatively low cost. Finally, even though estimates in this paper are based solely on the volume of tweets from each location for a given period of time, more information can be potentially extracted from the content of each post. Topic modeling, sentiment analysis, use of language, spelling, share of posts coming from locals and tourists and other information can be extracted from posts to improve estimates.

There are several papers which explore the use of social media data to measure economic outcomes. Antenucci et al. (2014) analyze tweets to build a model that accurately predicts unemployment insurance claims in the US. In a similar manner, Llorente et al. (2015) use Twitter data to estimate regional unemployment rates in Spain. Glaeser et al. (2017) use Yelp data to predict changes in the number of overall establishments and restaurants across zip codes in the US. As far as I am aware, this is the first paper to explore the use of social media data to estimate economic activity across the globe.

## 2 Estimating GDP (and its problems)

GDP is the most widely used measure of a country's economic performance. Given its importance, national income measurements are governed by the

United Nations System of National Accounts (SNA), which sets a global standard that allows for international comparisons of economic activity across countries. However, adherence to these standards is entirely voluntary, and cannot be rigidly enforced. While some countries adhere to the most recent standards set by the revised SNA 2008, many developing countries have still not adopted the previous 1993 SNA standards and are still using 1968 SNA methodologies[2].

As an illustration of the degree of measurement error, Johnson et al. (2013) study the revision in the Penn World Tables (PWT), a popular dataset frequently used for making comparisons of GDP across countries. The authors find that the standard deviation of the change in countries average growth over the period $1970-1999$ was 1.1 percent per year when comparing data in version 6.1 to version 6.2[3]. Given that the average growth rate is 1.56 percent, this is a relatively large discrepancy. In fact, Dawson et al. (2001) claim that some results in the economic literature based on PWT are purely a product of measurement error in the data[4].

While these examples show the complications involved when measuring GDP, these problems tend to be accentuated in developing countries where statistical offices tend to have fewer resources for constructing this immensely intricate task. Jerven (2013) explains that in African countries the informal sector is so large that it cannot be left out of GDP estimates, and thus a variety of innovative accounting practices are implemented to take them into consideration. This leads to a series of pragmatic decisions to be made within statistical offices, which are subject to the availability of trustworthy data, financial resources and political instructions. In turn,

[2]According to Jerven (2013), as of 2010, in Africa only Cameroon and Lesotho were upgrading their system to incorporate SNA 2008 standards, while three African countries still used the 1968 SNA methodologies.

[3]Version 6.1 of the PWT was released in 2002, while version 6.2 was released in 2006.

[4]In particular, they find that the empirical link between output and volatility and the cross-country test in the Permanent Income Hypothesis are driven by measurement error in the data.

Jerven (2013) believes that GDP statistics from African countries are "best guesses of aggregate production".

A second concern is that given the importance surrounding official GDP estimates, both in terms of market fluctuations as well as public perception of politicians' performances, governments can find short-term benefits in manipulating these estimates. In less developed countries, higher growth estimates of real GDP per-capita which have later been revised downward have been associated with a higher probability of reelection (Brender and Drazen (2005)). Even though as we have seen GDP estimates are subject to measurement error, Kerner and Crabtree (2018) find that there are non-random variations in official macroeconomic estimates. Furthermore, given that GDP estimates are produced by countries' statistics agencies via surveys that are not publicly available, it is difficult for non-governmental organizations, international organizations and the public at large to corroborate these estimates.

Official GDP estimates also have limitations in terms of lags and geographical aggregation. At best, advanced GDP estimates are available 30-45 days after the end of the reference quarter. After this release, revised estimates which incorporate information from additional surveys are disclosed up to three years later. Finally, GDP estimates are geographically bound to political and geographical borders. Even in the case of many developed countries which produce estimates for metropolitan areas, these estimates are geographically rigid and cannot be combined across country borders. For many developing countries, there are no sub-national estimates of economic activity.

These concerns and limitations surrounding official GDP estimates have motivated efforts in finding proxies that may be able to estimate economic activity. The bulk of this literature has centered around efforts to use satellite night-light images to estimate GDP at the country level (Donaldson and Storeygard (2016)). I contrast and compare Twitter and night-lights as proxies for GDP in section 5. This paper adds to this literature by exploring whether social media can be used as a proxy for GDP both at the country level as well as for sub-national regions. On the one hand, these proxies could be used by countries' statistic agencies as supplements that improve the accuracy of their estimates. On the other hand, these proxies could also be used as a tool for non-governmental agencies and international organizations to corroborate official GDP estimates.

## 3 Twitter and Twitter Data

Twitter is a social media application which allows users to post short messages of any subject of their choosing. These messages are known as *tweets*. Twitter started in 2006 and by 2012 had 140 million global users and 340 million tweets per day. Initially, tweets were limited to 140 characters, because Twitter was originally designed as an SMS mobile phone-based platform. In its early days, 140 characters were the limit that mobile carriers imposed with the SMS protocol standard so Twitter was simply creatively constrained. As Twitter eventually grew into a web platform, the 140-character limit remained as a matter of branding.[5] Unless restricted by the user, tweets are publicly available and can be read via the application or on a web browser.

Twitter initially did not allow users to share images, videos or other sorts of media in their tweets. This changed in August 2011, when Twitter rolled out a platform that allowed users to add images to their tweets. Until September 2013, image tweets did not present the image, but instead included a link where your image could be viewed. In September 2013, images could be previewed directly on the tweet.

The dataset used in this paper contains all geo-tagged image tweets posted on Twitter between January 1, 2012 and December 31, 2013.[6] This

[5]In 2017 however, Twitter expanded this limit to 280 characters per tweet.

[6]According to Weidemann and Swift (2013) roughly 20 percent of tweets are geo-

dataset was provided directly by Twitter, through a Twitter Data Grant submission in 2014.[7] The total dataset contains 140 million tweets from all around the world. Each tweet contains information on: i) a unique identifier for each individual Twitter user; ii) the latitude and longitude (with 5 decimal points, which gives a precision of 1.1 meters) of where the tweet was sent from[8]; iii) the date and time in which the tweet was sent; iv) the image tweeted; and v) any accompanying text[9].

The underlying mechanism which relates the volume of tweets and economic activity is not clear, but there may be several factors at play. On one hand, part of the cross-country differences seems to be explained by smart-phone penetration, which is in essence an indicator of a economic development. Even though one can post tweets from a desktop or other technological devices, it is safe to assume that the large majority of tweets are sent from smart-phones.

For within country differences, particularly among developed countries where smart-phone penetration is saturated, the underlying mechanism is different. Even though there are several reasons why social media users post on social networks, a substantial amount of posts seem to display the consumption of luxury goods and services (i.e.: food from a renowned restaurant, the latest trend in shoes, exclusive sporting events). In a way, social

located. Lee (2015) finds that 42 percent of tweets contain an image, but this study was based on 1 million tweets sent solely from US West Coast users, which likely biases the results. In December 2018, I queried the Twitter API on multiple occasions, collecting a dataset of 10,000 random Tweets. Among this dataset, I find that 4.9 percent of tweets are geo-located and 22.8 percent have images.

[7]The grant was awarded to the Cultural Analytics Lab, directed by Lev Manovich.

[8]The latitude and longitude from these posts are generated automatically via very reliable hardware and software and thus I am reasonably certain of their accuracy. Nonetheless, throughout this paper, I hardly need such level of precision, except in the exercise carried out in subsection 8.2.

[9]Given the concern of bots (i.e.: an autonomous program that can interact with computer systems or users) biasing the dataset, I do identify bots as users which at some point sent more than five tweets in a span of one minute. Removing these users from the data does not significantly change the estimates presented in section 4 and are thus left in our sample.

media has become the ideal medium to exhibit what Veblen (1899) coined as conspicuous consumption: when the utility consumers attain stems more from revealing their wealth and income to others, rather than from the direct utility derived from the good itself. To the conspicuous consumer, the public display of wealth or income is a means to attaining or sustaining a given social status. In order to preserve this status, they must constantly exhibit these consumption patterns online. I examine these hypothesis in section 7.

Table 1 summarizes this Twitter data by year and by country income groups (using the World Bank's classification). We see that the average number of tweets per country was roughly 100,000 in 2012 and increased to 500,000 in 2013. This is an indication of the growth of image tweets over this period. The breakdown of average tweets per income group shows that countries in higher income groups have more tweets, although the growth rates from 2012 to 2013 are larger among lower income countries.

Table 2 provides information on the distribution of tweeted images across the 10 countries with most tweets over this time period. The US tops the list of tweets in 2012 and 2013, with 31.7 percent of the share of total tweets over this period. The US is also the largest country in the world in terms of GDP, with roughly 22.4 percent of the share of global GDP. While some countries, like Japan, have a smaller share of tweets than of global GDP, others, like the UK, have a larger share of tweets than GDP.

## 3.1 Visual examples of what Twitter data reflect

Figure 1 shows that there are some clear visual patterns to the location and distribution of tweets worldwide that seem to represent economic activity and population density. The location from where each image tweet was sent is represented by a small light blue point. There are clusters of these light blue points both in areas that are more densely populated as well as areas with higher levels of per capita income. For example, in the United States, the largest concentration of image tweets seem to be centered along the coastal areas, but not so in the less-densely-populated South West and Rocky Mountain States. South America has a cluster of tweets mainly surrounding big cities in Ecuador, Colombia and Venezuela in the north and Brazil, Argentina, Uruguay and Chile further south. In Africa, image tweets tend to be concentrated in richer countries: Morocco, Algeria and Egypt, and in Sub-Saharan Africa in South Africa, Nigeria and Kenya. Western Europe seems to be mostly lit up and the concentration of tweets becomes sparser as we move east into Ukraine, Belarus, Latvia, Estonia and ultimately into Russia. The case of Australia is also telling: tweets are concentrated around large cities off the west coast like Melbourne, Sydney and Canberra, as well as Perth on the east coast.

Figure 2 shows a more detailed view of East Asia that depicts the clear cutoff in the number of tweets being sent from South Korea and North Korea, respectively. The cluster of points in the north-west part of South Korea corresponds to tweets sent from Seoul. Hardly any tweets are sent from the other side of the border, which corresponds to North Korea. Internet access is strictly limited in North Korea, and primarily used by the government and foreigners[10].

A closer look at image tweets sent from Europe in Figure 3 lets us see in closer detail how tweets are in fact clustered around capitals and big cities, where population densities and economic activity tend to be higher. We can also depict the roads and highways that connect these large cities. Interestingly, several research papers (Banerjee et al. (2012) and Ghani et al. (2016)) have shown that economic development spreads along network routes; this same pattern seems to be reflected in the location of tweets as well.

Similar patterns emerge in Figure 4 for the US West Coast. First of

[10]As of 2016, the use of Twitter and several other social media applications has been banned in North Korea.

all we notice that large cities like Los Angeles and San Francisco are easily visible by the large and extended cluster of tweets that surrounds them. But we can also identify major highways connecting these cities easily depicted.

Moving to the Middle East and North-East Africa, Figure 5 shows that the number of tweets is smaller and solely concentrated in a few big cities. Interestingly, in Egypt we can see a cluster of tweets flowing down the Nile River: this pattern resembles the concentration of population and economic development surrounding the river.

# 4 Empirical Estimates and Methods

## 4.1 Using Twitter to Estimate GDP

The main goal of this paper is to study whether Twitter data is a valid proxy for estimating current GDP in USD at the country level. First, I aggregate the number of tweets by country and year. I use the precise location (latitude-longitude) to geocode the country of origin where each post was sent from. I then aggregate the volume of tweets by country per year.

In order to assess the validity of tweets as a proxy for GDP at the country level, I estimate:

$$\ln GDP_{i,t} = \beta_0 + \alpha_t + X'_{i,t}\gamma + \beta_1 \ln Tweets_{i,t} + \varepsilon_{i,t}, \tag{1}$$

where the explained variable is the natural log of GDP of country $i$ in year $t$. The vector $X_{i,t}$ is composed of country characteristics including population, the share of the population with access to the internet and continent to which it belongs. The coefficient we are most interested in is $\beta_1$ which shows the relevance of the number of image tweets taken from that country in each of those years for estimating GDP. In Equation 1, year fixed effects ($\alpha_t$) control for any differences in the use of Twitter from one year to the other as well as changes in global economic conditions.

The corresponding estimates are reported in Table 3. There are 184 countries in the dataset with data on GDP, Twitter and population for both years[11]. In column 1, I regress the natural log of GDP solely on the number of image tweets sent from each country. This is the baseline regression. The coefficient of interest on $ln(Tweets)$ is highly significant and the $R^2 = 0.78$. When the population of the country is included in column 2, the coefficient on $ln(Tweets)$ is reduced, but remains statistically significant at the 1 percent level. The $R^2$ increases to 0.87 and the partial $R^2$ of $ln(Tweets)$ is 0.66. This is an important result given that Figure 1 seems to reflect that, unsurprisingly, the clusters of tweets are concentrated in areas with large populations. Thus the results in column 2 indicate that even when controlling by population, the volume of tweets still holds information for estimating economic activity. Column 3 includes categorical dummies for the continents in which each country is situated. This captures economic development as well as cultural differences in image sharing on social media platforms that exist between regions. Neither the coefficient of interest or the goodness of fit greatly change. Column 4 adds the share of the population with access to the internet. The number of observations are reduced to 180 countries per year because the World Bank does not have data on the share of the population with access to internet for six countries[12]. The coefficient of interest on $ln(Tweets)$ is reduced, but remains statistically significant at the 1 percent level. The $R^2 = 0.94$ and the partial $R^2$ of $ln(Tweets)$ is 0.29. The $R^2$ without including tweets would be 0.78, similar to the $R^2$ in Column 1 with only $ln(Tweets)$ as independent variable.

Table 3 shows that the number of image tweets sent in a year is a good measure for estimating GDP at the country level, being able to explain 78 percent of the cross-country variation in GDP on its own. In all specifications, the coefficient on the number of image tweets is statistically significant.

[11] I also remove countries in which Twitter was banned for a period of time during any of these years: China and Iran

[12] These are: Libya, Kosovo, Curacao, Palau, South Sudan and San Marino.

Figure 6 is a visual representation of these estimates for our baseline model in Column 1: the estimates lay pretty closely around the 45 degree line. There are a few exceptions that stand out; most notably Cuba where internet is restricted. Figure 7 plots the residuals of Equation 1 against the fitted values, allowing us to study the distribution of the residuals; which seems to be randomly distributed around zero (i.e.: no clear pattern emerges).

As a robustness check, I run Equation 1 independently for each income group, as per the World Bank's classification[13]. I run the most complete model which includes the number of tweets, the population and the percentage of population with access to internet (i.e.: the model that corresponds to Column 4 in Table 3). The corresponding estimates are reported in Table 4 and show that running the model separately for each income group has little effect on the goodness of fit and relevant coefficients of the model. The measures of $R^2$ in each of the different specifications are similar and vary between 0.87 for Low-income countries and 0.92 for Lower-middle income countries. The coefficients on $ln(Tweets)$ vary somewhat for each income group, but are statistically significant at the 1 percent level for all groups. More importantly, the partial $R^2$ of $ln(Tweets)$ is between 0.46 and 0.52.

Given that these groups are based on income levels, Table 4 shows that the number of tweets is able to accurately estimate relatively small differences in GDP levels across countries. Furthermore, Table 4 shows that tweets can be used to measure economic activity in developing countries where estimating GDP is particularly complicated.

[13]As of 1 July 2013, low-income economies are defined as those with a GNI per capita, calculated using the World Bank Atlas method, of $1,045 or less in 2012; middle-income economies are those with a GNI per capita of more than $1,045 but less than $12,746; high-income economies are those with a GNI per capita of 2,746 or more. Lower-middle-income and upper-middle-income economies are separated at a GNI per capita of $4,125.

### 4.2 Data Quality Issues

Section 2 showed that GDP estimates have been criticized for being inaccurate, particularly in developing countries. If this is true, it could be the case that as I am trying to estimate the GDP reported by countries and not necessarily the *true* GDP, it is possible that my estimates are off due to measurement error on official GDP figures. In this scenario, data from tweets could be used by statistical agencies as a complementary measure to produce more accurate estimates.

To examine this, I incorporate a measure of data quality developed by the World Bank. The World Banks Statistical Capacity Indicator is a composite score assessing the capacity of a country's statistical system. It is based on a diagnostic framework assessing the following areas: methodology, data sources, and periodicity and timeliness. The overall score is a simple average of all three area scores on a scale of 0-100, where higher values indicate better data quality assessment.

Given that the World Bank works solely with Low-income, Lower-middle income and Upper-middle income countries, the data available for such measures are restricted to these countries. There are 140 countries for which there is an indicator on the quality of the data, as well as GDP, Twitter, population and percent of population with access to the internet. Hence, I run the same regression in Equation 1 for the subset of countries for which this data is available for 2012 and 2013. As can be seen in Columns 1-4 of Table 5, the overall estimates are very similar for this subset of countries as for our complete data presented in Table 3, both in terms of the coefficient on $ln(Tweets)$ as well as the $R^2$.

I then collect the residuals of Equation 1 using the subset data and run the following regression:

$$Residuals_{i,t}^2 = \beta_0 + \beta_1 DataQuality_{i,t} + \varepsilon_{i,t}, \tag{2}$$

where I regress the squared residuals for country $i$ in year $t$ on the data quality index and GDP. The coefficient of interest is $\beta_1$: a negative and statistically significant coefficient would indicate that the residuals in my baseline model in Equation 1 are larger for countries with low data quality, and vice versa.

Column 5 of Table 5 shows that the data quality indicator coefficient is in fact negative and statistically significant at the 1 percent confidence level.

This shows that GDP estimates under the baseline model are more accurate for countries with high quality data, and vice versa. The negative coefficient on the data quality index in Equation 2 suggests that there is information to be captured from Twitter data that could help close the gap between estimated GDP and the *true* GDP. Social media data could thus be used as a complement to survey data to increase the accuracy of GDP estimates.

## 5 Twitter vs Lights

The use of visible light emanating from earth as captured by weather satellite images has been widely suggested as a good proxy for measuring economic activity. Different studies have shown that night lights can be used to measure GDP estimates at the country level (Pinkovskiy and Sala-i-Martin (2016)), GDP growth at the country level (Henderson et al. (2012)) and GDP for sub-national regions (Doll et al. (2006), Ghosh et al. (2010), Henderson et al. (2012) and Sutton et al. (2007)). According to these studies, the intensity of artificial night-lights highly correlates with GDP and thus can be used to estimate economic activity for different geographic regions. This paper clearly draws a lot from this literature.

Satellite night-light images come from the US Air Force. Several of their weather satellites circle the earth 14 times per day, recording the intensity of earth-based lights. Each satellite observes every location on the planet (between 65 degrees S latitude and 65 degrees N latitude) every night at some time between 20:30 hs and 22:00 hs. The intensity of lights is measured for every 30-second output pixel and is averaged across all valid evenings in a year. The raw data are heavily processed to correct for several atmospheric and visual distortions. These tasks include the identification of clouds, removal of glare, identification of intense natural light during summer months, etc. The objective is to leave only man-made light visible. They then average all valid images over the year and report the intensity of light for approximately every 0.86 square kilometer. Intensity of night lights reflects outdoor and some indoor use of lights.

Among the main advantages of using night-lights to estimate economic activity is that satellite data are available more readily than official GDP estimates, that they can be more reliable for some developing countries and that they can be used to estimate economic activity at sub-national regions for which official statistics are not available.

Social media data have these advantages as well. In fact, social media data have a higher-frequency and more granular geographic scope. While luminosity data from satellite images must be averaged over time in order to have a more accurate measure free from cloud and glare, the stream of social media posts is continuous. Furthermore, while intensity of nigh-lights is captured in 0.86 square km pixels, each social media post has its corresponding latitude and longitude coordinate, precise to the nearest 1.1 meters; this allows for unconstrained and more flexible geographical aggregations.

However there are several shortfalls with night-light data. Chen and Nordhaus (2010) show that satellite night-light data suffers from substantial measurement error, both in the time-series and the cross-section aspect of the data. The authors regress the logarithm of luminosity for the same year across images from different satellites and find standard errors in the range of 0.20. They also find slightly smaller standard errors when regressing year-to-year variations in luminosity from images taken by the same satellite. This means that the measurement error for individual grid cells is in the order of

20 logarithmic percent. They also find that images from specific satellites are consistently dimmer than other satellites. This leads the authors to question the reliability of the night-lights data and conclude that it "is at best a noisy indicator" (Chen and Nordhaus (2010)).

Doll (2008) finds that satellite recording of nightime light density has a tendency to overestimate the true extent of lit area on the ground. This occurs because these images tend to attribute light generated at a particular site to nearby sites as well. This effect is referred to as overglow. In part, this is explained by the procedure in which satellites capture night-lights, but it is also due to the large overlap that exists between pixels.

Furthermore, luminosity at the pixel or grid level is bottom coded at 0 and top coded at 63. As a consequence, in certain geographically small and rich countries, more than 3 percent of the pixels are topped off at 63 (i.e.: Netherlands and Belgium). Thus inhibiting the estimator to parse out differences at the top end of the distribution. When using social media data there is no top coding since we can have an infinite number of posts coming from the same location. This could prove useful when trying to detect small differences at the top end of the income distribution.

At the moment, the main advantage for the use of night-light images over social media data, is that satellite images are available since 1992, allowing for research on evolution of GDP and medium-term growth paths. Henderson et al. (2012) use this to estimate economic growth of countries and regions over the last 30 years. Currently, this sort of analysis could not be done using social media data. Since social media became widely popular in the mid-2000s, it is not suitable to perform historical research. Furthermore, the availability of long time period of satellite night-light images has led researchers to show that night-lights have been a valid proxy for measuring economic activity for several decades. This is not the case with social media data, which may have very different uses in some years time and invalidate its usefulness as a proxy for GDP.

Given that the Twitter data used in this paper correspond to 2012 and 2013, I collect Operational Linescan System data on night lights intensity from the United States Air Force Defense Meteorological Satellite Program for those same years. These two datasets will be used to compare the potential use of each of them as proxies for estimating GDP at the country level. As a first approach, I map the distribution and intensity of both night-lights and Twitter data[14]. Figure 8 presents Twitter posts on the top and satellite night-light images on the bottom. The two maps show a striking resemblance, as the bulk of tweets and night-light intensity clusters around large cities and capitals around the world. Figure 9 focuses on Europe and again shows the similarity between the location and volume of tweets relative to the intensity of night-lights emanating from earth.

In order to compare and contrast the validity of night-lights and tweets as a proxy for GDP at the country level, I estimate:

$$\ln GDP_{i,t} = \beta_0 + \alpha_t + X'_{i,t}\gamma + \beta_1 \ln Lights_{i,t} + \beta_2 \ln Tweets_{i,t} + \varepsilon_{i,t}, \quad (3)$$

where the explained variable is the natural log of GDP of country $i$ in year $t$. As in Equation 1, the vector $X_{i,t}$ is composed of country characteristics including population, the share of the population with access to the internet and continent to which it belongs. The two coefficients of interest are $\beta_1$ and $\beta_2$, which respectively show the relevance of the intensity in night-lights and the number of image tweets for estimating GDP. Year fixed effects ($\alpha_t$) control for any differences in the use of Twitter from one year to the other and changes in global economic conditions.

Column 1 of Table 6 includes only night-light intensity and the model explains 85 percent of the variation in GDP at the country level. Column 2 includes all variables in $X'_{i,t}$ which comprises population, continent dummies and percent of population with access to the internet, and night-light data

[14]This exercise was in fact the way I became truly excited about the idea that social media data could be used as a proxy for economic activity.

explains roughly 88 percent of the variation in GDP. Including both Twitter data as well as night-lights in Column 3 of Table 6, the coefficient on both of these variables are statistically significant, reflecting that they both provide distinct information for estimating GDP and suggesting that ideally both measures could be used together to make a more accurate proxy of GDP. An important result for the validity of social media data as a proxy is that when both proxies are included in the model, the partial $R^2$ of lights is 0.22 whereas for tweets it is 0.32.

Another way to compare the use of night-lights and Twitter data as a proxy for estimating economic activity is to analyze the estimated GDP using each of the two proxies. Figure 10 plots both of these estimates for year 2013. We see that the bulk of the estimates lie closely around the 45-degree line which shows that the two proxies not only estimate a significant portion of the variation in GDP, but that the individual estimates for each country are similar using each of these two proxies. This validates the use of Twitter lights as a proxy[15].

## 6 Estimating Changes in GDP

The previous sections have established the validity of social media data as a proxy for estimating the level of economic activity. An obvious extension is to see whether social media data can be used to estimate within country changes in economic activity from one period to the next.

This exercise has already been carried out for satellite night-lights with diverse success. Henderson et al. (2012) show that night-light luminosity does a reasonable job at predicting annual fluctuations in GDP growth for countries over a 25 year period. Although they find that luminosity has pre-

[15] Section 6 extends this analysis and shows that both Twitter data as well as night-lights can also be used to estimate within-country variation in GDP from one year to the other. By including country fixed effects to Equation 3 the model estimates the within country changes in GDP to 2012 and 2013 and Table 7 shows that the coefficient on both tweets and night-lights is positive and statistically significant

dictive power on estimating short-term annual changes in economic activity, the proxy is more accurate at estimating long-term growth between 1992 and 2007.

Due to the weakness in predicting annual changes, the authors study whether light intensity is prone to a ratchet issue: given that light growth seems to be reflecting installation of new capacity, light intensity is nondecreasing and thus does not capture decreases in economic activity. Although the authors do not find such a ratchet effect, the volume of social media posts seem to be more sensitive to short-term fluctuations in economic activity (see section 7) and thus could serve as a more accurate proxy for estimating changes in economic activity. In this sense, I argue that social media data is better equipped as a proxy for estimating short-run variations in economic activity, particularly in recessionary periods.

In order to study whether tweets have predictive power in estimating annual fluctuations in GDP and how it compares to night-lights, I estimate:

$$\ln GDP_{i,t} = \beta_0 + \alpha_i + \alpha_t + X'_{i,t}\gamma + \beta_1 \ln Lights_{i,t} + \beta_2 \ln Tweets_{i,t} + \varepsilon_{i,t}, \quad (4)$$

which is equivalent to Equation 3 with country fixed effects ($\alpha_i$). In essence this measures the annual variation in GDP for country $i$ between years 2012 and 2013. The coefficients of interest are $\beta_1$, which indicates the relationship between changes in GDP and changes in light intensity and, more importantly, $\beta_2$, which shows the relationship between changes in GDP and changes in the volume of tweets. The corresponding estimates are reported in Table 7.

Column 1 shows results when Equation 4 includes only tweets as explanatory variable; the coefficient is positive and statistically significant, with an $R^2$ of 0.25. Column 2 includes several country characteristics and the coefficient on Tweets remains positive and statistically significant, at the 5 percent level. Columns 3 and 4 mimic these specifications for night-lights: regardless of the specification, the coefficient on night-lights is positive and statistically significant (at the 5 percent and 10 percent level, respectively), with an $R^2$ between 0.28 and 0.31. Finally columns 5 and 6 include both tweets and night-lights together. The coefficient on both proxies are positive and statistically significant in both specifications (at the 5 percent level for night-lights when full set of country characteristics is included). This indicates that these two measures may not be capturing exactly the same aspects of economic activity and the explanatory power improves with the inclusion of both.

Albeit not being very precise, the relationship between annual changes in social media posts and annual changes is GDP are informative. Furthermore, it seems that tweets and night-light luminosity could be used together to produce a more accurate estimate of annual variations in economic activity.

## 7 Conspicuous Consumption, Leisure and Tweets

The previous sections have shown diverse ways in which the volume of image tweets is related to economic activity. Nonetheless, it is still not clear what the underlying mechanism that relates social media posts and economic activity[16].

One possible explanation is related to smart-phone penetration, which is essentially an indicator of economic development. Under this hypothesis, more developed countries have more smart-phones per individual, which in essence increases the ratio of the population with social media accounts (which are free of charge to the user), and thus increases the number of social media posts per capita. Notice that under this hypothesis, the number of posts per user does not vary across individuals, but the aggregate number of posts per country increases as a result of more individuals with smart-

[16]Given that this paper is studying the validity of social media as a proxy for economic activity and not trying to establish any causal effect between the two, a thorough understanding of their relationship is not necessary. Nevertheless, a more clear understanding of the underlying mechanism that drives the two is helpful for understanding why the proxy works and more importantly, when it might stop being accurate.

phones and hence social media accounts. This explanation works pretty well to explain cross-country variations, especially between developed and developing countries.

Nevertheless, this hypothesis does not explain why the volume of social media posts explains within-country differences (as shown in subsection 8.2). Notably in developed countries, where smart-phone penetration has been saturated for some time, there must be another underlying mechanism that drives more tweets in one region than another. Although there are likely several factors at play, one possible hypothesis is that social media applications serve as a medium to showcase consumption of goods and services among the network of users.

This behavior falls in line with what Veblen (1899) coined as conspicuous consumption: when the utility consumers attain stems more from revealing their wealth and income to others, rather than from the direct utility derived from the good itself. Under this hypothesis, individuals display wealth to their network as a means to attaining a given social status which is what they ultimately desire. In order to preserve this status, they most constantly exhibit the consumption patterns online. In this scenario, social media represents the ideal medium through which individuals may showcase a stream of consumption of goods and services to friends and acquaintances[17].

According to Becker (1965), consumption and leisure are complementary goods; thus we would expect both of them to occur at the same time[18]. To test whether leisure tweets are better predictors of economic activity than

[17] Han et al. (2019) discuss extensively the implications this representation of consumption has at an aggregate level. They present a theoretical model in which consumption is more salient than nonconsumption in the social transmission process which affects consumption behavior along the network. This visibility bias causes people to perceive that others are consuming heavily and thus increases aggregate consumption.

[18] If this were the case, we would expect a larger share of images on social media posted during non-working hours to display the consumption of goods and services relative to images posted during working hours. In Appendix B I use a computer vision API to show that in fact images posted during leisure hours portray more consumption related topics than images posted during working hours.

tweets sent during work hours, I divide all image tweets sent from the US in 2012 in two groups: (i) those sent during working hours (Monday to Friday between 9:00-17:00 hs); and (ii) those sent during leisure time, defined as non-working weekday hours (Monday to Friday after 17:00 hs and before 9:00 hs) and weekends (Saturday and Sunday any time).

Under this hypothesis, I expect the coefficient on leisure tweets to be positive and statistically significant, while the coefficient on tweets during working hours to be either negative or not statistically significant. In order to directly study the hypothesis that tweets are a byproduct of consumption, I will see if the volume of tweets are a valid proxy for personal consumption expenditure in the US. In order to test this, I aggregate the number of tweets in each state which were sent during working and leisure times and run the following regression:

$$\begin{aligned} \ln Consumption_i = \beta_0 + \beta_1 \ln Population_i + \beta_2 \ln LeisureTweets_i + \\ \beta_3 \ln WorkTweets_i + \varepsilon_i \end{aligned} \tag{5}$$

where I estimate personal consumption expenditure for state $i$ in year 2012. The coefficients of interest are $\beta_2$ and $\beta_3$ which show the relevance of the number of image tweets taken in each of those time slots for estimating consumption at the state level.

The corresponding estimates are reported in Table 8. Columns 1 and 2 introduce the volume of all tweets, irrespective of when they were taken. The coefficient is positive and statistically significant (at the 10 percent level) even when controlling for population in column 2. Once again, population is included as a control to test whether the volume of image tweets is capturing something more than just population. Given that the coefficient remains positive and statistically significant indicates that social media posts captures something else besides population, and the explanatory power is greatly improved by the inclusion of both. Column 3 includes only leisure hour tweets, and the coefficient is positive and statistically significant even when controlling for population. Columns 5 and 6 introduce both leisure and work tweets: while leisure tweets are positive and statistically significant, work tweets are negative (although not statistically significant).

These results seem to support the hypothesis that Twitter serves as a medium which enables users to share their consumption habits throughout their network. This hypothesis could explain one of the underlying mechanisms that relates the volume of tweets and economic activity, with image tweets being a byproduct of consumption. Given that consumption represents roughly 2/3 of overall GDP in the US since 2000, this explains why image tweets are a good proxy for estimating economic activity at large.

# 8 Exploiting Time and Space Continuity of Social Media Data

## 8.1 Estimating GDP for Short Time Intervals

One of the main advantages of using social media data in estimating GDP is that given the continuous aspect of social media posts, one can gauge the economic activity of a specific geographical region in any specific time interval one considers relevant. This is not the case with official GDP estimates, which are estimated quarterly for developed countries and only yearly for most developing countries. The infrequent nature of official GDP estimates deems them of little value for businesses and individuals when evaluating financial decisions and investments. Thus, being able to produce weekly or monthly GDP estimates using social media data could prove extremely useful and is a major advantage of social media data.

In order to exemplify this possibility, I estimate weekly and monthly GDP using social media data for the US for the year 2012. Clearly, this exercise could be replicated for any country and for any time period. For this, I use a subset of the original dataset which consists of all image tweets posted in the US during the 2012. This dataset includes roughly 8 million tweets. Given that the popularity of Twitter has been growing over time as shown in Table 1, there is a clear upward trend in the number of tweets over time. The non-stationary nature of tweets over this time period force me to detrend the data.

In order to estimate GDP at the monthly level, I use the coefficients from the preferred model depicted in Equation 1 divided by twelve:

$$\ln GDP_i = (\beta_0/12) + (\beta_1/12)\ln Tweets_i + (\beta_2/12)Population_i + \varepsilon_i, \quad (6)$$

where I estimate GDP for country $i$ in year 2012. I then collect $(\beta_0/12)$, $(\beta_1/12)$ and $(\beta_2/12)$ and use the detrended sub-sample of US tweets to estimate monthly measures of GDP for the US in 2012. In doing this, I assume that the relationship between tweets and economic activity is constant throughout every month of the year. This assumption is debatable: it is easy to imagine particular events (i.e.: TV shows, sporting events, natural disasters, etc.) that boost the use of Twitter without necessarily increasing economic activity. Nonetheless, given the unpredictable nature of many of these events, this assumption is not skewing my results in a predictable manner[19].

Given the lack of monthly GDP estimates for the US, there are a series of leading indicators that economists and investors use to gauge economic activity in finer time intervals. I gather two of the most respected monthly indicators: the Leading Index for the US (USSLIND) put together by the Federal Reserve Bank of Philadelphia and the Composite indicator for the US put together by the Organization for Economic Co-operation and Development (OECD).

The USSLIND uses housing permits, state initial unemployment insurance claims, delivery times from the Institute for Supply Management manufacturing survey, and the interest rate spread between the 10-year Treasury

[19] Eventually, with several years of Twitter data, a clear periodical or seasonal cycle in the volume of tweets could be studied and incorporated into the model.

bond and the 3-month Treasury bill to construct an index that captures economic activity in the country. The OECD's Composite indicator are constructed to predict cycles in a reference series chosen as a proxy for economic activity. As defined by the OECD, the variation in economic output relative to its long term potential represent the fluctuations in economic activity.

Figure 11 plots these two monthly leading indicators together with the constructed monthly GDP estimate using Twitter data. The USSLIND estimate for 2012 shows strong economic performance in the first months of the year, with a slowdown in economic activity beginning in April. There is a strong rebound in August that lasts until October; November shows a strong decrease in economic activity but there is a slight recovery in December. The OECD indicator on the other hand, shows what seems to resemble an elephant: economic activity gains momentum in the first months of the year, but then begins to slowdown, reaching its minimum in August, to then begin a smooth upward cycle till the end of the year. My estimates using Twitter data combine certain aspects of these two indicators: until August the economic cycle resembles the USSLIND pretty closely, but after that there seems to be a pretty smooth increase in economic activity that is more in line with the OECD leading indicator.

An important observation to point out is that although the two leading indicators do share similar overall trends, there are visible differences between the two. Leading indexes are proxies and by no means exact or accurate. They depict the overall trend of the state of the economy in short-time intervals. The fact that the Twitter based estimates follow the overall economic cycle of these indicators is more important than if it is able to closely trace either of the two indicators shown here.

Given that social media posts are uploaded on a continuous basis, one can estimate economic activity at a higher frequency. I exploit this advantage and in a similar fashion to the way I construct monthly estimates, I estimate economic activity at the weekly level for the US in 2012. The only difference is that for this I divide the coefficients in Equation 1 by 52 (roughly the number of weeks in a year):

$$\ln GDP_i = (\beta_0/52) + (\beta_1/52)\ln Tweets_i + (\beta_2/52)Population_i + \varepsilon_i, \quad (7)$$

where I estimate GDP for country $i$ in year 2012. In a a similar manner as when estimating GDP at the monthly level, I now use $(\beta_0/52)$, $(\beta_1/52)$ and $(\beta_2/52)$ in Equation 7 on the detrended subset of US tweets to estimate weekly measures of GDP for the US in 2012.

Figure 12 plots these weekly estimates of GDP using this methodology. The weekly estimates trace a similar economic cycle to the monthly estimate, but with more short-term variations and jolts.

To the best of my knowledge, there are no publicly available and respected leading economic indicators at the weekly level for the US. Thus I do not offer here alternative indicators to contrast and compare my estimates with. The high-frequency of social media posts clearly offers a unique tool to produce a weekly indicator of economic activity.

## 8.2 Estimating GDP for Sub-National Regions

Another advantage of social media data in estimating GDP lies in the geographic detail of social media posts. Exploiting the accuracy of the location from where social media posts are sent allows one to aggregate posts at any sub-national geographical level one deems interesting. This includes aggregating data between areas that are not officially bound together and thus fabricate meaningful areas of study that are not possible with official datasets.

Given that National accounts data tends to be aggregated at the country-level it is not suitable for this type of analysis. Even in the cases when they do offer regional level estimates, it is not possible to use estimates from different countries in the same analysis because each survey is conducted independently with different questions and methodologies and thus cannot be taken together.

Several papers attempt to do this sort of analysis using night-light data (Pinkovskiy (2017) and Michalopoulos and Papaioannou (2018)). But satellites recording nightime light density have a tendency to overestimate the true extent of lit area on the ground, given that they tend to attribute light generated at a particular site to nearby sites as well. This effect is referred to as overglow (Doll (2008)). In part this is due to technical issues in the procedure satellites use to capture night-lights, but also due to the large overlap that exists between pixels. Social media data may be more suitable for this exercise since geo-tagged posts are given with latitude and longitude coordinates with 5 decimal points, which gives a location within approximately a 10 meter radius. Thus social media posts may suffer from measurement error, but do not suffer from an overglow phenomenon that night-light satellite images do have.

In order to corroborate the validity of Twitter data as a proxy for GDP at the sub-national level, I first estimate GDP for individual US states and compare that to official state level estimates provided by the BEA. For this exercise, I collect the subset of all tweets from 2012 taken in the US and use the latitude and longitude from each post to assign them to the state from where they originate. Using the total number of tweets sent from each state, I use the coefficients from Equation 1 and collect GDP estimates for 2012 for each US state. Figure 13 plots both the BEA as well as the estimates from the Twitter data. The bulk of the estimates lie closely around the 45-degree line which shows that Twitter data is a valid proxy for estimating GDP at the state level. Furthermore, the correlation coefficient between these two estimates is 0.985.

The geographic granularity of social media data allows us to go beyond estimating economic activity at the state level: with this in mind, I plot a density map of economic activity using 2012 Twitter data for the US. Figure 14 shows a 2D density map of economic activity in the US for 2012. Density maps divide the plotted area in a multitude of small fragments and represent the count of a particular variable (in this case economic activity based on use of Twitter) in each fragment. The contours depict areas with larger economic activity with areas shaded in red showing the highest concentration of economic activity. According to these estimates, economic activity in the US is concentrated in large cities along the East Coast. There is also a concentration in California, in the area surrounding San Francisco. According to this Figure, the area around the city of New York seems to bear the largest concentration of economic activity.

Another exercise that can be done using Twitter data is to estimate economic activity in areas around national borders. It is not possible to estimate economic activity in these areas across national borders using official GDP estimates because even though some countries do publish regional GDP estimates at the sub-national level, each survey is conducted independently with different questions and methodologies and thus cannot be taken together. Social media data, in this case Twitter, can help us around this issue: we can aggregate all the Twitter posts from a specific geographic area and take them together to use as a proxy for economic activity.

The border between Mexico and the US represents an interesting example to carry out this exercise as they are important trade partners and a there is a great deal of economic synergy between the two countries. In 2012, US exports to Mexico totaled $215,875 million, which represents 14 percent of total exports in the US and 49 percent of total imports for Mexico. On the other hand, Mexico exports to the US totaled $323,026 million, which represents 74 percent of total exports for Mexico and 14 percent of total imports in the US.

Figure 15 shows a density map of economic activity in the US-Mexican border based on tweets. There are clear economic regions that go beyond the international border, particularly in the region close to the Gulf Coast.

This shows that the area around Houston, San Antonio and Austin in Texas concentrate the largest level of economic activity which extends south of the border and into important cities in Mexico, like Monterrey and San Luis Potosí. The same phenomenon is visible in the Pacific Ocean side of the border, with a cluster of economic activity integrating US cities close to the border like Los Angeles and San Diego, with border cities in Mexico like Tijuana. These exercises are useful not only for businesses trying to make decisions in the region, but also for scholars who are interested in studying cross-border spillover effects in border-cities[20].

## 8.3 Local Economic Shock: Vaca Muerta, Argentina

The previous exercises have showcased the advantages of using Twitter data to estimate economic activity in terms of the continuous time and the geographic granularity of social media posts. In some scenarios, both characteristics can be combined to produce a timely and geographically detailed estimate of economic activity. This is extremely useful for studying local economic shocks, such as factory openings (Greenstone et al. (2010)) and closings, discovery of natural resources or natural disasters (Kocornik-Mina et al. (2015)). For example, nowcasting the impact of such events could be extremely useful for the government in estimating how much recovery aid they need to send to a certain affected area. Official data is not well equipped for such tasks because of delays in estimates and because many times they are not geographically granular enough to perceive local economic shocks.

One recent such case of a discovery of natural resources that has transformed a local economy occurred in the providence of Neuquén, Argentina. The Vaca Muerta Formation is a geologic formation which has had oil production since 1918. But in July 2011, a large oil discovery was made, with resources estimated at 16 billion barrels of oil and 308 trillion cubic feet of gas. If exploited, the proven reserves of the country would increase more

[20]For example, see Hanson (2001) and Coronado et al. (2015).

than eight times.

Beyond the impact this caused in the energy market of the country at large, the local region was particularly disrupted by the migration, investment, changes in infrastructure and overall economic shock that such a discovery had on the small local economy. The closest town to Vaca Muerta is called Añelo, which in 2010 had 2,500 inhabitants. Since then, informal estimates have reported the population to more than double and rent prices to skyrocket. There are newspaper articles claiming that prices of sq/ft in Añelo are comparable to the high end neighborhoods in Buenos Aires. But none of this has been corroborated or certified by official statistics. Argentina does produce provincial GDP estimates, but the local economic effects Vaca Muerta has on a small town like Añelo do get dampened by changes in economic activity in larger cities when looking at provincial estimates. This is where we can exploit the advantage that social media offer granular geographic data, which allows us to aggregate tweets at any region of interest, even particularly small ones.

In order to estimate the growth of economic activity in the region surrounding Vaca Muerta, I identify all social media posts being sent from that region in 2012 and 2013. Figure 16 shows the distribution of tweets in Argentina for each of the two years[21]. This figure highlights how the distribution of tweets in Argentina from one year to the other solely differ in the region surrounding Vaca Muerta. In order to get an estimate of the change in economic activity in this region between 2012 and 2013, I use the elasticity between tweets and GDP from Table 7. Based on the difference in volume of tweets originating from the region, I find that economic activity grew by 38.1 percent from one year to the other. As far as I am aware, this is the first time the local economic effect of Vaca Muerta on the region has been quantified.

[21] Given that the number of tweets grew substantially between 2012 and 2013 (see Table 1), I restrict the 2013 tweets to a random sample of tweets equivalent to the number of tweets sent out in 2012.

## 9 Conclusion

The main goal of this paper is to understand whether social media data from Twitter could be used as a proxy for estimating GDP at the country level. Using 140 million geo-located tweets from across the world, I find that the volume of tweets is an accurate predictor of economic activity. I then find that the residuals from my preferred model are negatively correlated to a data quality index provided by the World Bank which suggests that my estimates are more accurate for countries which are considered to have more reliable GDP data. These two findings taken together suggest that social media data could be used as a complement to survey data to increase the accuracy of GDP estimates or as a tool for international organizations to corroborate official estimates.

A supplementary goal of this paper is to demonstrate some of the advantages of estimating economic activity using social media data. Given that social media posts are available in real time and with precise geographical information about the location of the user, one can calculate the economic activity for a particular geographical region in any specific time interval. To showcase this, I first produce weekly and monthly GDP estimates for the US which show seasonal variations in economic activity which seem to be in line with other leading indicators. I then exploit the geographic granularity of social media posts to create GDP estimates at the sub-national level for the US and also create density maps for economic activity in the US-Mexican border. But the exercise that combines all the benefits and clearly provides an example of how local authorities could use this proxy to fill in gaps in their official estimates, is the example of estimating the local economic effects of Vaca Muerta in the region. Given how constrained the area of impact is, traditional economic indicators at the provincial level are not able to suited to perceive the economic effects.

Neither of these exercises could have been carried out using official national accounts data: at best, they are available quarterly and at state or province level. But even then, the lagged nature of official GDP estimates deems them of little value for businesses and individuals when evaluating financial decisions and investments.

An important concern with the use of proxies is that many times they lack an underlying theoretical model that demonstrates the relationship between the proxy and the variable of interest. This issue was made apparent with the Google Flu Trends project, which analyzed changes in flu-related search queries to accurately map the incidence of flu. Although the model was highly accurate the first couple of years, it became very imprecise as peoples' behaviour changed in the wake of panic fuelled by media reports. This led to people that did not have flu-like symptoms to search for flu-related queries which caused the model to wildly over-predict flu outbreaks. In this paper, I closely study the underlying relationship between tweets and economic activity. Preliminary results seem to suggest that image tweets are a byproduct of consumption, with social media platforms being the ideal medium to showcase consumption among the network of users. This would explain why the volume of image tweets accurately predicts GDP. Nonetheless, more research needs to be done in addressing this relationship before such a model is used and incorporated to estimate economic activity. In this sense, this paper is just a first step in a an ongoing research to unleash the full potential of social media data in economic research.

For many developing countries, GDP estimates are produced annually and only at the country level. Being able to produce sub-national estimates at precise time-intervals of economic activity could be tremendously helpful for country officials and international organizations to evaluate certain policies. For example, measuring the economic impact of *maquiladoras* in special economic zones, or studying the effects of unanticipated discoveries of natural resources in a particular region. The data needed to evaluate these crucial questions might already be available.

## References

**Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher R, and Matthew D. Shapiro**, "Using Social Media to Measure Labor Market Flows," NBER Working Papers 20010, National Bureau of Economic Research, Inc March 2014.

**Askitas, Nikolaos and Klaus F. Zimmermann**, "Nowcasting Business Cycles Using Toll Data," *Journal of Forecasting*, July 2013, *32* (4), 299–306.

**Banerjee, Abhijit, Esther Duflo, and Nancy Qian**, "On the Road: Access to Transportation Infrastructure and Economic Growth in China," Working Paper 17897, National Bureau of Economic Research March 2012.

**Becker, Gary S.**, "A Theory of the Allocation of Time," *The Economic Journal*, 1965, *75* (299), 493–517.

**Brender, Adi and Allan Drazen**, "Political budget cycles in new versus established democracies," *Journal of Monetary Economics*, 2005, *52* (7), 1271 – 1295.

**Chen, Xi and William D. Nordhaus**, "The Value of Luminosity Data as a Proxy for Economic Statistics," Cowles Foundation Discussion Papers 1766, Cowles Foundation for Research in Economics, Yale University August 2010.

**Coronado, Roberto, Marycruz De Len, and Eduardo Saucedo**, "So Close to Mexico: Economic Spillovers along the Texas-Mexico Border," in "Ten-Gallon Economy," Palgrave Macmillan, New York, 2015, pp. 183–198.

**Dawson, John W., Joseph P. Dejuan, John J. Seater, and E. Frank Stephenson**, "Economic information versus quality variation in cross-country data," *Canadian Journal of Economics/Revue canadienne d'conomique*, 2001, *34* (4), 988–1009.

**Doll, Christopher**, "CIESIN Thematic Guide to Night-time Light Remote Sensing and its Applications," 01 2008.

**Doll, Christopher N. H., Jan-Peter Muller, and Jeremy G. Morley**, "Mapping regional economic activity from night-time light satellite imagery," *Ecological Economics*, April 2006, *57* (1), 75–92.

**Donaldson, Dave and Adam Storeygard**, "The View from Above: Applications of Satellite Data in Economics," *Journal of Economic Perspectives*, November 2016, *30* (4), 171–98.

**Ghani, Ejaz, Arti Grover Goswami, and William R. Kerr**, "Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing," *The Economic Journal*, 2016, *126* (591), 317–357.

**Ghosh, Tilottama, Rebecca L. Powell, Christopher D. Elvidge, Kimberly E. Baugh, Paul C. Sutton, and Sharolyn Anderson**, "Shedding Light on the Global Distribution of Economic Activity," *The Open Geography Journal*, December 2010, *3* (1).

**Glaeser, Edward L, Hyunjin Kim, and Michael Luca**, "Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity," Working Paper 24010, National Bureau of Economic Research November 2017.

**Greenstone, Michael, Richard Hornbeck, and Enrico Moretti**, "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings," *Journal of Political Economy*, 2010, *118* (3), 536–598.

**Han, Bing, David Hirshleifer, and Johan Walden**, "Visibility Bias in the Transmission of Consumption Beliefs and Undersaving," Working Paper 25566, National Bureau of Economic Research February 2019.

**Hanson, Gordon H.**, "U.S.Mexico Integration and Regional Economies: Evidence from Border-City Pairs," *Journal of Urban Economics*, 2001, *50* (2), 259–287.

**Henderson, J. Vernon, Adam Storeygard, and David N. Weil**, "Measuring Economic Growth from Outer Space," *American Economic Review*, April 2012, *102* (2), 994–1028.

**Jerven, Morten**, *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*, Cornell University Press, 2013.

**Johnson, Simon, William Larson, Chris Papageorgiou, and Arvind Subramanian**, "Is newer better? Penn World Table Revisions and their impact on growth estimates," *Journal of Monetary Economics*, 2013, *60* (2), 255–274.

**Kerner, Andrew and Charles Crabtree**, "The IMF and the Political Economy of Data Production," *SocArXiv*, January 2018.

**Kocornik-Mina, Adriana, Thomas K.J. McDermott, Guy Michaels, and Ferdinand Rauch**, "Flooded cities," GRI Working Papers 221, Grantham Research Institute on Climate Change and the Environment December 2015.

**Lee, Kevan**, "What analyzing 1 million tweets taught us," Retrieved from https://thenextweb.com/socialmedia/2015/11/03/what-analyzing-1-million-tweets-taught-us/, November 2015. Accessed on Oct. 2, 2018.

**Llorente, Alejandro, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro**, "Social Media Fingerprints of Unemployment," *PLOS ONE*, 2015, *10* (5), 1–13.

**Masood, Ehsan**, *The Great Invention: The Story of GDP and the Making (and Unmaking) of the Modern World*, Saqi Books, September 2014.

**Michalopoulos, Stelios and Elias Papaioannou**, "National Institutions and Subnational Development in Africa," *The Quarterly Journal of Economics*, 2018, *129* (1), 151–213.

**Pinkovskiy, Maxim and Xavier Sala-i-Martin**, "Lights, Camera Income! Illuminating the National Accounts-Household Surveys Debate," *The Quarterly Journal of Economics*, 2016, *131* (2), 579–631.

**Pinkovskiy, Maxim L.**, "Growth discontinuities at borders," *Journal of Economic Growth*, 2017, *22* (2), 145–192.

**Sutton, Paul C., Christopher D. Elvidge, and Tilottama Ghosh**, "Estimation of Gross Domestic Product at Sub-National Scales using Nighttime Satellite Imagery," *International Journal of Ecological Economics and Statistics*, May 2007, *8* (S07), 5–21.

**Veblen, Thorstein**, *The Theory of the Leisure Class*, McMaster University Archive for the History of Economic Thought, 1899.

**Weidemann, C and Jennifer Swift**, "Social media location intelligence: The next privacy battle - An ArcGIS add-in and analysis of geospatial data collected from Twitter.com," *International Journal of Geoinformatics*, 2013, *9* (2), 21–27.

Table 1: Twitter Data Summary Statistics: Mean and S.D.

| | 2012 | 2013 |
|---|---|---|
| Tweets | 109,678.1 | 528,694.9 |
| | (354,724.1) | (2,397,003.8) |
| *By Income Group* | | |
| High (62) | 211,993.9 | 1,126,937.3 |
| | (541,334.7) | (4,048,721.5) |
| Upper-middle (51) | 89,123 | 406,004.9 |
| | (201,367.9) | (836,887.4) |
| Lower-middle (43) | 37,394.5 | 209,938.9 |
| | (106,230.9) | (929,686.9) |
| Low (28) | 735.9 | 3,602.5 |
| | (898.9) | (4,370.9) |

**Notes:** Top row shows the mean number of tweets per country and standard deviation in brackets. The bottom half of the table shows the mean number of tweets and the standard deviation sent by each country divided by income groups, where the number of countries per income group is shown in brackets.

Table 2: Top 10 Countries in Tweets 2012-2013

| | 2012 Tweets | 2013 Tweets | 2012-2013 Share total Tweets (%) | 2012-2013 Share world GDP (%) |
|---|---|---|---|---|
| Top 10 | | | | |
| U.S.A. | 8,260,789 | 29,376,135 | 29.3 | 22.4 |
| U.K. | 3,093,930 | 8,775,427 | 8.7 | 3.6 |
| Indonesia | 5,44,329 | 6,401,877 | 6.4 | 1.3 |
| Spain | 2,121,122 | 6,210,364 | 6.2 | 1.9 |
| Japan | 1,825,886 | 5,128,888 | 5.5 | 8.3 |
| Saudi Arabia | 1,121,424 | 5,198,632 | 3.1 | 0.9 |
| Brazil | 707,881 | 3,064,005 | 3.1 | 3.4 |
| Turkey | 630,344 | 3,112,159 | 3.0 | 1.2 |
| Mexico | 956,809 | 3,028,057 | 2.6 | 1.6 |
| Russia | 110,5047 | 2,560,921 | 2.5 | 2.4 |

**Notes:** Top 10 countries in terms of most tweets shared in total between January 2012 and December 2013. Columns 1-2 shows the total number of tweets sent out from each country for each year, Column 4 shows each country's share of total tweets over this time period and Column 5 shows each country's share of world GDP over 2012-2013.

Table 3: Estimating Country GDP

| Dep. var.: ln(GDP) | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| ln(Tweets) | 0.67*** | 0.49*** | 0.46*** | 0.17*** |
| | (0.02) | (0.02) | (0.02) | (0.02) |
| ln(Population) | | ✓ | ✓ | ✓ |
| Continent | | | ✓ | ✓ |
| Internet | | | | ✓ |
| $R^2$ | 0.78 | 0.87 | 0.89 | 0.94 |
| Adj. $R^2$ | 0.78 | 0.87 | 0.88 | 0.94 |
| $R^2$ w/o Tweets | 0.00 | 0.60 | 0.75 | 0.78 |
| Partial $R^2$ Tweets | 0.78 | 0.66 | 0.56 | 0.29 |
| Fixed-effects | Year | Year | Year | Year |
| Num. obs. | 368 | 368 | 368 | 356 |
| RMSE | 1.08 | 0.85 | 0.78 | 0.56 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

**Notes:** The dependent variable in all columns 1-4 is the log of GDP. This is using data from 2012 and 2013. All specifications have year fixed effects. Column (4) has fewer observations because the World Bank data does not have data on the percent of population that uses the internet for four countries.

Table 4: Estimating Country GDP: By Income Group

| Dep. var.: ln(GDP) | Low | Low-middle | Upper-middle | High |
|---|---|---|---|---|
| ln(Tweets) | 0.44*** | 0.42*** | 0.37*** | 0.47*** |
| | (0.09) | (0.04) | (0.05) | (0.04) |
| ln(Population) | ✓ | ✓ | ✓ | ✓ |
| Continent | ✓ | ✓ | ✓ | ✓ |
| $R^2$ | 0.87 | 0.92 | 0.89 | 0.90 |
| Adj. $R^2$ | 0.84 | 0.91 | 0.88 | 0.89 |
| $R^2$ w/o Tweets | 0.53 | 0.80 | 0.85 | 0.82 |
| Partial $R^2$ Tweets | 0.46 | 0.48 | 0.53 | 0.52 |
| Fixed-effects | Year | Year | Year | Year |
| Num. obs. | 56 | 78 | 104 | 132 |
| RMSE | 0.95 | 0.72 | 0.78 | 0.79 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

**Notes:** The dependent variable in all columns is the log of GDP. These estimates correspond to estimating Column (4) of Table 3 independently for each income group. The first column is for the subset of Low-income countries, second column for Low-middle income countries, third column for Upper-middle income countries, and the fourth column for High-income countries. These categories are based on The World Bank classifications. This is using data from 2012 and 2013. All specifications have year fixed effects.

Table 5: Data Quality Issues

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dep. var.: | ln(GDP) | ln(GDP) | ln(GDP) | Residual$^2$ |
| ln(Tweets) | $0.60^{***}$ | $0.37^{***}$ | $0.39^{***}$ | |
| | (0.02) | (0.02) | (0.03) | |
| Data Quality | | | | $-0.01^{***}$ |
| | | | | $(< 0.01)$ |
| ln(Population) | | ✓ | ✓ | |
| Continent | | | ✓ | |
| Internet | | | ✓ | |
| R$^2$ | 0.76 | 0.90 | 0.91 | 0.04 |
| Adj. R$^2$ | 0.76 | 0.90 | 0.91 | 0.03 |
| Num. obs. | 240 | 240 | 240 | 240 |
| RMSE | 1.01 | 0.65 | 0.62 | 1.30 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

**Notes:** The dependent variable in columns 1-4 is the log of GDP; and in column 5 the residual squared. Using the World Bank classification, columns 1-4 estimate the log of GDP for the subset of Low-income, Low-middle and Upper-middle income countries combined. In column (5) I collect the squared residuals of the estimation in column 4 and regress them on the data quality index and the log of tweets.

Table 6: Lights vs Twitter for Estimating GDP

| Dep. var.: ln(GDP) | (1) | (2) | (3) |
|---|---|---|---|
| ln(Lights) | 0.88*** | 0.31*** | 0.26*** |
| | (0.02) | (0.03) | (0.03) |
| ln(Tweets) | | | 0.08*** |
| | | | (0.03) |
| ln(Population) | | ✓ | ✓ |
| Continent | | ✓ | ✓ |
| Internet | | ✓ | ✓ |
| $R^2$ | 0.85 | 0.88 | 0.95 |
| Adj. $R^2$ | 0.85 | 0.88 | 0.95 |
| Partial $R^2$ Lights | 0.85 | 0.33 | 0.22 |
| Partial $R^2$ Tweets | – | – | 0.32 |
| Num. obs. | 350 | 342 | 342 |
| RMSE | 0.84 | 0.63 | 0.53 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

**Notes:** The dependent variable in all columns 1-3 is the log of GDP. This is using data from 2012 and 2013. All specifications have year fixed effects. Columns 2-3 have fewer observations because the World Bank data does not have data on the percent of population that uses the internet for some countries.

Table 7: Country Fixed Effects

| Dep. var.: ln(GDP) | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| ln(Tweets) | 0.02*** | 0.01** | | | 0.02*** | 0.02*** |
| | (< 0.01) | (< 0.01) | | | (< 0.01) | (< 0.01) |
| ln(Lights) | | | 0.06** | 0.04* | 0.09*** | 0.07** |
| | | | (0.03) | (0.03) | (0.02) | (0.07) |
| ln(Population) | | ✓ | | ✓ | | ✓ |
| Continent | | ✓ | | ✓ | | ✓ |
| Internet | | ✓ | | ✓ | | ✓ |
| $R^2$ | 0.25 | 0.28 | 0.28 | 0.31 | 0.35 | 0.37 |
| Adj. $R^2$ | 0.20 | 0.22 | 0.21 | 0.23 | 0.29 | 0.31 |
| Fixed-effects | country | country | country | country | country | country |
| Num. obs. | 364 | 304 | 356 | 298 | 352 | 298 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

**Notes:** The dependent variable in all columns is the log of GDP. These estimates come from including country fixed-effects to Equation 3. This is using data from 2012 and 2013. Columns 2 and 4 have fewer observations becuase satellite night-light data is available for fewer countries than tweets in our dataset. Columns 4 and 6 have even fewer observations because the World Bank does not have internet access data for some countries.

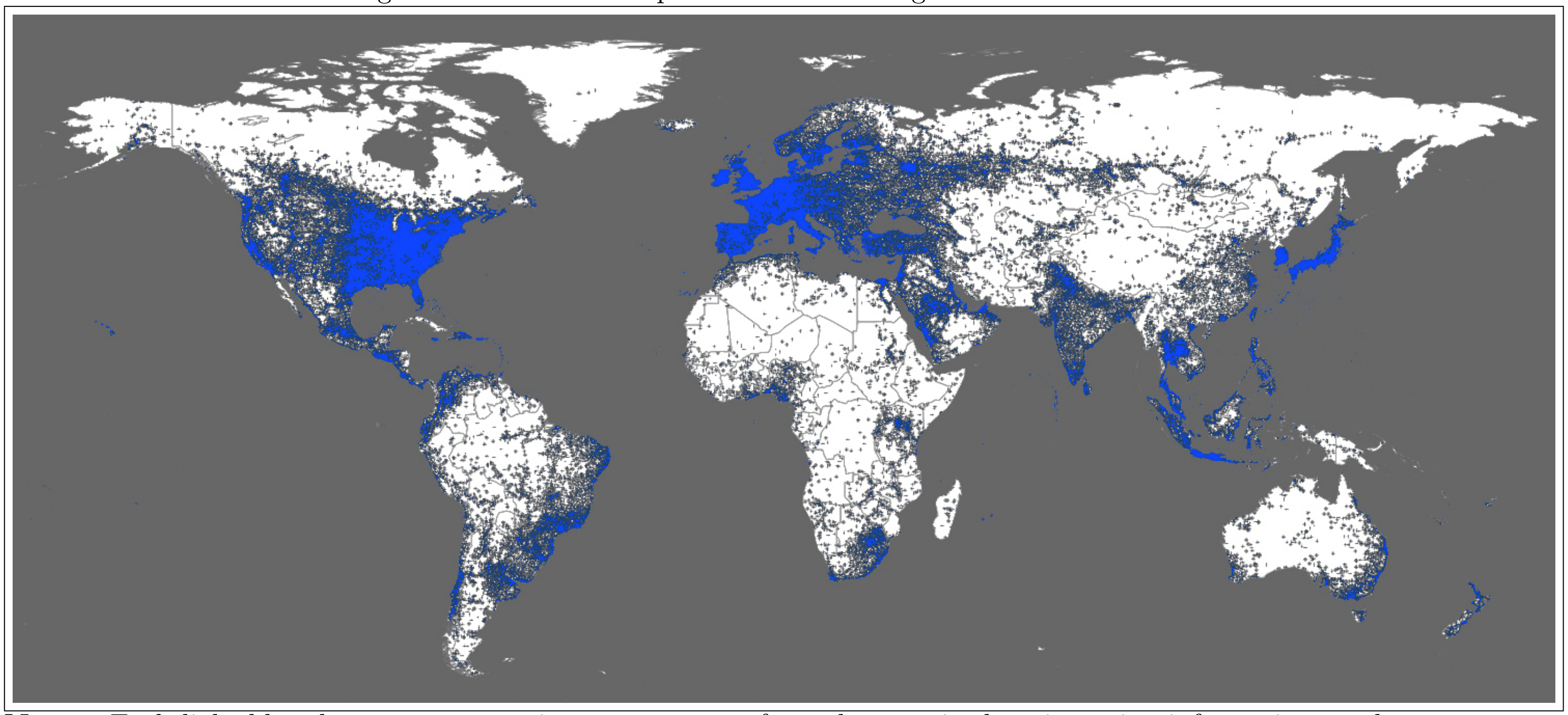
Table 8: Mechanism Between Tweets and Economic Activity

| Dep. var.: ln(Consumption) | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| ln(All Tweets) | $0.51^{***}$ | $0.15^{*}$ | | | | |
| | (0.07) | (0.02) | | | | |
| ln(Leisure Tweets) | | | $0.51^{***}$ | $0.22^{**}$ | $0.93^{**}$ | $0.56^{***}$ |
| | | | (0.07) | (0.02) | (0.44) | (0.08) |
| ln(Work Tweets) | | | | | $-0.42$ | $-0.09$ |
| | | | | | (0.43) | (0.08) |
| ln(Population) | | $1.04^{***}$ | | $1.04^{***}$ | | $1.03^{***}$ |
| | | (0.03) | | (0.03) | | (0.03) |
| $R^2$ | 0.54 | 0.99 | 0.55 | 0.99 | 0.56 | 0.99 |
| Adj. $R^2$ | 0.53 | 0.99 | 0.54 | 0.99 | 0.54 | 0.99 |
| Partial $R^2$ Leisure Tweets | – | – | 0.55 | 0.42 | 0.55 | 0.44 |
| Num. obs. | 48 | 48 | 48 | 48 | 48 | 48 |
| RMSE | 0.68 | 0.12 | 0.67 | 0.12 | 0.67 | 0.12 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

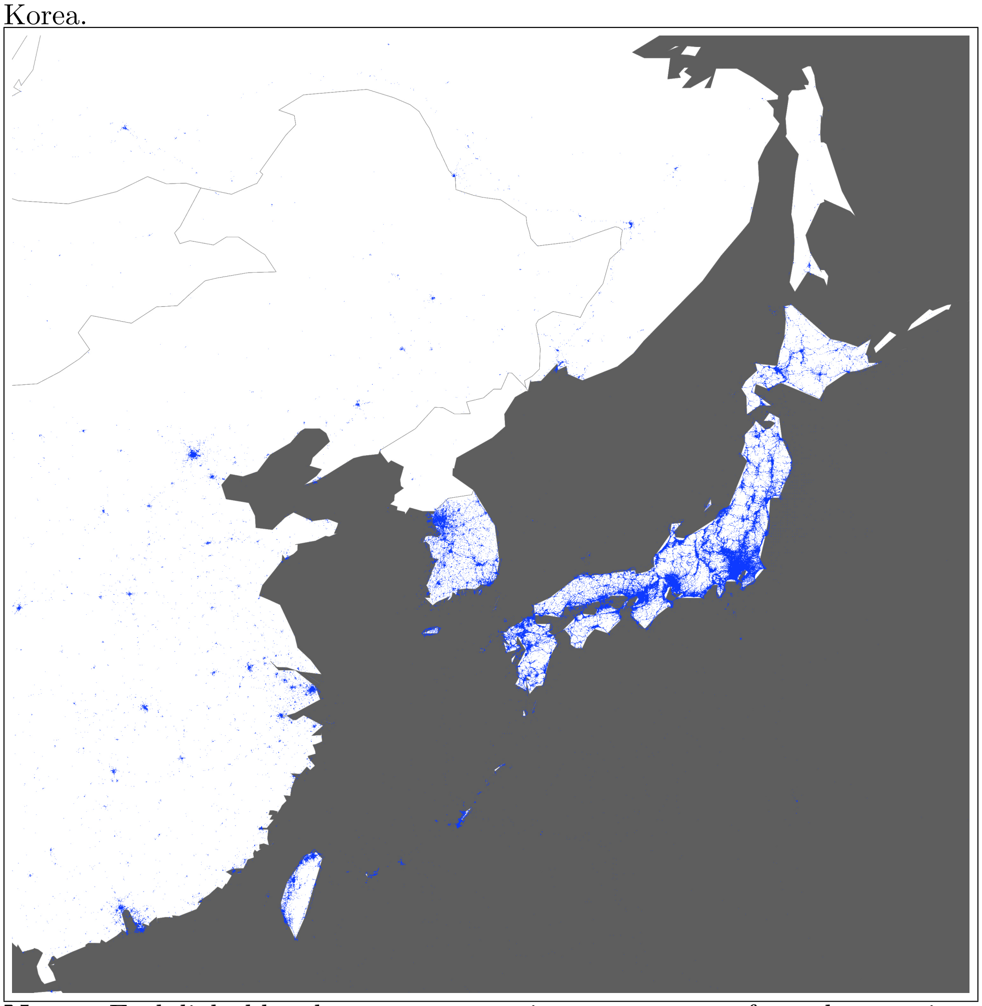
**Notes:** This table uses all image tweets sent from the US in 2012. Tweets are aggregated at the state level and categorized into one of two labels depending on the time in which they were posted. Work tweets includes all tweets sent during weekdays between 8am and 5pm. Leisure tweets includes tweets sent during the weekday before 9am and after 5pm as well as tweets sent on weekends, irrespective of time.

Figure 1: Worldwide Map of Location of Image Tweets.



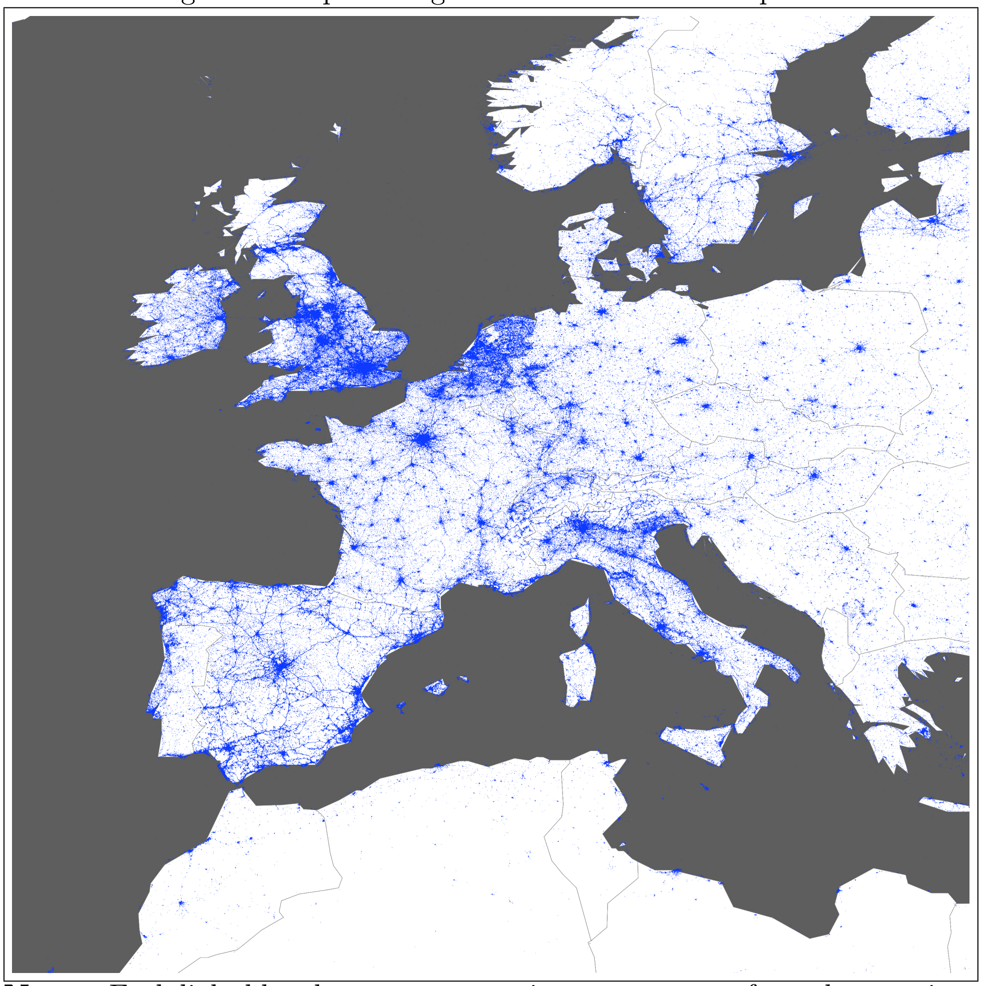
**Notes:** Each light blue dot represents an image tweet sent from that precise location using information on the latitude-longitude. This is a subset of 100 million random tweets from the complete sample for Jan. 2012 - Dec. 2013.

Figure 2: Map of Image Tweets sent from Japan, South Korea and North Korea.



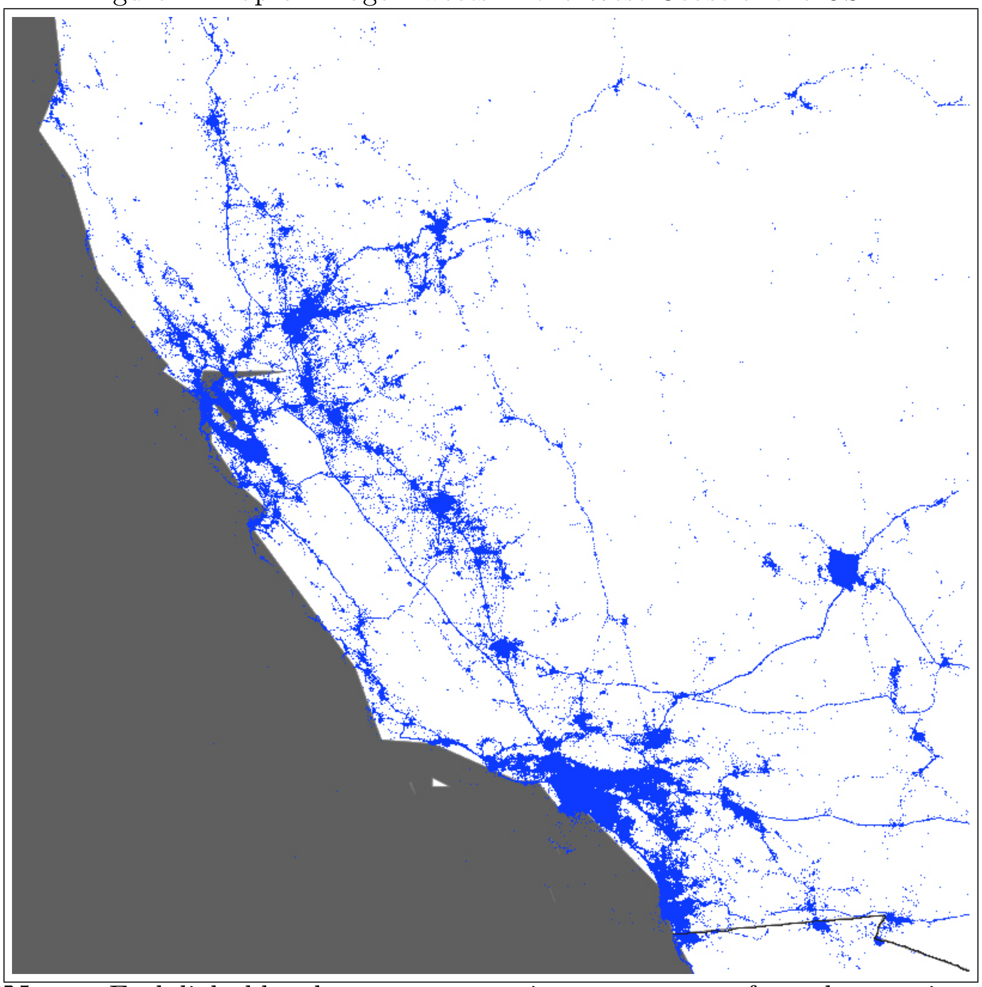
**Notes:** Each light blue dot represents an image tweet sent from that precise location using information on the latitude-longitude. This is a subset of 100 million random tweets from the complete sample for Jan. 2012 - Dec. 2013.

Figure 3: Map of Image Tweets sent from Europe.



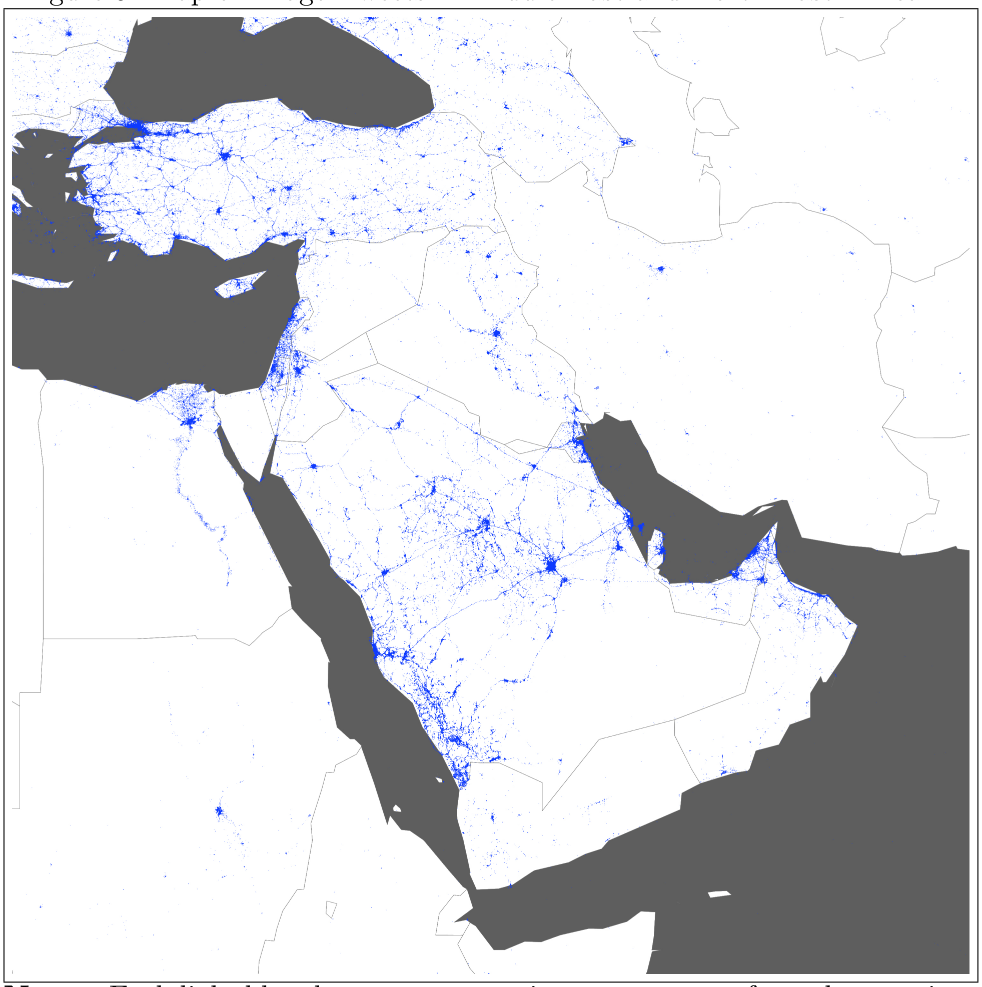
**Notes:** Each light blue dot represents an image tweet sent from that precise location using information on the latitude-longitude. This is a subset of 100 million random tweets from the complete sample for Jan. 2012 - Dec 2013.

Figure 4: Map of Image Tweets in the West Coast of the US.



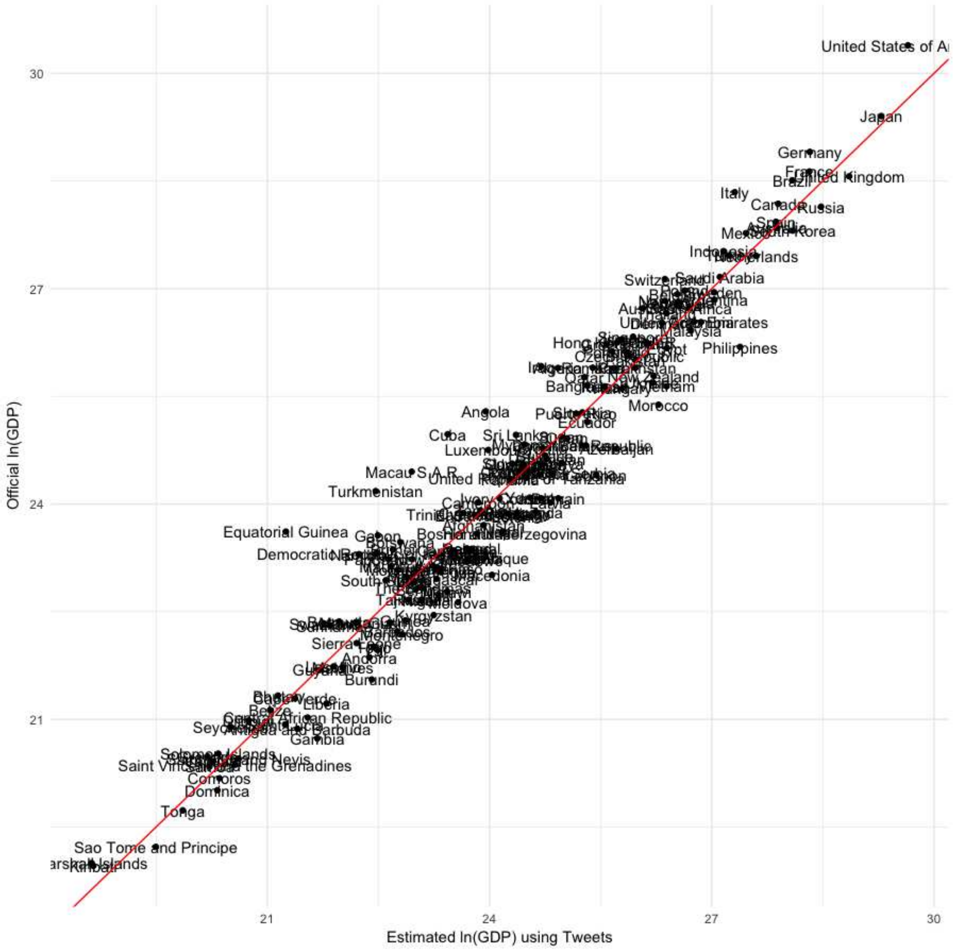
**Notes:** Each light blue dot represents an image tweet sent from that precise location using information on the latitude-longitude. This is a subset of 100 million random tweets from the complete sample for Jan. 2012 - Dec. 2013.

Figure 5: Map of Image Tweets in Middle East and North-East Africa.



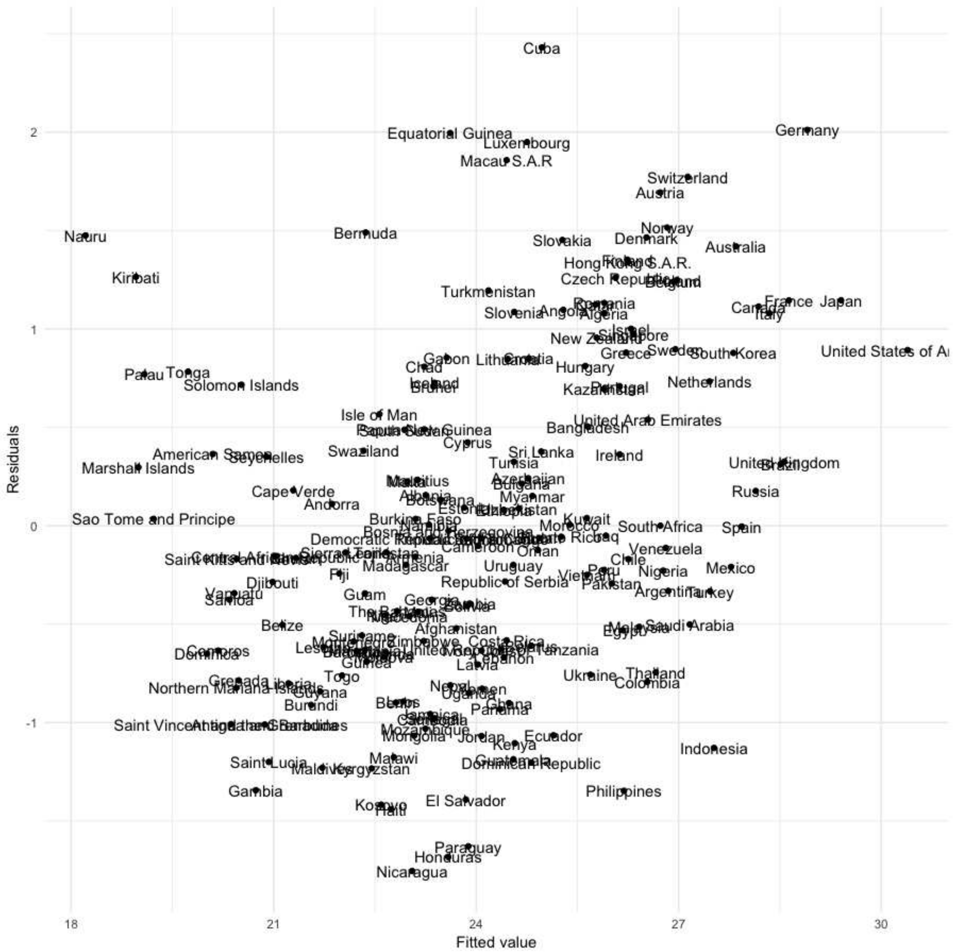
**Notes:** Each light blue dot represents an image tweet sent from that precise location using information on the latitude-longitude. This is a subset of 100 million random tweets from the complete sample for Jan. 2012 - Dec. 2013.

Figure 6: Estimated vs Official GDP for 2013



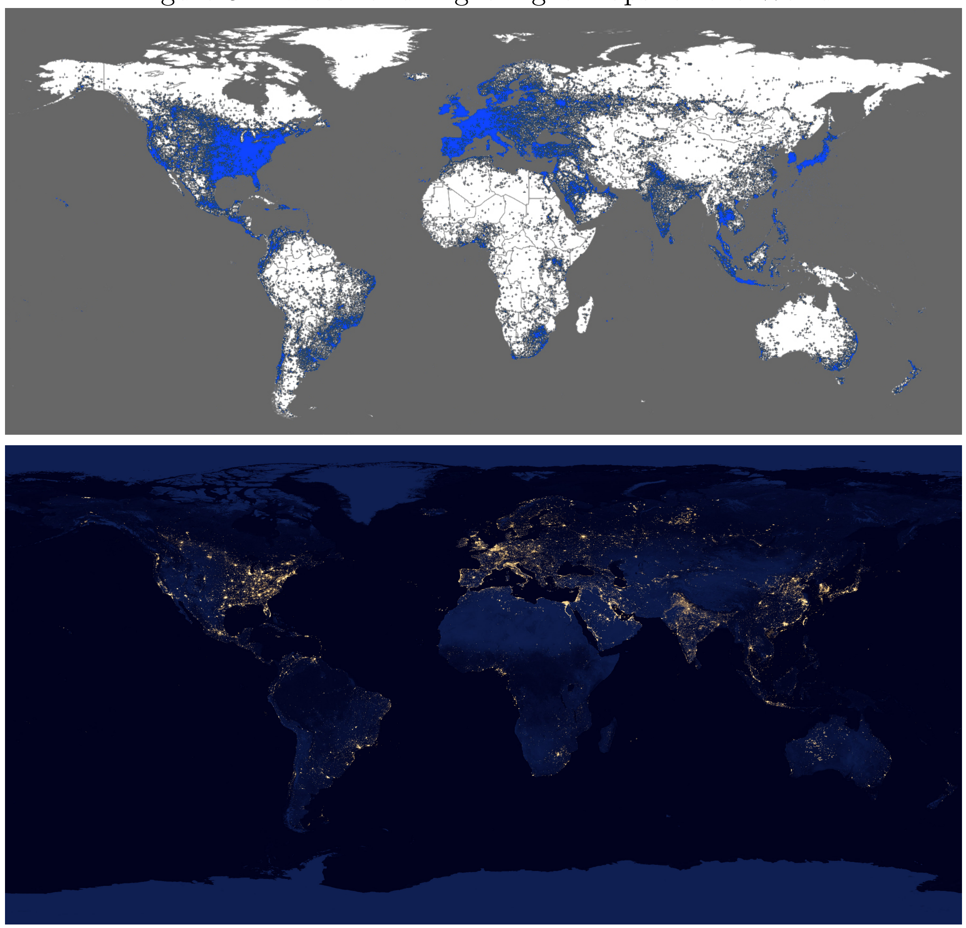

**Notes:** Red line represents 45 degree line, which indicates where GDP estimates based on Tweets per country and official estimates for GDP are equal. This is for year 2013 and 178 countries included in the sample.

Figure 7: Residual and Fitted Value for 2013 GDP Estimates



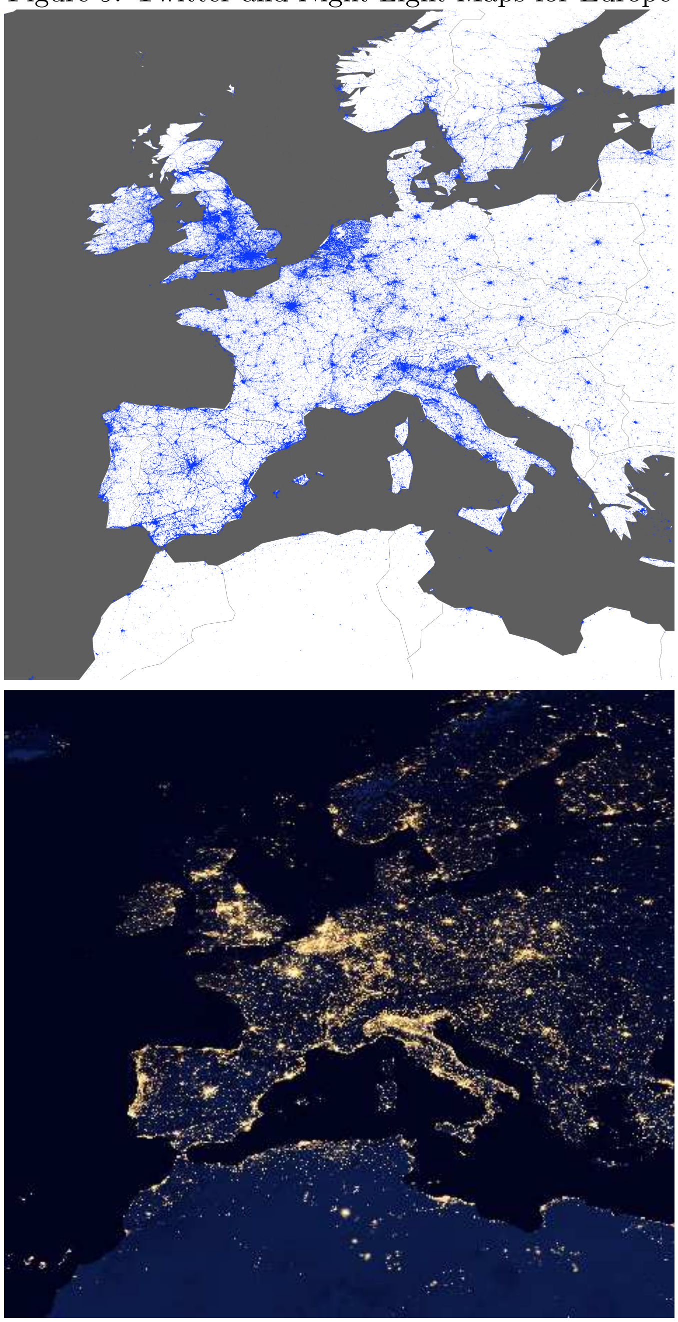

**Notes:** Plot shows the distribution of the residuals of preferred model against the fitted values. This is for year 2013 and 178 countries included in the sample.

Figure 8: Twitter and Night-Light Maps for the World



**Top:** Twitter posts. Each light blue dot represents an image tweet sent from that precise location using information on the latitude-longitude. This is a subset of 100 million random tweets from the complete sample for Jan. 2012 - Dec. 2013.
**Bottom:** Satellite night-light image for 2012.
Source:`https://earthobservatory.nasa.gov/images/79765/night-lights-2012-flat-map`

Figure 9: Twitter and Night-Light Maps for Europe

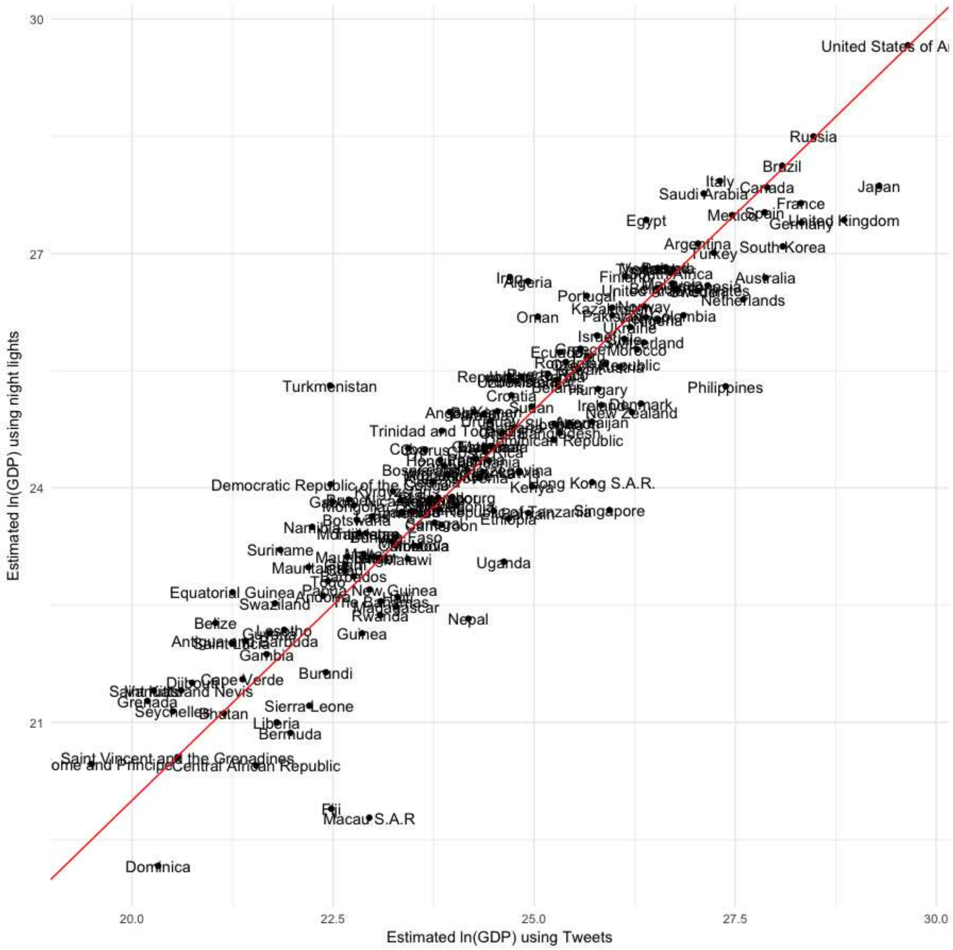


**Top:** Twitter posts. Each light blue dot represents an image tweet sent from that precise location using information on the latitude-longitude. This is a subset of 100 million random tweets from the complete sample for Jan. 2012 - Dec. 2013.

**Bottom:** Satellite night-light image for 2012.

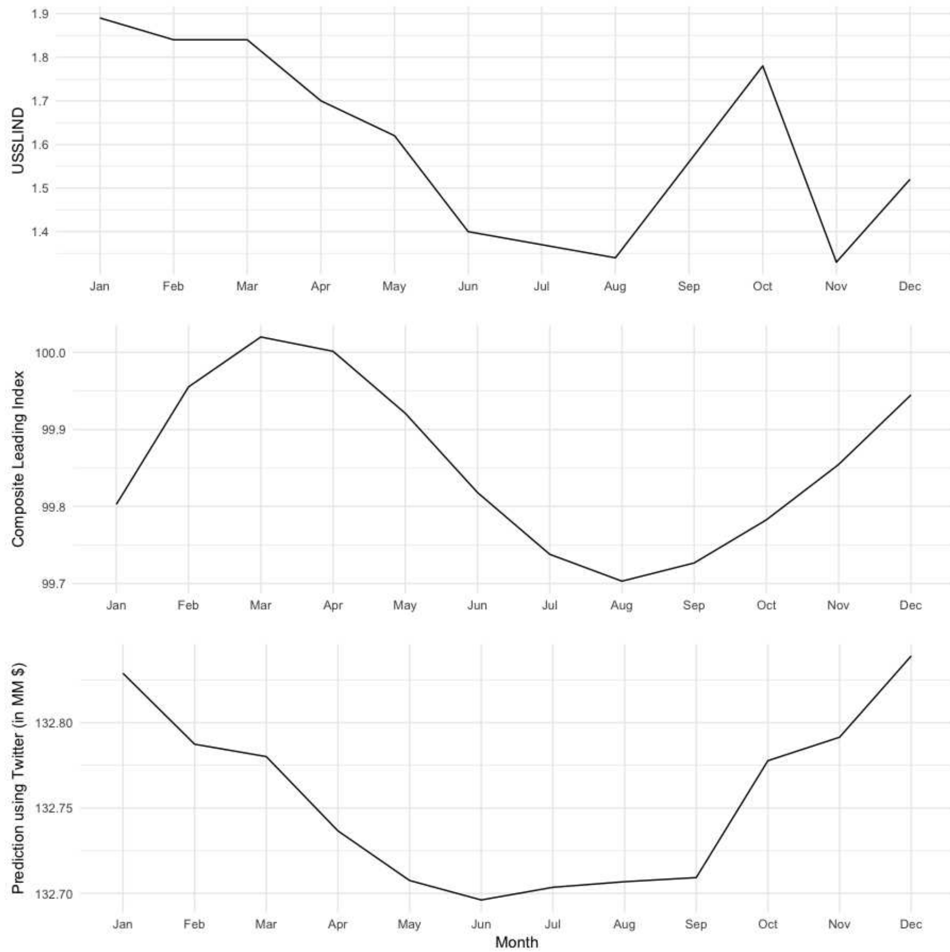
Source:`https://earthobservatory.nasa.gov/images/79765/night-lights-2012-flat-map`

Figure 10: Night-Lights vs Twitter GDP Estimates for 2013



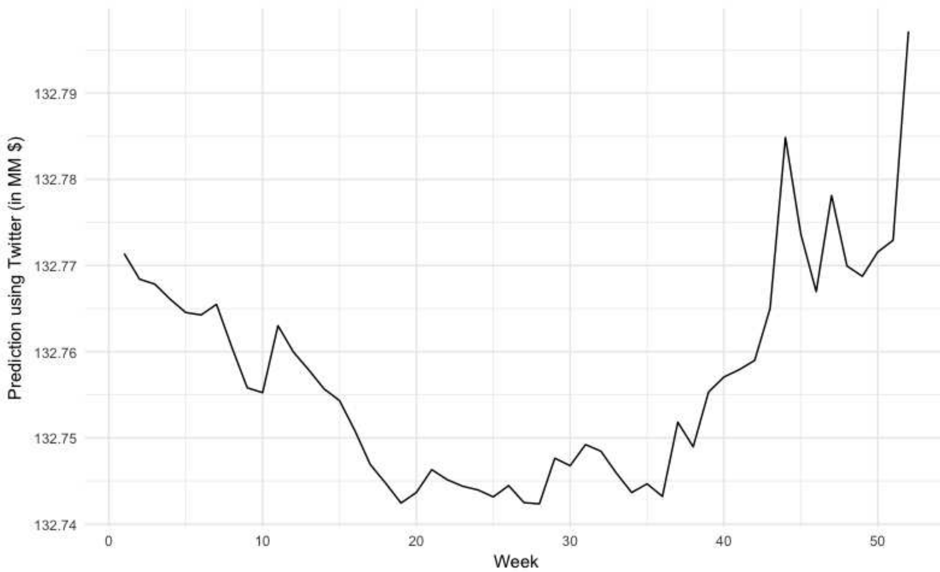

**Notes:** Red line represents 45 degree line, which indicates where GDP estimates using Twitter and night-lights are equal. This is for year 2013 and 171 countries included in the sample.

Figure 11: Monthly Indicators of USA Economic Activity for 2012



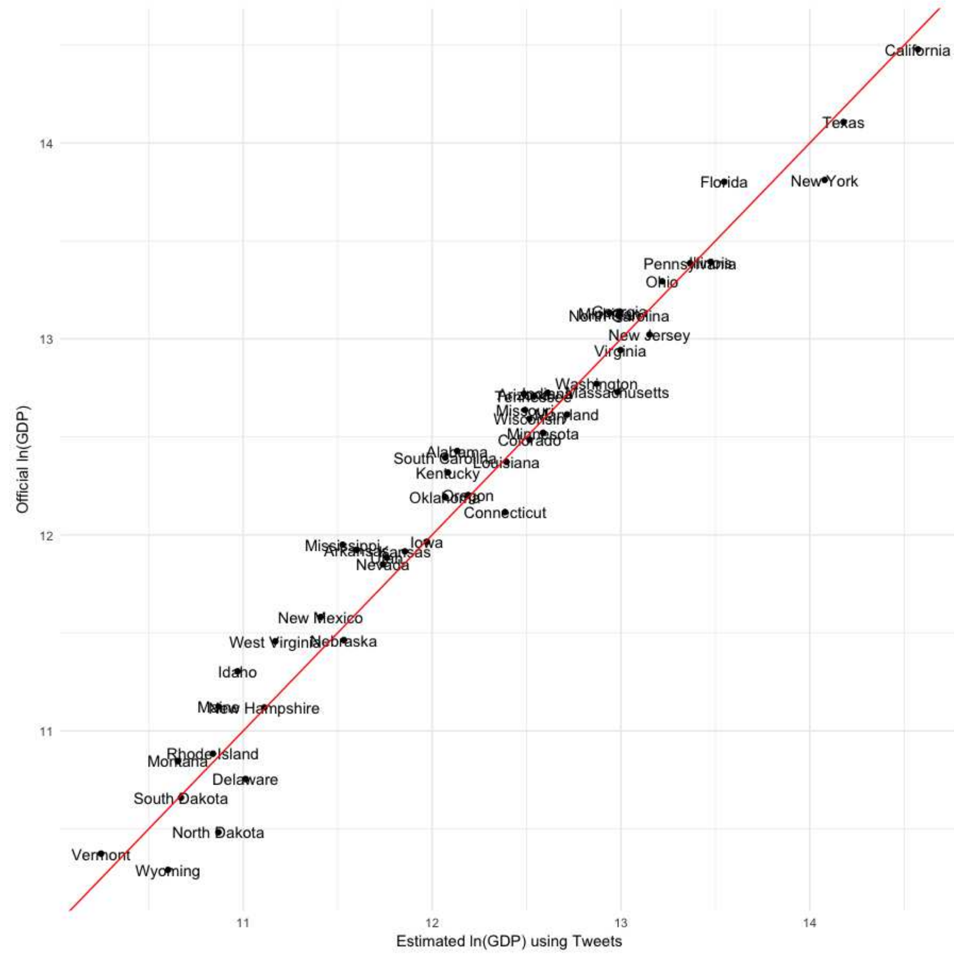

**Top:** Leading Index for the US (USSLIND) by the Federal Reserve Bank of Philadelphia.
**Middle:** Composite leading indicator by the OECD.
**Bottom:** Monthly GDP estimates using volume of tweets posted each month in the US in 2012.

Figure 12: Weekly Estimate of USA Economic Activity for 2012



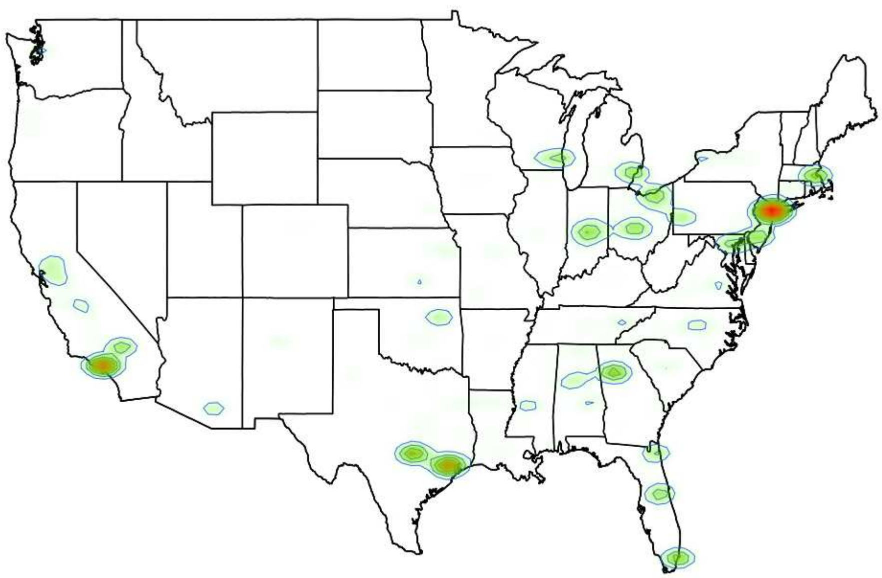

**Notes:** Weekly GDP estimated from number of tweets posted each week in the US in 2012.

Figure 13: Estimated vs Official GDP at the State Level for 2012



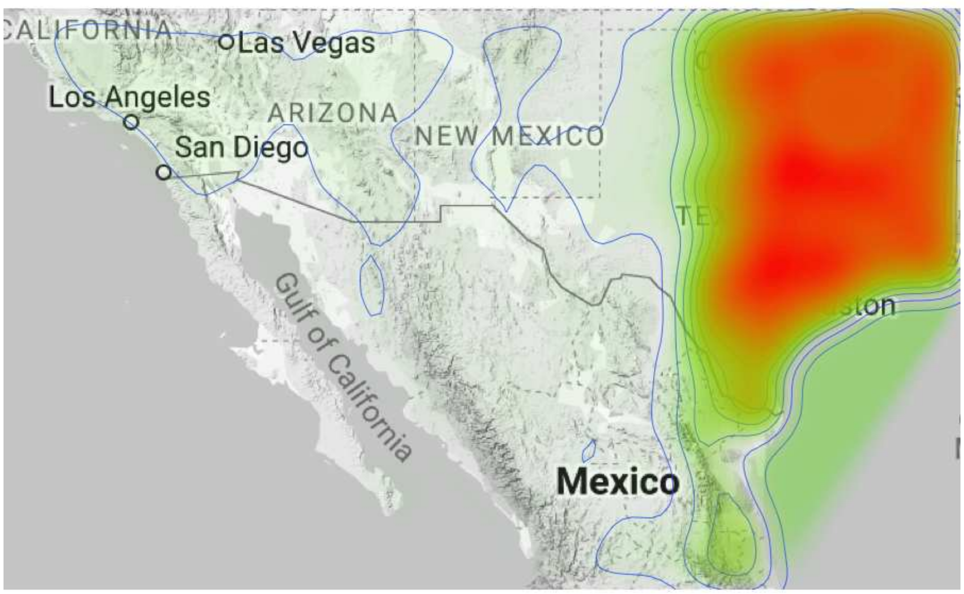

**Notes:** Red line represents 45 degree line, which indicates where estimates based on volume of tweets and official estimates for GDP at the US state level are equal. This is for year 2012.

Figure 14: Density Map for Economic Activity in the US in 2012



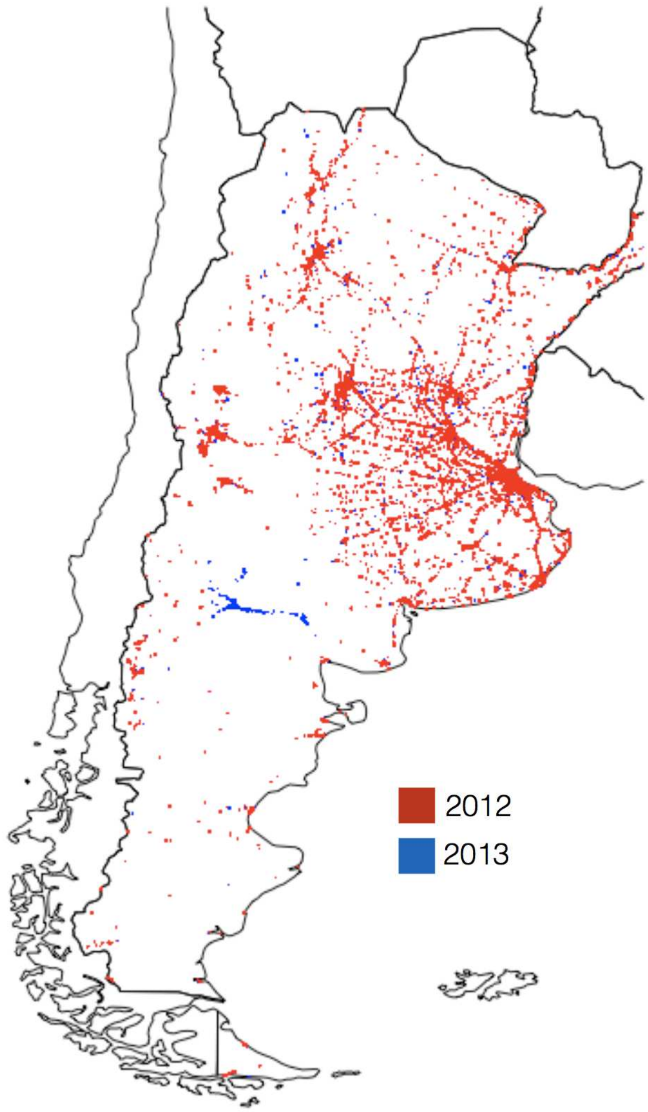
**Notes:** Contours contain areas with more economic activity with areas shaded in red representing the highest concentration of economic activity.

Figure 15: Density Map for Economic Activity in the US-Mexico Border in 2012




**Notes:** Contours contain areas with more economic activity with areas shaded in red representing the highest concentration of economic activity.

Figure 16: Tweets in Argentina in 2012-13




**Notes:** Figure shows distribution of tweets in Argentina for 2012 and 2013. Given that there was a large increase in the number of Tweets from 2012 to 2013 nationwide, I randomly selected a number of tweets for 2013 equivalent to that of 2012. Tweets from 2012 plotted appear on top to show the difference in distribution of tweets between the two years that is present in the area around Vaca Muerta.

# Appendix

## A Supplemental data

### A.0.1 Socio-economic data

The World Bank provides freely and publicly available data on various relevant socio-economic indicators at the country level. Given that one of the main objectives of this paper is to provide an effective proxy for estimating GDP that allows for more transparency in official statistics, it is important that all the data used in this paper is publicly available and thus could be replicated by individuals and institutions. Besides from current GDP in USD, I also obtain total population for each country from the World Bank database.

Another indicator I obtain from the World Bank is the percent of the population that uses the internet. Given that Twitter requires internet service access to establish a connection, the penetration of the internet in a given country is a useful variable to include in our baseline regression.

As described in section 2 official GDP estimates suffer several issues, particularly in developing countries. If this is in fact the case, I could be estimating a reported GDP that is not in fact the *true* GDP. Thus if the estimates using Twitter are imprecise, it could be in part because of measurement error in the GDP calculations I am trying to estimate in the first place. The World Bank produces a composite score assessing the capacity of a country's statistical office. In particular they focus on three specific areas: methodology, data sources, and periodicity and timeliness. The overall score is a simple average of all three area scores on a scale of 0-100, where higher values indicate higher quality data. In subsection 4.2 I use these data quality scores to see if the discrepancies in our estimates are larger for countries with inferior data quality, as assessed by the World Bank.

### A.0.2 Night-time light satellite data

Given that much of the previous literature on alternative ways of estimating GDP has come from night-lights from satellite data, I contrast and compare the performance of satellite and Twitter data as a proxy for GDP (see section 5). The United States Air Force Defense Meteorological Satellite Program satellite orbits the earth roughly 14 times a day, taking images that record the luminous intensity radiated from the earth. Although the main purpose of this task is to detect moonlit clouds, a useful byproduct is that they also capture lights emitted from human settlements.

Scientists then process these images and perform a series of tasks (i.e.: remove intense sources of natural light during summer months, auroral activity, days where cloud cover obstructs the earth's surface, etc.) that leave only man-made light visible. They then average all valid images over the year and report the intensity of light for approximately every 0.86 square kilometer. The intensity figure is an integer between 0 and 63, where higher values indicate more light. These datasets are made publicly available by the National Oceanic and Atmospheric Administration's National Geophysical Data Center. These are also the datasets used in the majority of the papers referred to in section 5 that use night-light as a proxy for GDP. I downloaded these datasets for 2012 and 2013, the same years for which I have complete Twitter data.

# B Computer Vision API to Assess Content of Images

If leisure and consumption are complementary goods, as suggested by Becker (1965), we would expect both of these to occur at the same time. In order to test this hypothesis via social media, I divide all image tweets sent from the US in 2012 in two groups: (i) those sent during working hours (Monday to Friday between 9:00-17:00 hs); and (ii) those sent during leisure time, defined as non-working weekday hours (Monday to Friday after 17:00 hs and before 9:00 hs) and weekends (Saturday and Sunday any time).

I then collect a random sample of 1,000 images from each subset and find the concepts displayed in each of these images by querying a computer vision API. I use Clarifai API that returns different concepts (which includes objects, themes and moods) associated to the image together with a probability score of the likelihood that each concept is truly present in the image. For each image, I collect all concepts above a 0.9 probability score and tally the frequency of concepts for each of the two datasets.

Two examples of such images and the accompanying attributed concepts can be found in

Table 9 shows the five most frequent concepts for each dataset and compares the frequency between them. The table indicates that there are substantial differences in the concepts most widely showcased in the images posted during leisure and working hours. More importantly though, it would seem that several of the top concepts present in the images posted during

Table 9: Top 5 Concepts in Leisure and Work Tweets

| Concept | Leisure tweet (Freq.) | Work tweet (Freq.) | P-value |
|---|---|---|---|
| **Top concepts for leisure tweets** | | | |
| Shopping | 174 | 55 | <0.001 |
| Merchandise | 172 | 62 | <0.001 |
| Food | 162 | 157 | 0.807 |
| Recreation | 151 | 48 | <0.001 |
| Alcohol | 148 | 59 | <0.001 |
| **Top concepts for work tweets** | | | |
| Food | 162 | 157 | 0.807 |
| Cup | 91 | 75 | 0.224 |
| Coffee | 89 | 72 | 0.188 |
| Tea | 75 | 68 | 0.602 |
| Desk | 32 | 55 | 0.015 |
| Observations | 1,000 | 1,000 | |

leisure time are related to consumption (i.e.: shopping and merchandise). This indicates that leisure tweets tend to showcase consumption more than image tweets posted during working hours; and would validate the idea that consumption and leisure are complementary.

Figure 17: Example of Computer Vision API

| Concept | Prob. Score |
|---|---|
| food | 0.989 |
| refreshment | 0.963 |
| plate | 0.952 |
| lunch | 0.951 |
| vegetable | 0.944 |
| bread | 0.943 |
| hot | 0.941 |
| meal | 0.939 |
| no person | 0.939 |
| tortilla | 0.935 |
| delicious | 0.931 |
| grow | 0.924 |
| drink | 0.922 |
| dinner | 0.918 |
| restaurant | 0.913 |
| meat | 0.904 |
| party | 0.886 |
| sauce | 0.879 |
| dish | 0.878 |
| table | 0.875 |

| Concept | Prob. Score |
|---|---|
| telephone | 0.998 |
| coffee | 0.994 |
| computer | 0.983 |
| laptop | 0.983 |
| business | 0.979 |
| office | 0.974 |
| desk | 0.971 |
| no person | 0.967 |
| screen | 0.95 |
| technology | 0.947 |
| cup | 0.942 |
| dawn | 0.933 |
| money | 0.929 |
| internet | 0.928 |
| paper | 0.914 |
| composition | 0.913 |
| still life | 0.911 |
| organization | 0.899 |
| drink | 0.895 |

**Top:** Example of leisure tweet: image tweet taken in the US during Monday to Friday after 17:00 hs and before 9:00 hs) or weekend (Saturday and Sunday any time).
**Bottom:** Example of a work tweet: image taken in the US on Monday to Friday between 9:00-17:00 hs