2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8316

From Twitter to GDP: Estimating Economic Activity From Social Media

Indaco, Agustín
The Graduate Center CUNY, New York, U.S.A

Abstract

This paper shows how the use of data derived from Twitter can be used as a proxy for measuring GDP at the country level. Using a dataset of 270 million geo-located image tweets shared on Twitter in 2012 and 2013, I find that: (i) Twitter data can be used as a proxy for estimating GDP at the country level and can explain 94 percent of the variation in GDP; and (ii) that the residuals from my preferred model are negatively correlated to a data quality index which assesses the capacity of a country's statistical system. This suggests that my estimates for GDP are more accurate for countries which are considered to have more reliable GDP data. Taken together, these findings show that institutions and individuals could use social media data to corroborate official GDP estimates; or alternatively for government statistic agencies to incorporate social media data to complement and further reduce measurement errors.

Keywords: *National Accounts, Big Data.*

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 the economic performance of countries (Masood (2014)). One of the many problems with estimating GDP is that its measurement is often complicated and expensive to produce, particularly for developing countries. This could lead to measurement errors that in turn mislead policy evaluation and recommendations. 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. In light of this, a lot of 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 will argue for the use of data derived from social media posts as a proxy for measuring GDP. By locating and analyzing the content of hundreds of million social media posts, I will show that one can estimate economic activity at the country level.

Social media can contribute greatly to economic research in this regard, as vast information can be extracted from the location and content of their posts. Measures taken from social media can serve both as substitutes as well as complements to traditional survey data. In particular, social media data has several properties that result beneficial when estimating economic measures. First, social media data is publicly available and has a low cost of obtaining and storing. 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 statistic agencies and the public at no cost (other than the necessary computing power and data storage). The public aspect of this data also allows for more transparency in official statistics, as the estimates could be replicated endlessly by individuals and institutions all over the world. Second, social media data is available in real time, which allows for economic estimates that are currently produced in annual or quarterly intervals to be produced at shorter time intervals. This would allow for clearer foresight for companies and individuals when making economic decisions. Third, 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 politicaly and thus fabricate meaningful areas of study that are not possible with official datasets.

Recently, 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 in a series of papers. 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), Henderson et al. (2012) and Sutton et al. (2007)). These studies have shown that the intensity of artificial night-lights highly correlates with GDP and thus can be used to estimate economic activity for different geographic regions.

In this paper I propose using posts from popular social media applications, in this case Twitter, as a measure that has all the same benefits as night-lights, but can be a more accurate estimator of GDP and has several other advantages.

For this paper I have all geo-located image tweets shared on Twitter for the years 2012 and 2013. I have two main findings: (i) that social media data can be used as a proxy for estimating GDP at the country level, as shown by the preferred model explaining 94 percent of the variation in GDP; and (ii), I find a negative correlation between the residuals of my model and a data quality score put together by the World Bank which suggests that my estimates for GDP are more accurate for countries which are considered to have more reliable GDP data. This is a strong result that suggests that social media data could be used as a complement to survey data to increase the accuracy of GDP estimates.

2. Data

2.1 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 emerged in 2006 and by 2012 it had 140 million global users which sent out 340 million tweets per day. Unless restricted by the user, tweets are publicly available and can be read via the application or on a web browser. Created as a text-only platform, 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.

The dataset used in this paper contains all geo-tagged image tweets posted on Twitter for years 2012 and 2013. This dataset was provided directly by Twitter, through a Twitter Data Grant submission in 2014 by the Cultural Analytics Lab. The total dataset contains 270 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 (5 decimal points) of where the tweet was sent from; iii) the date and time in which the tweet was sent; iv) the image tweeted; and v) any accompanying text.

Table 1 summarizes this Twitter data by year and by country income groups (using the World Bank's classification). The breakdown of average tweets per income group shows that countries in higher income groups have more tweets.

Figure 1 shows that there is 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. Clusters of light blue points can be found both in areas that are more densely populated as well as areas where we know have higher levels of per capita income. For example, in the United States, the largest concentration of image tweets seem to be centered in the coastal areas, but not so in the less-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.

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 (122)	211,993.9	1,126,937.3
	(541,334.7)	(4,048,721.5)
Upper-middle (100)	89,123	406,004.9
	(201,367.9)	(836,887.4)
Lower-middle (86)	37,394.5	209,938.9
	(106,230.9)	(929,686.9)
Low (51)	735.9	3,602.5
	(898.9)	(4,370.9)

Note: number of countries per income group in brackets

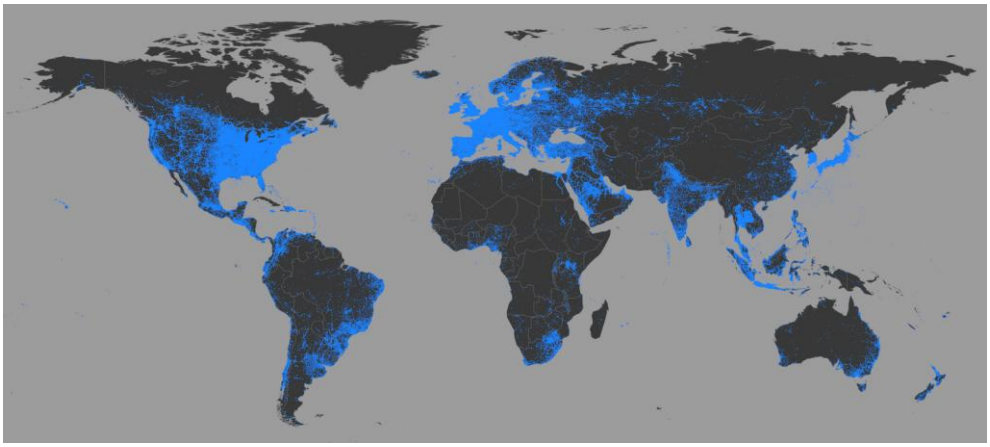


Figure 1. Map of image Tweets shared around the world in 2012 and 2013.

2.2 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 GDP, 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 use the internet. Given that Twitter requires internet service access to establish a connection, the penetration of internet in a given country is a useful variable to include in our baseline regression.

The World Bank also 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 Section 3.1 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.

3. Using Twitter to estimate GDP

The main goal of this paper is to see whether Twitter data is a valid proxy for estimating GDP. I will estimate GDP at the country level using tweets as the main variable of interest for panel data for years 2012 and 2013. For this I will estimate:

$$ln(GDP)_{i,t} = \beta_0 + \alpha_t + X'_{i,t} + \beta_1 \ln(Tweets)_{i,t} + \varepsilon_{i,t} \quad (1)$$

where I estimate GDP for country i in year t. The vector Xi,t is composed of country characteristics including population, the share of the population with access to internet and continent to which it belongs. The coefficient of relevance to us is β1 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 (αt) control for any differences in use of Twitter from one year to the other. All of these are publicly available data.

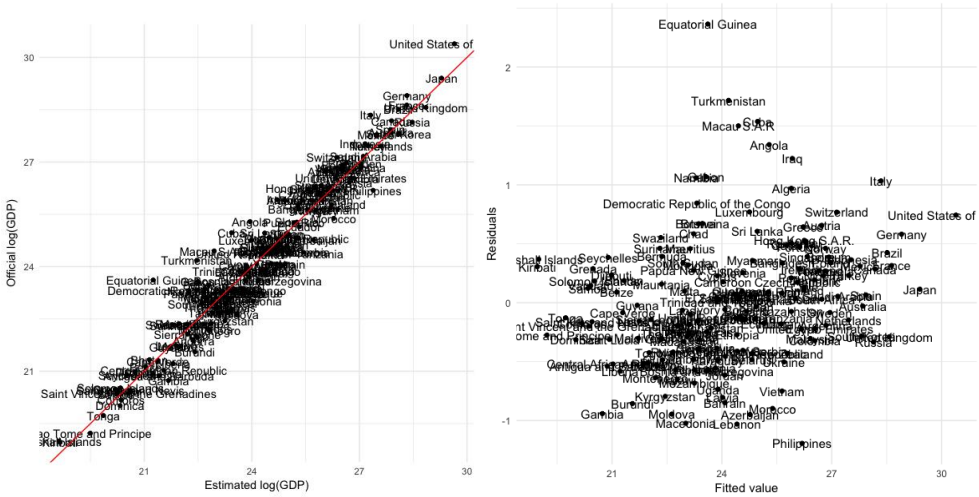
The corresponding estimates are reported in Table 2. There are 184 countries in our dataset for which I have data on GDP, Twitter and population, for both years. In column 1, I regress the natural log of GDP solely on the number of image tweets sent from each country. This is the baseline regression. The coefficient of interest on ln(Tweets) is highly significant and the R2 is 0.78. When the population of the country is included in column 2, the coefficient on ln(Tweets) is reduced, but remains statistically significantly different from zero at the 1 percent level, and R2 increases to 0.87. In column 3, I add categorical dummies for the continents in which each country is situated in. This captures the cultural differences in image sharing on social media platforms that exists between regions. Neither the coefficient of interest or the goodness of fit change greatly. Column 4 adds the share of the population that has access to 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 (these are: Libya, Kosovo, Curacao, Palau, South Sudan and San Marino). The coefficient of interest on ln(Tweets) is again reduced, but remains statistically significantly different from zero at the 1 percent level, and R2 increases to 0.94. These measurements are slightly larger than those obtained by similar studies using night-lights (Doll (2006) and Sutton (2007)).

Table 2 shows that the number of image tweets sent in a year is a pretty good measure for estimating GDP at the country level, being able to explain 78 percent of the variation in GDP on its own, and up to 94 percent when introducing other variables that are readily and publicly available. In all specifications, the coefficient on the number of image tweets is statistically significantly different from zero. Figure 2(a) is a visual representation of these estimates: the estimates lay pretty closely around the 45 degree line. There are a few exceptions that stand out; most notably Equatorial Guinea. Figure 2(b) 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).

Table 2. Estimating Country GDP

Dep. Var.: ln(GDP)	(1)	(2)	(3)	(4)
ln(Tweets)	0.66***	0.49***	0.45***	0.18***
	(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
Num. obs.	368	368	368	356

Note: ***p<0.01, **p<0.05, *p<0.10



Figures 2 (a) Estimated vs Actual GDP for 2013 and (b) Residual and Fitted Value for 2013 GDP Estimates

3.1 Data quality issues

While the previous section showed that image tweets could be used to estimate GDP at the country level, Jerven (2013) 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 and thus that the model's estimates are off because of measurement error on the official GDP estimates. If this is the case, data from tweets could be useful as an additional measure at the national level to produce more accurate estimates.

In order to analyze this, I will incorporate a measure of data quality put together 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 Upper-middle income, Lower-middle income and Low-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 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 3, the overall estimates are very similar for this subset of countries as for our general model presented in Table 2, both in terms of the coefficient on ln(Tweets) as well as the R2. I then collect the residuals of equation 1 and run the following regression:

$$|Residuals|_{i,t} = \beta_0 + \beta_1 DataQuality_{i,t} + \beta_2 \ln(Tweets)_{i,t} + \beta_3 \ln(GDP)_{i,t} + \varepsilon_{i,t} \qquad (2)$$

where I regress the absolute value of the residuals for country i in year t on the data quality index, the number of tweets and GDP. The coefficient of interest is β1: a negative and statistically significant coefficient would indicate that our baseline model in Equation 1 more accurately estimates GDP for countries which have more reliable national account estimates. Column 5 of Table 3 shows that the data quality indicator coefficient is in fact negative and statistically significantly different than zero at the 1 percent confidence level. Column 6 includes countries' GDP estimate to control for the possibility that the model's estimates are more accurate for countries with larger economies. Adding this coefficient makes the coefficient on the data quality variable slightly more negative and still statistically significant. This shows that GDP estimates using our baseline model are more accurate for countries with high quality data, and vice versa. Given the long literature showing that official GDP estimates are inaccurate, it is important to acknowledge that it is possible that the GDP estimates we are trying to fit the model to are in fact inexact. 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. This is a strong result that suggests that social media data could be used as a complement to survey data to increase the accuracy of GDP estimates.

Table 3. Data Quality Issues

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	**ln(GDP)**	**ln(GDP)**	**ln(GDP)**	**ln(GDP)**	**Abs. Resid.**	**Abs. Resid.**
ln(Tweets)	0.60***	0.37***	0.39***	0.24***	✔	✔
	(0.02)	(0.02)	(0.03)	(0.03)		
Data Quality					-0.04**	-0.05***
					(<0.01)	(<0.01)
ln(Population)		✔	✔	✔		
Continent			✔	✔		
Internet				✔		
ln(GDP)						✔
R^2	0.76	0.90	0.91	0.93	0.03	0.07
Adj. R^2	0.76	0.90	0.91	0.93	0.03	0.06
Num. obs.	240	240	240	236	240	240

Note: ***p<0.01, **p<0.05, *p<0.10

4. Conclusion

The main goal of this paper was to investigate whether social media data from Twitter could be used as a proxy for estimating GDP at the country level.

For this, I have all geo-located image tweets shared on Twitter for the years 2012 and 2013. The first finding of this paper is that Twitter data can be used as a proxy for estimating GDP at the country level, and my preferred model can explain 94 percent of the variation in GDP. The coefficient on the number of tweets sent from each country is statistically significant.

I then go on to study the relationship between the residuals of my preferred model and a data quality score computed by the World Bank. Given the numerous concerns related to the accuracy of the official GDP estimates, particularly in less-developed countries, it could be the case that the figures we are trying to estimate are not in fact the true GDP. This could in turn cause our estimates to be imprecise because of the measurement error in the official GDP estimate we are trying to calibrate our model to in the first place. I find that the residuals from my model are in fact negatively correlated to the data quality index; which suggests that my estimates for GDP 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.

These findings lead us to conclude that social media data contains useful information that can be used to estimate GDP at the country level. Potentially, this could be used for institutions and individuals to corroborate official GDP estimates, or alternatively for

government statistic agencies to incorporate social media data to complement and further reduce measurement errors.

References

Doll, C. N. H., J.-P. Muller, and J. G. Morley (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics*, 57(1):75–92.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012). Measuring Economic Growth from Outer Space. *American Economic Review*, 102(2):994–1028.

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

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

Pinkovskiy, M. and X. Sala-i Martin (2016). Lights, camera income! illuminating the national accounts- household surveys debate. *The Quarterly Journal of Economics*, 131(2):579–631.

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