CARMA 2018

From Twitter to GDP: Using Social Media to Estimate Economic Activity

Agustín Indaco
@agustindaco

July 12, 2018

Measuring GDP (and its problems)

- Despite incessant debate about its ability to accurately measure the state of the economy, GDP is the most widely used measure of a country's economic performance.

- But these estimates:
  1. are **complicated & expensive:** particularly for developing countries.
  2. are subject to **measurement error:** s.d. of change in PWT countries' average growth over the period 1970–1999 was 1.1 percent per year between version 6.1 & 6.2, given average growth rate of 1.56 percent (Johnson et al. (2013)). Kerner and Crabtree (2018) find that there are non-random variations in official macroeconomic estimates.
  3. use **outdated standards:** while some countries adhere to the most recent standards set by the revised SNA 2008, many developing countries have still not adopted the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.

Efforts to estimate GDP using proxies

- Inflow and outflow of **daily postal data records** between countries have been used to estimate GDP (Hristova et al., 2016).

- **Toll data** from Germany has been used to create leading index of economic activity (Asiktas & Zimmermann, 2013)

- **Satellite night-light images:**
  1. economic growth over time (Henderson et al., 2012)
  2. improve estimates of economic output, particularly for countries with poor statistical systems (Chen & Nordhaus, 2010)
  3. GDP at subnational level (Doll et al., 2006 and Sutton et al., 2007).

Advantages of Social Media Data

1. **Publicly available and has a low cost** of obtaining and storing
    - increased transparency as could be replicated
2. Available in **real time**
    - allows for short-time interval estimations
3. Geo-tagged social media posts can be geographically assigned to a **precise location**
    - can estimate at sub-national level or aggregate data between areas that are not politically bound together

Data

- **270 million geo-tagged image Tweets** sent worldwide for 2012, 2013 and first 6 months of 2014 (Twitter grant, 2014).

- Corresponding information: unique user id, latitude and longitude, time of post, accompanying text.

Data

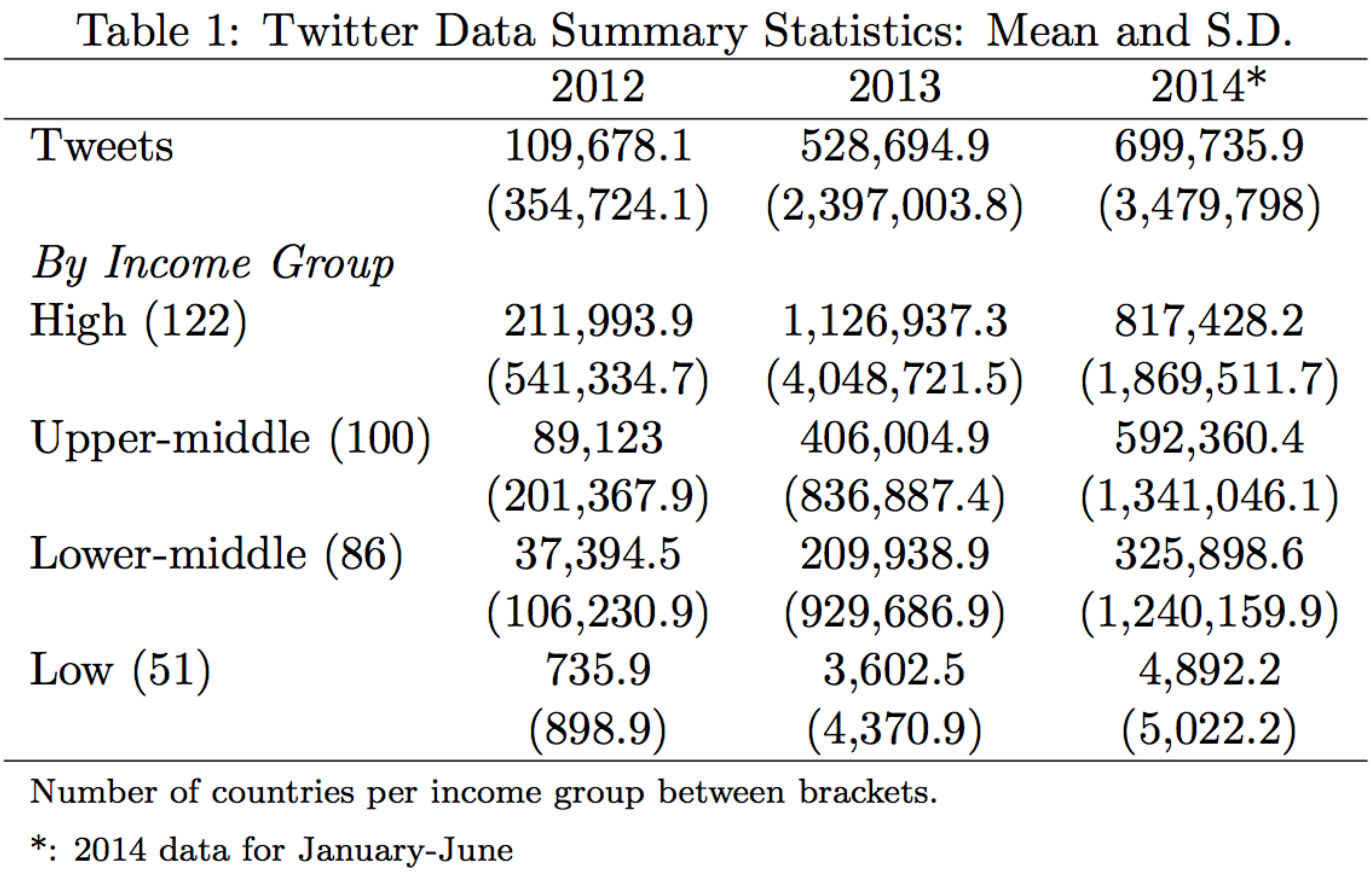


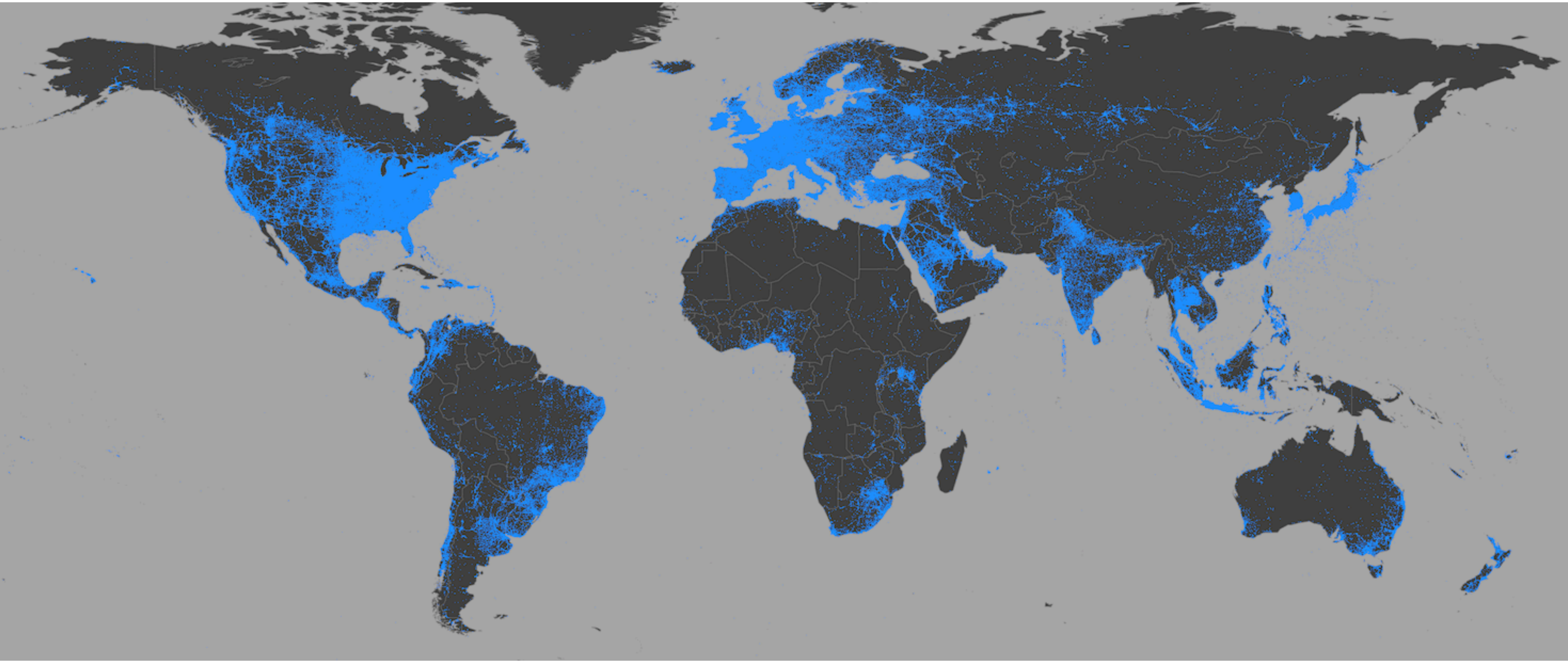
Table 1: Twitter Data Summary Statistics: Mean and S.D.

	2012	2013	2014*
Tweets	109,678.1	528,694.9	699,735.9
	(354,724.1)	(2,397,003.8)	(3,479,798)
By Income Group			
High (122)	211,993.9	1,126,937.3	817,428.2
	(541,334.7)	(4,048,721.5)	(1,869,511.7)
Upper-middle (100)	89,123	406,004.9	592,360.4
	(201,367.9)	(836,887.4)	(1,341,046.1)
Lower-middle (86)	37,394.5	209,938.9	325,898.6
	(106,230.9)	(929,686.9)	(1,240,159.9)
Low (51)	735.9	3,602.5	4,892.2
	(898.9)	(4,370.9)	(5,022.2)

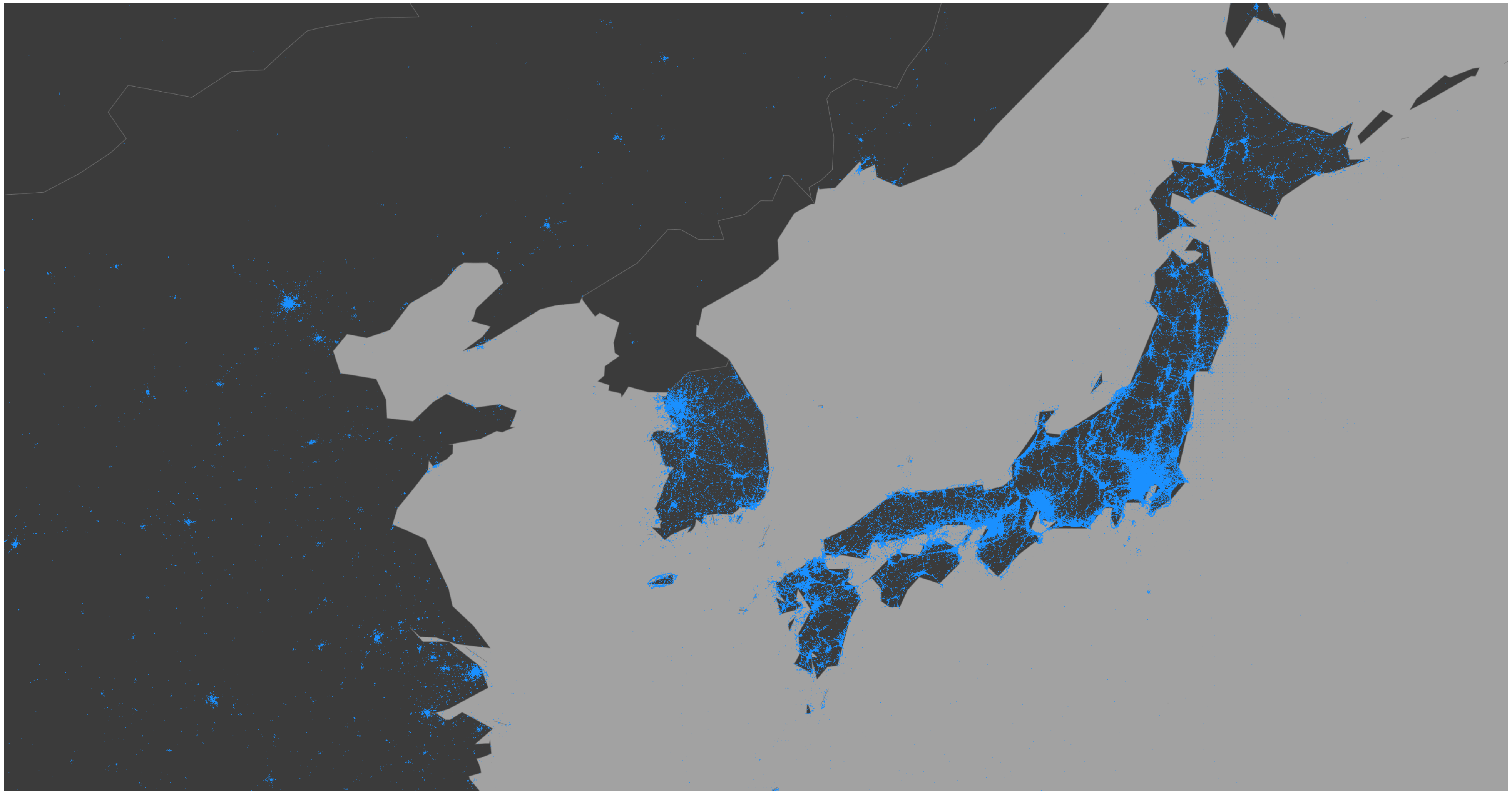
Number of countries per income group between brackets.

*: 2014 data for January-June

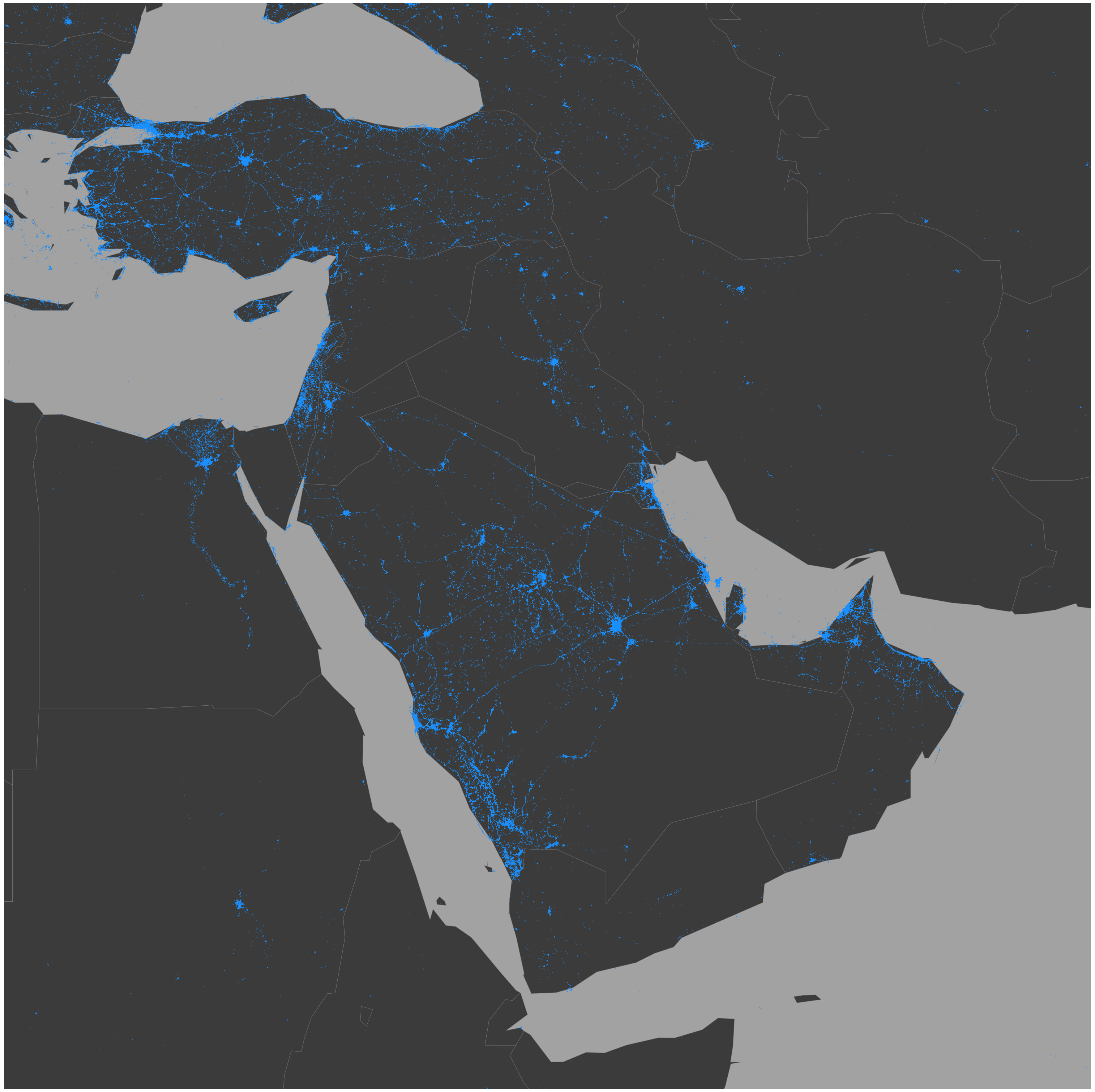
Twitter map



South Korea and North Korea



Middle East and Northeast Africa



Data: Supplemental data

- **GDP (constant 2010 US$):** converted from domestic currencies using 2010 official exchange rates.

- **Population:** total

- **Individuals using the Internet (% of population)**

- **Data Quality:** 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.

Model

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

α_t : time fixed effects

$X'_{i,t}$: population, internet access and continent dummies for country i in year t

Results

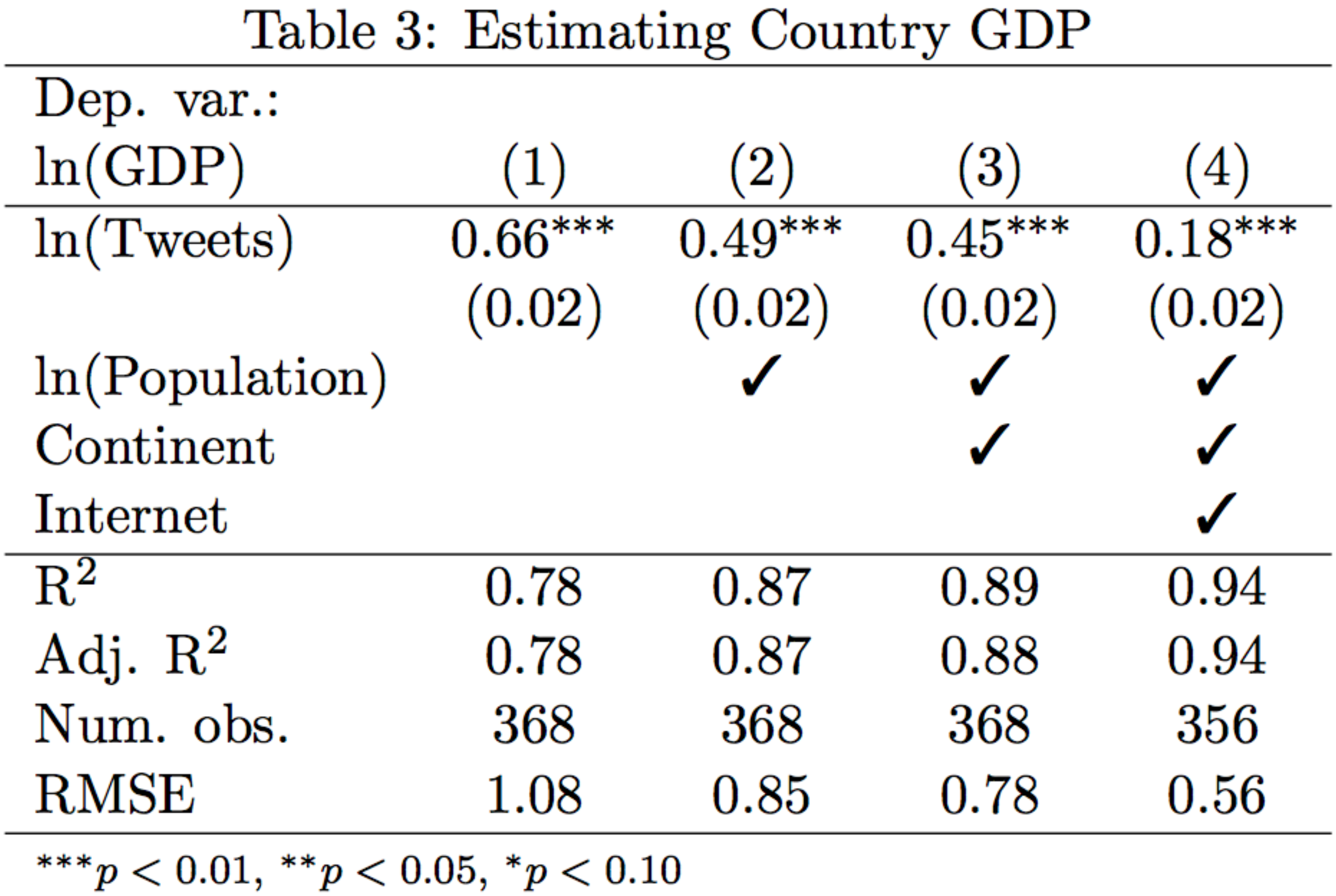
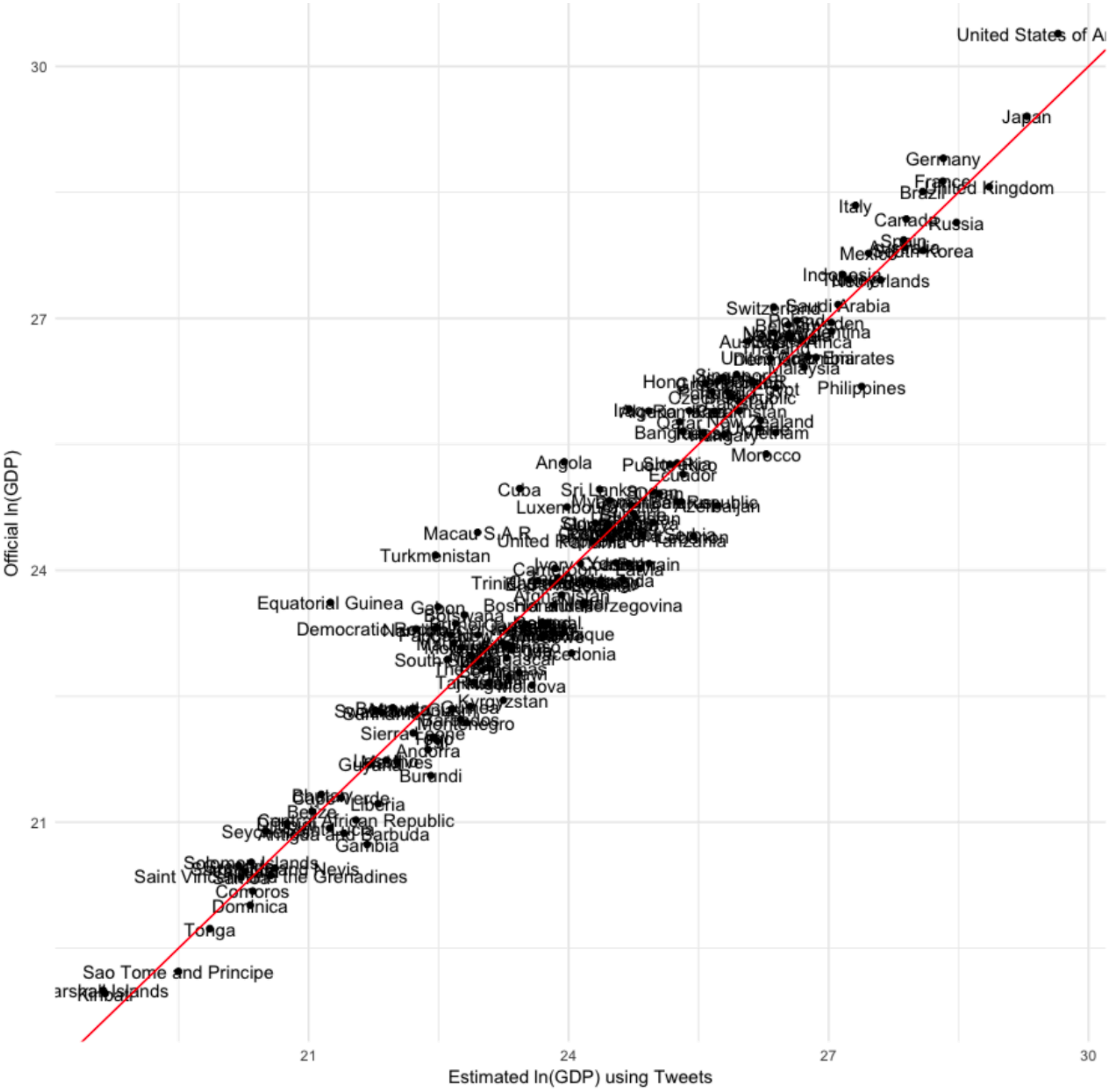


Table 3: Estimating Country GDP

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
RMSE	1.08	0.85	0.78	0.56

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

Results

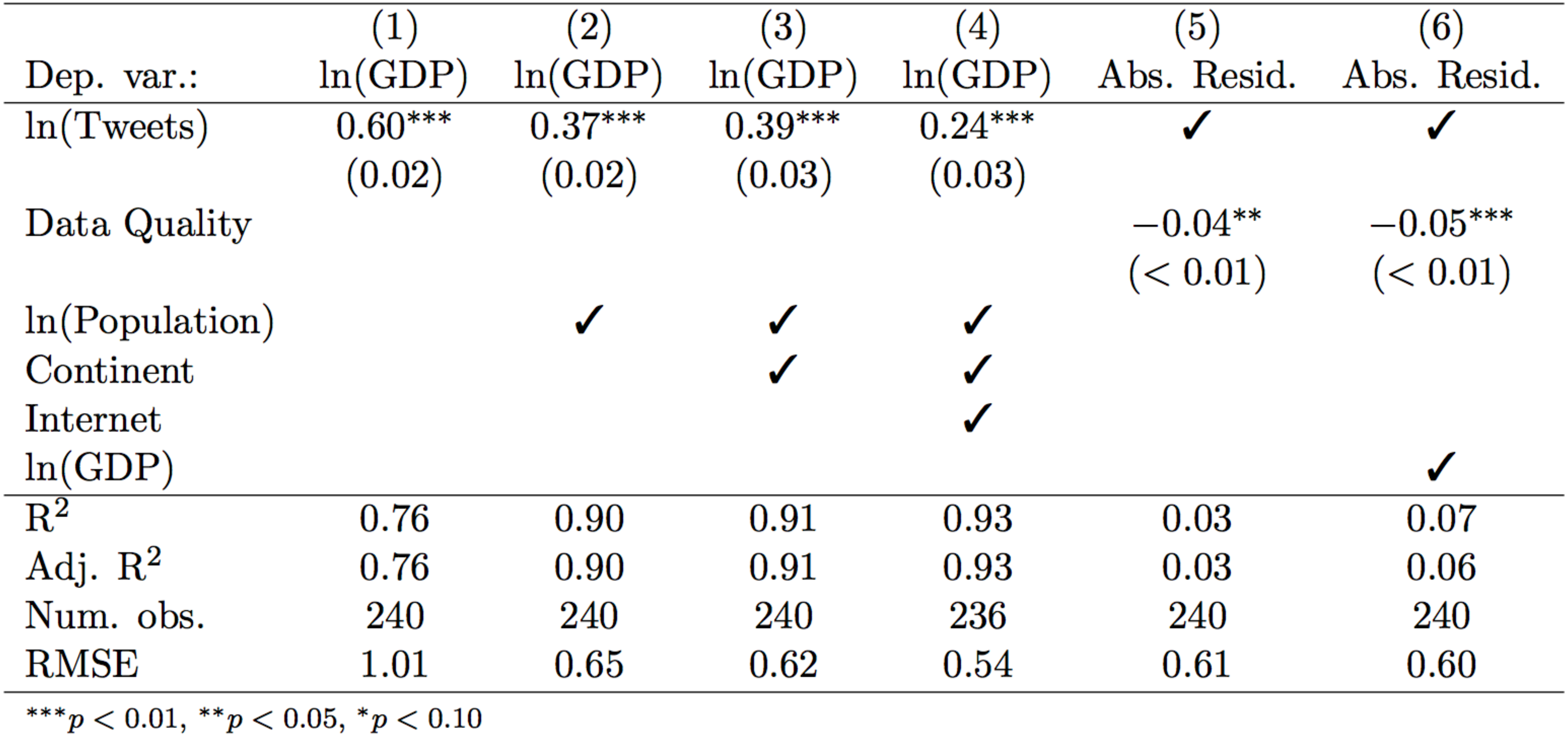

Official ln(GDP)
Estimated ln(GDP) using Tweets
30
27
24
21
21
24
27
30
United States of A
Japan
Germany
France
Brazil
United Kingdom
Italy
Canada
Russia
Spain
Mexico
Korea
Indonesia
Netherlands
Switzerland
Saudi Arabia
Sweden
Argentina
Africa
United Arab Emirates
Malaysia
Hong Kong
Philippines
Czech Republic
Egypt
Iraq
Qatar
New Zealand
Kazakhstan
Bangladesh
Vietnam
Morocco
Angola
Puerto Rico
Ecuador
Slovakia
Cuba
Sri Lanka
Luxembourg
Azerbaijan
Macau S.A.R.
United Republic of Tanzania
Turkmenistan
Cameroon
Latvia
Bahrain
Trinidad
Equatorial Guinea
Gabon
Botswana
Bosnia and Herzegovina
Democratic
Mozambique
Macedonia
South
Madagascar
Moldova
Kyrgyzstan
Montenegro
Sierra Leone
Andorra
Guyana
Maldives
Burundi
Cape Verde
Belize
Liberia
Central African Republic
Seychelles
Antigua and Barbuda
Gambia
Solomon Islands
Saint Kitts and Nevis
Saint Vincent and the Grenadines
Comoros
Dominica
Tonga
Sao Tome and Principe
Marshall Islands
Kiribati

Data Quality: Corroboration

- Many countries, particularly developing countries, have **inaccurate GDP estimates.**
- I may be trying to estimate the GDP reported by countries and **not necessarily the *true* GDP.**
- **I want to see if GDP estimates are more accurate for countries which are considered to have more reliable GDP data.**
- World Bank Data Quality score: assesses capacity of country's statistical office. Score is a simple average of three area scores on a scale of 0-100, where higher values indicate higher quality data.
    1. methodology
    2. data sources
    3. periodicity and timeliness.

Data Quality: Model

$$|Residuals|_{i,t} = \beta_0 + \beta_1 DataQuality_{i,t} + \beta_2 \ln Tweets_{i,t} + \beta_3 \ln GDP_{i,t} + \varepsilon_{i,t}$$



	(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
RMSE	1.01	0.65	0.62	0.54	0.61	0.60

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

Conclusion

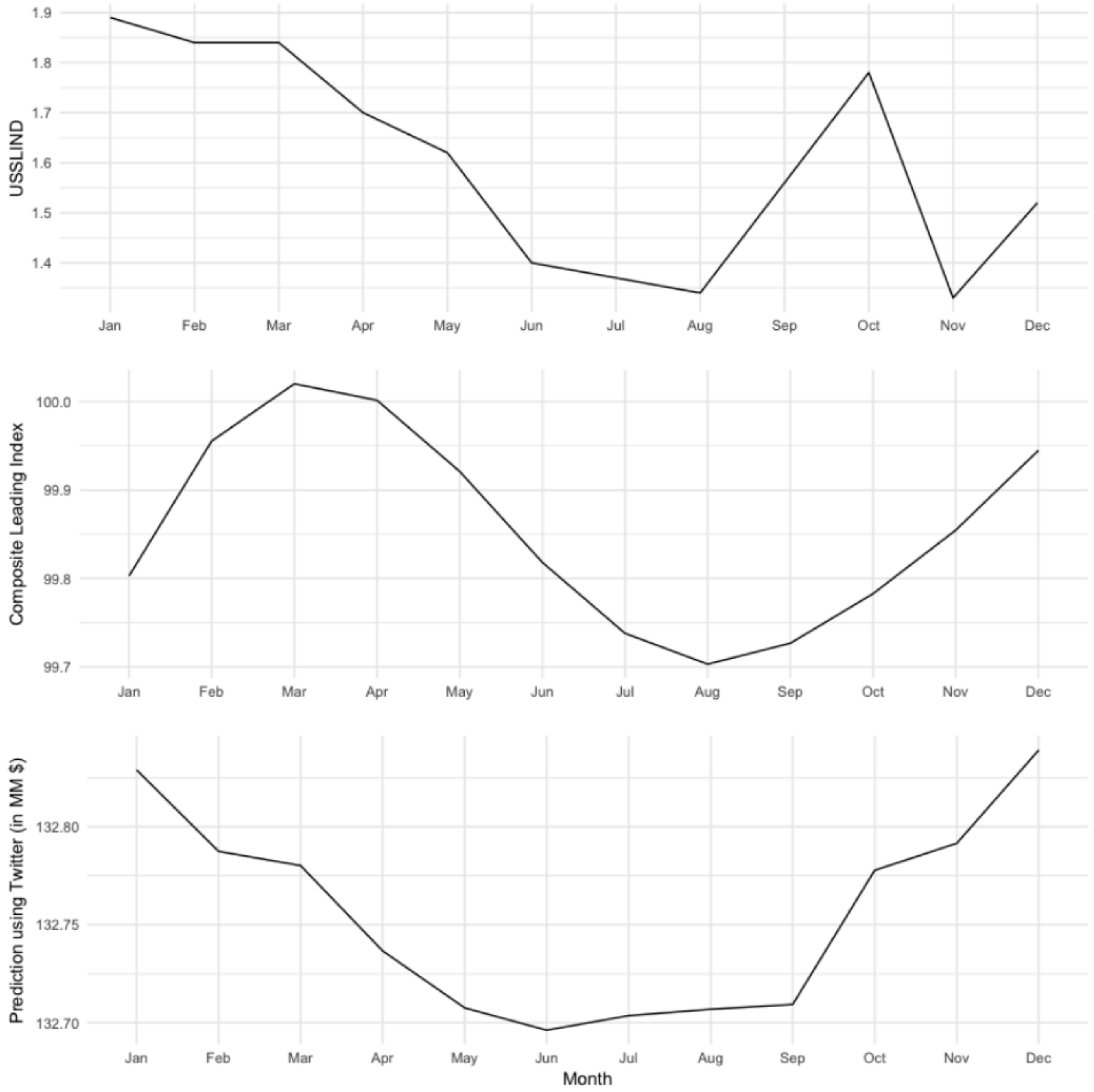
- Twitter data can be used as a proxy for estimating GDP at the country level, and my **preferred model can explain 94 percent of the variation in GDP.**

- Residuals from model are negatively correlated to data quality index: suggests that **GDP estimates are more accurate for countries which are considered to have more reliable GDP data.**

- Taken together, this suggests that **social media data could be used as a complement to survey data to increase the accuracy of GDP estimates.**

Other uses: short-time interval

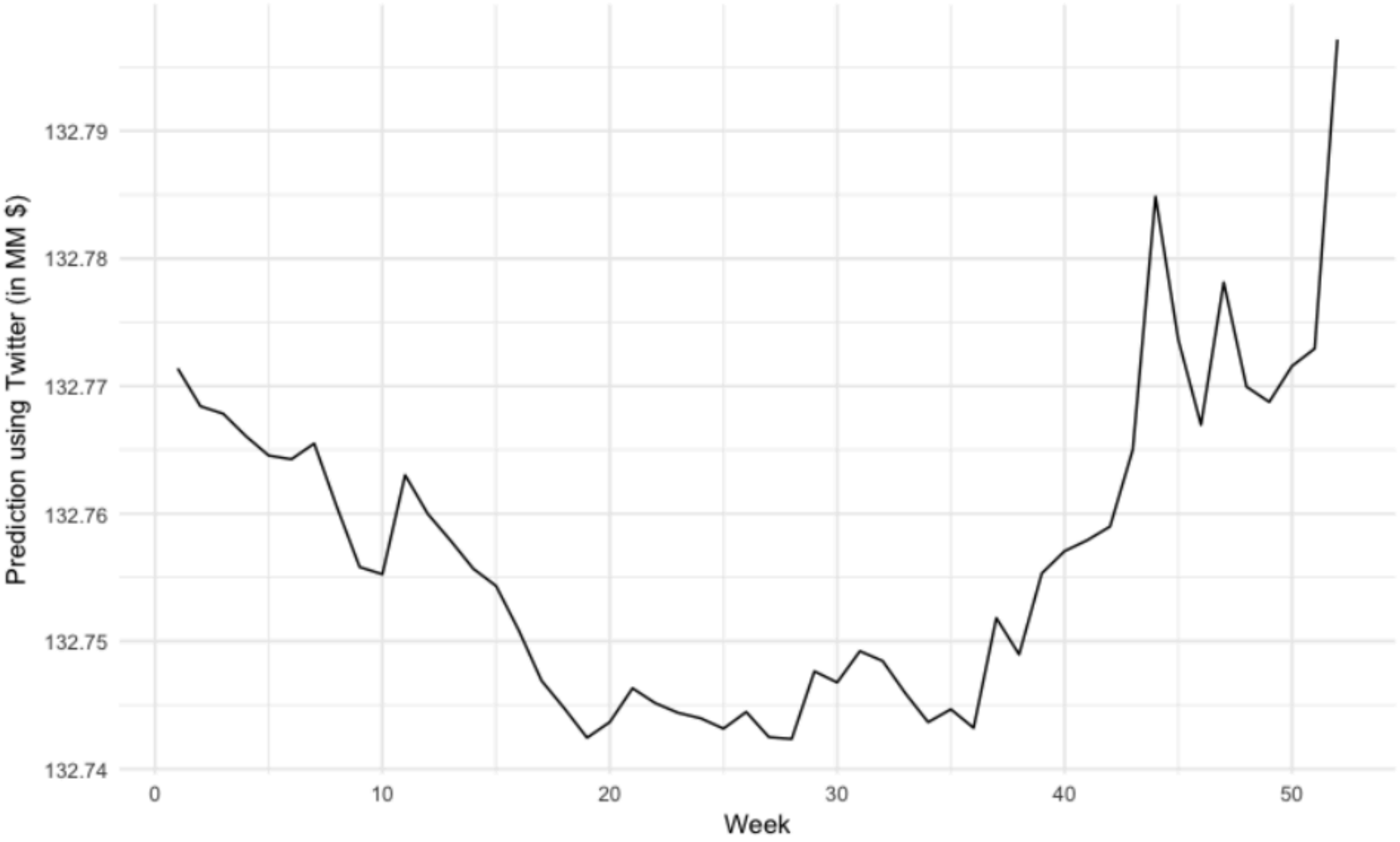
Figure 10: Monthly Indicators of USA Economic Activity for 2012



Top: Leading Index for the US (USSLIND) by the Federal Reserve bank of Philadelphia.
Middle: Composite leading indicator by the OECD.
Bottom: GDP estimates using Twitter data.

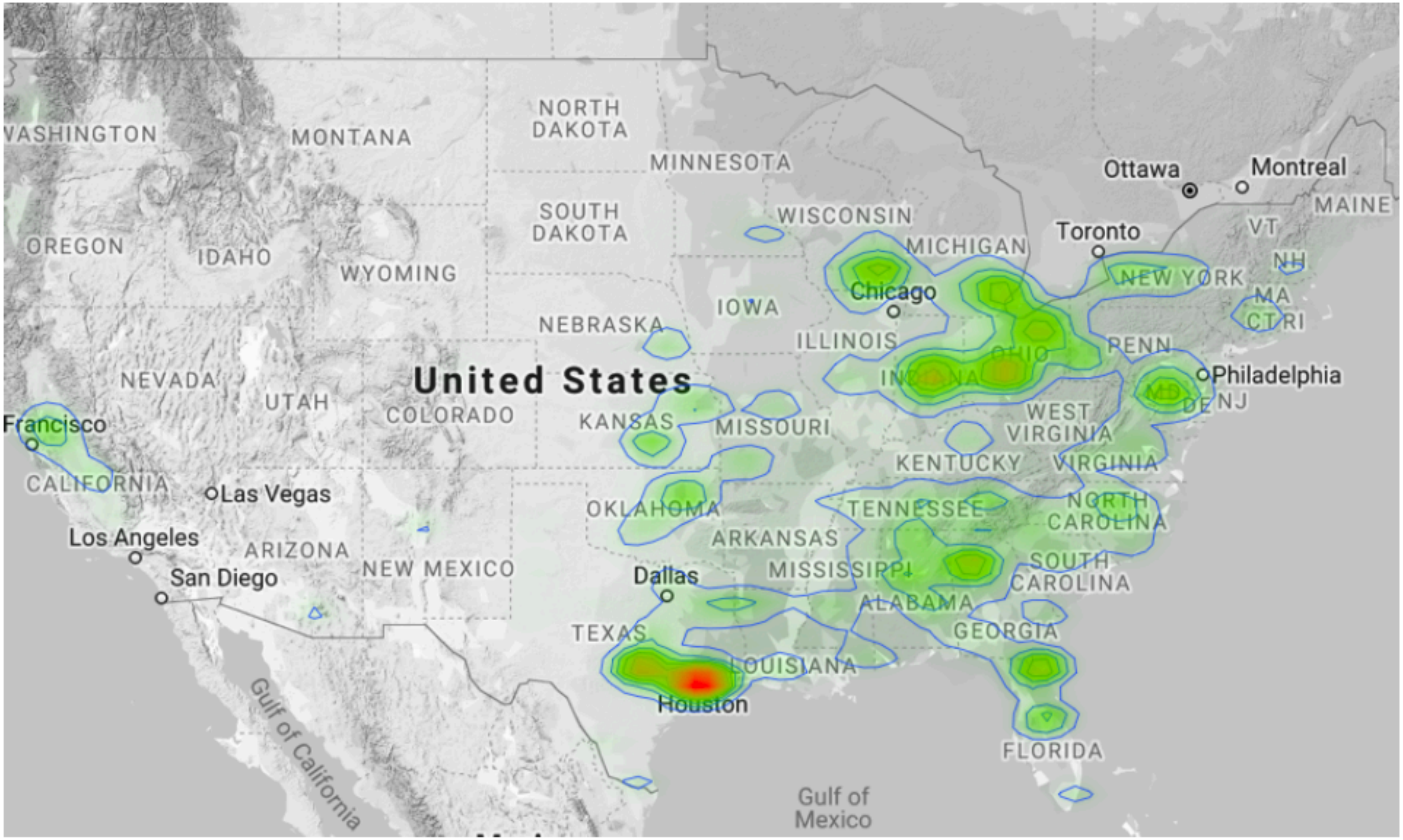
Other uses: short-time interval

Figure 11: Weekly Estimate of USA Economic Activity for 2012



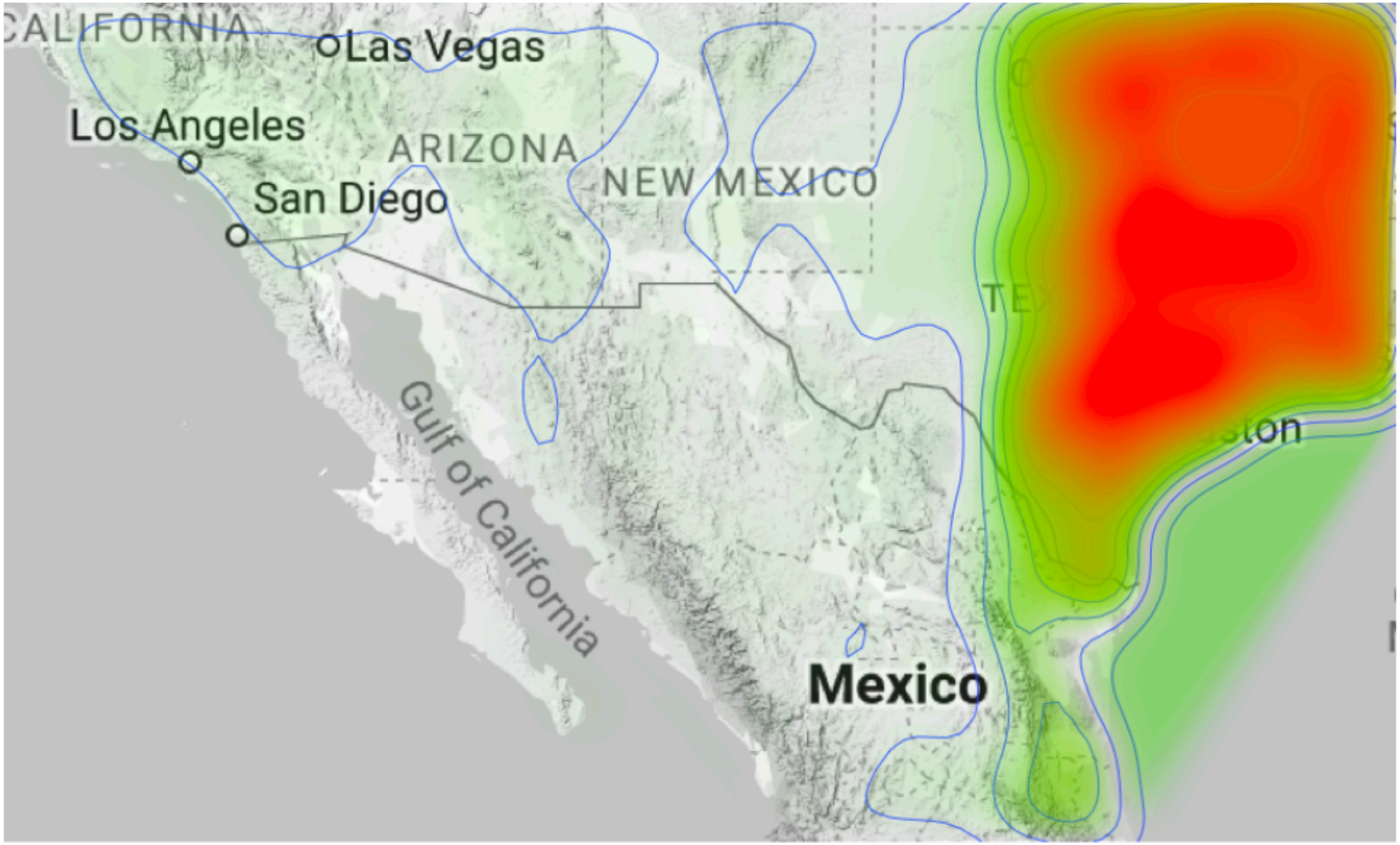
Other uses: geographic granularity

Figure 13: Density Map for Economic Activity in the US in 2012



Other uses: geographic granularity

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



CARMA 2018

Thank you!

Appendix: GDP estimates by income group

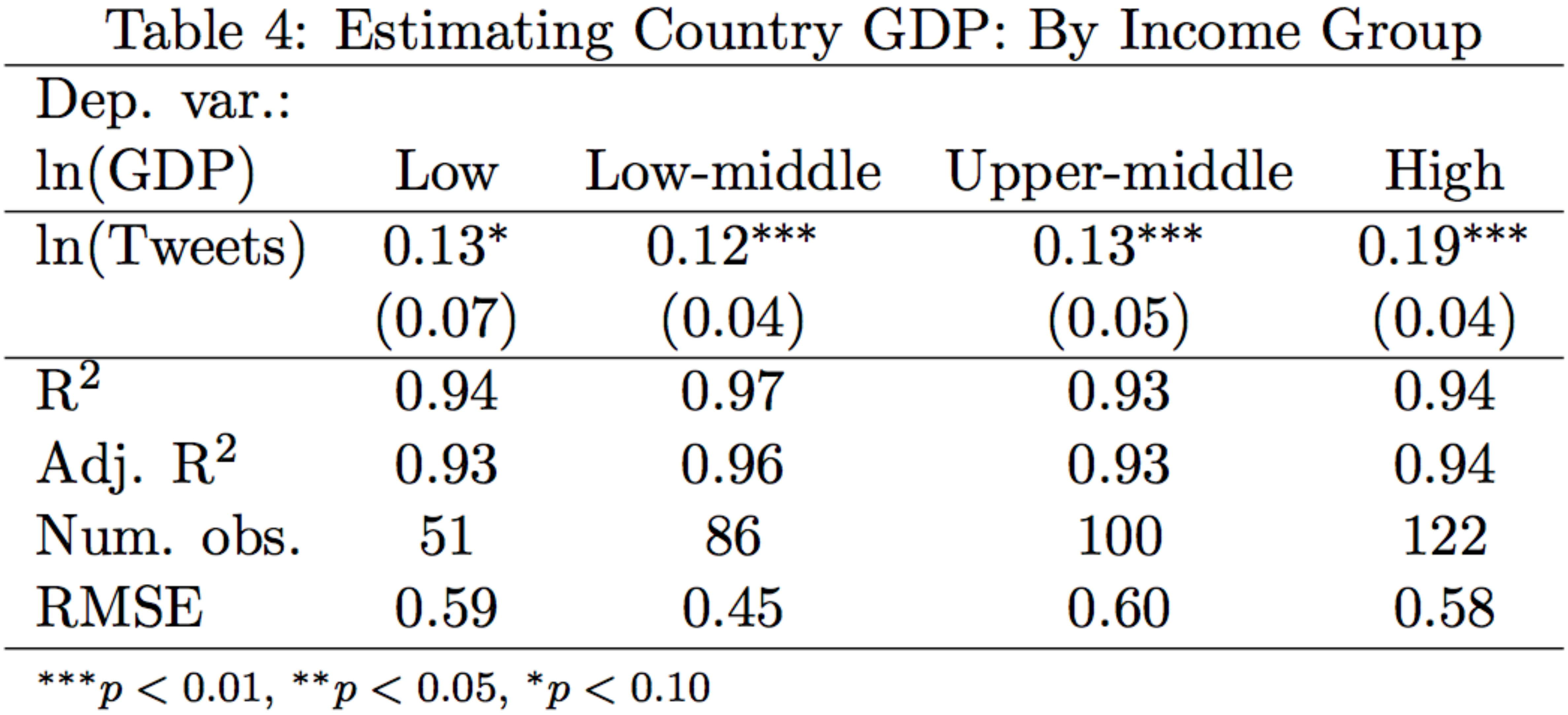


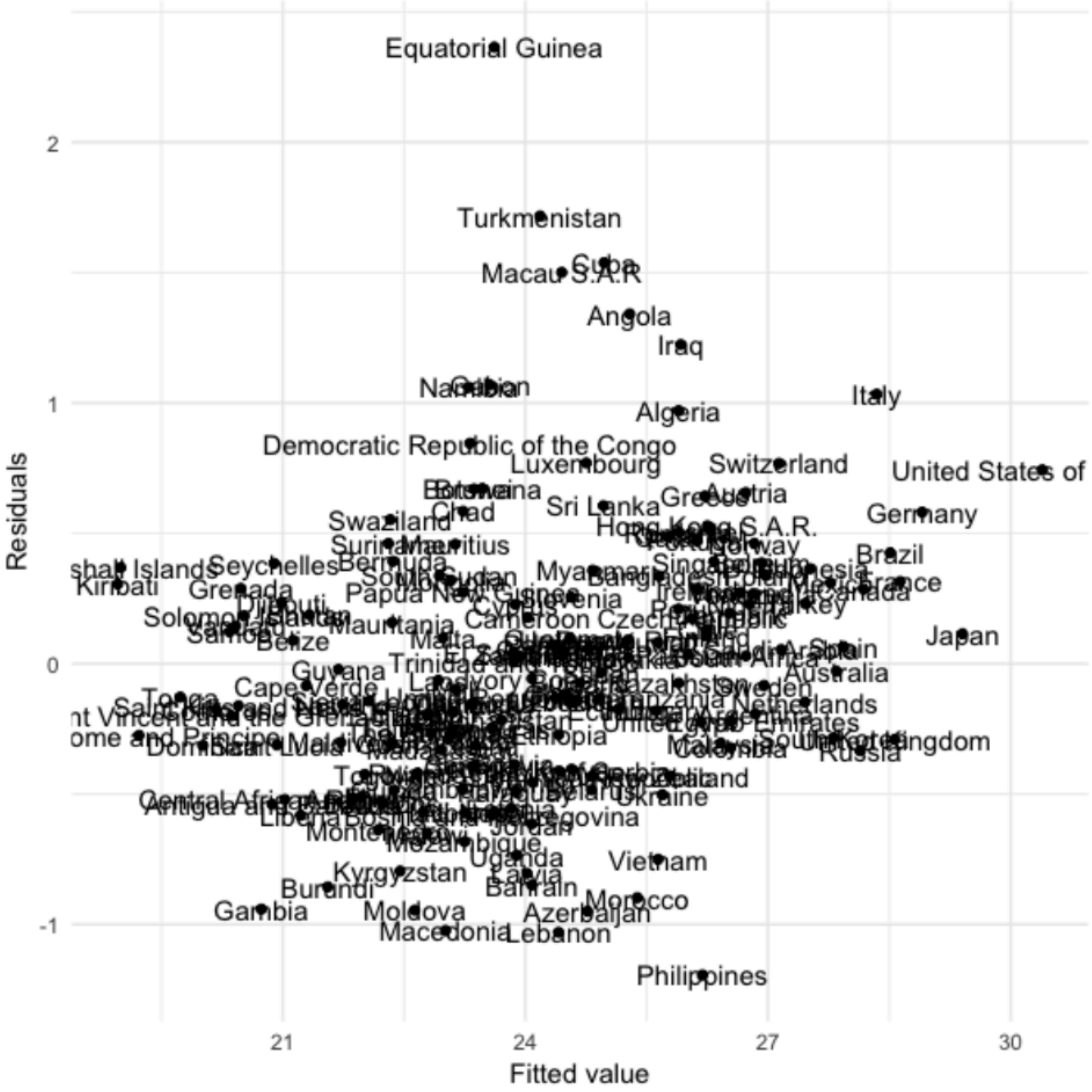
Table 4: Estimating Country GDP: By Income Group

Dep. var.: ln(GDP)	Low	Low-middle	Upper-middle	High
ln(Tweets)	0.13*	0.12***	0.13***	0.19***
	(0.07)	(0.04)	(0.05)	(0.04)
R^2	0.94	0.97	0.93	0.94
Adj. R^2	0.93	0.96	0.93	0.94
Num. obs.	51	86	100	122
RMSE	0.59	0.45	0.60	0.58

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

Appendix: Residuals of GDP Estimates

Figure 7: Residual and Fitted Value for 2013 GDP Estimates


Residuals
Fitted value
Equatorial Guinea
Turkmenistan
Cuba
Macau S.A.R
Angola
Iraq
Namibia
Gabon
Italy
Algeria
Democratic Republic of the Congo
Luxembourg
Switzerland
United States of
Botswana
Greece
Austria
Sri Lanka
Chad
Germany
Swaziland
Hong Kong S.A.R.
Norway
Suriname
Mauritius
Brazil
shall Islands
Seychelles
Bermuda
Myanmar
Singapore
Bangladesh
Kiribati
Grenada
Sudan
Papua New Guinea
Slovenia
Cyprus
Ireland
Canada
France
Turkey
Djibouti
Solomon Islands
Samoa
Mauritania
Cameroon
Czech Republic
Belize
Malta
Japan
Spain
South Africa
Guyana
Trinidad and Tobago
Australia
Cape Verde
Tonga
Kazakhstan
Sweden
Netherlands
nt Vincent and the Grenadines
United Arab Emirates
Ethiopia
ome and Principe
Dominica
Saint Lucia
Malaysia
Colombia
United Kingdom
Russia
Serbia
Finland
Ukraine
Central African Republic
Antigua and Barbuda
Liberia
Bosnia and Herzegovina
Jordan
Montenegro
Mozambique
Uganda
Vietnam
Kyrgyzstan
Latvia
Burundi
Bahrain
Morocco
Gambia
Moldova
Azerbaijan
Macedonia
Lebanon
Philippines