Who tweets (and where) ?
Social, political and environmental determinants of Twitter use in France

Data Science in the Alps Workshop
Grenoble, March 20th 2018

Gilles Bastin[1] Etienne Dublé[2] Sophie Kuegler[3]

[1]Sciences Po Grenoble / UMR Pacte (gilles.bastin@iepg.fr)

[2]LIG (etienne.duble@imag.fr)

[3]UMR Pacte (sophie.kuegler@umrpacte.fr)

1. The Problem
2. Why it Matters
3. Understanding the social, political and environmental determinants of Twitter use in France using geolocalised tweets
4. What's Next ?

The Problem

The promised land and the sociologist's nightmare

What we know (the promised land)

- Total Number of Monthly Active Twitter Users (worldwide - Jan. 2018) : 330 M
- Total Number of Tweets sent per Day(worldwide - Dec. 2017) : 500 M
- Percentage of Twitter users who tweet on Mobile : 80%
- Number of Monthly Active users in France : 21.8 M

What we know (the promised land)

Le Top 30 des Marques*

Rang	Marques B	Visiteurs uniques par mois
1	Google	34 517 000
2	Facebook	32 673 000
3	YouTube	24 779 000
4	Twitter.com	12 140 000
5	Apple	12 021 000
6	Orange	11 222 000
7	Leboncoin.fr	11 029 000
8	Amazon	10 383 000
9	Instagram	9 786 000
10	LinkedIn	9 616 000
11	PagesJaunes	8 936 000
12	Wikipedia	8 713 000
13	SFR	8 297 000
14	Snapchat	8 025 000
15	King	7 620 000
16	Tele Loisirs	7 248 000
17	Le Monde	7 074 000
18	AccuWeather.com	6 839 000
19	Shazam	6 817 000
20	France Televisions	6 596 000
21	Waze	6 087 000
22	L Equipe	5 968 000
23	20minutes.fr	5 852 000
24	Le Figaro	5 782 000
25	AlloCine	5 632 000
26	Dailymotion	5 629 000
27	WhatsApp	5 533 000
28	Yahoo	5 504 000
29	Skype	5 463 000
30	vente-privee	5 454 000

Figure 2: Twitter monthly users on mobile devices in France - March 2016

What we know (the sociologist's nigtmare)

- 44% of Twitter users never sent a Tweet. Only 8% of users have sent more than 50 tweets.
- Some biases are known (Twitter data & Mediametrie 2016) :

1. A gender bias
2. Young users are over-represented
3. 34 % of users are CSP+ (29% of the web users / 25% of the population)

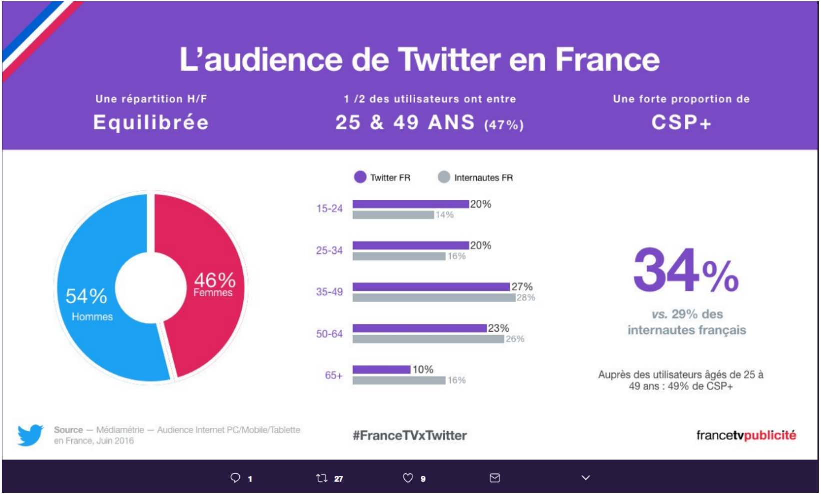

Figure 3: Twitter users in France : some well known biases

The users attributes inference literature (1)

- Profile information, tweeting behavior, linguistic content of tweets, social network information (RT)
- Used to infer gender (Rao et. al., 2010, Liu & Ruths, 2013), age (Schler et. al., 2006 ; Al Zamal et. al., 2011), occupation and social class (Sloan et. al., 2014 ; Preotiuc-Pietro et. al., 2015 ; Mac Kim et. al., 2016), location (Jones et. al., 2007), political orientation (Thomas et. al., 2006 ; Rao et. al., 2010), ethnicity (Pennacchiotti & Popescu, 2011 ; Rao et. al., 2011)
- Supervised Machine Learning

The users attributes inference literature (2)

- Using twitter accounts lists to infer profession (Ke et. al., 2016)
- Using external data such as websites visitors demographics (Goel et. al., 2012 ; Culotta et. al., 2015)
- Using geotagged tweet to retrieve localized demographics from census data : US County data (Mislove et. al., 2011)

Why it Matters

Why it Matters

- Twitter based research has to be aware (and control) biases at the user level : tweets are sent by a very biased part of the overall population
- It has also to know more about tweeting behavior as a social behavior : people do not tweet randomly during the day or wherever they are
- Mapping tweets geographically can help us better understand issues such as how connected the virtual public sphere is to actual physical environments (equipments, urban segregation, political participation. . .)

Understanding the social, political and environmental determinants of Twitter use in France using geolocalised tweets

The Data

- 32.8 M Tweets sent from France between 2014 and 2017 with GPS geolocalisation
- Every Tweet was attributed to the IRIS zone it was sent from (+/- 2.500 inhabitants)
- Census (and other) data were collected to describe every IRIS
- A dataset with 47.484 IRIS counting at least one tweet between 2014 and 2017
- Some information being only available at the town level (e.g. political participation) another dataset was created with 33.881 towns (most of them being small/very small towns that count only one IRIS)

Geotagging : a bias ?

- The share of users who enable location services (at least once) = 41 % of worldwide users
- the share of tweets with geographic information = 2.5 % of tweets
- The share of geotagged tweets (latitude / longitude) has been estimated at 0.85 % of all tweets (Sloan et al. 2013)
- The share of users who ever geotagged a tweet at 3.1% (Sloan & Morgan, 2015)
- no clear sociological bias among those users (gender, Age, Class)
- but a linguistic/national bias (interface language matters : 8.8 % of Turkish users geotagged tweets ; 2.6 % of French ones ans 0.3 % of Korean ones)
- Changes in the Twitter settings have reduced the number of geotagged tweets (users have to opt in)
- Using the API and Bounding Boxes limitations is very efficient to gather all geotagged tweets in France (2.5%*3% = 0.075%)

The geographical structure of Tweeting

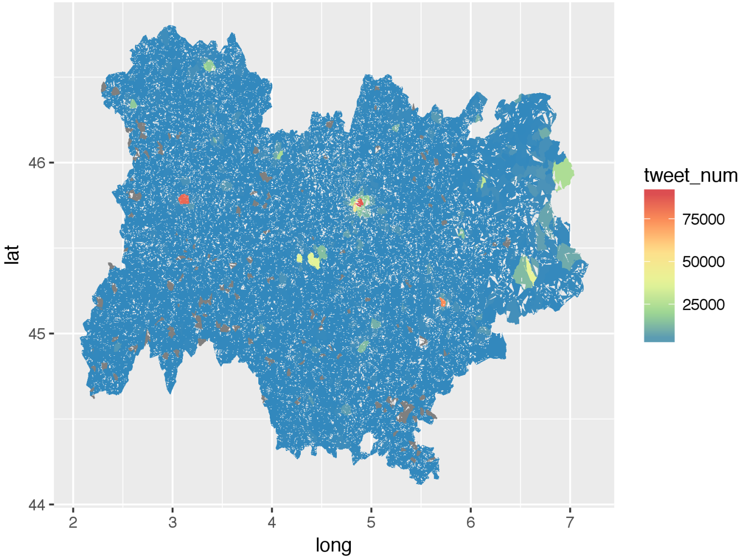

Figure 4:

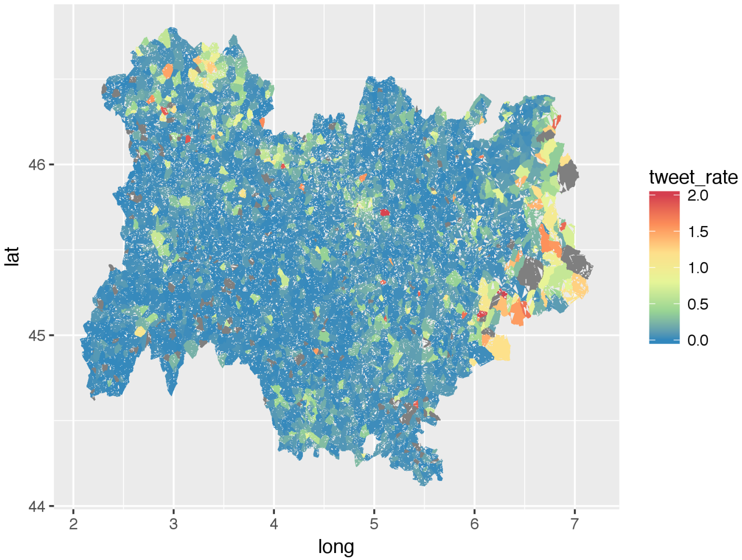

Figure 5:

Modeling tweeting behavior (1)

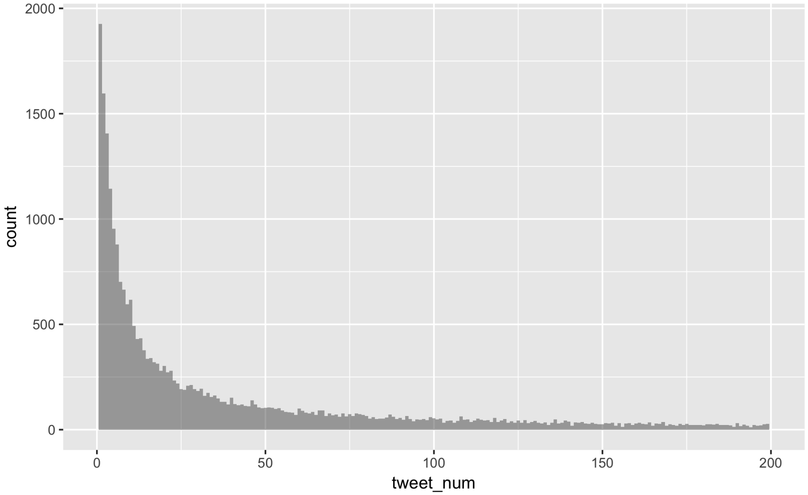

Figure 6: Distribution of tweet_num

Modeling tweeting behavior (2)

- A subset containing only towns with more than 1.000 inhabitants
- A log-linear regression model explaining log(tweet_num) by the following predictors :

1. urban environment (number of inhabitants ; density)
2. tourism (hotels/pop ; tourism information points / pop ; airports/pop ; museums/pop)
3. demographics (men/pop ; young people/pop ; foreigner/pop)
4. social classes (white collars/pop ; highly educated / pop)
5. wealth (income ; unemployment)
6. activity (high schools/pop ; shops/pop ; businesses/pop)
7. political participation (participation rate at the 2014 local elections)

Modeling tweeting behavior (3)

```
Call:
lm(formula = sqrt(tweet_num) ~ pop2014 + densite + hotel + tourisme +
    depl1 + p1529 + pH + etr + pcs3 + dipl3 + chom + tcom + musee +
    gare + aerop + lycee + entcom + gdesent + revenu + participation,
    data = commune.big)

Residuals:
    Min      1Q  Median      3Q     Max
-313.59  -10.41   -3.18    6.98  507.48

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.502e+01  8.698e+00   5.176 2.32e-07 ***
pop2014        1.562e-03  1.995e-05  78.308  < 2e-16 ***
densite        3.682e-03  1.954e-04  18.844  < 2e-16 ***
hotel         -4.274e-03  3.235e-02  -0.132 0.894897
tourisme      -2.266e-01  8.749e-02  -2.590 0.009600 **
depl1         -3.997e-02  2.297e-02  -1.740 0.081828 .
p1529          1.202e+00  9.550e-02  12.583  < 2e-16 ***
pH            -8.414e-01  1.556e-01  -5.407 6.57e-08 ***
etr            1.734e-01  7.432e-02   2.333 0.019683 *
pcs3          -2.538e-01  1.147e-01  -2.212 0.027013 *
dipl3          7.836e-02  7.937e-02   0.987 0.323584
chom           4.384e-01  7.493e-02   5.851 5.05e-09 ***
tcom           1.308e+00  4.649e-02  28.127  < 2e-16 ***
musee         -2.775e-01  2.624e-01  -1.057 0.290420
gare          -5.266e-01  1.180e-01  -4.461 8.27e-06 ***
aerop          4.042e-01  8.018e-01   0.504 0.614208
lycee          5.848e-01  2.411e-01   2.425 0.015308 *
entcom         1.921e-02  1.049e-03  18.303  < 2e-16 ***
gdesent        3.655e-03  6.796e-03   0.538 0.590679
revenu        -5.501e-04  1.477e-04  -3.725 0.000196 ***
participation -1.474e-01  2.873e-02  -5.131 2.93e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.13 on 9442 degrees of freedom
  (290 observations deleted due to missingness)
Multiple R-squared:  0.7877,    Adjusted R-squared:  0.7873
F-statistic:  1752 on 20 and 9442 DF,  p-value: < 2.2e-16
```

Figure 7:

Modeling tweeting behavior (4)

```
Call:
lm(formula = sqrt(tweet_num) ~ pop2014 + densite + hotel + tourisme +
    depl1 + p1529 + pH + etr + pcs3 + dipl3 + chom + tcom + musee +
    gare + aerop + lycee + entcom + gdesent + revenu + participation,
    data = commune.big)

Residuals:
    Min      1Q  Median      3Q     Max
-313.59  -10.41   -3.18    6.98  507.48

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.502e+01  8.698e+00   5.176 2.32e-07 ***
pop2014        1.562e-03  1.995e-05  78.308  < 2e-16 ***
densite        3.682e-03  1.954e-04  18.844  < 2e-16 ***
hotel         -4.274e-03  3.235e-02  -0.132 0.894897
tourisme      -2.266e-01  8.749e-02  -2.590 0.009600 **
depl1         -3.997e-02  2.297e-02  -1.740 0.081828 .
p1529          1.202e+00  9.550e-02  12.583  < 2e-16 ***
pH            -8.414e-01  1.556e-01  -5.407 6.57e-08 ***
etr            1.734e-01  7.432e-02   2.333 0.019683 *
pcs3          -2.538e-01  1.147e-01  -2.212 0.027013 *
dipl3          7.836e-02  7.937e-02   0.987 0.323584
chom           4.384e-01  7.493e-02   5.851 5.05e-09 ***
tcom           1.308e+00  4.649e-02  28.127  < 2e-16 ***
musee         -2.775e-01  2.624e-01  -1.057 0.290420
gare          -5.266e-01  1.180e-01  -4.461 8.27e-06 ***
aerop          4.042e-01  8.018e-01   0.504 0.614208
lycee          5.848e-01  2.411e-01   2.425 0.015308 *
entcom         1.921e-02  1.049e-03  18.303  < 2e-16 ***
gdesent        3.655e-03  6.796e-03   0.538 0.590679
revenu        -5.501e-04  1.477e-04  -3.725 0.000196 ***
participation -1.474e-01  2.873e-02  -5.131 2.93e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.13 on 9442 degrees of freedom
  (290 observations deleted due to missingness)
Multiple R-squared:  0.7877,    Adjusted R-squared:  0.7873
F-statistic:  1752 on 20 and 9442 DF,  p-value: < 2.2e-16
```

Figure 8:

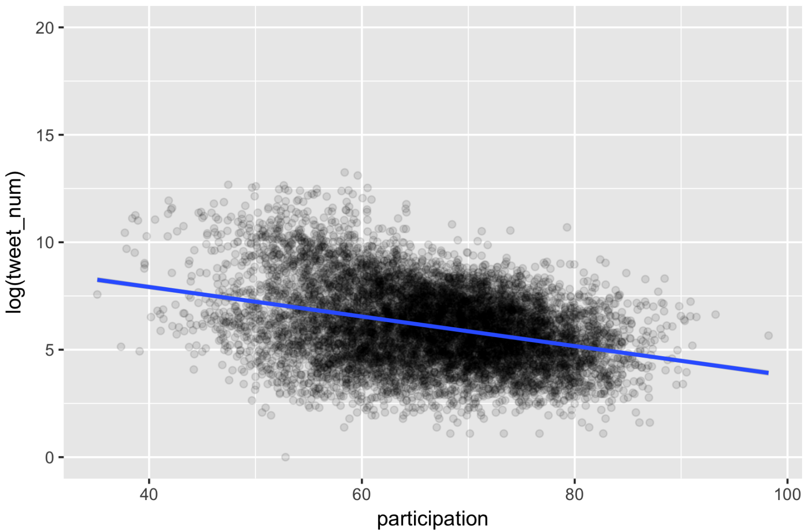

Figure 9: A surprise

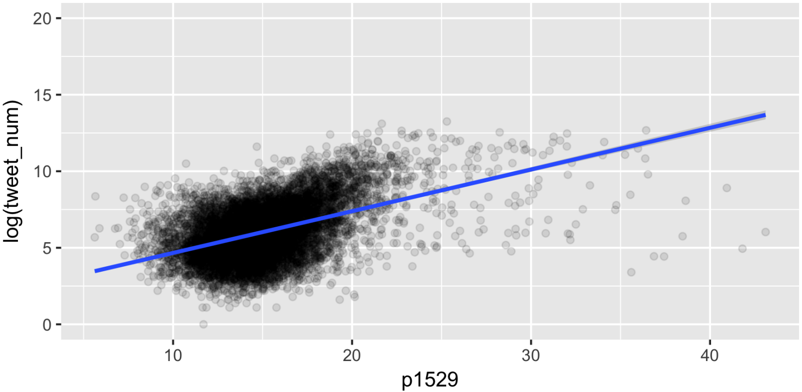

Figure 10:

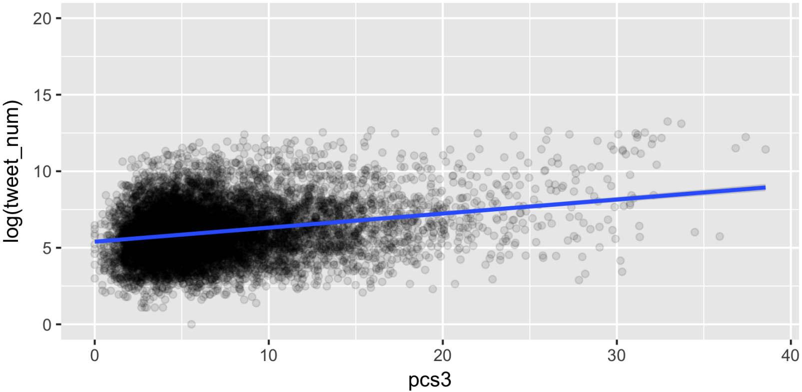

Figure 11:

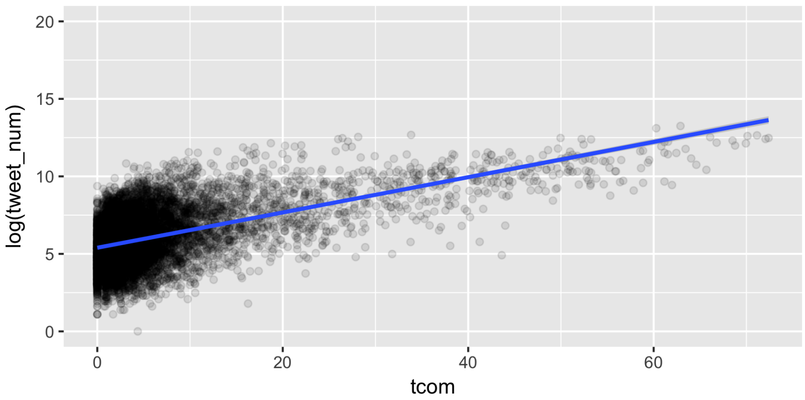

Figure 12:

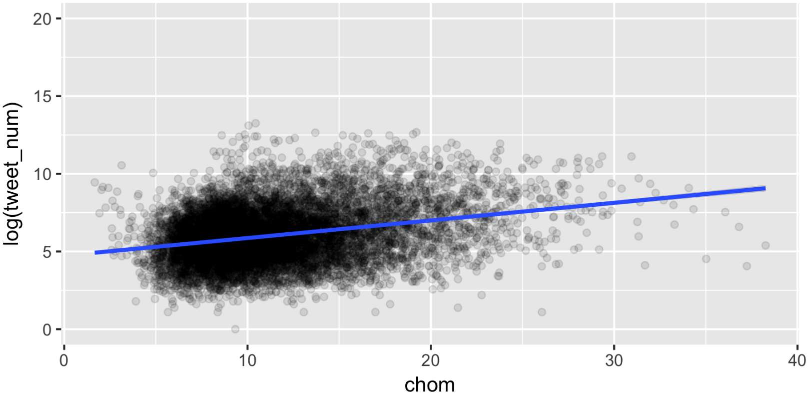

Figure 13:

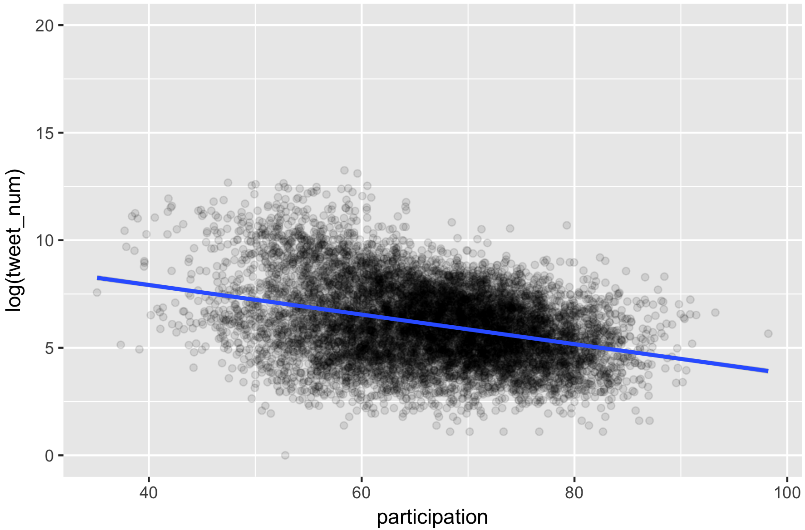

Figure 14: A surprise

What's Next ?

What's Next ?

- We need to address the ecological fallacy risks, notably identify and isolate tweets that are sent by tourists and commuting people
- We need to look at the content of the tweets to expand the scope of the questions we will be able to ask (such as : is there a relation between the content of the tweets posted in a specific town and participation levels ?)