University of Zurich[UZH]

Master Program in Biostatistics
www.math.uzh.ch/biostat

Does the Blue Bird Get the Flu?

Using Twitter for Flu Surveillance

Author: Servan Grüninger (*servan.grueninger@gmail.com*)

Supervision: Prof. Dr. Reinhard Furrer (UZH) & Prof. Dr. Marcel Salathé (EPFL)

Version of July 1, 2017

Contents

1 Introduction

We all know it and we all hate it: The common flu. What may be a mere nuisance for some, can have deadly consequences for others. Every year, between 112'000 and 275'000 patients in Switzerland seek medical care because of influenza-like symptoms, several hundred of which eventually succumb to the disease (für Gesundheit, 2016a).

However, these numbers represent just the tip of the proverbial ice-berg. Studies have shown that only a minority of the people suffering from influenza or influenza-like symptoms actually seek medical care (Goff et al., 2015). This puts traditional influenza surveillance methods, which are usually based on data from healthcare providers acting as sentinels, at a certain disadvantage, because they are more likely to catch the more severe flu cases while underestimating the overall magnitude of the flu. Also, traditional influenza surveillance systems such as "Sentinalla" in Switzerland (für Gesundheit, 2016b) or the "U.S. Outpatient Influenza-like Illness Surveillance Network" (ILINet) in the USA (for Disease Control and Prevention, 2016) only publish their reports with a lag of one to two weeks due to the time it takes to gather and aggregate the available information from the surveillance sentinels.

Hence, novel methods might be suitable to complement traditional epidemiological information in order to make influenza surveillance faster, more exhaustive and more reliable.

1.1 Complementary epidemiology

Epidemiologists have always used a wide range of information to study the transmission and propagation of disease - from simple counts of disease incidences and patient statistics up to very sophisticated disease models and biomedical parameters. So it should not come as a suprise that the advent of powerful genetic screening techniques and cheap computing power have added a lot of new weapons to the epidemiolgist's arsenal.

There exist new participatory disease surveillance system in which patients regularly fill out a short survey about their health status and possible diseases symptoms. Other surveys go even further and ask the patients to send in saliva or sputum samples for analysis so that the pathogen causing the symptoms can be identified accurately.

A different branch of complementary epidemiology uses digital data to make statistical inferences about disease transmission or disease spread. Thanks to vast amount of digital footprints each one of us leaves behind, these data sources do not have to be medical one in order to be epidemiological useful. For example, Google tried to use search queries in order to predict the spread and the intensity of the flu in certain countries. Others are using Wikipedia page views or tweets to detect a surge in influenza activity.

Tweets are an especially rich source of information due to the ease-of-access to the Twitter-API as well as due to the fact, that tweets contain a direct expression of sentiment of some sort. With millions of tweets sent out every day, the source of information is incredibly rich, so it appears straightforward to fit a model using the content of those tweets as independent variables and the official influenza data dependent variable. There is one catch, however: This approach is prone to overfitting, i.e. to picking up signals that do not indicate that the user has the flu, but that are caused by other, unrelated characteristics, which just happen to correlate with the flu season, for example. Google Flu initially fell prey to these kinds of overfitting, linking search terms such as "High School Basketball" to flu disease state - just because the basketball season coincided is in winter which happens to coincide with the flu season. The Google researchers rooted out these kind of correlations, but the main problem

remains in all approaches which use large data sets to infer flu states: How to prevent overfitting if the set of independent variables (e. g. the tweets) is in the billions, while your dependent variables (the official flu information) is in the thousands?

One approach to mitigate this problem is to restrict the tweets used to those, which clearly indicate that the user or somebody in her surroundings fell sick to the flu. If somebody tweets: "stuffy nose, headache and fever - #flu sucks!" or "nobody at work - everybody's taking a #flu leave", then these tweets indicate a clear presence of a flu infection - either within the tweeter him or herself or within the people in surrounding him or her. Hence, we can use these tweets to get an estimate over the amount of twitter users that are currently tweeting about the flu or influenza like symptoms - and thereby over the distribution of flu in the areas where the tweeters are located at .

However, even with the powerful methods from natural language processing, the identification of tweets that indicate disease state (as opposed to general awareness of the flu, for example) is not trivial. There are several very promising approaches **(citations!)**, but most of them still depend on some sort of correlation with the official data - making the again prone to overfitting.

It would be prudent then, to validate any keywords that might indicate disease state by comparing them with the true disease state of the tweeter. Since it is implausible to do so with over 300 million active twitter users (https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/) - or even with the 70 million active users in the US (https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/), we need to aim for a smaller subset.

This is what (Bodnar et al., 2014) have done. They built a flu classification model on tweets from users of which they knew the disease state up to the temporal resolution of a month. I.e. they had the possibility to build their model knowing which one of their twitter users were sick and which weren't within a certain month. Hence, they did not only correlate Tweet content with population-level, but could directly assign a twitter user's timeline with his or her disease state.

This model has a different limitation, though: Is based on a very small data set consisting of of totally 104 twitter accounts generating a total of 37'599 tweets during the study perios. Out of this sample, 35 users fell sick during the study period and generated a total of 1609 tweets in the month in which they were sick. Furthermore, all twitter users stemmed from the same state (Pennsylvania) and belonged to approximately the same socio-economic group (young students of the Pennsylvania State University). Hence, one would assume that their tweeting behaviour might be different from that of the average twitter user.

Hence, we need to test the performance of the algorithm for different cities and states and compare the results with reliable epidemiological data.

In his dissertation (Bodnar, 2015), Todd Bodnar shows that above-described algorithm performs excellently on national, state-level and county-level ili-predictions (see Figure 1). This Master thesis is (unsuccessful) attempt to reproduce these results.

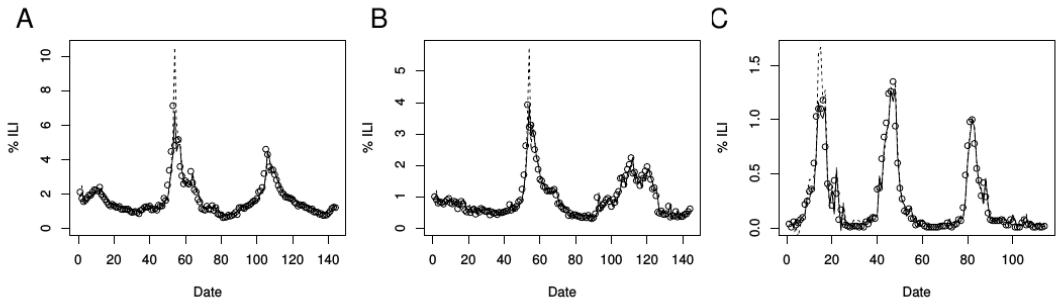


Figure 1: Comparison of Twitter's forecasting (dashed lines) and retroactive measurements (solid lines) to the CDC's reported Influneza rates (circles) for national (A), HHS Region 1 (B), and Seattle area (C) (taken from (Bodnar, 2015))

2 Description of the data set

In the following I will describe the basic characteristics of the data set used. If not mentioned otherwise, all manipulations of the data set were done using (**?**). I will cite each package in addition to Rbase I used, but since single packages are used for several functions, I will only cite when they are mentioned the first time. All functions and data sets are available at xy.

2.1 The starting point

At the beginning of my analysis, I was handed a data set with tweet ratings, subdivided into three different sets:

- **all_tweets** contains the whole set of rated tweets (2.8470397×10^9 rows)
- **one_hundred** contains the rated tweets of those users who sent at least 100 tweets (4.2611004×10^7 rows)
- **sick** contains the rated tweets of all those users who sent at least one sick tweet (4.13165×10^6 rows)

Each of the sets contains a row per tweet with the following six columns:

```
##          userID longitude latitude       time sick state
## [1,] 1000007198 -86.34844 39.63168 1424580963    0    30
## [2,] 1000007198 -86.34844 39.63168 1424580963    0    30
## [3,] 1000009051 -87.63464 24.39631 1409880397    0    56
## [4,] 1000009051 -87.63464 24.39631 1409880397    0    56
## [5,] 1000010509 -90.14008 29.86666 1394405061    0    36
## [6,] 1000010509 -90.13791 29.88957 1411750890    0    36
```

- **userID** a unique identifier of each Twitter user in the data set
- **longitude** geographical longitude in decimal degrees

- **latitude** geographical latitude in decimal degrees
- **time** UNIX timestamp marking the time when tweet was sent
- **sick** binary variable indicating whether tweet was labelled as "sick=1" or "healthy=0" by the Twitter rating algorithms
- **state** U.S. state (or District of Columbia) in which the tweet was sent

I ignored the *one_hundred* data set and only analysed the other two sets. All data sets were handled using (**?**) or (**?**).

2.2 Description of the *sick* data set

As mentioned above, the *sick_user* data set should contain all tweets from those users who had at least one of their tweets labelled as "sick" by the classifier.

First, I preprocessed and filtered the data set in order to remove all those tweets that were sent from outside the US mainland (e.g. from the northern Mexico or southern Canada) or were otherwise incorrectly geolocated (e.g. having coordinates which locate the tweeter in the middle of the ocean). To do so, I excluded all tweets lying outside a rough rectangular window with W -125° to W -66° representing the longitudinal and N 25° to N 50° representing the latitudinal expansion of the window. This way, a total of 42860 entries were removed with 4088790 entries remaining.

In the next step, I ran a custom-written function using a polygon lookup based on the coordinates of each tweet to determine the statename as well as to remove all those tweets which could not be assigned to any specific state. Of course, one might wonder why I did not just use the state code already present in the data set to assign each tweet to its respective state. There are two reasons for this: First, I did not have any reference table relating the state codes to the respective state names. Second, the polygon serves as an additionol control for the reliability of the data set. If state codes could not clearly be assigned to a specific state, this would mean that the codes could not be used as reference for future analyis. Luckily, this doesn't seem to be the case. Each state code could clearly be assigned to a specific US state with the sole exception of state code "56", which comprised all those tweets that came to lie on a state or country border, were sent from Mexico or Canada, or were geolocated to the ocean (see Figure 2a)

Most of them came to lie either a the coastline or the Canadian-US-border and the Mexican-US-border, respectively. I removed a total of 180290 tweets that were sent from either Canada or Mexico (see Figure 2b). In order to reassign the unassigned tweets from the coastline, I first changed the coordinates of the "border cases" by 0.1 degrees longitude and latitude towards the center of the US main land (e.g. if a tweet was sent from northeastern Canadian border, I added 0.1 degrees to its longitude and subtracted 0.1 degrees from its latitude before re-running the code). Those tweets that were still unassigned, received the same state name as majority of their neighbours within a 0.1x0.1 degree window. This way, a an additional 211511 tweets were removed, most of them at the coastline or from the ocean (see Figure 2b).

Afther pre-processing the *sick_user* data set was left with a with 3696989 tweets remaining. These tweets were sent by total of 2.13426×10^5, meaning that on average, each user sent 17.3221116 tweets. This is in stark contrast to the number of tweets reported in (Bodnar, 2015) (175.59 Tweets per user

over the whole study period). A slight decrease in the average tweet number should be expected due to the fact that I discarded those tweets outside the designate time or geographical window. However, a tenfold decrease in average tweet number seems suspect (the time window analysed in (Bodnar, 2015) was March 3rd 2011 to March 4th 2015, so only sligthly longer than in my case). Furthermore, the maximal number of tweets sent per user over the course of the 208 week period was 86, an incredibly low number given the fact that there twitter users out there who sent over a hundred tweets **per day** (see also Figure 3) for distribution of the number of tweets sent per user in the *sick_user* data set). Hence, I am led to believe that the *sick_user* data set does is not representative of rest of the data set, let alone the total corpus of tweets produced.

In addition, the large majority of the users within the *sick_user* data set never sent a tweet that was labelled as "sick" by the flu classifier (see Figure 4). In fact, only 2.0647×10^4 out of 2.13426×10^5 (or 9.6740791%) ever sent such a tweet.
Also, a total of 919 users *only* sent tweets that were labelled as sick - something that seems rather unlikely to happen. Finally, those 1.9728×10^4 users who sent both "sick" and "healthy" tweets had a significantly lower average tweet rate than those users who only sent "healthy" tweets (16.0096817 and 17.5342179, respectively. $p = 2.2723211 \times 10^{-83}$ using a Mann-Whitney U-Test). In fact, a Kolmogorov-Smirnov test indicates that the two subsets do not even follow the same probability distribution ($p = 0$), something that can also easily be seen in Figure 5.

Hence, it is unclear how exactly the *sick_user* data set was constructed, since it is neither a representative subset of the whole twitter data set (for that, the percentage for sick tweets is too high - see Section 2.3) nor does it exclusively contain tweets from users who had at least one of their tweets labelled as "1 = sick".

Nevertheless, I used this data set as a basis to develop a basic grasp of the data set as well as to develop functions to analyse the data set in depth and to compare it with official flu data. However, I do not report any more results based on this data set, since the exact selection criteria used for this set are unclear and hence the inferences from it are not to be trusted. All following graphs, calculations and statistics are based on the full Twitter dat set aggregated over weeks.

2.3 Description of the *all_tweets* data set

In order to analyse the complete data set with all 2.8470397×10^9, I transformed them into "big.matrix" objects using the (**?**) package, removed all tweets before 2011-03-05 and after 2015-07-11 and aggregated the remaining 2.764211×10^9 tweets with regard to states and weeks in which the tweets where sent. The cut-off date for each week corresponded to the dates the official CDC flu reports were published. All tweets within the seven day time window leading up to a specific data were assigned to said date, including the tweets sent on that date (For example, if tweet was sent on 2015-07-11, 2015-07-07 or 2015-07-05 it was assigned to 2015-07-11. However, if it was sent on 2015-07-04 it was assigned to the previous week ending on 2015-07-10).

Since there are a total a total of 208 weeks between 2011-03-05 and 2015-07-11 and a total of 50 different state labels in the original data set (48 labels for states on the US mainland, 1 label for the District of Columbia and 1 label for the tweets that could not be assigned to any of the former 49 areas), I received a data set with 10400 rows after aggregation (one for each state-week-pair). Each row has the following six columns:

```
##    week state sick  total healthy     sick_per
## 1:   23    34    1  86616   86615 1.154521e-05
## 2:  194    43    2  69629   69627 2.872366e-05
## 3:  155    16    0 140482  140482 0.000000e+00
## 4:   67    24  686 181757  181071 3.774270e-03
## 5:   40    28    0 101685  101685 0.000000e+00
## 6:   17     0    0  10142   10142 0.000000e+00
```

- **week** the week in which the aggregated tweets were sent
- **state** the state in which the tweet were sent
- **sick** total number of tweets that were labelled as "sick" in the given week and state
- **total** total number of tweets sent in the given week and state
- **healthy** total number of tweets that were not labelled as "sick" in the given week and state
- **sick_per** percentage of tweets labelled as sick among the total tweets sent in the given week and state

The complete data set consisted of a total of 2.8470397×10^9 tweets and hence was larger than the set reported by (Bodnar, 2015) which contained 2,732,174,105 tweets. This difference is simply due to the fac the tweets in my data set were collecte until July 2015, while the tweets analysed in (Bodnar, 2015) were only collected up to March 2015.

In a first step, I removed all tweets before 2011-03-05 and after 2015-07-11 as well as outside the rough geographical window around the US mainland (W -125°, W -66°, N 25° , N 50°) as described above, leaving 2.764211×10^9 tweets.

Next, I added the corresponding date to each week index and then aggregated the whole data set with regard to week and state code (i.e. calculated the number of tweets sent within a given week in a given state). In order to assign state names to state labels present in the data set, I used the label/name relationships established in the *sick_user* data set (see Subsection 2.2). Since tweets with state code "56" predominantly stemmed from the Mexico and Canada or other areas outside the U.S. mainland (see Figure 2a), I removed all corresponding state/week pairs from the aggregated data set (2.2335232×10^8 tweets in total), resulting in a data set with 10192 rows and 16 columns (see below), containing a total of 2.6147111×10^9 tweets aggregated over states and weeks.

```
##    week state sick  total healthy     sick_per            statename       date
## 1:   23    34    1  86616   86615 1.154521e-05            wisconsin 2011-08-13
## 2:  194    43    2  69629   69627 2.872366e-05             nebraska 2014-11-22
## 3:  155    16    0 140482  140482 0.000000e+00             delaware 2014-02-22
## 4:   67    24  686 181757  181071 3.774270e-03             kentucky 2012-06-16
## 5:   40    28    0 101685  101685 0.000000e+00            tennessee 2011-12-10
## 6:   17     0    0  10142   10142 0.000000e+00 district of columbia 2011-07-02
```

There were a total amount of 1.189809×10^6 tweets labelled as "sick", a number that is considerably larger than the 2.0894×10^4 tweets labelled as "sick" in the *sick_user* data set. This further

shows that that the latter does not contain the full subset of tweets labelled as "sick". Relatively speaking, 0.0455044 % of all tweets in the *all_tweets* data set were labelled as "sick" (as opposed to 9.6740791 % in the *sick_user data set*).

The total amount of users who have sent at least one tweet labelled as sick during the study period was 2.7052×10^4. Not that this is an upper estimate, since a user could be classified as "sick" more than once between 2011-03-05 and 2015-07-11. Since my analysis rests on weekly aggregated data, I would not be able to differentiate between a user who is classified as sick two times and two individual users who are classified as sick once. However, this is only a problem when assessing the total number of tweeters over the whole study period - it does not pose a problem when looking at the data categorised by weeks and/or states or at averaged data.

What is peculiar, however, is the fact that the total number of sick users found in the twitter data set is considerably lower than the number reported in (Bodnar, 2015) (182801 users labelled as sick), despite the former being an upper estimate of the total number of individual sick users. At the other hand, the average number of individual users found in the data set during the first year (2011) for the whole country is 1.7538177×10^5, while (Bodnar, 2015) only reports a total of 45086 users being active during this year. Using the information that only around 0.85% of all tweets are geotagged (Sloan and Morgan, 2015), we can estimate the total number of Twitter users ative in 2011 based on the two mentioned sample estimates, respectively. While the former sample estimate gives us an average of 2.0633149×10^7 active users in 2011, the latter estimate based on Bodnar (2015) only amounts to 5.3042353×10^6 total active users in 2011 - a number which is a far cry from the roughly 25 million monthly active twitter users officially reported by Twitter in 2011 (Twitter, 2013).

In addition, I could observe a very peculiar difference in the average tweet frequency between healthy and sick users. While healthy users sent an average of 31.4178087 tweets per week, sick users sent an average 45.9534869 tweets per week (see Figure 6), a difference that is highly significant (Mann-Whitney U-Test, $p = 7.3631306 \times 10^{-23}$). Also, the average tweet rate of all users combined (31.422503) is six times smaller than the average tweet rate per user reported in (Bodnar, 2015) (175.59). The median (32.5294692) tweet rate, however, is considerably higher than the median tweet rate reported in (Bodnar, 2015) (10).

These points give reason to believe that the data used to build the ILI models reported in (Bodnar, 2015) is not the same as the data I was analysing for this Master thesis.

3 Results

Figure 7 shows the total number of tweets sent per week relative to the total number of tweets sent in the study period. here is no obvious pattern discernible other than an increase in weekly tweets until the third quarter of 2014 when a sudden dip in tweet activity occurs. The activity pattern of the tweets labelled as "0 = healthy" is almost indiscernible from the temporal pattern of the complete data set. When looking at the weekly amount of tweets labelled as "1 = sick" one can see a different pattern: The weekly activity is fluctuating more strongly and shows clearly discernible peaks towards the end and the beginning of each year. This pattern turns out to be even more pronounced when correcting for the total amount of tweets sent per week.

A Kolmogorov-Smirnov test reveals that the weekly activity of the tweets labelled as "1 = sick" is in fact significantly different from the weekly activity of the tweets labelled as "0 = healthy" (p = 0.0264162 and p = 0 for the uncorrected and corrected weekly tweet counts). See Figure 8 for a side-by-side comparison of both the uncorrected and corrected weekly tweet activity.

Next, I looked at the total amount of tweets sent in each state. As can be seen in Figure 9a and in Figure 10a, the relative distribution per state largely follows the relative distribution of the state population. Notable exceptions are Maryland and New Jersey, which were the origin of many more tweets than expected, as well as New York, from where considerably fewer tweets originated than would be expected with regard to its population.

When comparing the relative number of tweets labelled as "0 = healthy" with those labelled as "1 = sick", we can see slight differences in the distribution, which become even more accentuated when normalising with the total number of tweets per state (Figure 11) However, the states with the most pronounced differences (District of Columbia, Montana, South Dakotam, North Carolina) are almost all states or districts, respectively, with a very low overall tweet count (North Carolina being the exception). A Chi-Squared Test for independence between the two distributions gives a p-value of 0.0800653. Repeating the calculations using number of Twitter users instead of number tweets yields similar results (Figure 9b, Figure 10b, Figure 12; p = 0.0822911)

3.1 Comparison with CDC data

In order to assess the validity of the ILI predictions provided by the flu classifier, I compared the normalised number of tweets labelled as "1 = sick" per week with the official ILI reports from the CDC on the national, regional and state level(extracted using the "cdcfluview" package (**?**)).

In a first step, I simply took the relative number of tweets labelled as "1 = sick" and plotted over time next to the official CDC ILI percentage data on the national level. In order to make them directly comparable to each other, I normalised both time series by the total sum of or relative tweets numbers and ILI percentages, respectively. Hence, the percentual values shown in Figure 13a do **not** represent weekly ILI percentage, but rather the percentual proportion of the relative number of tweets and the ILI percentages, respectively, of a given week within the whole 208 week study period (In other words: The percentages of each week add up to a 100%). Since the fluctuations in the Twitter data were very high, I plotted the data again after applying a two-week (Figure 13b) and four-week (Figure 13c) moving average smoother (using the "forecast" package (**?**)). This reduced the overall fluctuations a bit, but did not particularly improve the fit with the CDC curve. I did the same for each of the ten CDC flu surveillance regions (Figure 16). The situation improves slightly if we use the relative amount of sick **users** per week (as opposed to the relative amount of sick **tweets** per week), as can be seen from from Figures **??**. In both cases, however, the correlation between the relative ILI estimates based on Twitter data and the official CDC data were abysmal (Spearman's Rho was 0.0077319 and 0.0077319 for tweet- and user-based four-week average curves, respectively).

3.2 Comparison with regard to CDC activity levels

Next, I attempted to reduce the fluctations and increase the comparibility with the CDC data by grouping the percentual values into one of ten activity levels inspired by the CDC's same grouping

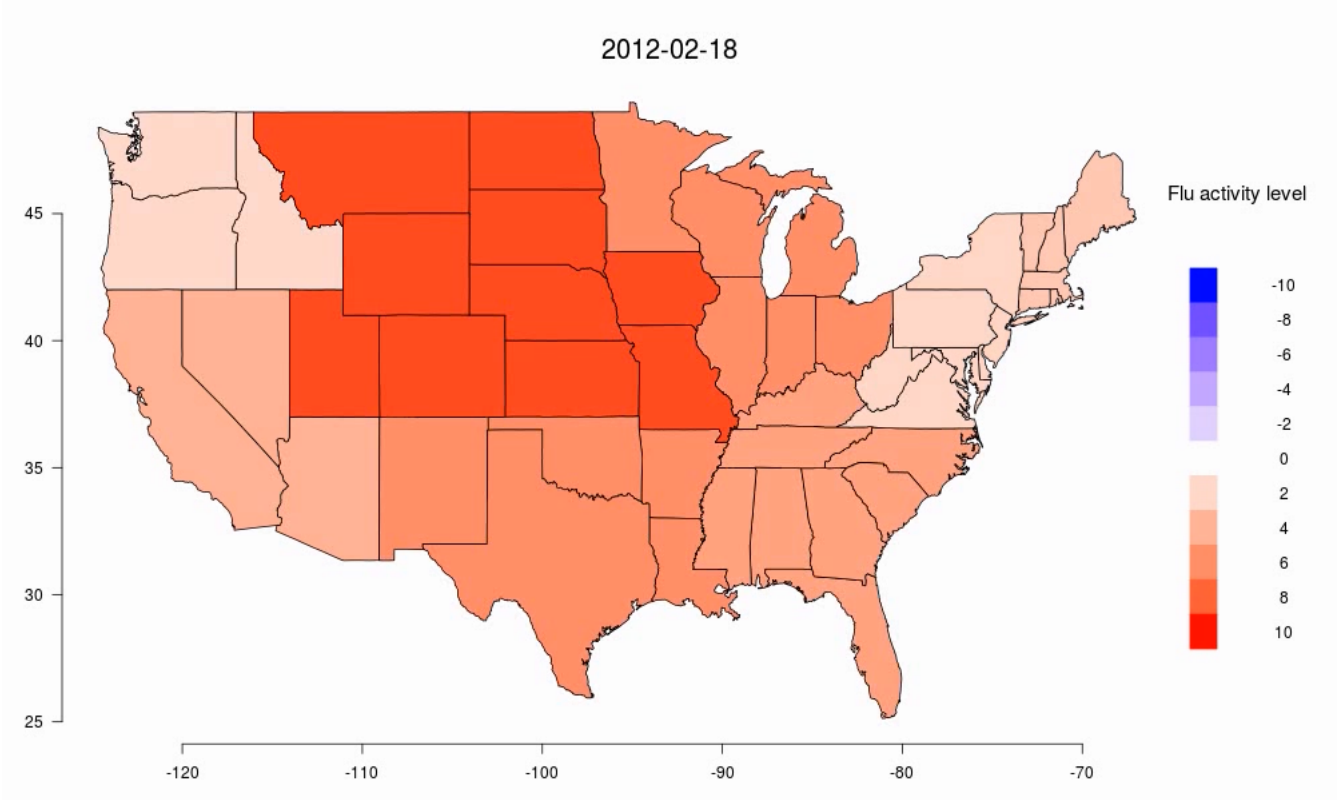
used for reporting.

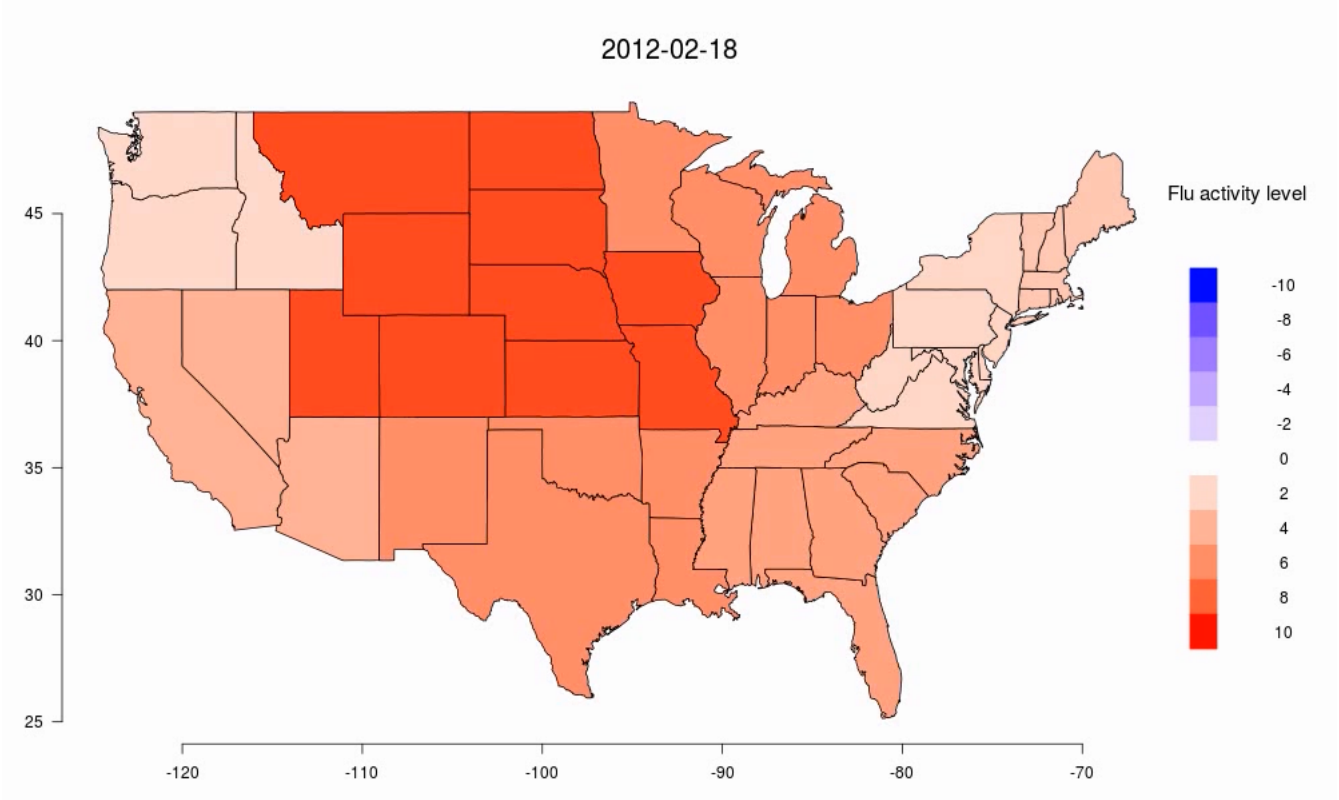
The CDC differentiates between ten different ILI activity levels which represent the deviation relative to the ILI baseline values. The activity levels compare the mean reported percent of visits due to ILI for the current week to the ILI baseline based on the number of reported ILI cases during non-influenza weeks which are defined as weeks with less than 2% of reported patient visits due to ILI (https://www.cdc.gov/flu/weekly/overview.htm) More precisely, the baseline is calculated by averaging the percentages of recorded ILI patients during non-influenza weeks for the previous three seasons and then adding two standard deviations (https://www.cdc.gov/flu/pastseasons/1314season.htm)

An activity level of 1 corresponds to values that are below the baseline, level 2 corresponds to an ILI percentage less than 1 standard deviation above the baseline, level 3 corresponds to ILI more than 1, but less than 2 standard deviations above the baseline, and so on, with an activity level of 10 corresponding to ILI 8 or more standard deviations above the baseline. (https://www.cdc.gov/flu/weekly/overview.htm)

Since a similar threshold does not exist for the Twitter data, I simply used the relative number of tweets labelled as "1 = sick" during weeks outside the flu season (June to September; seasonal flu activity can begin as early as October and continue to occur as late as May) as source to calculate yearly baseline values during off-season weeks. I then used these baseline values to calculate the weekly activity levels according to the rationale describe above. Figure 17a and Figure 18 shows the comparison on the national and regional level, respectively. I then did the same using the relative number of sick users (as opposed to sick tweets) instead (Figure 17b and Figures 19).

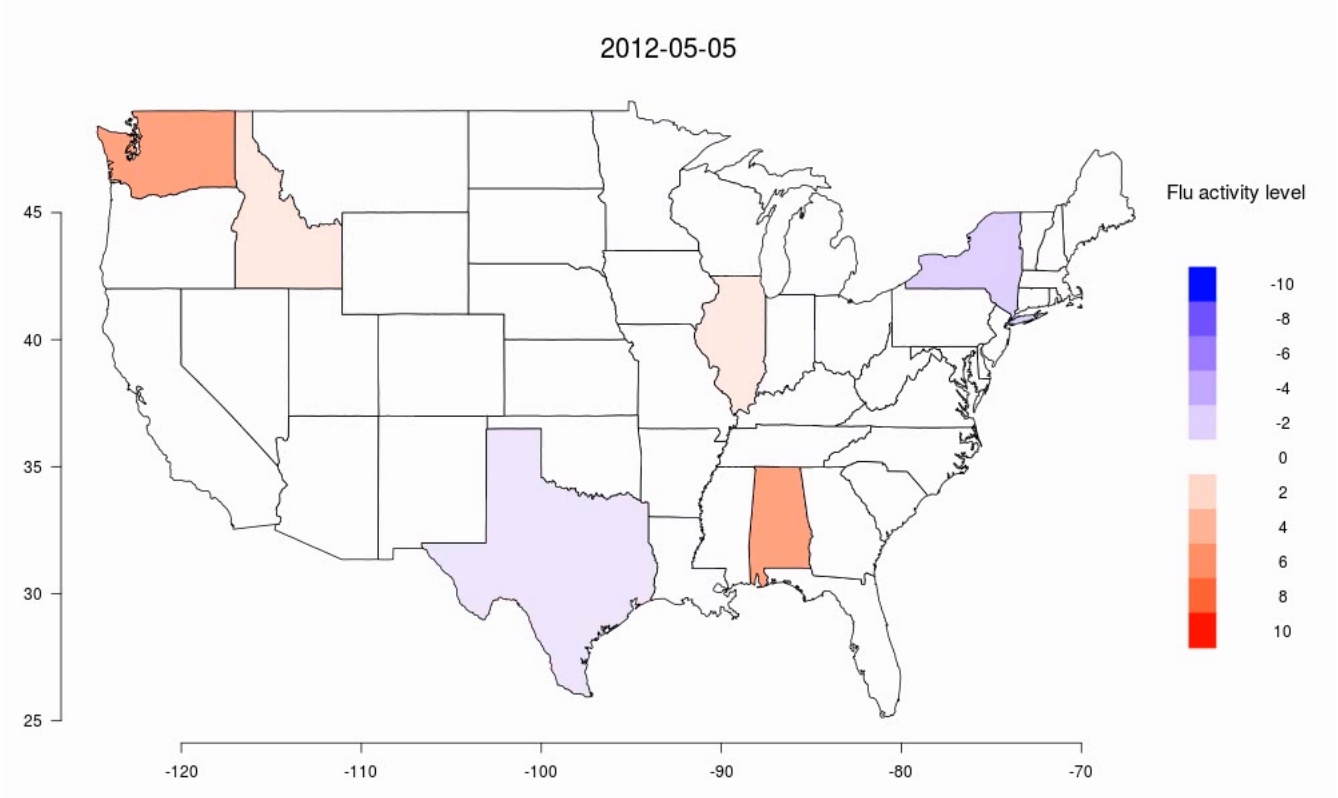
To get a better understanding of the spatio-temporal pattern of ILI activity levels, I built a small function that would take CDC ILI data as well as classified Twitter data and build a map of flu activity over time. The two videos below show the comparison of CDC and Twitter activity levels on a regional level. The first video shows the comparison of activity levels based on the relative number of tweets labelled as sick, the second video shows the comparison of activity levels based on the relative number of sick users per week. White means that CDC and Twitter activity levels are exactly the same, red means that the CDC reported higher ILI activity levels in a given state than those calculated from the Twitter data set, blue indicates the opposite. As can be seen, the Twitter ILI classifier hardly ever manages to emulate the CDC activity levels and when it does, it mainly happens during off-season weeks.

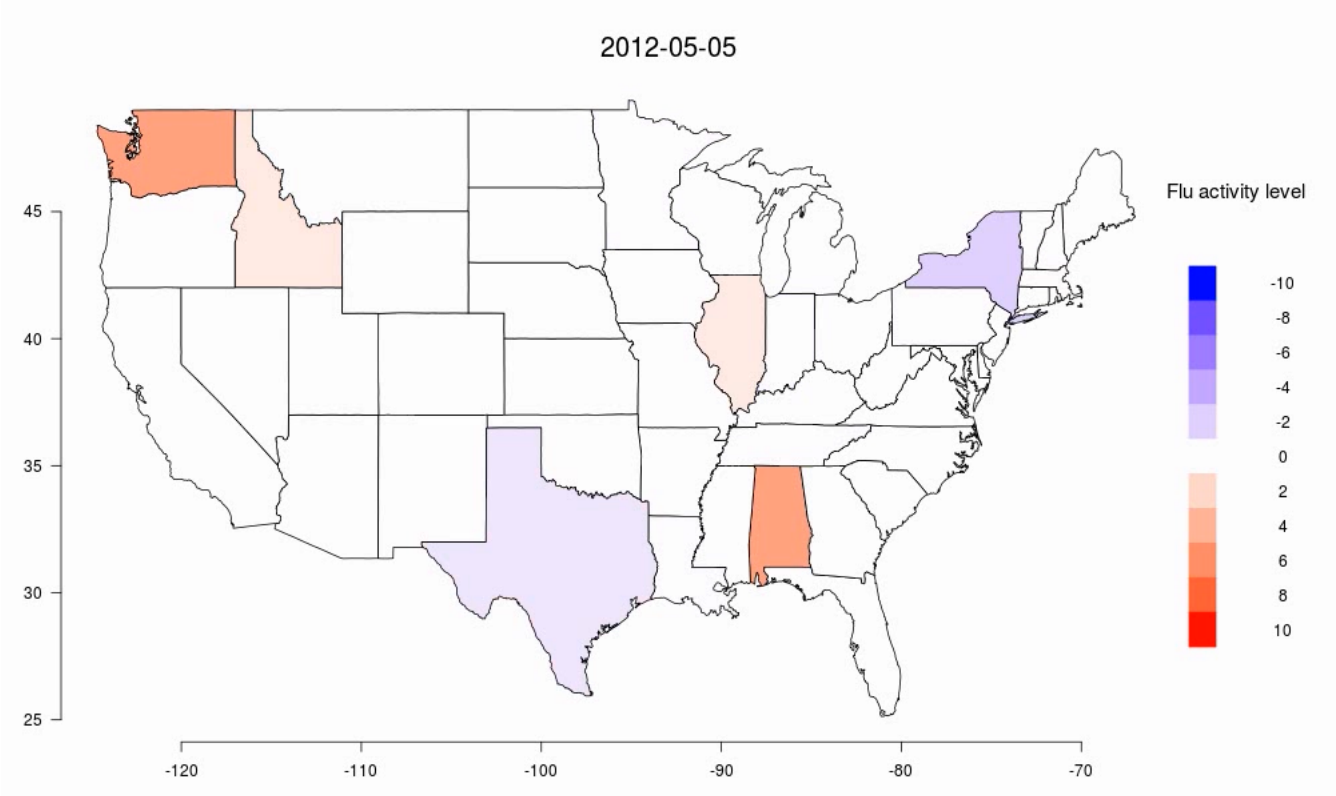

2012-02-18
Flu activity level
-10
-8
-6
-4
-2
0
2
4
6
8
10
45
40
35
30
25
-120
-110
-100
-90
-80
-70


2012-02-18
Flu activity level
-10
-8
-6
-4
-2
0
2
4
6
8
10
45
40
35
30
25
-120
-110
-100
-90
-80
-70

3.3 Comparison on the state level

In a next step, I tried to assess the performance of Twitter flu classifier by looking at state-level data. To do so, I again calculated activity levels for each state and week based on the relative amount of tweets labelled as "1 = sick" and the relative number of users classified as "1 = sick", respectively. Due to space reasons and since the fit with the CDC activity curves was worse than for the regional and national data, I did not include the individual time series to this report, but only the two videos showing the spatio-temporal differences in ILI activity levels. The first video contains the comparison with activity levels based on the relative amount of tweets labelled as "1 = sick" in each state, while the second video contains the comparison based on the relative number of users classified as "1 = sick" during as specific week and within a given state. Again, it is clear the activity levels based on the Twitter data fit the official CDC only poorly.

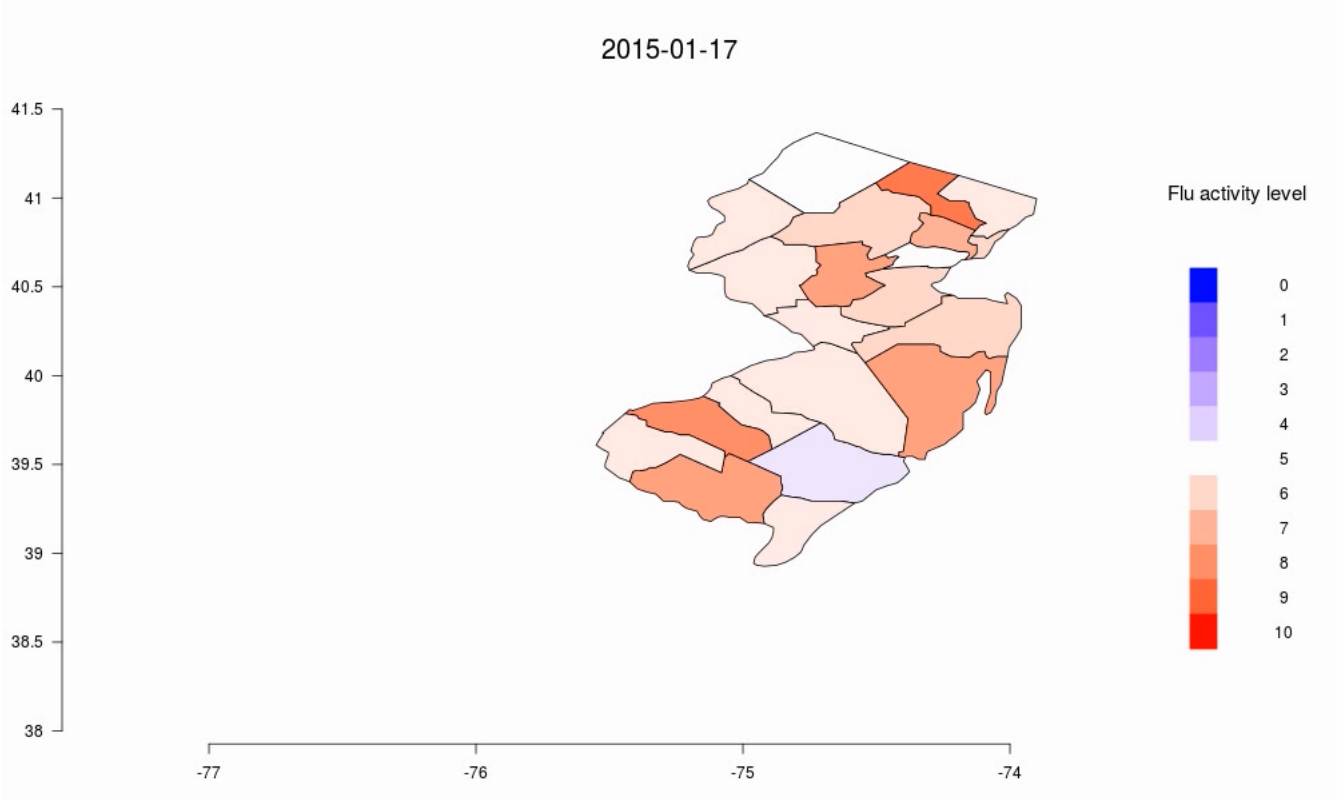


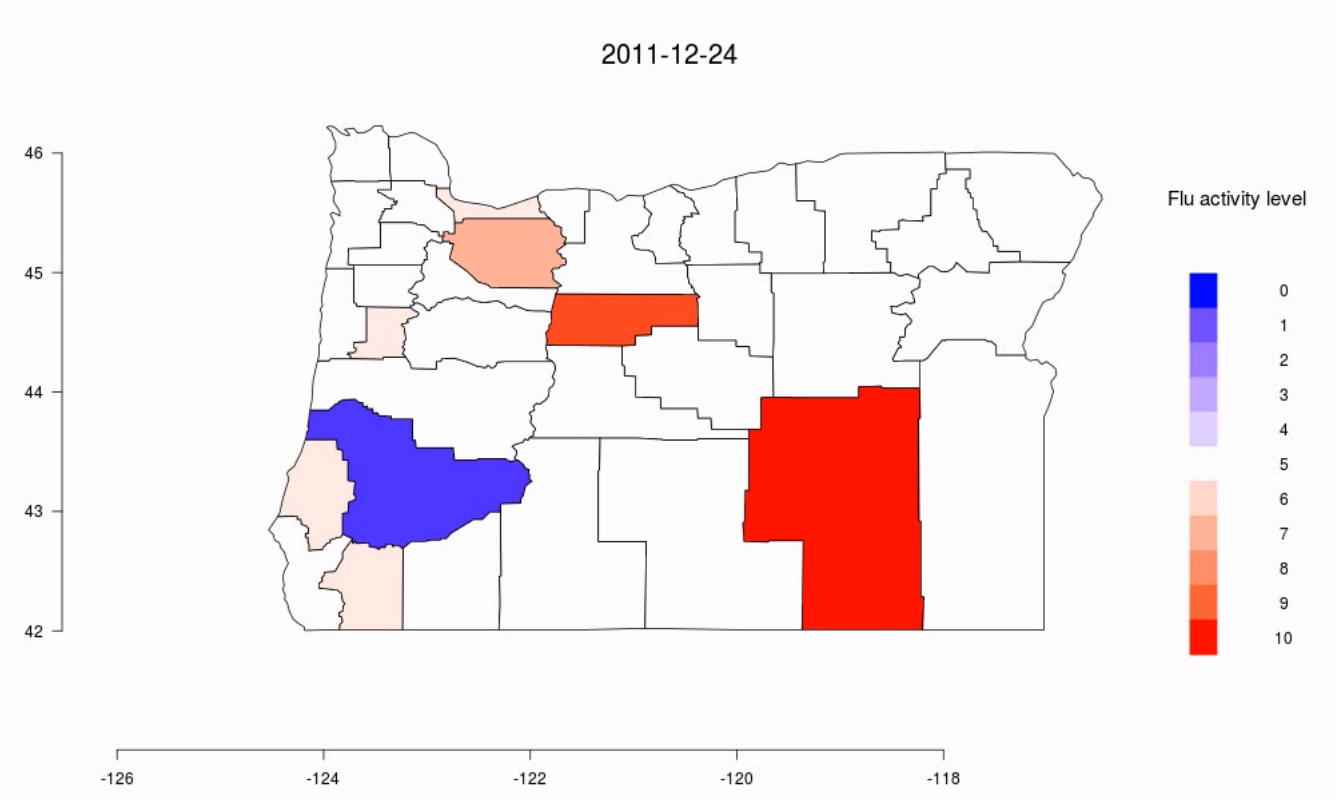


3.4 Comparison on the county level

In order to assess the performance of the classifier on the county level, I contacted 18 state health departments (Arkansas, California,Florida, Illinois, Iowa, Louisiana, Maine, Michigan, Minnesota, Mississippi, Missouri, New Jersey, New York, North Dakota, Oregon, South Dakota), asking them for their county-level or regional ILI data between 2011 and 2015. I only received the county-level data from the state of Oregon, while the state of California provided me with state-level data only (an official request for county-level data is pending). All other states did either not answer or declined my request. Luckily, the states of New Jersey and Mississippi provided (almost) complete county-level ILI data on their website for the requested time period. Since the data was only provided in pdf-format, I built a scraper for both states in order to retrieve the relevant information.

Below, you can see the spatio-temporal comparison of the performance of the Twitter flu classifier for the counties of New Jersey (top) and Oregon (below). Note, that Oregon did not provide ILI estimates for all counties (most of the northeastern counties are missing, for example).





4 Discussion

The failure to replicate the findings from (Bodnar, 2015) can have multiple reasons:

- Coding errors distoring the aggregation and analysis of the Twitter data set
- The findings from (Bodnar, 2015) are based on a different data set
- There are errors present in the Twitter classifier code
- The findings only be replicated by using the classified tweets as a starting point for more intricate models

I will address each one of these steps in the following passages and explain what I have done to addresse them during my thesis.

4.1 Errors in the aggregation and processing of the Twitter data set

This is the most obvious, but also most frequent source of errors to occur. Handling huge data sets does not only put a strain on the computer's hardware, but also on the computer user's software, since it requires a different way of handling, aggregating and manipulating data sets in order to prevent memory overflow errors or calculation that take until the end of the universe to finish.

It should not come as surprise, though, that very often in the course of this thesis I have been forced to rewrite various parts of my code or try to find a new a approach to a specific problem. It has occurred very often, too, that seemingly nonsensical code output could quickly be fixed by finding the misplaced column index or the redundant loop.

This also the reason, however, for which I am fairly confident that the results reported in this thesis do not contain any errors based on faulty code. For almost every step in the description, aggregation and analyis of the data I have usually chosen at least two different approaches (not all of them are reported in this thesis, but all the complete code source and all results are available on Github) Partially, because I usually encountered better methods along the way, partially because I wanted to have a control for my code in order to prevent any unintentional mistakes. Barring any obscure bugs in the packages I used (which seems extremely unlikely), the failure to reproduced the findings from (Bodnar, 2015) should not stem from any coding errors. Nevertheless, the code set I generated is comparable large and not at all as clean and simple as I wished it to be, thereby also increasing the probability for unwanted errors sneaking in. Hence, the in order to make this thesis as resproducible as possible and in order to facilitate any follow-up analysis, I will further clean up the code and the database structure.

4.2 The findings from (Bodnar, 2015) are based on a different data set

As written in the description section, there is ample evidence that the data set I used was not identical to the one used in (Bodnar, 2015). Basic statistical properties such as the average tweet rate, total number of sick users or total number of users were considerably different. One reason for this could be that the data set I analysed was processed by a different Twitter classifier. This is not too unlikely, since Todd Bodnar described several different flu classifiers in his thesis, of which apparently only one was able to fit the official CDC data as smoothly as shown in Figure 1. Nevertheless, personal e-mail correspondance with Todd Bodnar in fact confirmed that the data set described in his thesis and the one stored in the data base dumps of the Salathé research group were in fact the same. In this case, the only explanation would be that the data set was inadvertently

changed at some point after the end of this thesis.

4.3 There are errors present in the Twitter flu classifier

This is an idea that I entertained early on and that was in fact the original starting point of this thesis: The attempt to replicate the Twitter flu classifier analysis in a first step in order to improve in a second step.

However, in order to assess the quality of the Twitter flu classifier one would not only need access to it, but also get it up and running. A challenge that in end turned out to be too much for me.

In a first step, I tried to install the Twitte flu classifier (written in Java) based on the availabe code stored on Github. It quickly turned out, however, that the repository was incomplete, making it impossible to compile the code. After a series of exchanges with Todd Bodnar, I received additional code and packages he used to built his classifier.

After this, I encountered an additional problem, however. Since the Twitter classifier is heavily relying on an old version of the Java Amazon Web Services (AWS) API that has been updated without backwards compatibiltiy since the end of Todd Bodnar's thesis, I encountered a host of errors due to missing or deprecated functions.

Finally, after a short series of e-mail exchanges with Todd Bodnar, I was handed a compiled version of the Twitter flu classifier, allowing me to circumvent the necessity to debug the original code. Unfortunately, the jar-file encountered runtime erros when trying to analyse raw Twitter files, both on Ubuntu 16.04.2 LTS as well as on Windows 7.

Being a Java-novice I eventually abandoned the attempt to re-run the Twitter classifier. Should a re-implementation or update of the classifier be warranted for the future, one should either be proficient in Java to do so or rewrite the basic elements of it in another programming language (Python, for example)

4.4 The findings only be replicated by using the classified tweets as a starting point for more intricate models

In my eyes, this is the most likely explanation for the abysmal fit between the Twitter data and the official CDC data. In fact (Paul et al., 2015) report that autoregressive models of CDC data are very strong baseline models and in general better than twitter models alone. This shows that Twitter cannot predict CDC ILI rates on its own but should rather be used as an additional sourc of information to complement already existng estimates and reduce the error.
Also, (Paul et al., 2014) report that most disease models are using **revised** CDC data which are already corrected for mistakes. However, Twitter data might be much more useful when used with unrevised data instead. In addition, (Aramaki et al., 2011) report that Twitter data is most useful for the early detection of influenze epidemics, but could be rather rather sensitive to excessive news periods revolving around the flu.

Hence, using the Twitter data to fit a model to the official CDC data will certainly improve the fit between the Twitter flu predictions ans the official CDC ILI data. However, this would then again

open up the classifier to the perils of overfitting and would entirely defeat the purpose of building a classifier which is **independent** of population-level ILI data.

In fact, (Bodnar, 2015) fitted an SIR model to the results from the Twitter flu classifier in order to compare results. However, he reported that the model had to be refitted for every individual year based on the official CDC data (again, defeating the purpose of having a classifier independent of population level data). Also, the model fit was much lower (see Figure 20) and in no way comparable to the extraordinary fit shown in Figure 1). In the meantime, I received the code used to fit the SIR model as well as the data used to produce Figure 1. However, it as of yet unclear to me how exactly this data was produced based on the model parameters fitted in the SIR model. An answer to a follow-up e-mail to Todd Bodnar with regard to this is currently pending.

4.5 Missing parts

- Description of other methods to detect flu epidemics using digital data (for introduction and discussion)
- description of the rationale behind the Flu classifier (for introduction)
- Modelling a simple ARIMA model with lagged CDC data as external regressors (for results part - already started)
- Modelling a simple SIR model (for results part - waiting for raw data from Todd Bodnar)
- Formatting and citations
- Cleaning up code & repository (¿ any specific guidelines to follow for this?)

References

Todd Bodnar. Data science with social media for epidemiology and public health, 2015.

Bundesamt für Gesundheit. Saisonale Grippe (Influenza), December 2016a. URL https://www.bag.admin.ch/bag/de/home/themen/mensch-gesundheit/uebertragbare-krankheiten/infektionskrankheiten-a-z/grippe.html.

Jennifer Goff, Aaron Rowe, John S. Brownstein, and Rumi Chunara. Surveillance of Acute Respiratory Infections Using Community-Submitted Symptoms and Specimens for Molecular Diagnostic Testing. *PLoS Currents*, 2015. ISSN 2157-3999. doi: 10.1371/currents.outbreaks.0371243baa7f3810ba1279e30b96d3b6. URL http://currents.plos.org/outbreaks/?p=58984.

Bundesamt für Gesundheit. Saisonale Grippe - Lagebericht Schweiz, December 2016b. URL https://www.bag.admin.ch/bag/de/home/themen/mensch-gesundheit/uebertragbare-krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/saisonale-grippe---lagebericht-schweiz.html.

Centers for Disease Control and Prevention. Overview of Influenza Surveillance in the United States| Seasonal Influenza (Flu) | CDC, October 2016. URL https://www.cdc.gov/flu/weekly/overview.htm.

Todd Bodnar, Victoria C. Barclay, Nilam Ram, Conrad S. Tucker, and Marcel Salathé. On the ground validation of online diagnosis with Twitter and medical records. pages 651–656. ACM Press, 2014. ISBN 978-1-4503-2745-9. doi: 10.1145/2567948.2579272. URL http://dl.acm.org/citation.cfm?doid=2567948.2579272. bibtex: bodnar_ground_2014.

Luke Sloan and Jeffrey Morgan. Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE*, 10(11), November 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0142209. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4636345/.

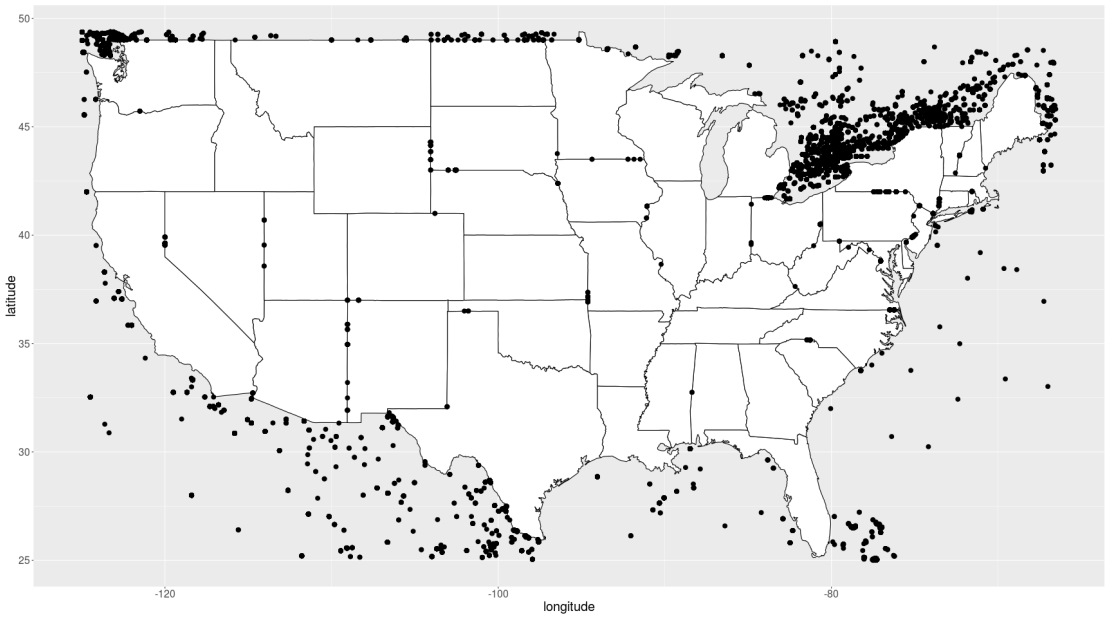
Twitter. Twitter_2013_annual_report, 2013. URL http://files.shareholder.com/downloads/AMDA-2F526X/4733143221x0x742484/A418947A-E065-4822-8BD4-00FA8EB4E795/Twitter_2013_Annual_Report_-_FINAL.pdf.

Michael Paul, Mark Dredze, David Broniatowski, and Nicholas Generous. Worldwide influenza surveillance through twitter. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*, 2015. URL https://pdfs.semanticscholar.org/6327/7acf07927625df96e668b8e812e6781f2a6b.pdf.

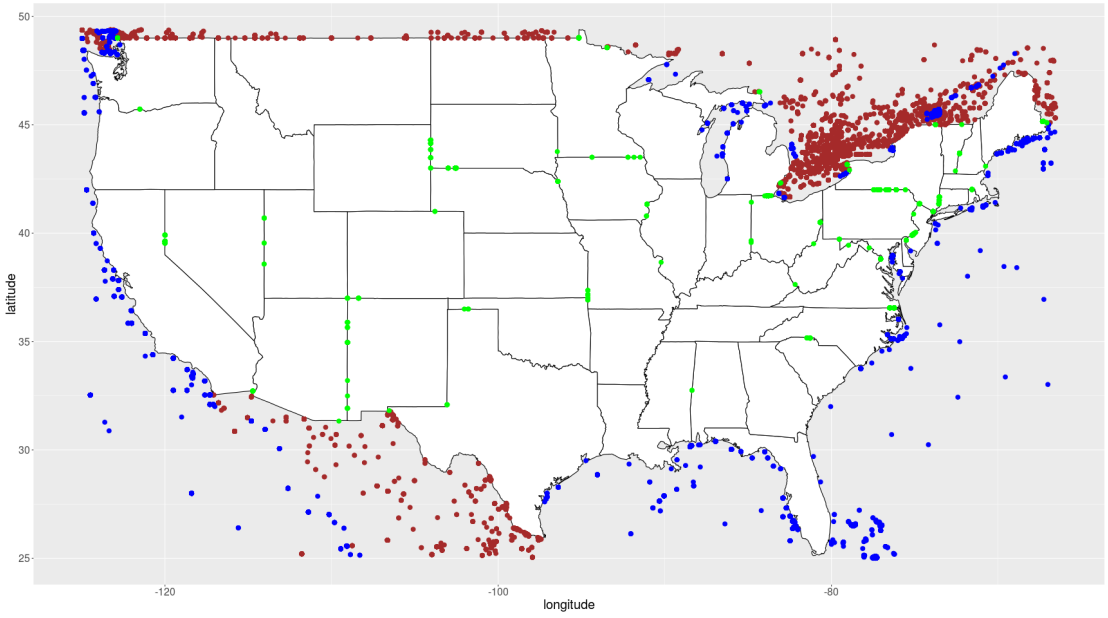
Michael J. Paul, Mark Dredze, and David Broniatowski. Twitter Improves Influenza Forecasting. *PLoS Currents*, 2014. ISSN 2157-3999. doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117. URL http://currents.plos.org/outbreaks/?p=39911.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011. URL http://dl.acm.org/citation.cfm?id=2145600.

R version and packages used to generate this report:
R version: R version 3.4.0 (2017-04-21)
Base packages: stats, graphics, grDevices, utils, datasets, methods, base
Other packages: data.table, knitr
This document was generated on Juli 01, 2017 at 03:55.



(a)



(b)

Figure 2: (a) All tweets having "state = 56" as code in the *sick_user* data set. These tweets could not be assigned to to any specific US state (b) Tweets whose origins was determined to be in Canada or Mexico (brown) or which could not be assigned to any US state (blue, mainly from the coastline or the ocean) and thus were removed from the *sick_user* data set. The green dots represent the tweets that had a "state = 56" as a code, i.e. that could not be assigned to any specific state in the original data set, but that could be recovered by the polygon lookup I performed. Note that the set of tweets shown in (b) is bigger than the set of tweets with state code "56". This is because some tweets were removed that did **not** have state code "56", but failed to be assigned to a state by the polygon lookup I performed.

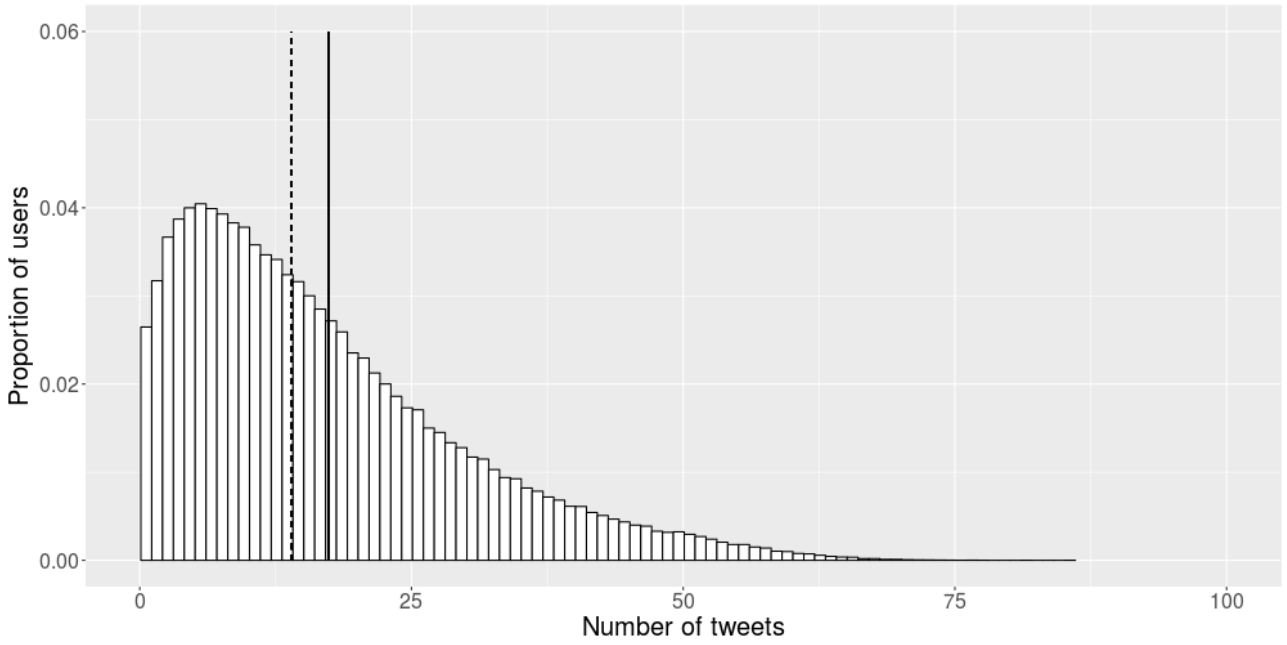
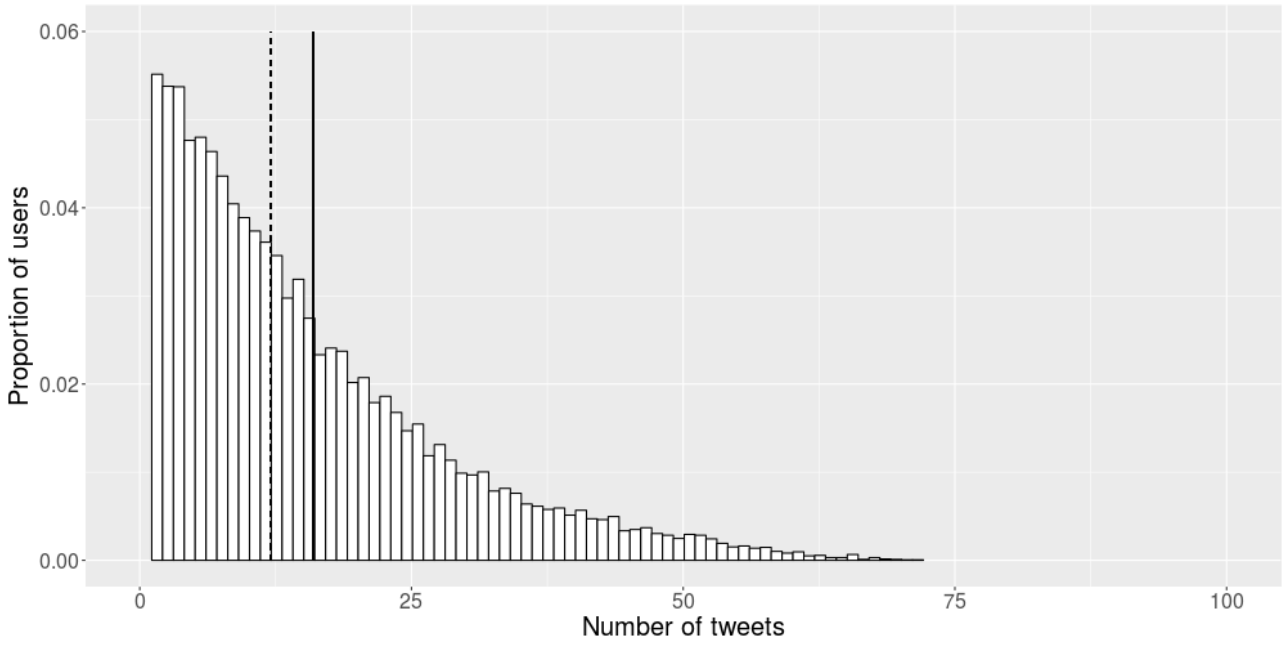
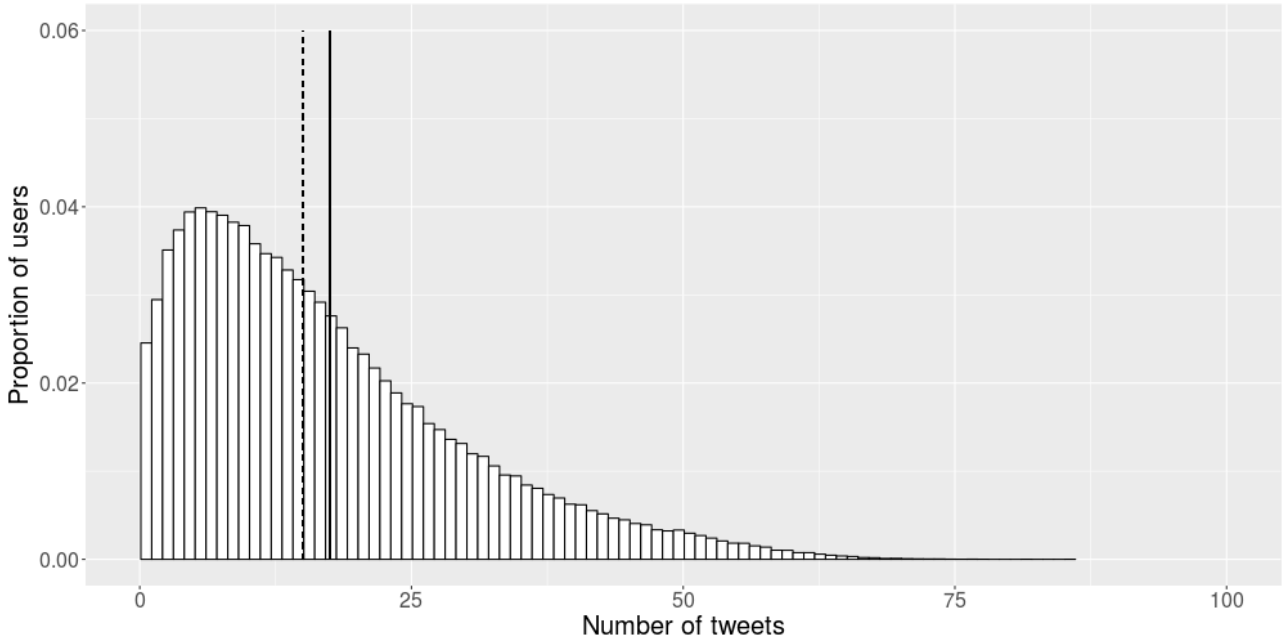


Figure 3: Histogram of the number of tweets sent per user in the *sick_user* data set during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1). As can be seen, many users only sent a handful of tweets during this time, whereas the user with the highest tweet activity sent 86 tweets. Mean = 17.3221116 (solid line); median = 14 (dashed line)

Figure 4: The total number of users who only sent tweets labelled as "1 = sick" (919), users who sent at least one tweet labelled as "1 = sick" and "0 = healthy" (1.9728×10^4), and users who only sent tweets labelled as "0 = healthy" (192779)



(a)



(b)

Figure 5: Histogram of numbers of tweets sent per user during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1). (a) contains only tweets from users who sent at least one tweet labelled as "1 = sick" and one tweet labelled as "0 = healthy". Mean = 16.0096817 (solid line); median = 12 (dashed line) (b) contains only tweets from users who never sent a tweet that was labelled as "1 = sick" by the classifier. Mean = 17.5342179 (solid line); median = 15 (dashed line). As can be seen, mode, median and mean of the number of tweets sent per user are significantly lower in (a) than in (b). Also, note that by construction (a) does not contain any user who only sent one tweet (since the users in this group are defined by having sent at last one tweet labelled "1 = sick" and one tweet labelled "0 = healthy")

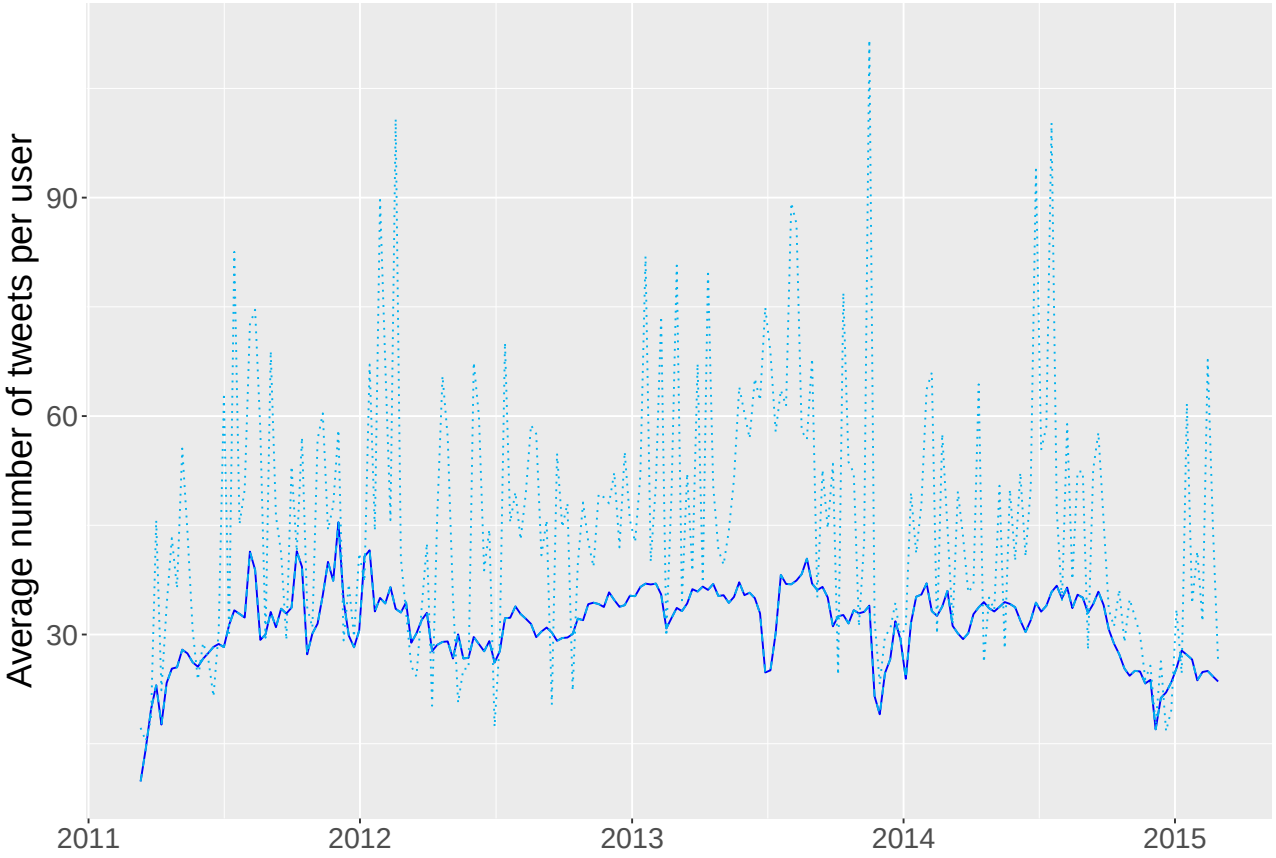


Figure 6: The average number of tweets per week sent by sick users (dotted light blue), healthy user (dashed light blue) and total users (solid blue). The average weekly tweet rate of sick users is significantly higher, while the tweet rate of healthy users is virtually indistinguishable from the total weekly tweet rate.

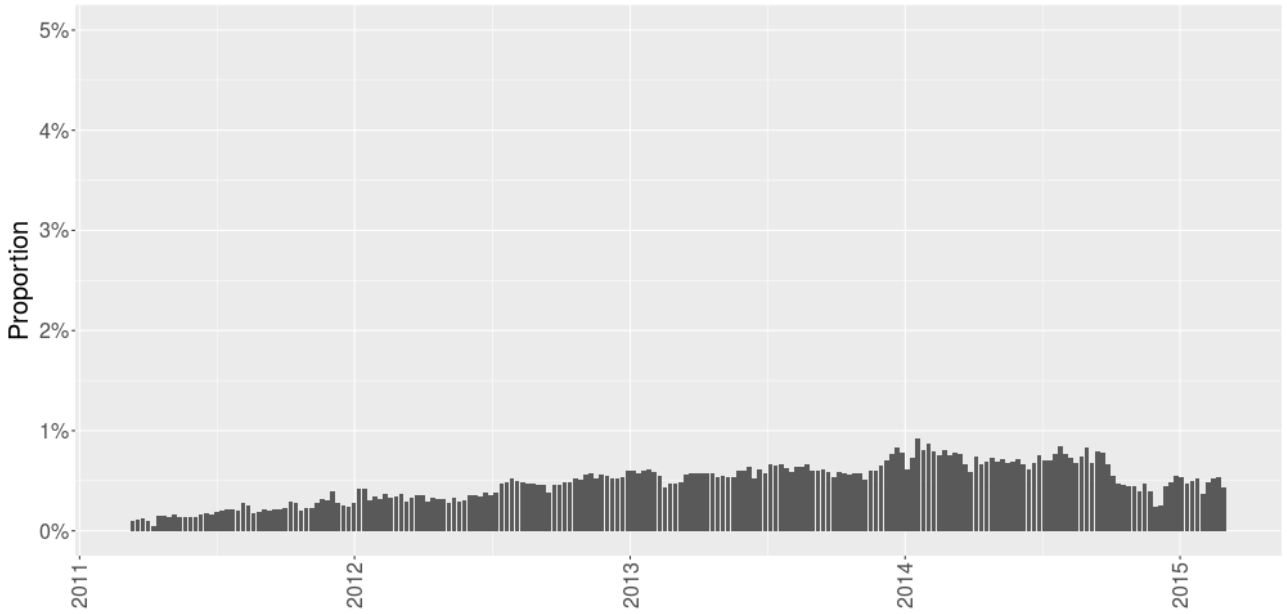


Figure 7: Relative number of tweets sent per week in the *all_tweets* data set between 2011-03-05 and 2015-07-11 (bin size = 1 week).

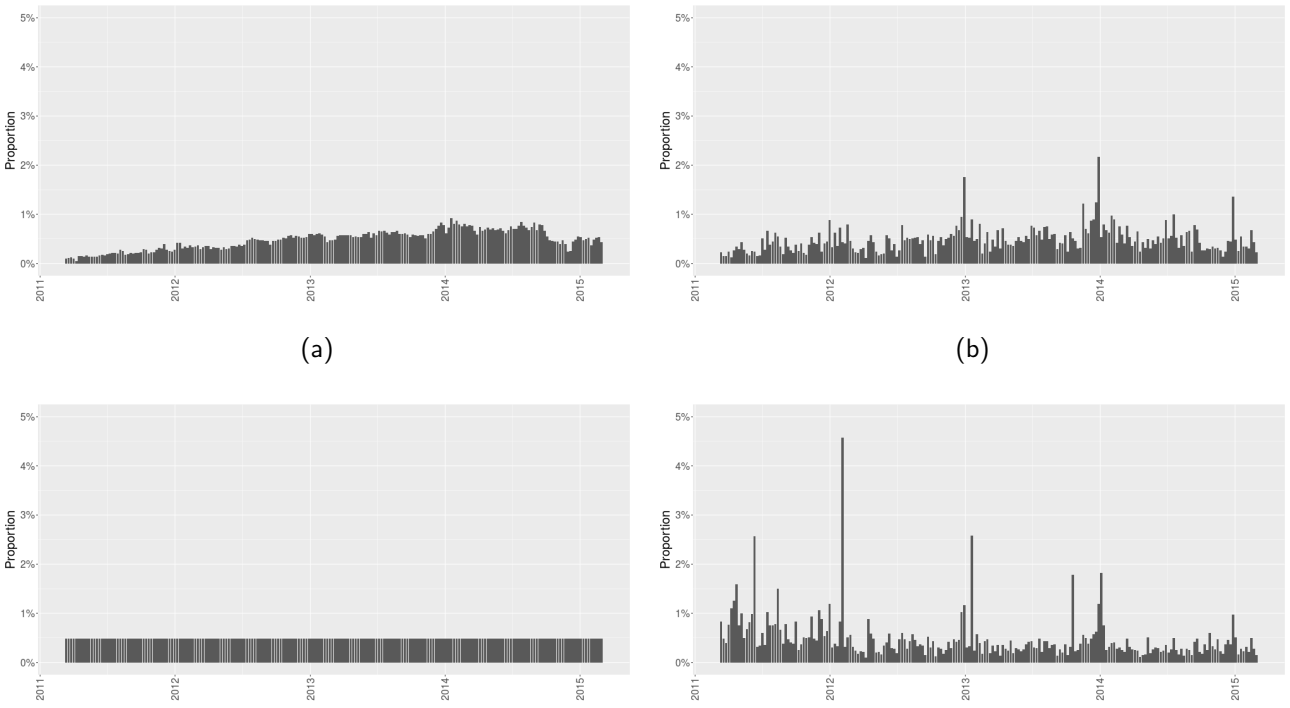
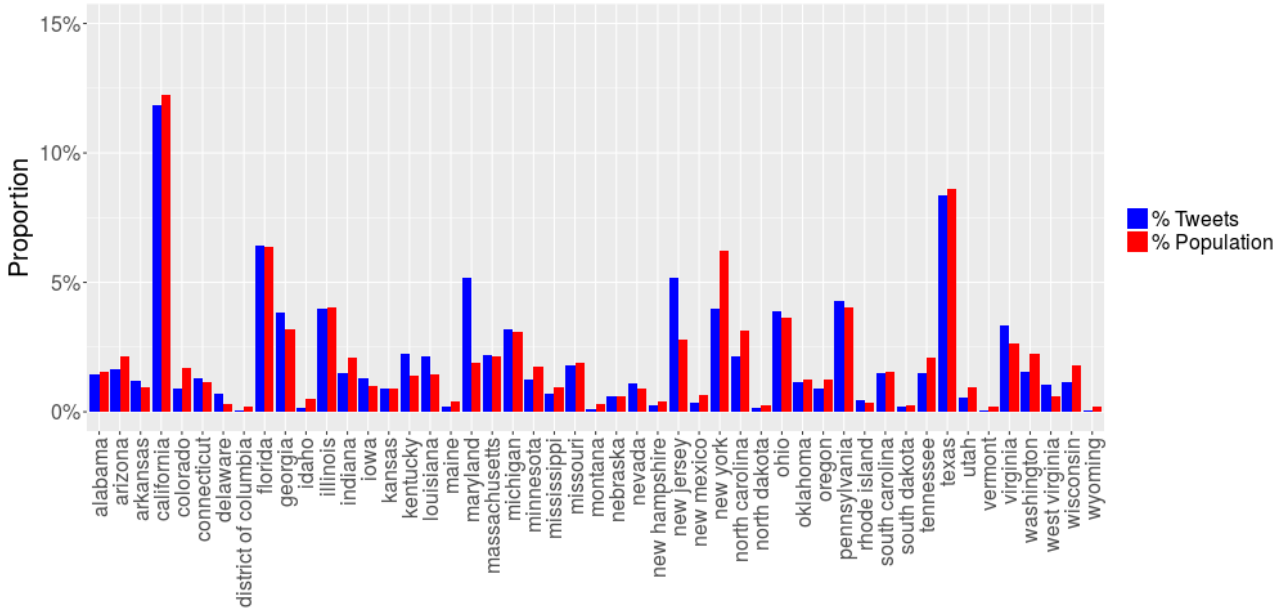
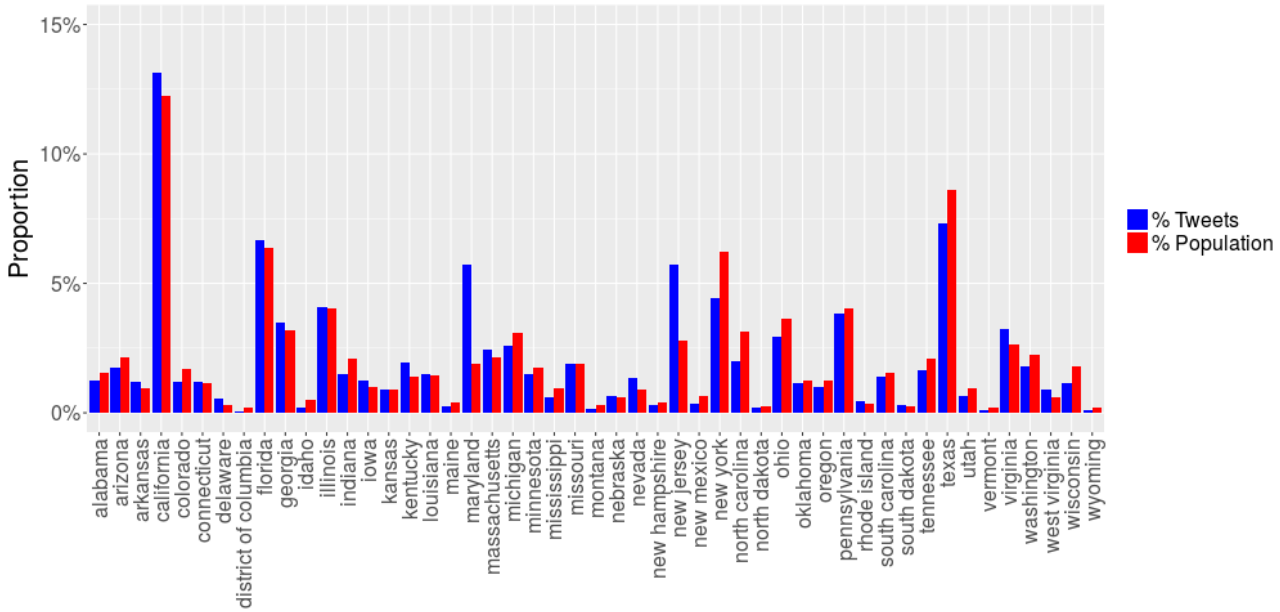


Figure 8: Histograms of numbers of the tweets sent per week during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1 week). (a) and (c) contain only tweets labelled as "0 = healthy"; (b) and (d) contain only tweets labelled as "1 = sick" by the classifier. The lower two histograms were normalised by the total amount of tweets sent per week. As can be seen, the tweets labelled as "1 = sick" follow a markedly different temporal pattern.

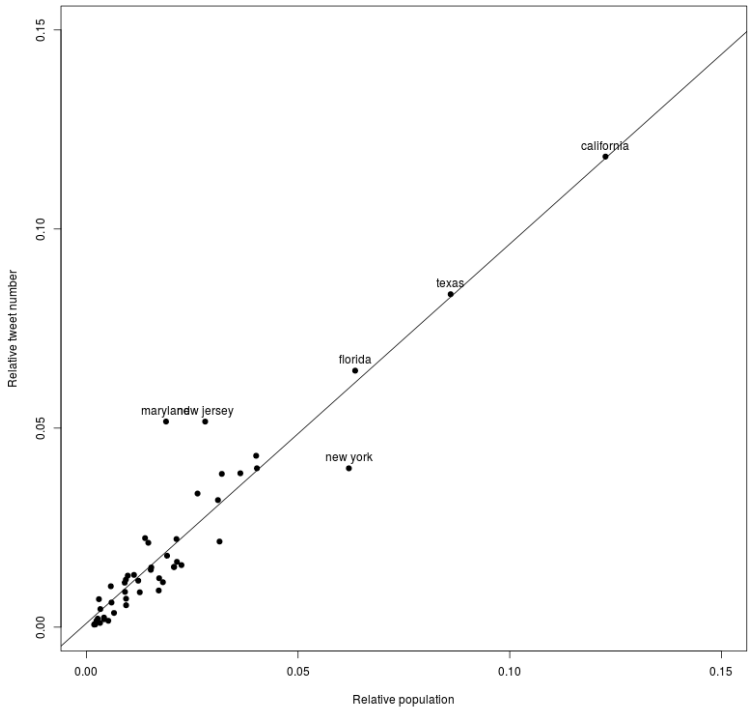


(a)

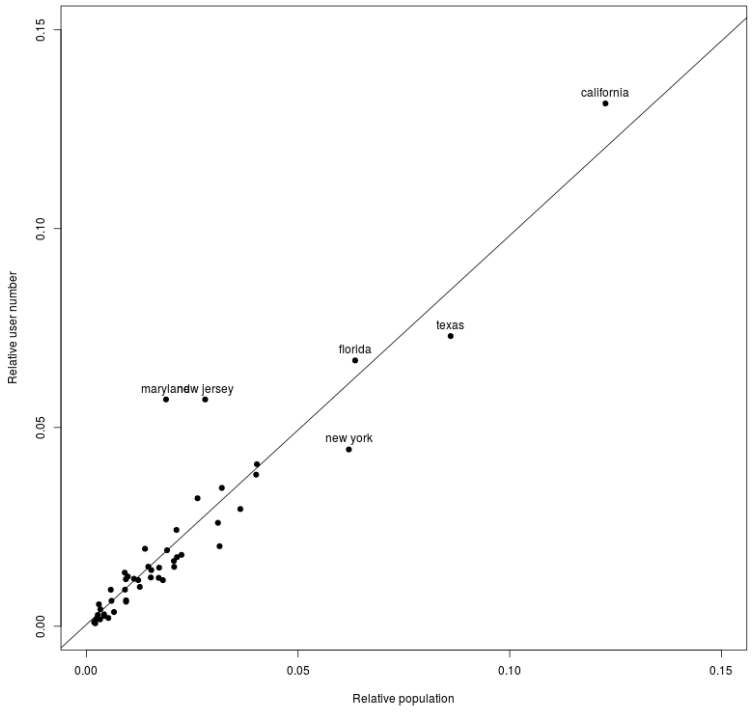


(b)

Figure 9: Relative number of tweets sent (a) and Twitter users (b) per state in the *all_tweets* data set between 2011-03-05 and 2015-07-11 compared to each state's relative population size.



(a)



(b)

Figure 10: Relative number of tweets sent per state (a) and Twitter users (b) in the *all_tweets* data set between 2011-03-05 and 2015-07-11 plotted against each state's relative population size

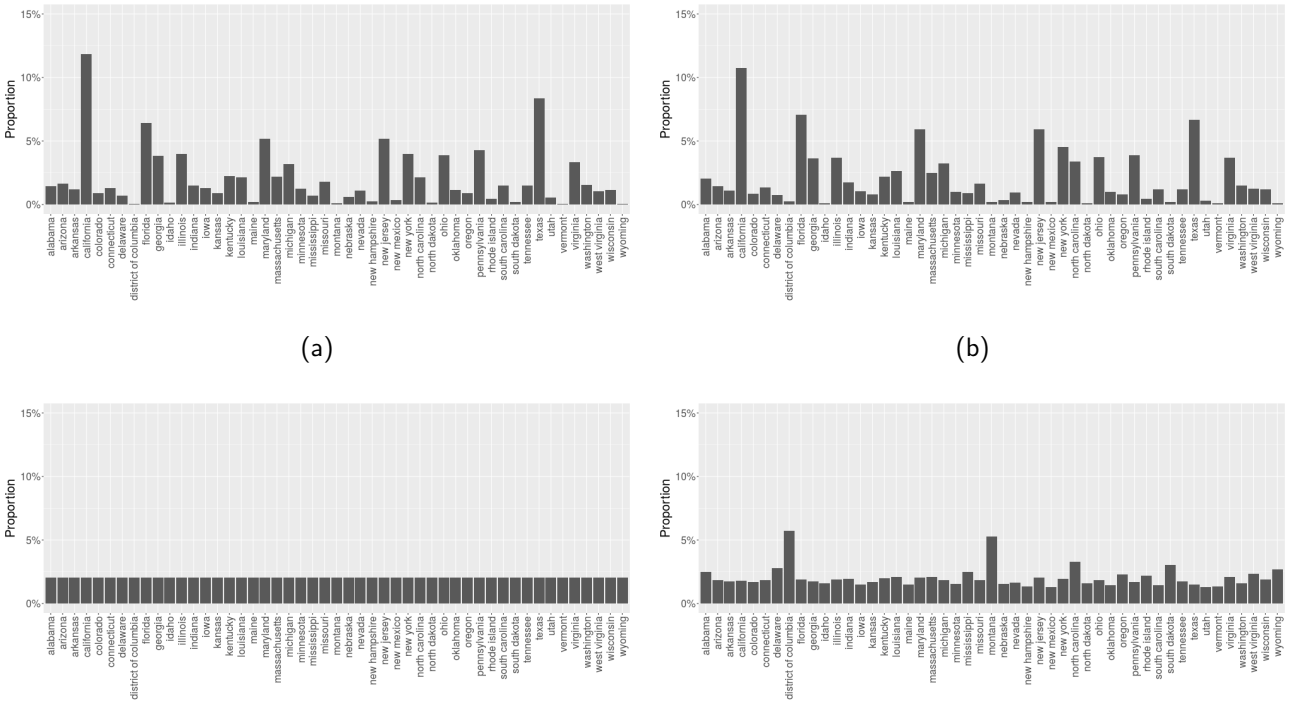


Figure 11: Histograms of numbers of the tweets sent in each state during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1 week). (a) and (c) contain only tweets labelled as "0 = healthy"; (b) and (d) contain only tweets labelled as "1 = sick" by the classifier. The lower two histograms were normalised by the total amount of tweets sent per state As can be seen, the tweets labelled as "1 = sick" follow a somewhat different spatial pattern.

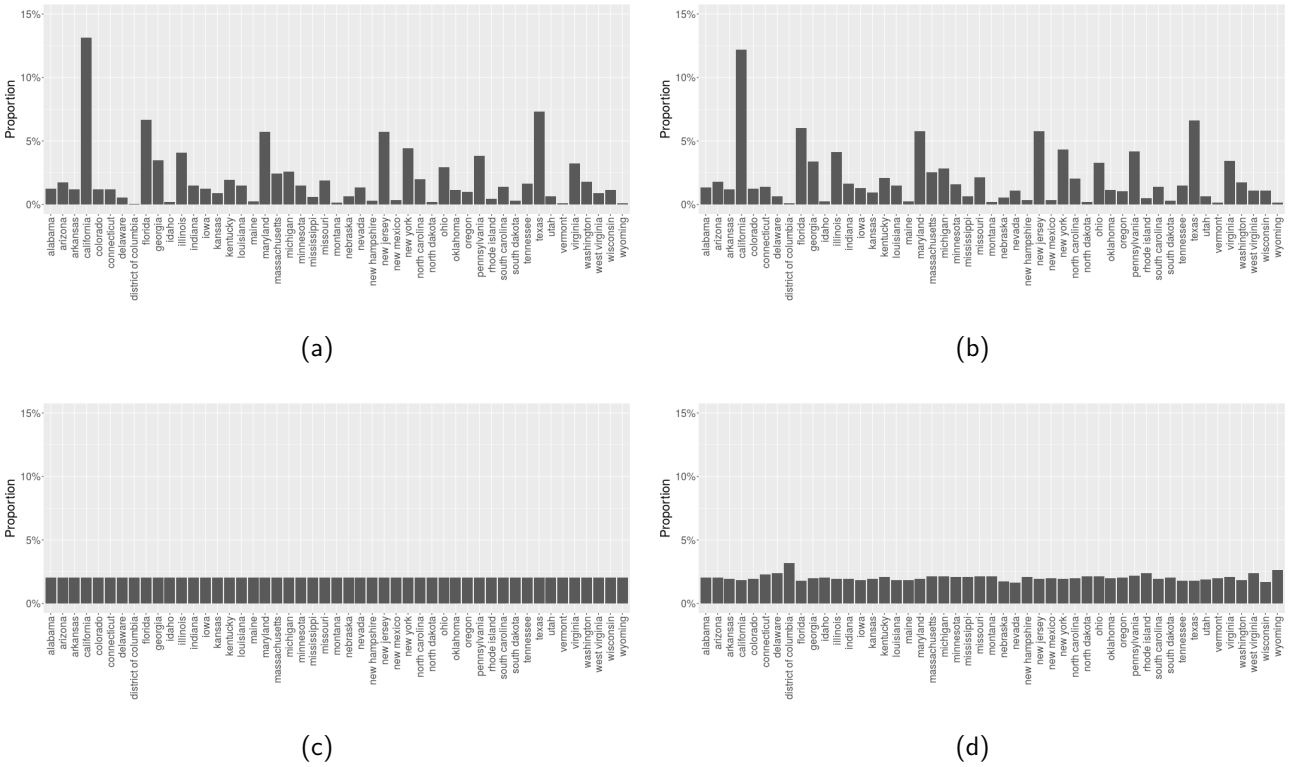
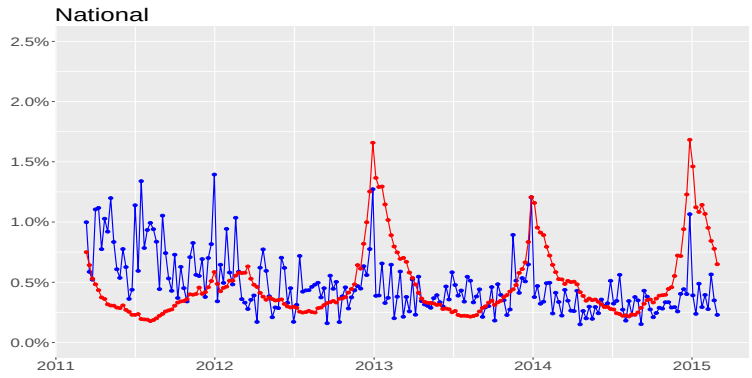
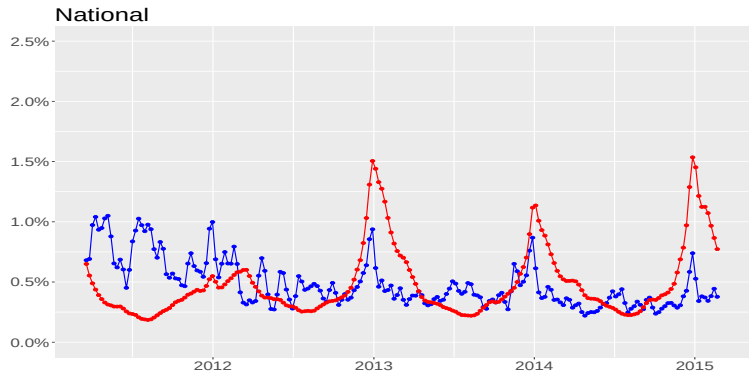


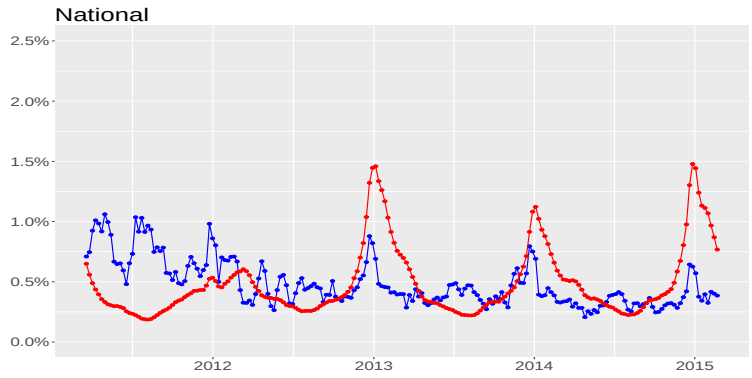
Figure 12: Histograms of number of users active in each state during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1 week). (a) and (c) contain only tweets labelled as "0 = healthy"; (b) and (d) contain only users with at least one tweet labelled as "1 = sick" by the classifier. The lower two histograms were normalised by the total number of user active per state. As can be seen, the users who were "diagnosed" as "sick" at some point by the classifier, follow almost the same spatial pattern.



(a)



(b)



(c)

Figure 13: Comparison between weekly CDC ILI rates (red) and the relative amount of tweets labelled as "1 = sick" from the Twitter flu classifier (blue). The data has been normalised in order to make them comparable, i. e. the percentages do not represent weekly ILI percentages, but instead sum up to a 100% over the whole time period. (a) without smoothing (b) after applying a two-week moving average smoother (c) after applying a four-week moving average smoother

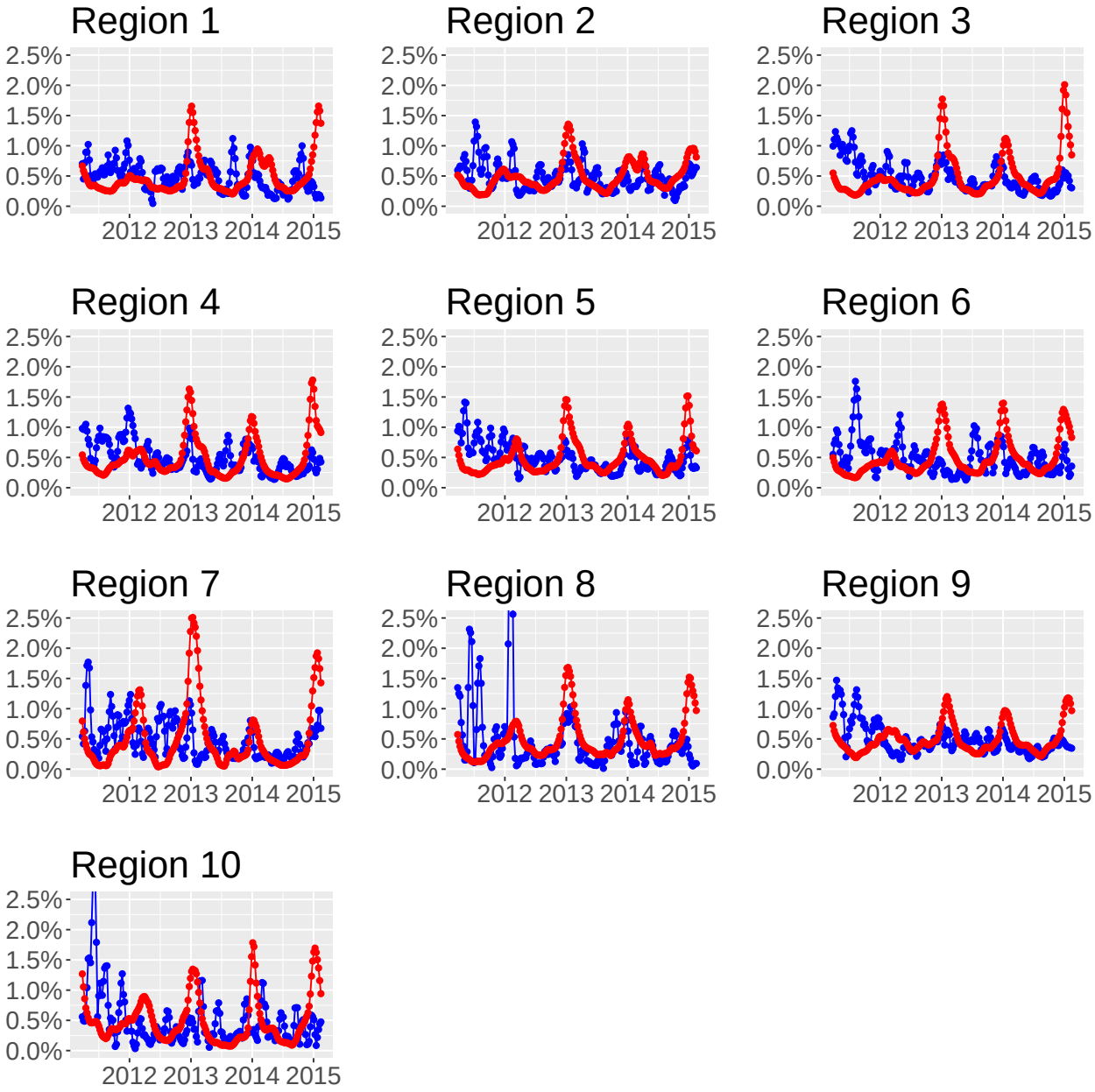
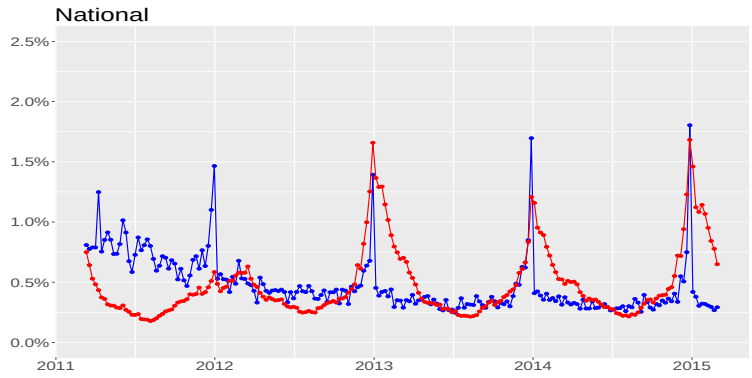
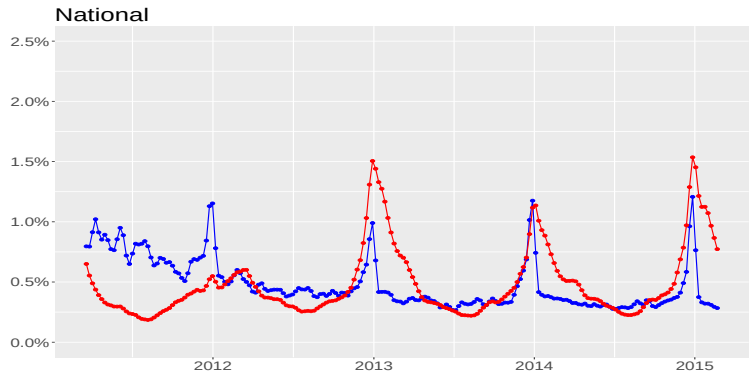


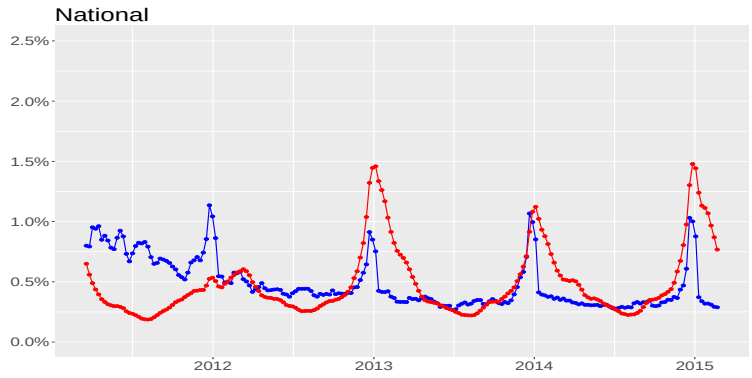
Figure 14: Relative number of tweets sent within each CDC ILI surveillance region per week (blue) compared with weekly ILI percentages in those regions (red). Data has been normalised and processed with a four-week moving average smoother. Note that Region 2 contains "Puerto Rico" and "Virgin Islands", Region 9 contains "Hawaii" and Region 10 contains "Alaska", all of which are missing from the Twitter data set.



(a)



(b)



(c)

Figure 15: Comparison between weekly CDC ILI rates (red) and the relative amount of users having sent at least one tweet classified as "1 = sick" from the Twitter flu classifier (blue) during a specific week. The data has been normalised in order to make them comparable, i. e. the percentages do not represent weekly ILI percentages, but instead sum up to a 100% over the whole time period. (a) without smoothing (b) after applying a two-week moving average smoother (c) after applying a four-week moving average smoother

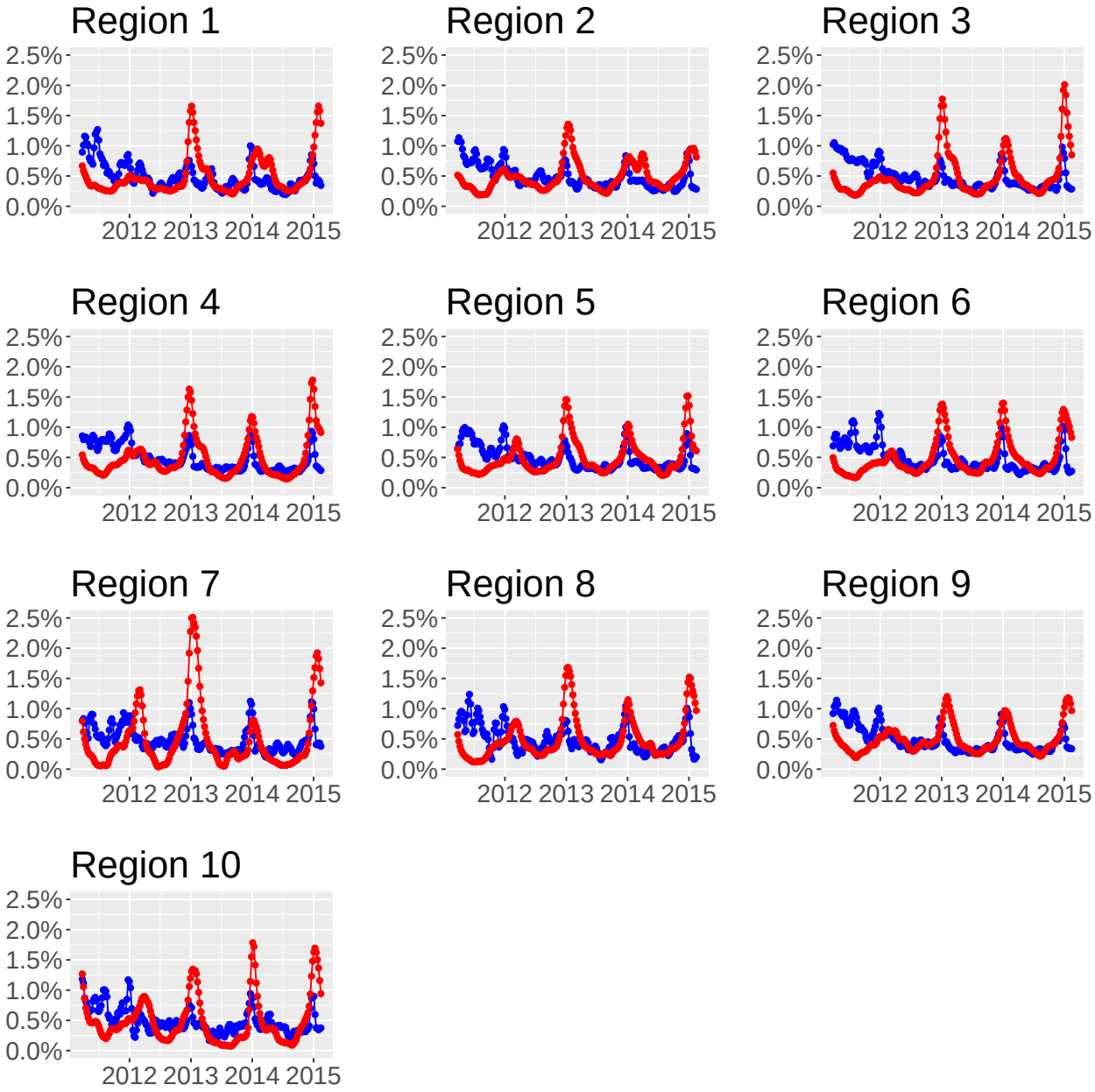
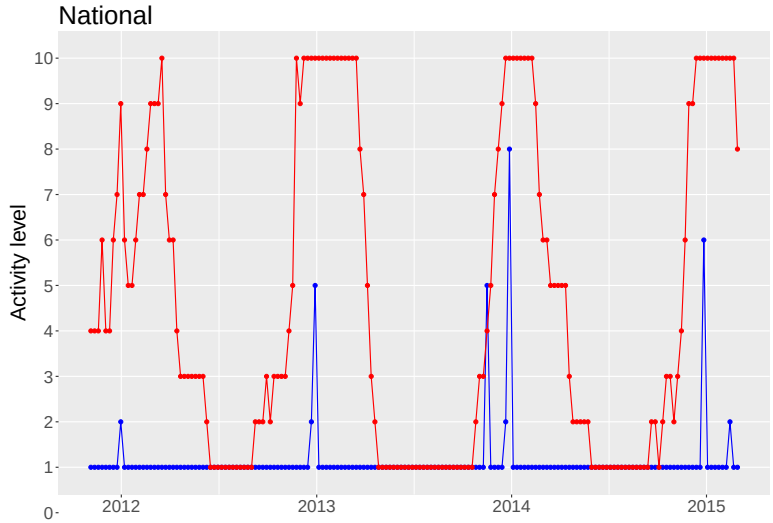
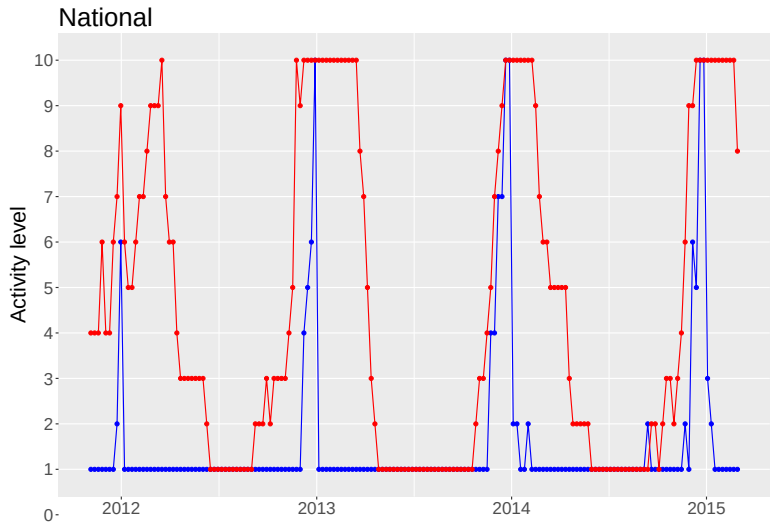


Figure 16: Relative number of users who sent at least one tweet labelled as "sick" within each CDC ILI surveillance region per week (blue) compared with weekly ILI percentages in those regions (red). Data has been normalised and processed with a four-week moving average smoother. Note that Region 2 contains "Puerto Rico" and "Virgin Islands", Region 9 contains "Hawaii" and Region 10 contains "Alaska", all of which are missing from the Twitter data set.



(a)



(b)

Figure 17: Comparison between weekly ILI activity levels reported by the CDC and calculated based on the Twitter data set. (a) activity levels based on relative number of sick tweets (b) activity levels based on relative number of sick users

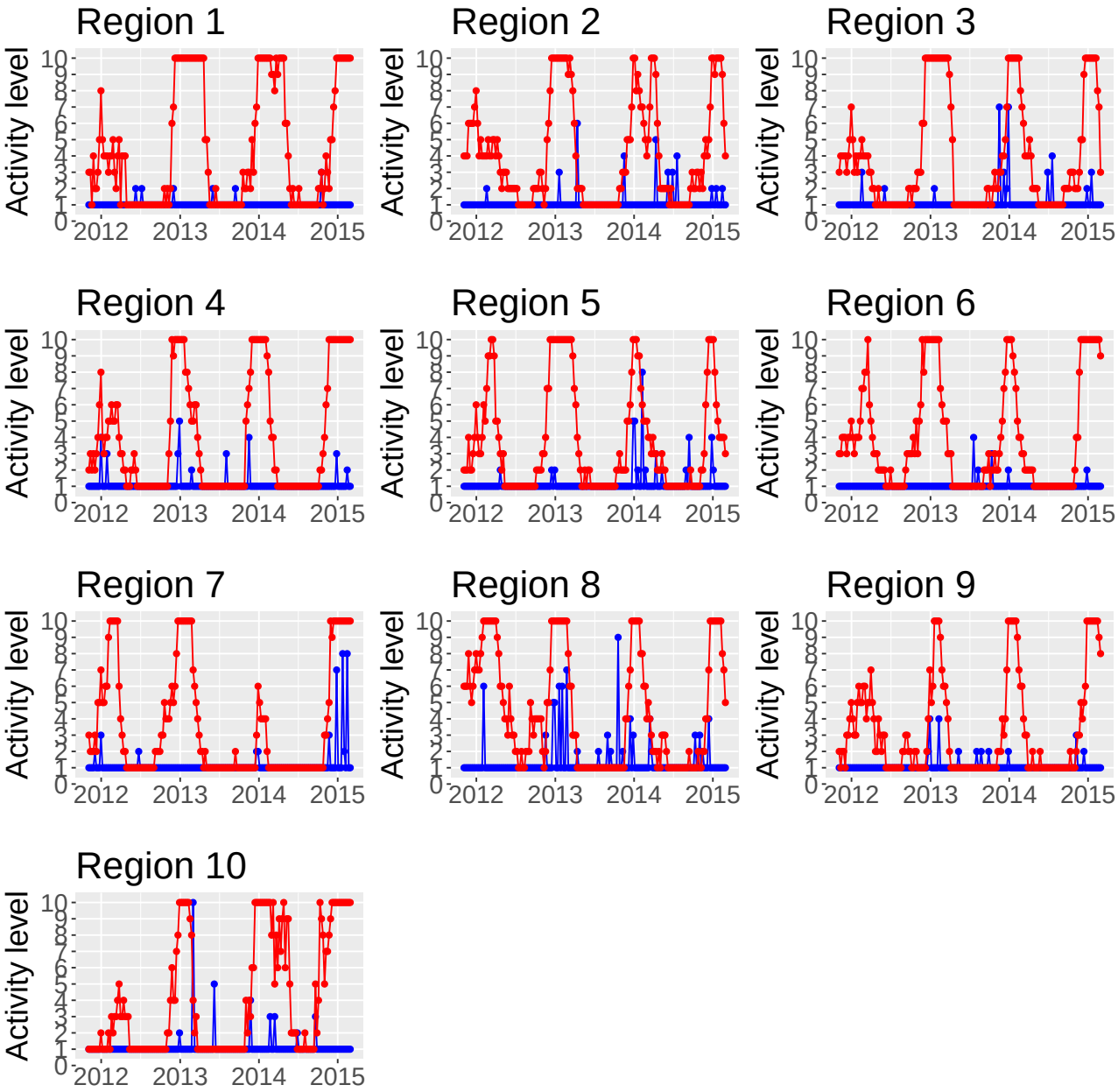


Figure 18: Comparison between regional weekly ILI activity levels reported by the CDC and calculated based on relative number of sick tweets per week

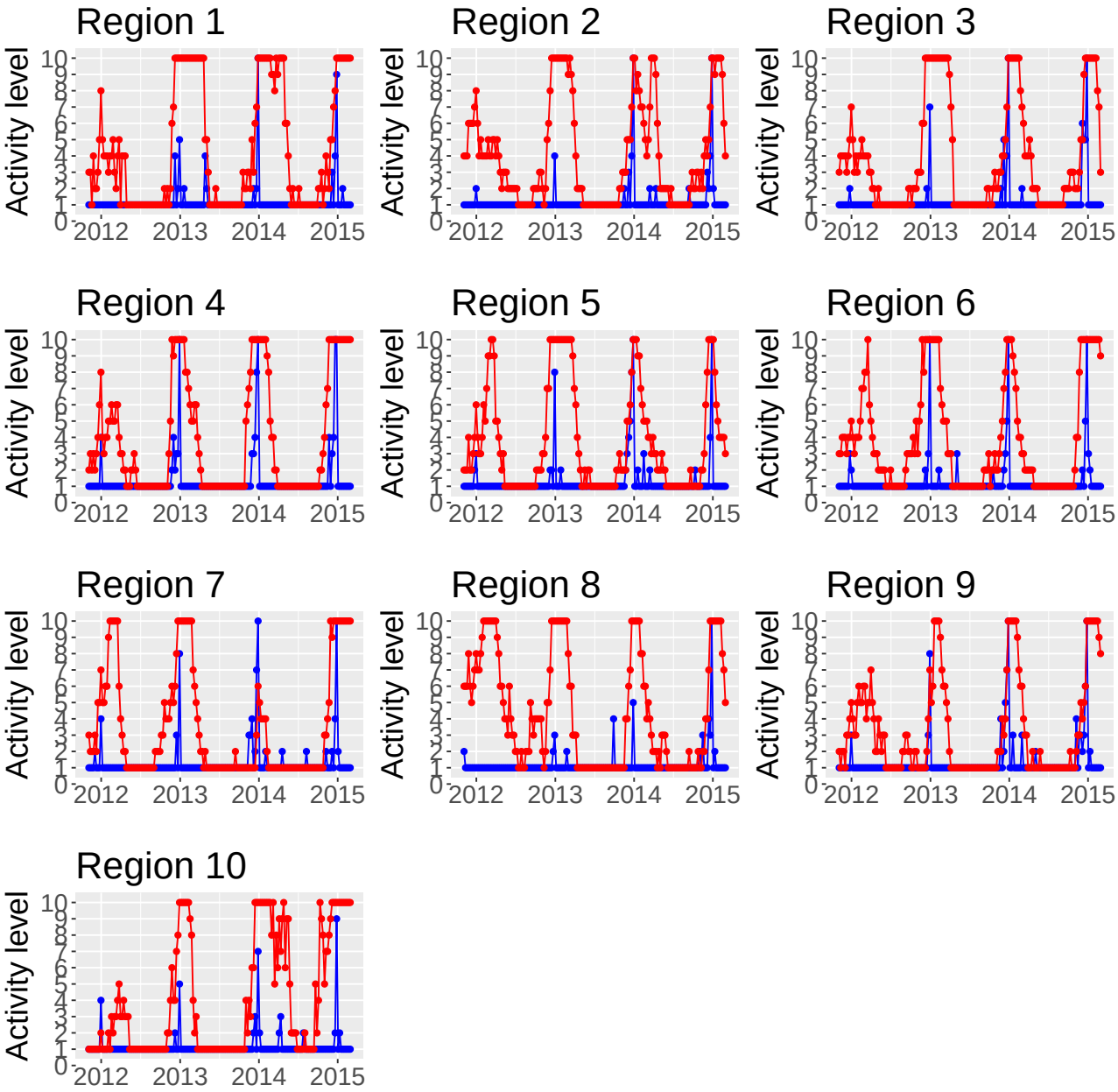


Figure 19: Comparison between regional weekly ILI activity levels reported by the CDC and calculated based on the relative number of sick users per week

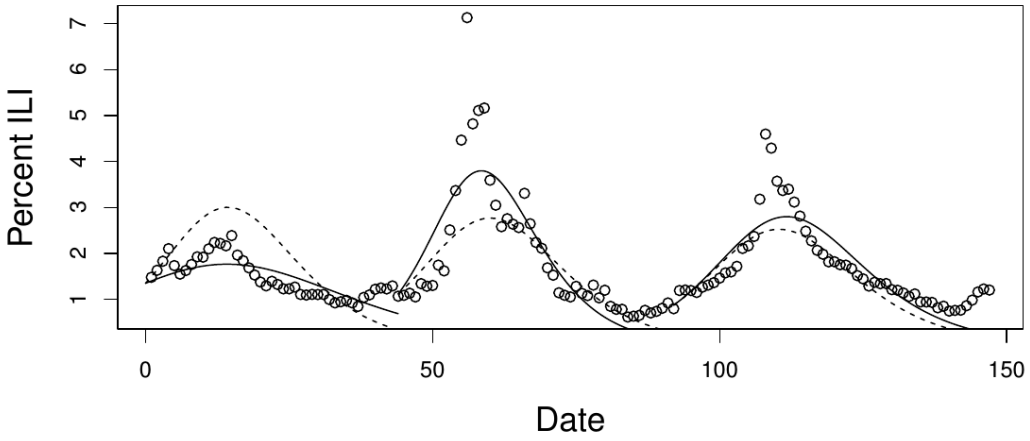


Figure 20: The CDC's estimates (circles) of influenza rates for a three year period compared to the best fit SIR models from the Twitter data using combined (dashed line) or yearly (solid line) parameters (taken from (Bodnar, 2015))