Does the Blue Bird Get the Flu?
Using Twitter for Flu Surveillance

Master Thesis in Biostatistics (STA495)

by

Servan Grüninger
09-737-040

supervised by

Prof. Reinhard Furrer
Prof. Marcel Salathé (EPFL)

Zurich, August 2017

Abstract

Coming soon...

Contents

Chapter 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 (Bundesamt für Gesundheit , 2017a). in the US, tens of thousands of people die from the flu each year and hundreds of thousands need to be hospitalised (Rolfes *et al.*, 2016).

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). In addition, both the Centers for Disease Control and Prevention (CDC) as well the Swiss Federal Office of Public Health only recommend patients to seek medical care if they belong to a risk group or if they show strong symptoms (Bundesamt für Gesundheit , 2016; Centers for Disease Control and Prevention, 2017).

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 (Bundesamt für Gesundheit , 2017b; Sentinella, 2017) or the "U.S. Outpatient Influenza-like Illness Surveillance Network" (ILINet) in the USA (Centers 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 to complement traditional epidemiological information might be needed in order to make influenza surveillance faster and more exhaustive. Luckily, the flu provides researchers with an excellent starting point to test new surveillance methods, due to the following

reasons:

Reliable data are available Most high-income countries have well-established Influenza surveillance systems which provide reliable data sources that can be used as "gold standard" with which the performance of new methods can easily be compared.

Spatio-temporal analysis possible Influenza leads to recurring epidemics each year all over the world, providing researchers with a vast amount of data to work with.

No stigma attached Influenza is "normal", in that Influenza patients are not stigmatised and it is not taboo to talk openly about it (in fact, getting sick with the flu is a very common content of small talk and news articles)

Relevant for public health Even though the flu is "normal", it is far from harmless. As outlined above, the flu poses a serious threat to the health of hundreds of thousands of people in Switzerland alone. Hence, better surveillance method would allow for better preventive and therapeutic measures.

Large body of pre-existing research Partially due to the reasons outlined above, there exists extensive research about about the prevention and treatment of the flu as well as about its etiology, transmission paths, pathophysiology, and virological characteristics. These information serve as excellent basis for epidemiological research

Thriving research community Many researchers are working on understanding, preventing, and combatting the flu on different levels, offering ample opportunities for collaborations.

1.1 Complementary epidemiology

Epidemiologists have always used a wide range of genetic, population and environmental information to study the transmission and propagation of disease - from simple counts of disease incidences, mortality or birth tables, and patient histories up to vast cohort studies, sophisticated disease models, and intricate clinical trials (Rothman, 2012; Koepsell and Weiss, 2014). The epidemiologist's goal is not only to discover where and when a disease occurs, but also why it does so and through which mechanisms. Hence, it should not come as a suprise that the advent of powerful genetic screening techniques, sophisticated algorithms, and cheap computing power have added a lot of new weapons to the scientific arsenal of the "disease detectives" (Bailey *et al.*, 2005; Khoury *et al.*, 2013; Gardy *et al.*, 2015) - additions that are sorely needed

to keep up with the ever-changing disease landscape.

Nowadays, epidemiological work is decidedly different in two particular ways: First, non-communicable disease are the primary cause of life-years lost in high-income countries and are on the rise world-wide (Lozano *et al.*, 2013). Second, infectious diseases can spread faster than ever before thanks to increased social and spatial mobility on a global scale (Hufnagel *et al.*, 2004).

For both challenges, however, novel epidemiological methods are rising to the task. They can help to track down the underlying causes of non-communicable disease as well as improve the speed and efficience of disease surveillance methods in order to keep up with emerging epidemics and even pandemics. In this regard, digital data sources appear to be especially promising to complement established epidemiological work (Salathe *et al.*, 2012; Simonsen *et al.*, 2016).

1.2 Digital epidemiology

In the digital age, everybody leaves traces. And thanks to the vast amount of digital footprints each one of us leaves behind, digital data sources do not have to be medical in nature in order to be epidemiological useful. "Digital epidemiology" can offer epidemiological insights that are very different from traditional surveillance systems and public health infrastructure. More importantly, data sources such as web queries, social networking sites, online news articles, or mobile devices and other wearables have the great advantages of being internationally available and accessible, offering fine-grained geospatial location of the users (or patients) and often allow for instantaneous feedback (Salathe *et al.*, 2012).

For example, Google tried to use search queries in order to predict the spread and the intensity of the flu in certain countries (Ginsberg *et al.*, 2009). Retrospective analysis of Google search queries and online media reports showed that these data sources could have been used to detect the Ebola epidemic of 2014/2015 quicker and with more sensitivity (Anema *et al.*, 2014; Milinovich *et al.*, 2015). And finally, Wikipedia page views proved to be reliable predictors to model the epidemiology of such different diseases as dengue fever, influenza, cholera, HIV/AIDS, or tuberculosis (Generous *et al.*, 2014).

Other services such as "Health Map" aggregate search engine queries, online news reports, official information from national and international health agencies as well as user eyewitness re-

ports in order to keep track of a wider range of disease outbreaks all over the world (Brownstein *et al.*, 2008; Freifeld *et al.*, 2008). Finally, there exist a vast range of participatory disease surveillance system, such as "FrontlineSMS", "Usahidi", "GeoChat", "Asthmapolis", "Outbreaks Near Me","Influenzanet", "FluTracking","Reporta", "Dengue na Web", "SaludBoricua", or "Flu near you", which use SMS, voice messages, smartphone apps, web forms or e-mail in order to collect data from and disseminate epidemiological information to afflicted populations (Freifeld *et al.*, 2010; Chunara *et al.*, 2013; Wójcik *et al.*, 2014; Chunara *et al.*, 2015). In addition, these forms of "participatory epidemiology" can also help in adding social and economic context to health-related data, defining the research goals and questions, improving the work-flow, or synthesising heterogenous sources of data (Bach *et al.*, 2017; Liu *et al.*, 2017).

And it does not stop with digital data alone. Many of the approaches outlined above can be combined with and extended by other biologically or clinically relevant data such as high-throughput sequence data, clinical visits, pharmaceutical prescriptions, or clinical symptoms, in order to allow for a more accurate description of the mechanisms of disease spread, the pathogens involved, and the treatments administered (Ray *et al.*, 2016).

For the scope of this thesis, I am focusing on one particularly potent source of information: social media data.

1.3 What can "Larry the Bird" tell us about the flu?

Social media offers the possibility to directly measure a user's sentiments and behaviours via his posts and interactions with other users - information that can be highly valuable from an epidemiological point of view. If somebody is suffering from the flu, her behaviour undoubtedly changes: She behaves more lethargicly, stays in bed, and might complain about her symptoms in the presence of family, friends, or work colleagues. Even people on Twitter or Facebook can exhibit "disease symptoms" which can be diagnosed. A tweet, in which somebody is complaining about having fever, joint aches, and a cough can be a tell-tale sign of the flu. Similarly, somebody who tweets that all his colleagues were absent from work due to the flu, gives researchers precious and above-all fast warning about an incoming flu wave.

Given that social media is often used to share information about one's personal well-being and/or feeling and takes up an increasing amount of time of many people (Bauer, 2016; Scott *et al.*, 2017; Asano, 2017), it is to be expected that behavioural changes due to a disease can also be detected by analysing the social media behaviour of the people afflicted. Either because

said people are explicitely informing their peers about their current affliction or because the frequency or other implicit aspects of their social media behaviour changes. All these changes can - in theory - be detected on the population level if researchers get access to data sets that are large enough.

However, using social media for epidemiological surveillance is deeply hampered by one crucial fact: Most social media plattforms do not offer access to user's profiles. The profiles of most social media users are either completely inaccessible to the public (e.g. Whatsapp, Snapchat, FB-Messenger) or only accessible if the user allows it (Facebook, Instagram). But even if profiles are publicly available, they are often not easily accessible and aggregatable for research purposes due to the strict rules of the respective application programming interfaces (APIs) provided by the respectie companies.

"Larry the Bird", as Twitter's trademark mascot is affectionately called by its creators (Rehak, 2014), is a notable exception to this. Due to the ease-of-access to the Twitter-API as well as due to the fact, that tweets often contain a direct expression of sentiment of some sort, researchers can gain access to millions of tweets sent out every day (Twitter, 2017). The source of information is so rich, in fact, that it has spawned a variety of studies in wide range of disciplines, such as political science (Tumasjan *et al.*, 2010, 2011; Stieglitz and Dang-Xuan, 2012; Newman, 2016), business (Swani *et al.*, 2014; Chae, 2015), economics (Bollen *et al.*, 2011b,a; Zhang *et al.*, 2012; Sul *et al.*, 2014), sociology (Poblete *et al.*, 2011; Himelboim *et al.*, 2013; McCormick *et al.*, 2015), communication science (Zhao and Rosson, 2009; Marwick and Boyd, 2011; Himelboim *et al.*, 2013; Hermida, 2013), psychology (Chen, 2011; Golbeck *et al.*, 2011; Qiu *et al.*, 2012; Eichstaedt *et al.*, 2015; Braithwaite *et al.*, 2016), nutrition science (Widener and Li, 2014; Vidal *et al.*, 2015; Abbar *et al.*, 2015), or medicine (Salathé *et al.*, 2013; Love *et al.*, 2013; Nwosu *et al.*, 2014; Adrover *et al.*, 2015; Eichstaedt *et al.*, 2015; Mowery *et al.*, 2017), even though studies from the computer and information sciences studies are clearly in the majority (Lee *et al.*, 2013; Zimmer and Proferes, 2014; Steiger *et al.*, 2015).

Hence, it seems straightforward to use Twitter data for epidemiological purposes as well and to fit a model using the content of those tweets as independent variables and the official influenza data as 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, for example, the flu season. Google Flu Trends (Ginsberg *et al.*, 2009) initially fell prey to this kind of overfitting, linking search terms such

as “High School Basketball” to flu disease state - just because the basketball season happened to be in winter which unsurprisingl coincided with the flu season. The Google researchers tried to root out these kind of correlations in order to improve the performance of their algorithms, but eventually had to admit defeat to their daunting task: The huge data masses coupled with changes of user behaviour, external influences from media reports, and the constant adaptions of Google’s search algorithm itself created too much noise to allow for reliable flu predictions (Olson *et al.*, 2013; Butler, 2013; Lazer *et al.*, 2014) Eventually, Google Flu Trends was discontinued in summer 2015 as a publicly available service, but the data are still accessible to and used by researchers all around the world (Google Flu Trends Team, 2015).

Google Flu trends was a pioneering attempt to use online data to make predictions and despite its (temporary?) failure, it provided researchers with many insights into the promises and perils of using big online data for epidemiological purposes. However, the main problem when using large data sets to infer flu states still remains: 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 relevant tweets 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 show a clear presence of a flu infection - either within the tweeter herself or within the people in surrounding her. Hence, we can use these tweets to get an estimate of the amount of twitter users that are currently tweeting about the flu or influenza like symptoms - and thereby of 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 methods to extract epidemiologically relevant data from tweets and creating descriptive or predictive models from them. These methods include simple keyword ratios (Lampos and Cristianini, 2010), partial-differential equations (Wang *et al.*, 2016), linear regression models (Culotta, 2010), autoregressive models (Achrekar *et al.*, 2011; Paul *et al.*, 2014, 2015), support vector machines (Paul and Dredze, 2011), probabilistic topic models (Paul and Dredze, 2011), sentiment detection (Aramaki *et al.*, 2011), and semantic text analysis (Lamb *et al.*, 2013). However, most of them still depend on some sort of correlation with official data from public health authorities. making them again prone to overfitting - up

to the point that even seemingly irrelevant tweets about zombies can "predict" flu outbreaks almost as good as tweets containing clinically relevant information (Bodnar and Salathé, 2013). Also, it has been shown that media reports can have a substantial influence on Twitter users' behaviour and thus on the content of their tweets - thereby creating a "news bias" in the Twitter models (Aramaki *et al.*, 2011).

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 the roughly 319 million of worldwide users who were active on a monthly base by the end of 2016 - or even with the 69 million monthly active users in the US (Twitter, 2017) - 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. That is, they had the possibility to build their model knowing which one of the observed Twitter users were sick and which were not within a specific 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. The first goal of this thesis is to assess the validity of the flu classifier they built based on these data (see Section 2.3) as well as to make suggestions on how its performance might be improved.

Chapter 2

The challenge of reproducibility

An important part of daily research practice is the evaluation and interpretation of other scientist's work. To do so, we mainly rely on information from scientific publications. With the plethora of new articles being published every day and with the increasing difficulty that even high quality journals have when it comes to ensuring the reproducibility of the experiments described in their articles, it is important that researchers know how to critically judge and evaluate papers in order to extract the necessary information they need without being led on the wrong track.

However, while everybody seems to believe that reproducibility is important, nobody seems to agree on what it exactly is. Hence, talks, papers and discussions about reproducibility are often rife with opportunities for misunderstandings: Some see reproducibility already achieved if a single confirmatory study shows the same overall effect as an exploratory study, others contend that a finding has only been reproduced if several independent studies could show the very same results as the experiment that is supposed to be replicated. In the following, I will give a quick overview over current challenges with regard to reproducibility and also explain how reproducibility plays a role in my thesis.

2.1 To replicate or to reproduce? Why semantics matter

One of the major sources of confusion in the discussion about (methods) reproducibility is its distinction from replicability (or "results reproducibility"). Many scientific articles tackling the so-called "reproducibility crisis" (Casadevall and Fang, 2010; Prinz *et al.*, 2011; Begley and Ellis, 2012; Begley, 2013; Begley and Ioannidis, 2015; Freedman *et al.*, 2015; Aarts *et al.*, 2015; Baker, 2016a,c; Crotty, 2014; Nosek and Errington, 2017) actually have the issue of replication

at heart (Schooler, 2014; Stroebe and Strack, 2014; Kullmann, 2015; Maxwell *et al.*, 2015; Earp and Trafimow, 2015; Camerer *et al.*, 2016; Loken and Gelman, 2017). So, what is the difference, then?

For Peng (2009), the replication of scientific studies hinges on "independent investigators, methods, data, equipment, and protocols" in order to evaluate the claims made in said study. However, replicating studies costs time and money and is sometimes neither feasible nor desirable. Nevertheless, "there is a need for a minimum standard that can fill the void between full replication and nothing." Aiming for reproducibility, defined as the verification of published results and conducting alternative analyses using "the data sets and computer code [...] made available to others", could provide such a minimal standard.

In other words, Peng (2009) sees reproducibility given if one can recreate the same results, statistical analyses, tables, and figures as those reported in the original study - not by recollecting the data and rewriting the code, but by using the very same data and code source used by the authors of the study that is supposed to be reproduced. Replication, however, depends on the ability to recreate the same results by independently collecting and evaluating the relevant data.

Cacioppo *et al.* (2015) offer very similar definitions. For them, "reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator". Replicability, on the other hand, "refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected".

These distinctions have also been adopted by the American Statistical Association, which calls a study reproducible, "if you can take the original data and the computer code used to analyze the data and reproduce all of the numerical findings from the study". Replicability is defined as "the act of repeating an entire study, independently of the original investigator without the use of original data (but generally using the same methods)" (Broman *et al.*, 2017).

However, Kenett and Shmueli (2015) differentiate between repeatability, reproducibility, and replicability and point out that the nomenclature can vary in meaning across different fields. They suggest "that these terms can be clarified by considering the intended generalisation of the study", since all three terms "are aimed at assuring generalizability, but the generalizability is typically of different types". Within this framework, a repeatable finding can be recreated by keeping every aspect of the original study or analysis constant (including exactly the same methods, codes, locations, experimentors etc.); reproducible findings can be recreated by different researchers in different locations, but by using the methods and data sources from the original

study; and replicable findings can be recreated by different researchers in different locations with different (i. e. recollected) data and sometimes by using different methods, too. For example, the PDF-version of this thesis can be exactly recreated by running the corresponding RNW-file (repeatability), while the findings described therein can be reproduced by using the data and codes provided in the Github-repository (reproducibility) or even by recollecting the data and rewriting the codes (replicability). In other words, repeatable research is least and replicable research is most generalisable.

Finally, Goodman *et al.* (2016) point out that the concept of reproducible research originally emerged in the computational sciences with a clear-cut definition, namely to "permit the reader of a paper to see the entire processing trail from the raw data and code to figures and tables", even though "a wide range of issues [are nowadays] subsumed under the rubric of reproducibility: design, reporting, analysis, interpretation, and corroborating studies (that is, replication)". However, instead of upholding the existing, but muddled distinction between repeatability, reproducibility, and replicability, they offer a new terminology in order to subsume everything under the concept of reproducibility. That is, they differentiate between methods reproducibility, results reproducibility, and inferential reproducibility.

Methods reproducibility corresponds to the most common interpretation of reproducibility, namely "the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results." Results reproducibility corresponds to replicability and is "the production of corroborating results in a new study, having followed the same experimental methods". Inferential reprodudicibility, however, is different from the concepts described so far and refers to "the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study".

I believe that the latter distinction is the most straightforward and exhaustive categorisation of different forms of reproducibility. Even the concept of repeatability can be subsumed under this categorisation, namely as an especially stringent case of methods reproducibility. Most importantly, however, this categorisation makes it clear that methods, results, and inferential reproducibility can occur independently from each other, something that is often forgotten in the discussions about reproducibility. In particular, methods reproducibility can be fulfilled even though results and inferential reproducibility are absent, e. g. if a finding based on Twitter user data can be perfectly recreated using the original raw data and codes, but fails replication when collecting data from a different set of Twitter users. The same holds true for results reproducibility, which can exist independently of methods reproducibility (e. g. if the raw

data and code of the original study are not provided, but replication is possible by recollecting the data using the methods described in the study) and inferential reproducibility (e. g. if the original results can be replicated but still warrant a different interpretation because of theoretical or methodological discrepancies). Finally, inferential reproducibility can be fulfilled even in the absence of methods and results reproducibility, e. g. when the raw data and code of the original study are missing, while the replication of the study based on the described methods shows an attenuated (or exarcebated) effect with the same directionality as the original study.

For rest of this thesis I will therefore adopt the definitons of Goodman *et al.* (2016) and distinguish between methods, results, and inferential reproducibility, respectively.

2.2 How to ensure reproducibility?

(Results) reproducibility is often called the "hallmark" (Aarts *et al.*, 2015; Munafò *et al.*, 2017) or "bedrock" (Casadevall and Fang, 2010) of science. After all, what good are scientific findings if they are only valid for the time and place they were originally created?

some of it is simply a matter of statistical necessity (**?**)

The ability to replicate research findings has always been the hallmark of reliable empirical research. A finding that couldn In recent years, discussions about the inability reproduce or replicate many research findings have become more prominente and in fact the papers written about it have steadily increased (Goodman *et al.*, 2016). The discussions revolving around the reproducibility of scientific findings have become so dominant, in fact, that even found their way into the news outlets and the broad media, causing a deluge of news article decrying the lack of replicability in different areas of the sciences (Lehrer, 2010; Carey, 2015a,b; Achenbach, 2015a,b; Yong, 2016; Engber, 2016; Baker, 2016b; The Economist, 2016; Feilden, 2017; Belluz, 2017; Meyer, 2017).

Despite the seemingly clear-cut distinction between methods, results, and inferential reproducibility, it is important to note,

reproducible epi research ¿ follow his recommend

A manifesto for reproducible science (Munafò *et al.*, 2017) ¿ more comprehensive approach

So this thesis has a second important goal: It is the attempt to reproduce *and* replicate the findings from

Finally, one should mention that replicability is neither a necessary nor a sufficient criterion for good science. There are many areas of science (evolutionary biology, paleontology, geology,

climatology) which make empirical claims that cannot be replicated - simply because it is impossible to replicate dinsoaurs marching on earth or vulcanoes erupting, for example. (German Research Foundation, 2017; **?**)

(Lehrer, 2010)

We explain and define different forms of reproducibility and what they mean for research (additional reading:(Mogil, Macleod, and others 2017; Reichlin, Vogt, and Würbel 2016; Ioannidis 2005; Kass et al. 2016), Lazic)

(Begley 2013))

One important aspect of ensuring reproducible research are well-designed studies as well as detailed and complete reports on the experimental procedures, the study design, potential confounders, techniques, and methods used, chemicals, and drugs used, study protocols, statistical designs and further important elements that might be different between different fields of research. (see for example (Kass et al. 2016; Nature 2017; Ayris, Group, and others 2013; Smith et al. 2017))

2.3 On the ground validation of online diagnosis with Twitter and medical records

In their study, conducted during the 2012-2013 Influenza season, Bodnar *et al.* (2014) analysed the tweeting behaviour of a group of 104 students from the Pennsylvania State University, of which they also had received medical records from the university's health services, telling whether a participant was sick with the flu during a given month or not.

The researchers collected a total of 37'599 tweets from the 104 accounts mentioned above ("seed accounts") as well as 30'950'958 tweets from the 913'082 accounts that were connected to those seed accounts (either by following one of them or by being followed by one of them). They then proceeded to divide the tweets from the seed accounts into two categories: Tweets that were sent during a month in which the user was sick and the rest. This way, a total of 1609 tweets from 35 users could be extracted. Then, they screened the tweets in both categories for the occurrence or absence of a set of keywords flu, influenza, sick, cough, cold, medicine, fever. The predictive power of these seven words were then tested by applying five different classification methods (Naive Bayes, random forest, C4.5, logistic regression, and support vector machine), which showed that the keywords were only poor predictors of a Twitter user's disease

status (see Figure 2.1).

In a second step, the researchers also applied simple bag-of-words techniques to identify relevant keywords, namely by finding the 12'393 most common keywords, ranking them according to their predictive power with regard to Influenza and finally choosing the top 10, 100, or 1000 keywords on this ordered list. The predictive power of the keywords (and thus the ordering of the list) was calculated by classifying the users as either being "sick" or "non sick" in a given month and then comparing said classification with the real disease status. The classification was again done using the five methods mentioned above, whereas the Naive Bayes classifier performed best with a classification accuracy of 89.72% and an AUC of 0.8544 when using the top 100 keywords (see Figure 2.1).

Even when doing everything correctly, the right way to interpret and analyse statisical information is still a challenge (Reichlin, Vogt, and Würbel 2016) . Scientists must overcome dichotomous thinking when it comes to analysing their data. No experiment is completely successful or unsuccessful, there are always shades of grey (see (Amrhein, Korner-Nievergelt, and Roth 2017; McShane and Gal 2016; Wasserstein and Lazar 2016; Colquhoun 2014; Ziliak 2016a; Ioannidis 2016; IntHout, Ioannidis, and Borm 2016; Benjamin et al. 2017)).

Participants learn to differentiate between Bayesian and frequentist type of thinking and learn their importance for reproducibility (see also (Christensen 2005)) Participants learn different ways to report and represent their data that improve analysis and prevent false interpretations (e.g. see the p-value you can't buy, the earth is flat, Participants learn how they can improve the interpretations and analysis of their experiments in their reports (see also ((Ziliak 2016b) Participants learn how their study design and experimental setup directly influences their statistical and scientific interpretation of the results (see also (Gelman 2016) Participants understand the differences between accuracy, specificity and sensitivity and its implications on reproducibility (see also (Baratloo et al. 2015)

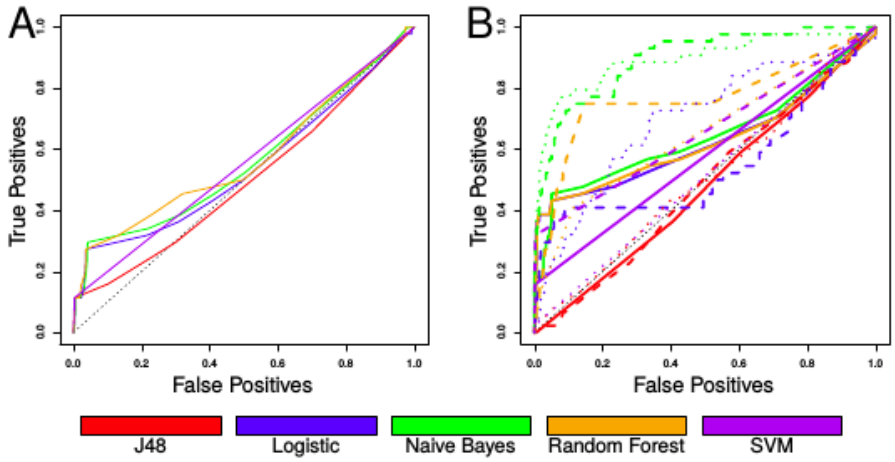


Figure 2.1: The ROC of classifiers that use hand chosen key words (a) and algorithmically chosen keywords (b) to determine if an individual is ill. The top 10 (solid line), 100 (dashed line), and 1000 (dotted line) were selected as features (taken from Bodnar *et al.* (2014))

This model has its limitations as well, though: It was based on a very small data set consisting of only 104 twitter accounts generating a total of 37'599 tweets during the study period. 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 use. In addition, the exact time of disease of the twitter users is not know. Due to privacy concerns, the researchers were limited to know in which month a specific twitter user was diagnosed with the flu. Finally, the model was built based on the tweeting behaviour and medical records of only one flu season (2011-2012). All these points either reduce the models' temporal resolution or add a considerable amount of bias to it.

Hence, it is necessary to test the performance of the Naive Bayes classifier describe above for different cities and states and compare the results with reliable epidemiological data. In his dissertation, Bodnar (2015) performed such tests on the level of counties, states, and the complete US mainland.

To do so, he applied the classifier to the tweets of each user within a 4-week sliding window

with a one-day step-size. "The classifier assigns score (sic!) to the day where the sliding window begins based on the Tweets the user has posted within the window. For example, when the sliding window first encounters a user's Tweet that says 'I am getting sick with the flu', the classifier will heavily lean toward her being sick. Later, the user may Tweet 'I am no longer sick' which will give a strong signal that the user is no longer sick which will tend to outweigh the user's previous "sick" Tweet even if they both occur in the same window. Of course, it is rare that such strong signals are in the data, so the classifier is built on an amalgamation of many weaker signals—mentioning going to a party asnot-sick signal, for example—which, while weaker, are more prevalent. We chosestep size of one day in order to increase the temporal granularity of the classifier. Users that are inactive for more than 30 days are not included for any analysis during that time window."[1]

Bodnar applied the algorithm in the above-described fashion to a set of 15'560'328 users who sent a total of 2'732'174'105 geotagged tweets between March 3^{rd} 2011 to March 4^{th} 2015.[2] In a next step, Bodnar created a Kermack-McKendrick SIR model based on the results of the classifier (Martcheva, 2015):

$$\frac{dS}{dt} = -SI\beta, \qquad \frac{dI}{dt} = -SI\beta - I\gamma, \qquad \frac{dR}{dt} = I\gamma.$$

S, I, and R represent the relative frequencies of susceptible, exposed, and recovered individuals, respectively, whereas β and γ are the transition probabilites from being susceptible to having

[1]Note that this way of assigning diseas status has one problematic implication, namely that information from future tweets is incorporated while information from past tweets are ignored. This insofar counterintuitive as a Twitter user who tweeted about being sick on day 1, will most certainly still be sick on day 2 - regardless of what she tweets about. However, the algortihm will not take into consideration any information from past days, so when shifting the onset of the 4-week-sliding window to day 2, all information from day 1 will be lost. If there are no additional flu keywords to be found in the tweets sent within the following 4-week-sliding window, the algorithm will mistakenly classify the user as being "non sick". In addition, the algorithm is bound to classify tweets too early as being sick. For example, if particularly strong "flu" signal turns up, the algorithm might classify the user as being "sick" up to four weeks before the actual onset of the sickness. In other words; instead of assuming that a user remains sick for some time after she has tweeted about being sick, the algorithm might give the faulty impression that the user has been sick up to the point where she tweeted about it - but not afterwards. There are no indications in the data set that something like this has happened, but since the overall number of users identified as being sick at any given point during the whole study time is very low, this does not mean much.

[2]Note that tweets from users who tweeted less than 10 times during this period as well as tweets that could not be attributed to a specific state on the US mainland were discarded. However, the total number of tweets analysed was not given in any of the documentations available to me.

the disease and from having the disease to recovering from it, respectively. To determine the values of β and γ, Bodnar fitted the SIR model based on the data from the Twitter classifier to the official ILI results from the Center for Disease Control (CDC) in Atlanta using a multi-grid search method. Values for β and γ were chosen such that the corresponding SIR model minimised the residual sum of squares (RSS):

$$\text{RSS} = \sum_t (I_{\gamma,\beta}(t) - I_{\text{CDC}}(t)).$$

Here, t denotes the time in weeks, while I denotes the percentage of people showing ILI symptoms based on the SIR model ($I_{\gamma,\beta}$ and the official CDC data (I_{CDC}), respectively. By doing so, he calculated the optimal values of β and γ for each flu season as well as for the whole study period combined (see Table 2.1). Note that in Bodnar (2015) it is written that an optimal $S(0)$ was also estimated using the multi-grid search method. However, in the R-code provided to me $S(0)$ was simply defined as $1 - I(0)$, where $I(0)$ is the relative number of Twitter users classified as "sick" during the first week of the time window the SIR-model was built on.

Table 2.1: National best-fit parameters for each year from the CDC's data (white) and Twitter data (grey); taken from Bodnar (2015)

Year	γ	β	RSS
2011-2012	0.1732	0.1749	0.0001047
	0.1176	0.1195	0.0001323
2012-2013	0.7715	0.9626	0.0009402
	0.7317	0.9020	0.0009492
2013-2014	0.6054	0.7288	0.0003114
	0.6046	0.7264	0.0003026
Combined	0.6998	0.8225	0.003719
	0.6765	0.7935	0.003252

Based on these values he could then calculate yearly ILI estimates for the flu seasons of 2011–2012, 2012–2013, and 2013–2014 (see Figure 2.2)

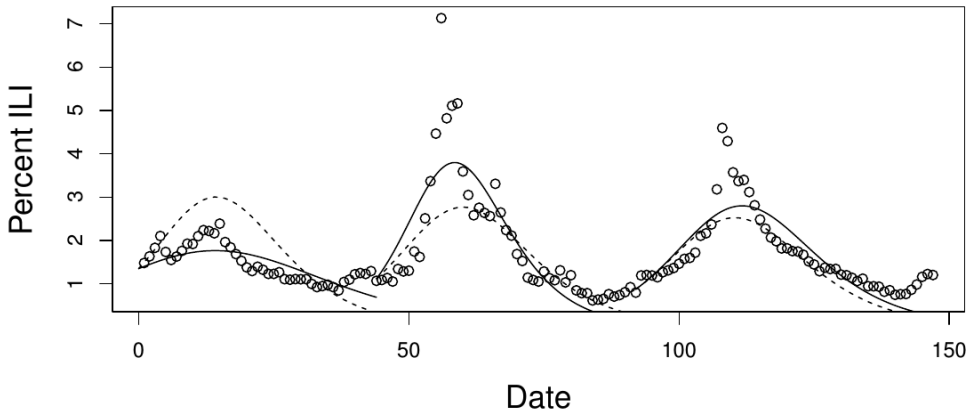


Figure 2.2: The CDC's estimates (circles) of influenza rates for a three year period compared to the best fit SIR models from the Twitter data using combined (dashed line) or yearly (solid line) parameters (taken from Bodnar (2015))

In addition, he fitted a autocorrelation model using the results from the Twitter base model and the official CDC data:

$$I_{\text{full}}(t+1) = a \cdot I_{\text{CDC}}(t-1) + b \cdot I_{\text{CDC}}(t) + c \cdot I_{\text{Twitter}}(t) + d.$$

Here, t denotes the time in weeks, I_{CDC} depicts the official ILI percentages from the CDC (lagged by two and one weeks, respectively), and I_{Twitter} denotes the ILI percentages received from the Twitter base model. [3]

Applying the above-described autocorrelation model, Bodnar was able to achieve a very close fit to the official Twitter data (see Figure 2.3).This Master thesis is the (unsuccessful) attempt to reproduce these results.

[3]Personal communication by Todd Bodnar. Note that it is unclear whether the "Twitter base model" consists of the raw results from the Twitter classifier or consists of these raw results combined with additional information. An e-mail requesting clarification of this has been sent on June 24th, with three more follow-up emails sent in June and August. However, a clear description of how the Twitter base model was calculated is still missing. Also, I am still missing the values for a, b, and c, since I only received a file with the model results, but not the paramater specifications.

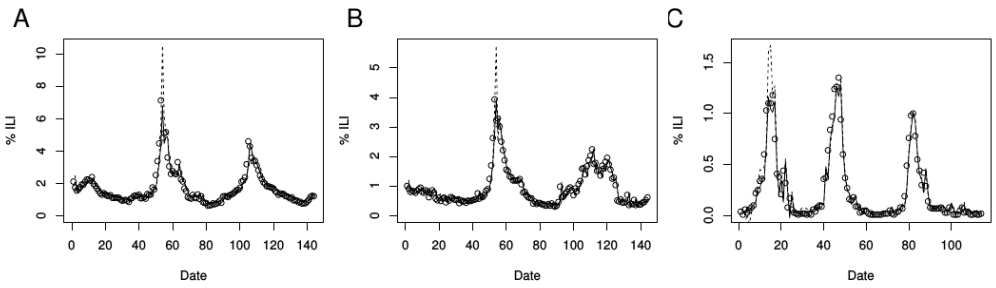


Figure 2.3: Comparison of Twitter's forecasting (dashed lines) and retroactive measurements (solid lines) to the CDC's reported Influneza rates (circles) for national (A), HHS Region 1 (B), and Seattle area (C) (taken from (Bodnar, 2015))

Chapter 3

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.

3.1 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 (**?**).

3.2 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 3.1a)

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 3.1b). 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 3.1b).

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 3.2) 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 3.3). 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 3.4.

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 3.3) 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 data set aggregated over weeks.

3.3 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 section 3.2). Since tweets with state code "56" predominantly stemmed from the Mexico and Canada or other areas outside the U.S. mainland (see Figure 3.1a), 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 19 (March 2011) to 30 (December 2011) million of 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 3.5), 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.

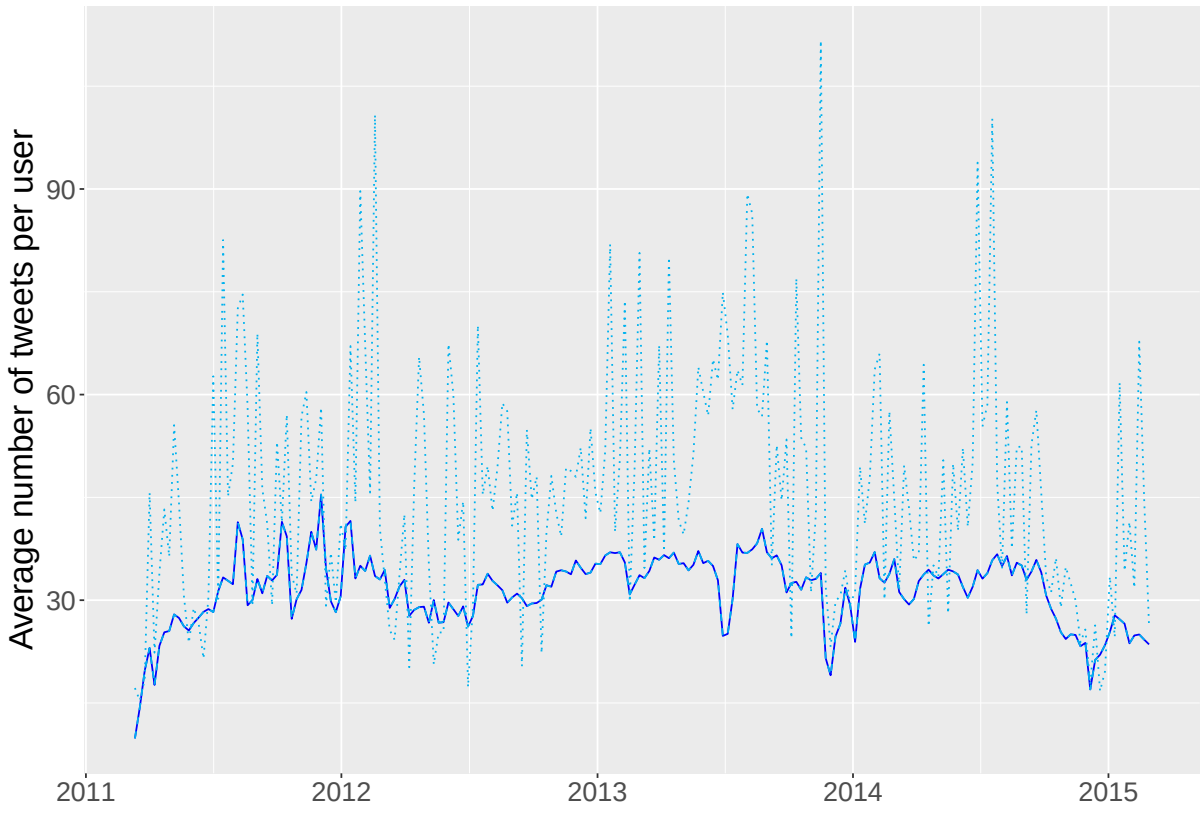
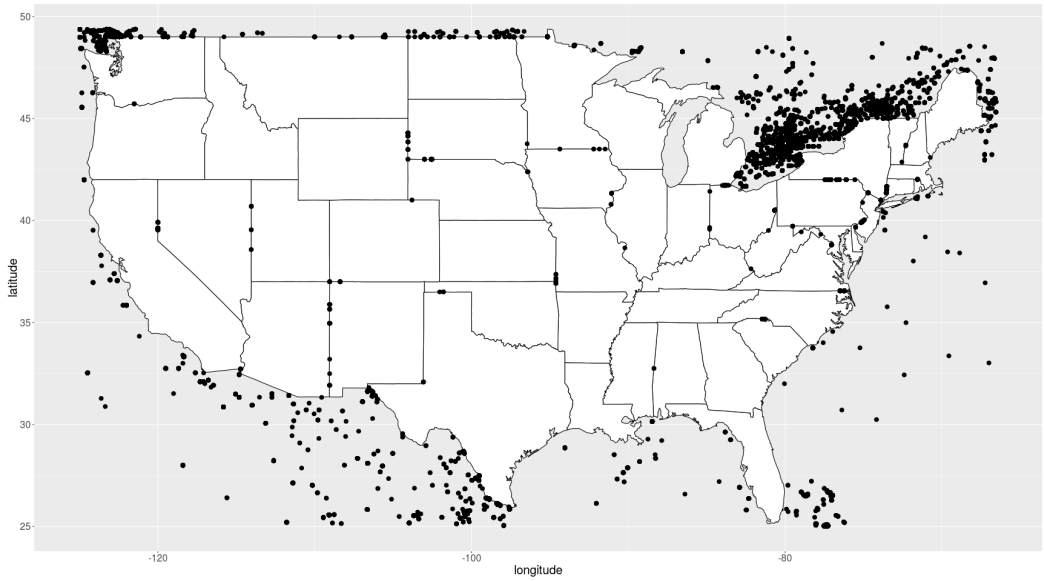
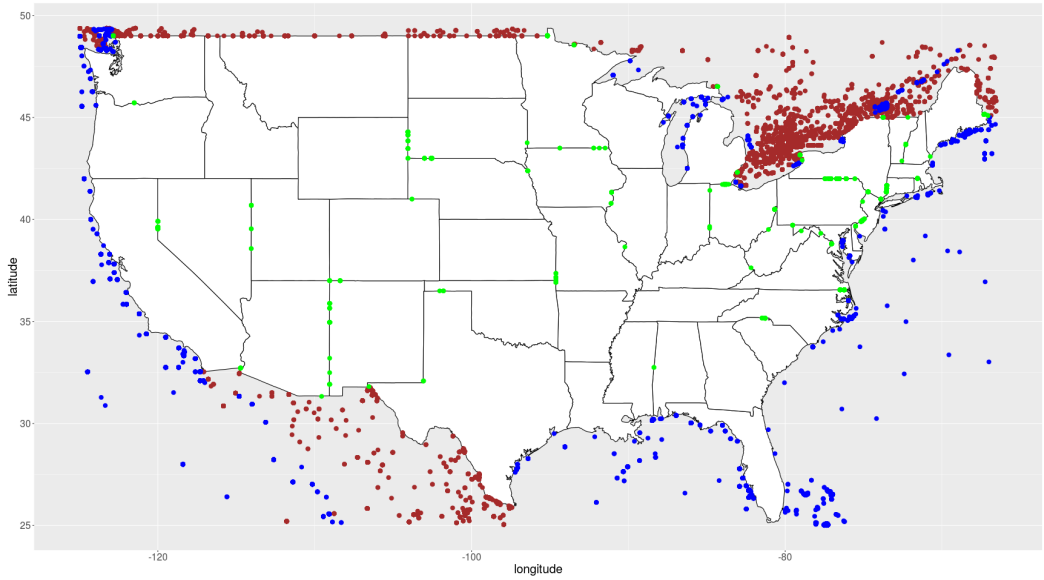


Figure 3.5: The average number of tweets per week sent by sick users (dotted light blue), healthy user (dashed light blue) and total users (solid blue). The average weekly tweet rate of sick users is significantly higher, while the tweet rate of healthy users is virtually indistinguishable from the total weekly tweet rate.



(a)



(b)

Figure 3.1: (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.

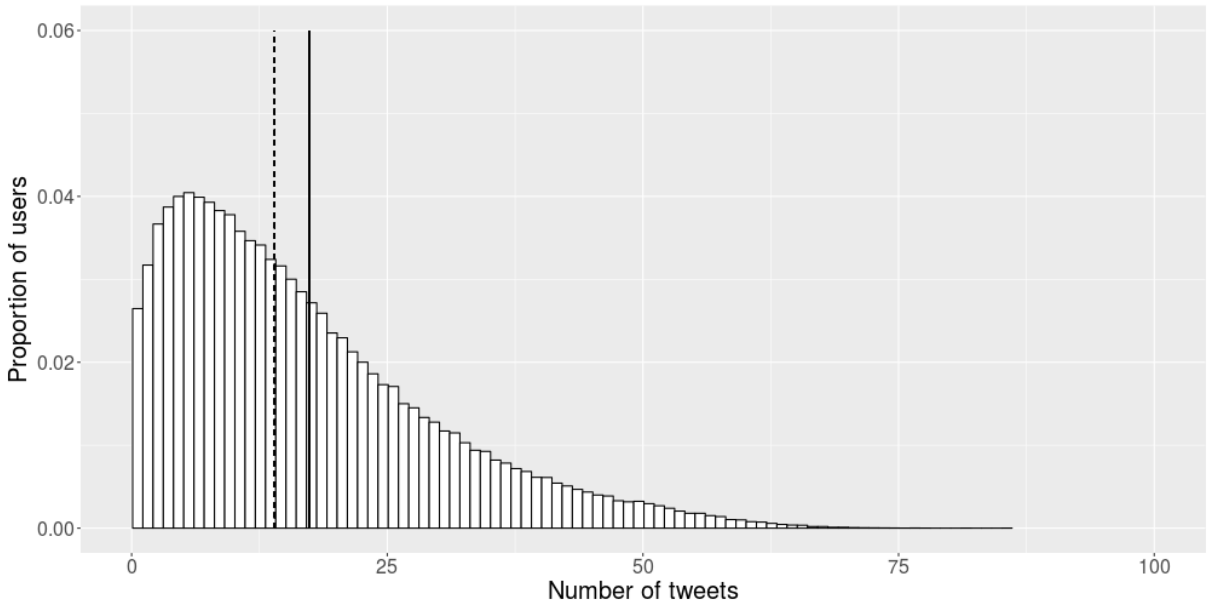
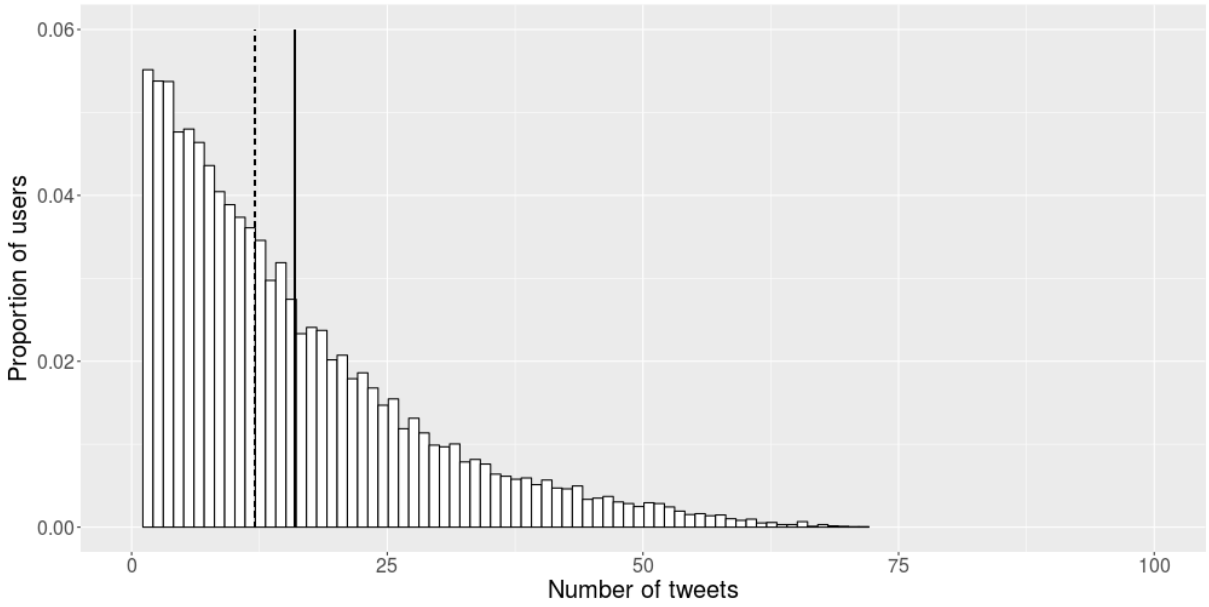
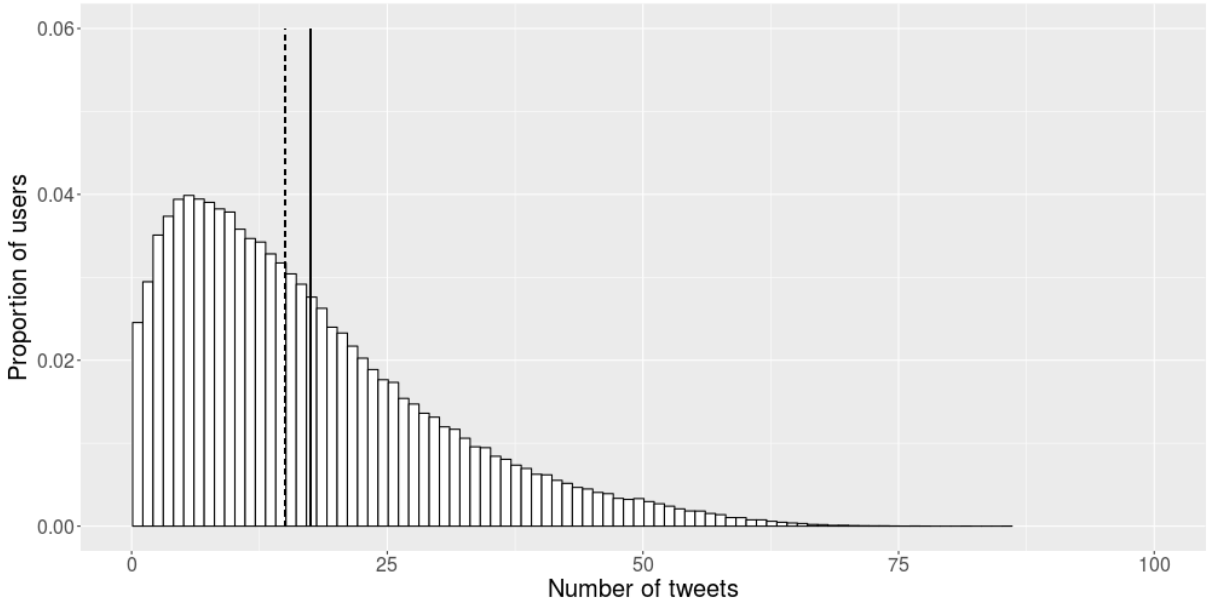


Figure 3.2: Histogram of the number of tweets sent per user in the *sick_user* data set during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1). As can be seen, many users only sent a handful of tweets during this time, whereas the user with the highest tweet activity sent 86 tweets. Mean = 17.3221116 (solid line); median = 14 (dashed line)

Figure 3.3: 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 3.4: 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")

Chapter 4

Results

Figure 4.1 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 4.2 for a side-by-side comparison of both the uncorrected and corrected weekly tweet activity.

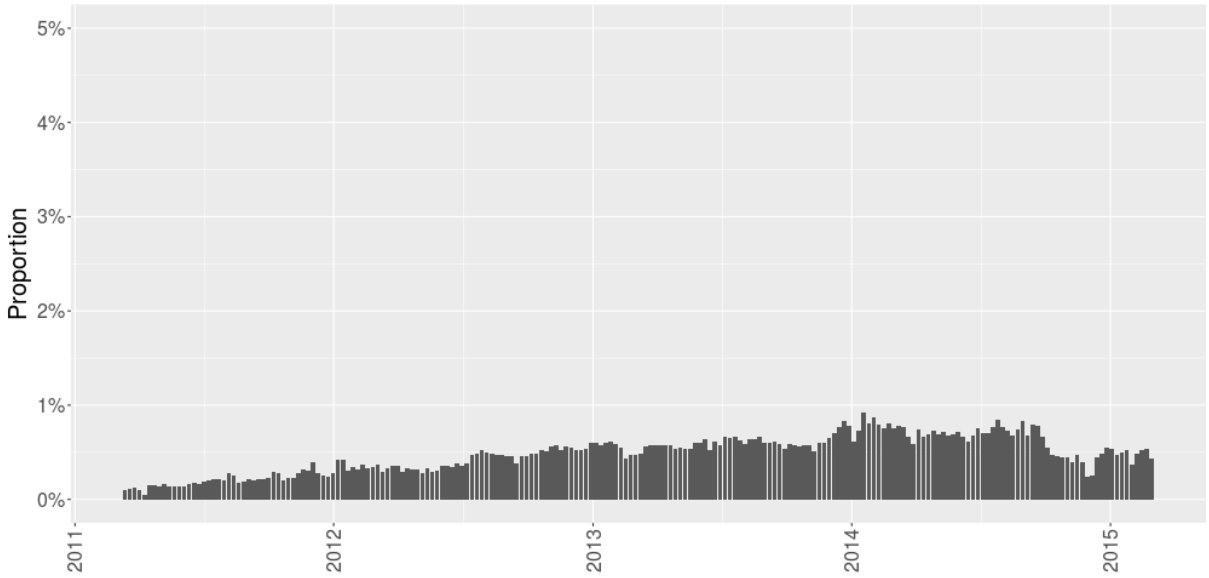
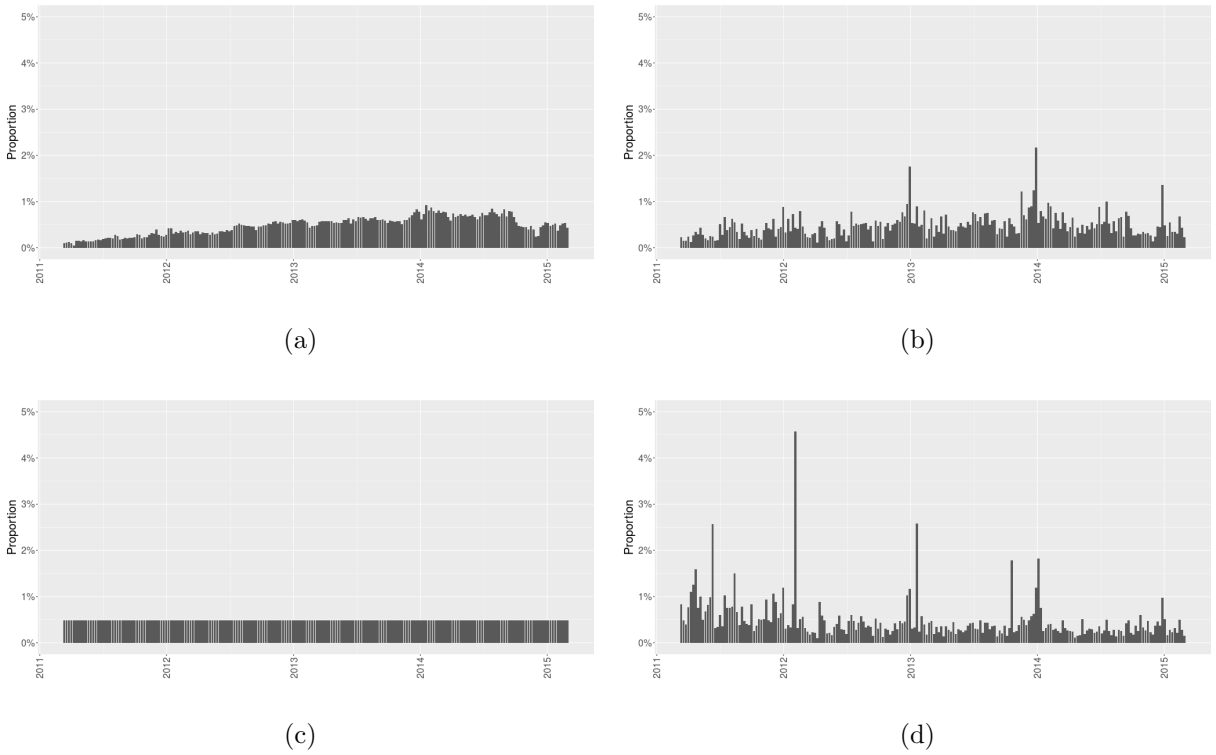


Figure 4.1: Relative number of tweets sent per week in the *all_tweets* data set between 2011-03-05 and 2015-07-11 (bin size = 1 week).



(a) (b)

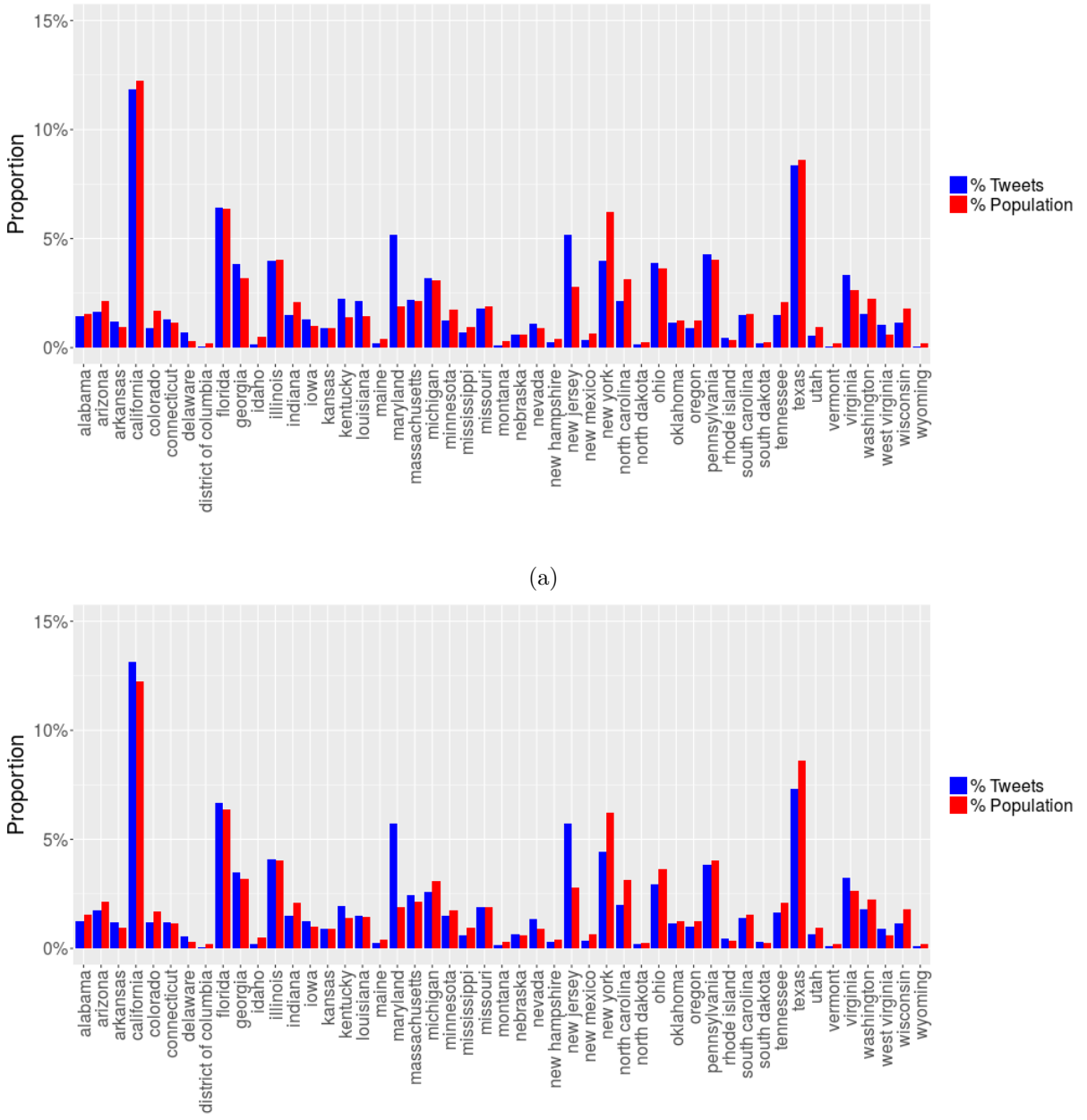
(c) (d)

Figure 4.2: 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.

Next, I looked at the total amount of tweets sent in each state. As can be seen in Figure 4.3a and in Figure 4.4a, 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 4.5) 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

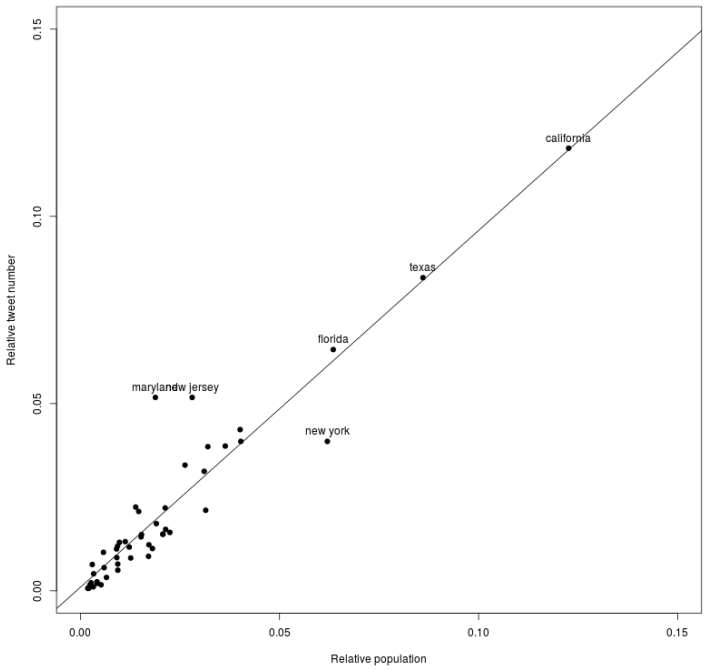
users instead of number tweets yields similar results (Figure 4.3b, Figure 4.4b, Figure 4.6; p = 0.0822911)



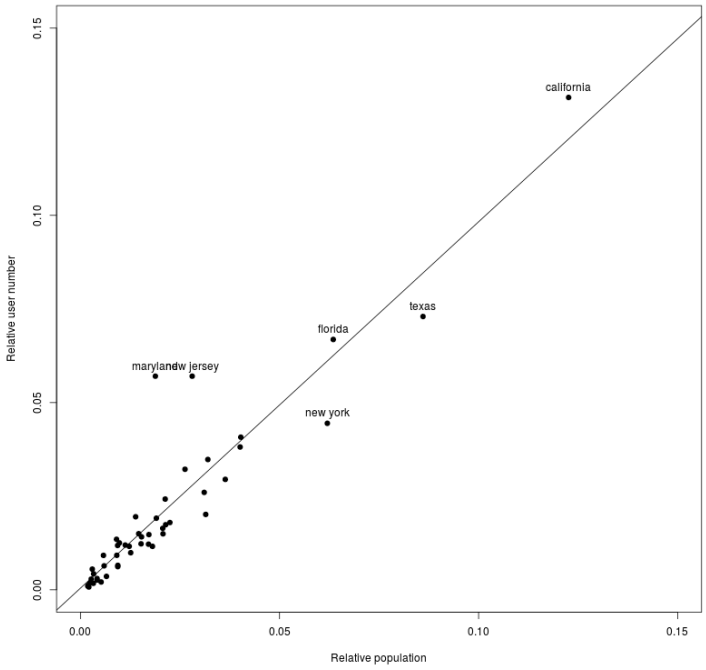
(a)

(b)

Figure 4.3: 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.

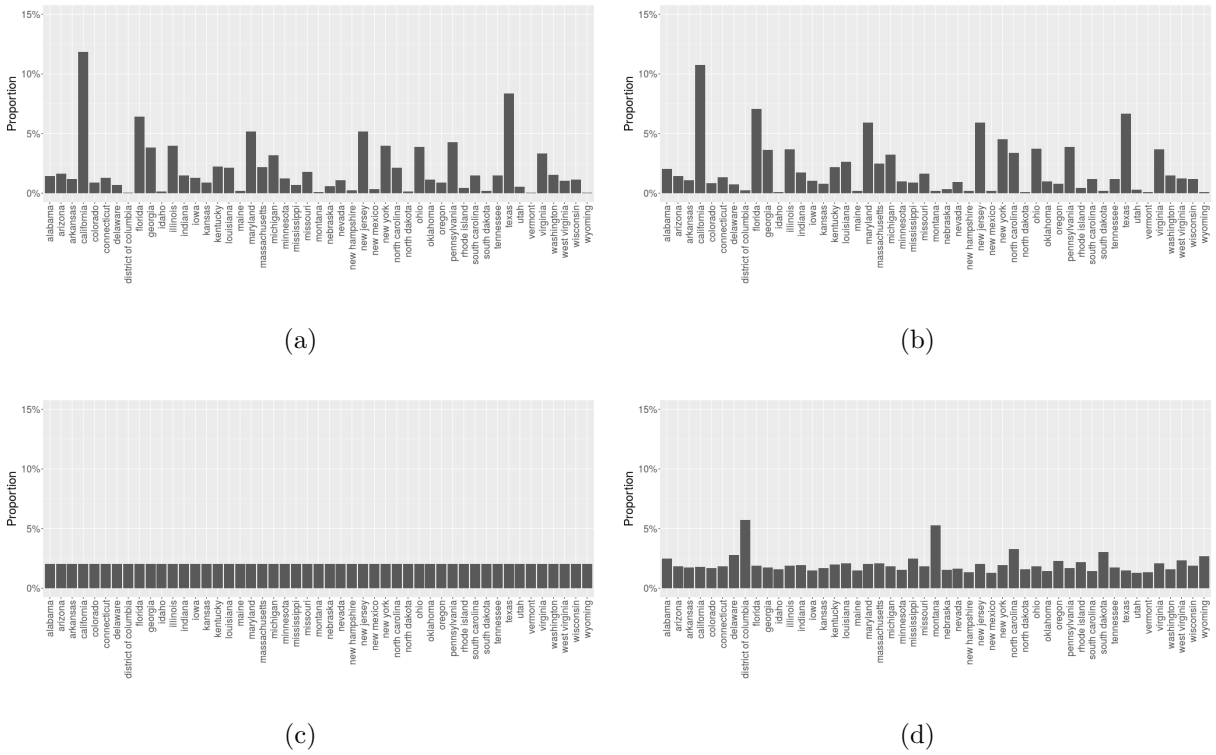


(a)



(b)

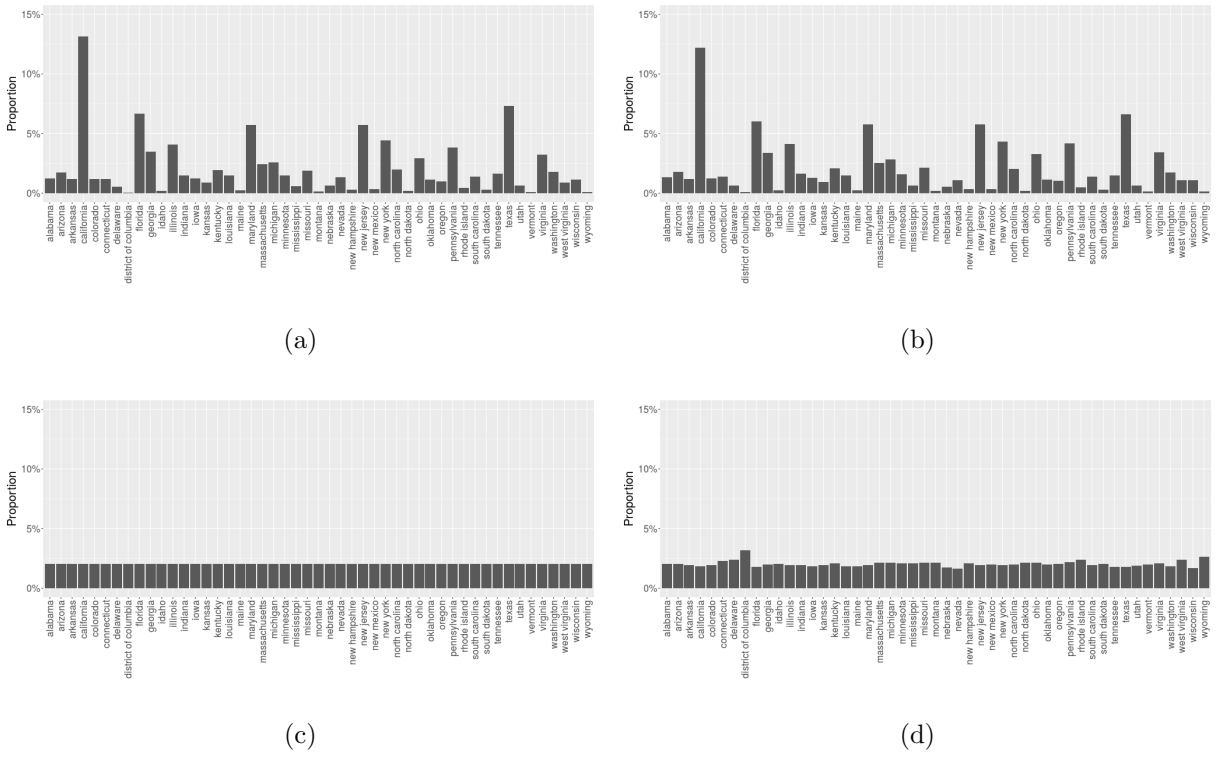
Figure 4.4: Relative number of tweets sent per state (a) and Twitter users (b) in the *all_tweets* data set between 2011-03-05 and 2015-07-11 plotted against each state's relative population size



(a) (b)

(c) (d)

Figure 4.5: 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.



(a) (b)

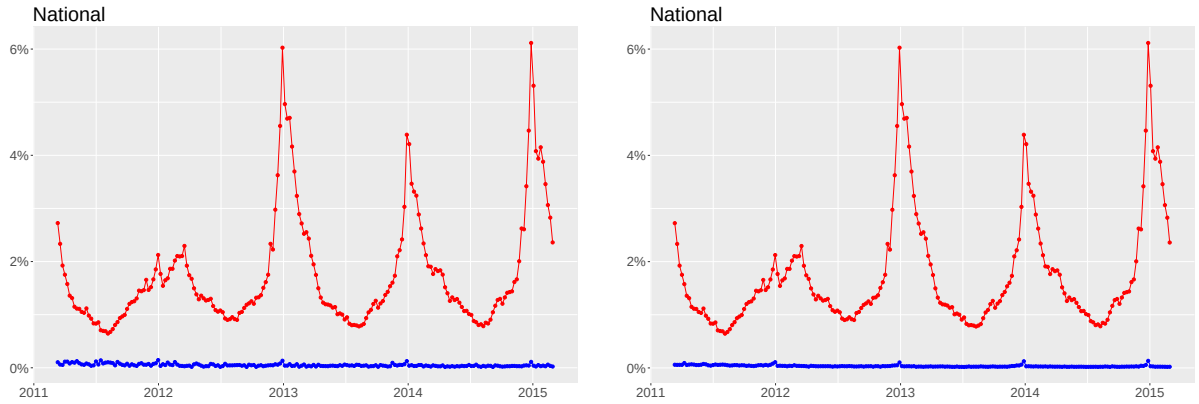
(c) (d)

Figure 4.6: 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.

4.1 Comparison with CDC data

In order to assess the validity of the ILI predictions provided by the flu classifier, I compared the the results from the Twitter classifier with the official ILI reports from the CDC on the national, regional and state level(extracted using the "cdcfluview" package (**?**)).

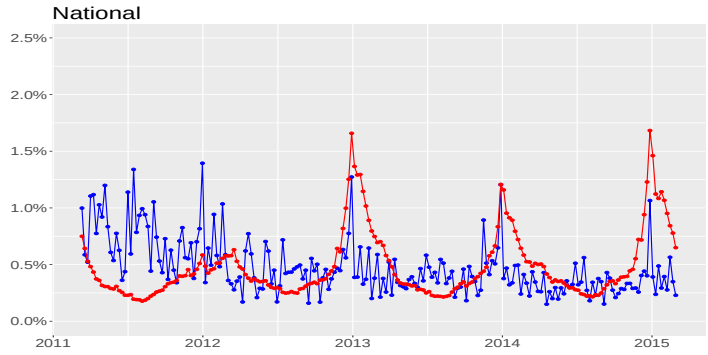
In a first step, I simply compared the official CDC ILI percentage data on the national level with the the relative number of tweets labelled as "1 = sick" per week and the relative number of sick users per week, respectively. As can be seen from Figure 4.7, the relative results from the Twitter classifer are an order of magnitude smaller than the official CDC data.



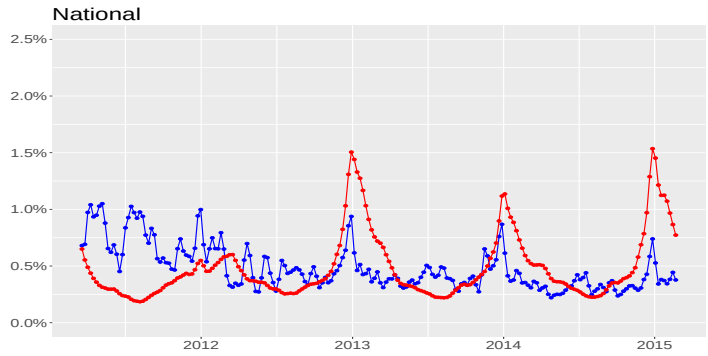
(a) (b)

Figure 4.7: Comparison between weekly CDC ILI rates (red) and the results from the Twitter classifier (blue). (a) shows the relative amount of tweets labelled as sick during a given week. (b) shows the relative amount of users labelled as "sick" by the classifier. As can be seen from both figures, the relative results from the Twitter classifer are an order of magnitude smaller than the official CDC data

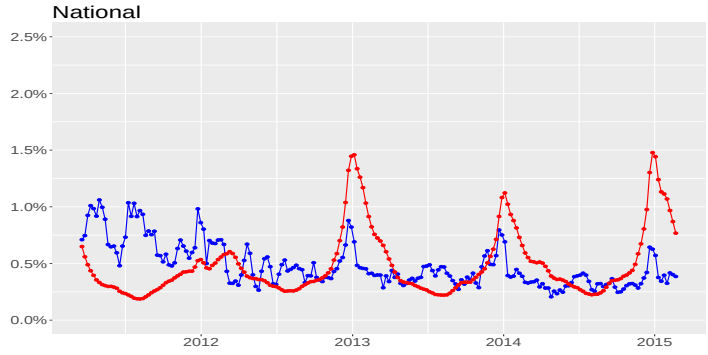
In order to make them directly comparable to each other, I normalised both time series by the total sum of relative tweets numbers and ILI percentages, respectively. Hence, the percentual values shown in Figure 4.8a do **not** represent weekly ILI percentage, but rather the percentual proportion of the 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 4.8b) and four-week (Figure 4.8c) 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 4.11). 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 Figures 4.10 and 4.11. 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).



(a)



(b)



(c)

Figure 4.8: 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

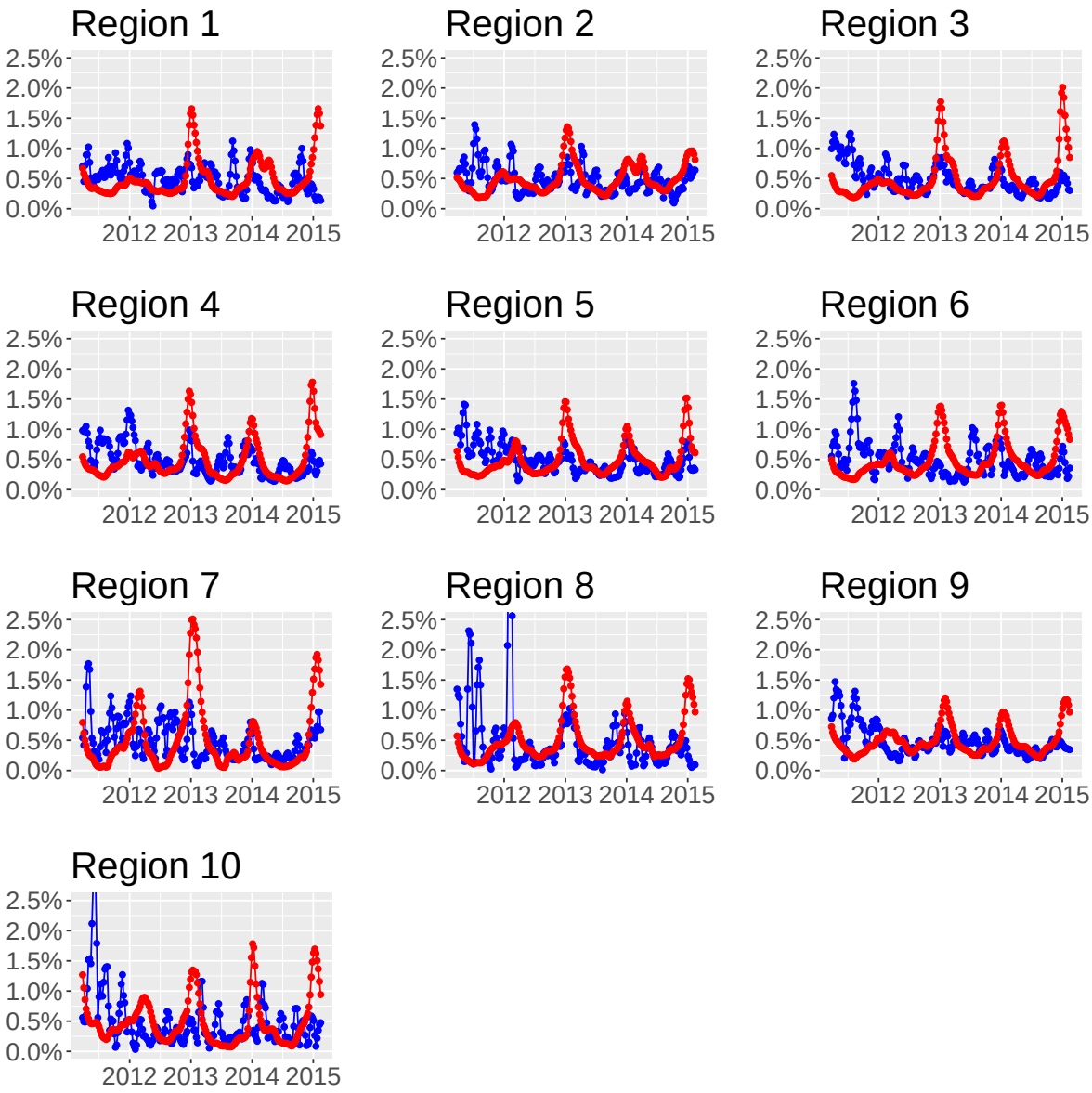
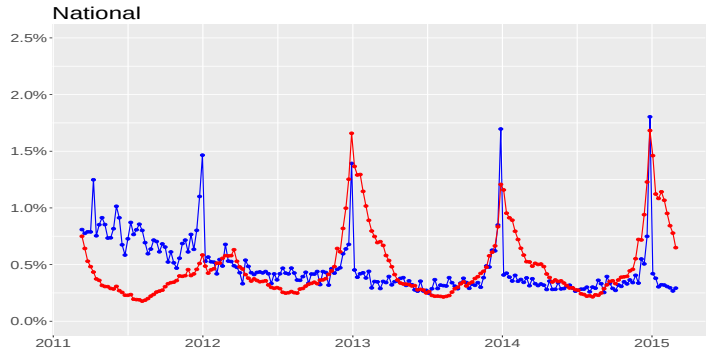
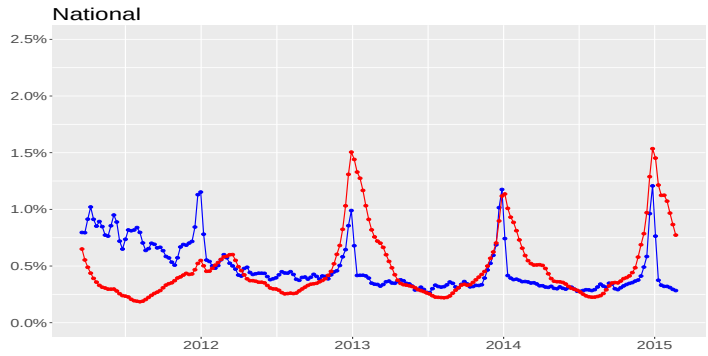


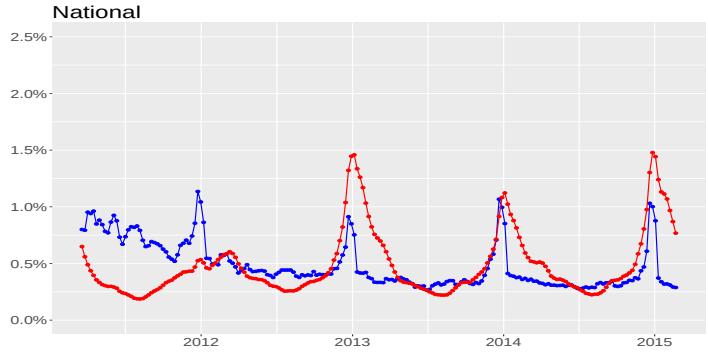
Figure 4.9: Relative number of tweets sent within each CDC ILI surveillance region per week (blue) compared with weekly ILI percentages in those regions (red). Data has been normalised and processed with a four-week moving average smoother. Note that Region 2 contains "Puerto Rico" and "Virgin Islands", Region 9 contains "Hawaii" and Region 10 contains "Alaska", all of which are missing from the Twitter data set.



(a)



(b)



(c)

Figure 4.10: Comparison between weekly CDC ILI rates (red) and the relative amount of users who 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

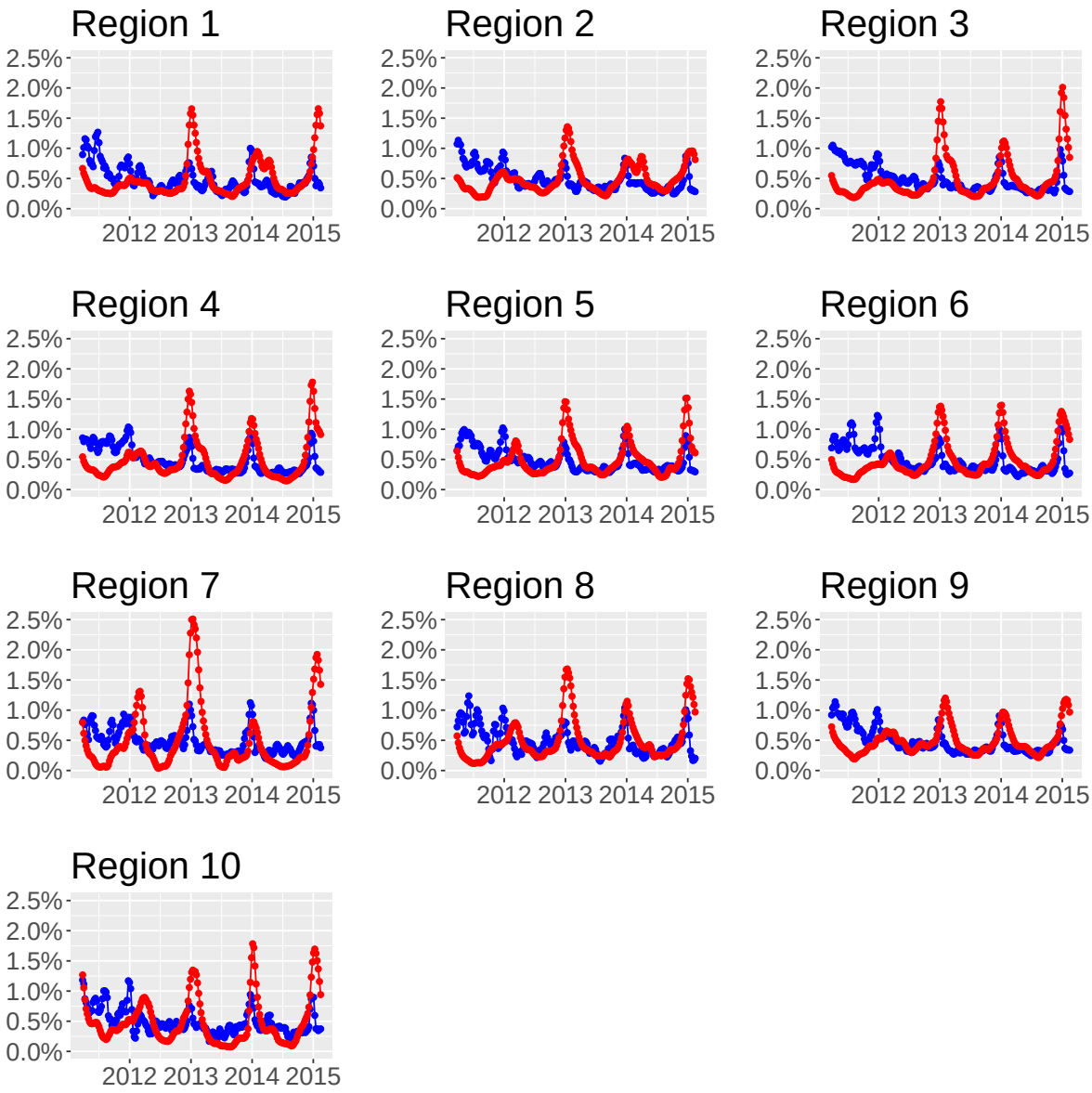


Figure 4.11: Relative number of users who sent at least one tweet labelled as "sick" within each CDC ILI surveillance region per week (blue) compared with weekly ILI percentages in those regions (red). Data has been normalised and processed with a four-week moving average smoother. Note that Region 2 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.

4.2 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

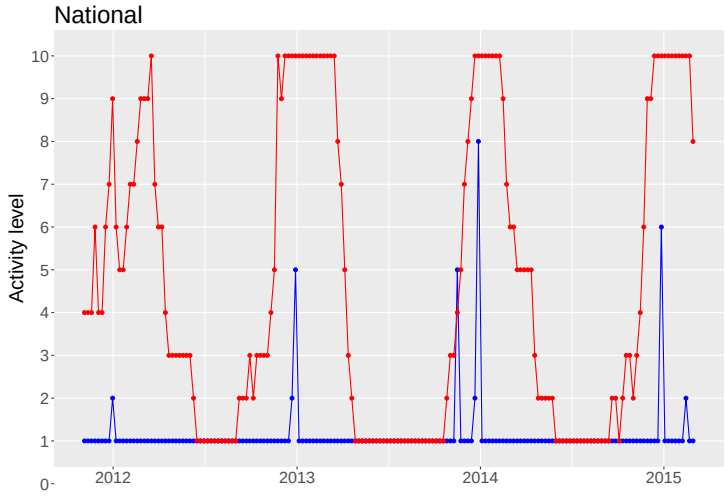
grouping used for reporting.

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. 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 (Centers for Disease Control and Prevention, 2016).

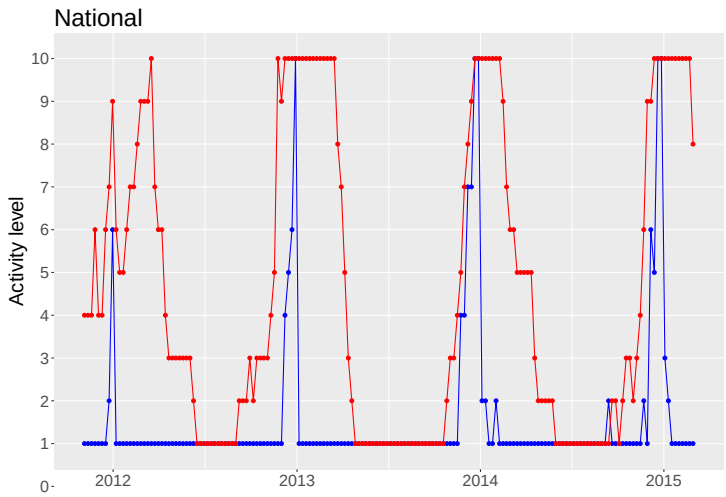
An activity level of 1 corresponds to values that are below the baseline, level 2 corresponds to an ILI percentage less than 1 standard deviation above the baseline, level 3 corresponds to ILI more than 1, but less than 2 standard deviations above the baseline, and so on, with an activity level of 10 corresponding to ILI 8 or more standard deviations above the baseline (Centers for Disease Control and Prevention, 2016).

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 4.12a and Figure 4.13 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 4.12b and Figures 4.14)[1].

[1]Note that the Twitter data available to me only spanned the time period between 2011 and 2015. In order to data comparable with the official CDC data, I only calculated the baseline based on the off-season weeks directly preceding a specific flu season (as opposed to calculating the baseline based on the off-season weeks of the three preceding weeks)



(a)



(b)

Figure 4.12: 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

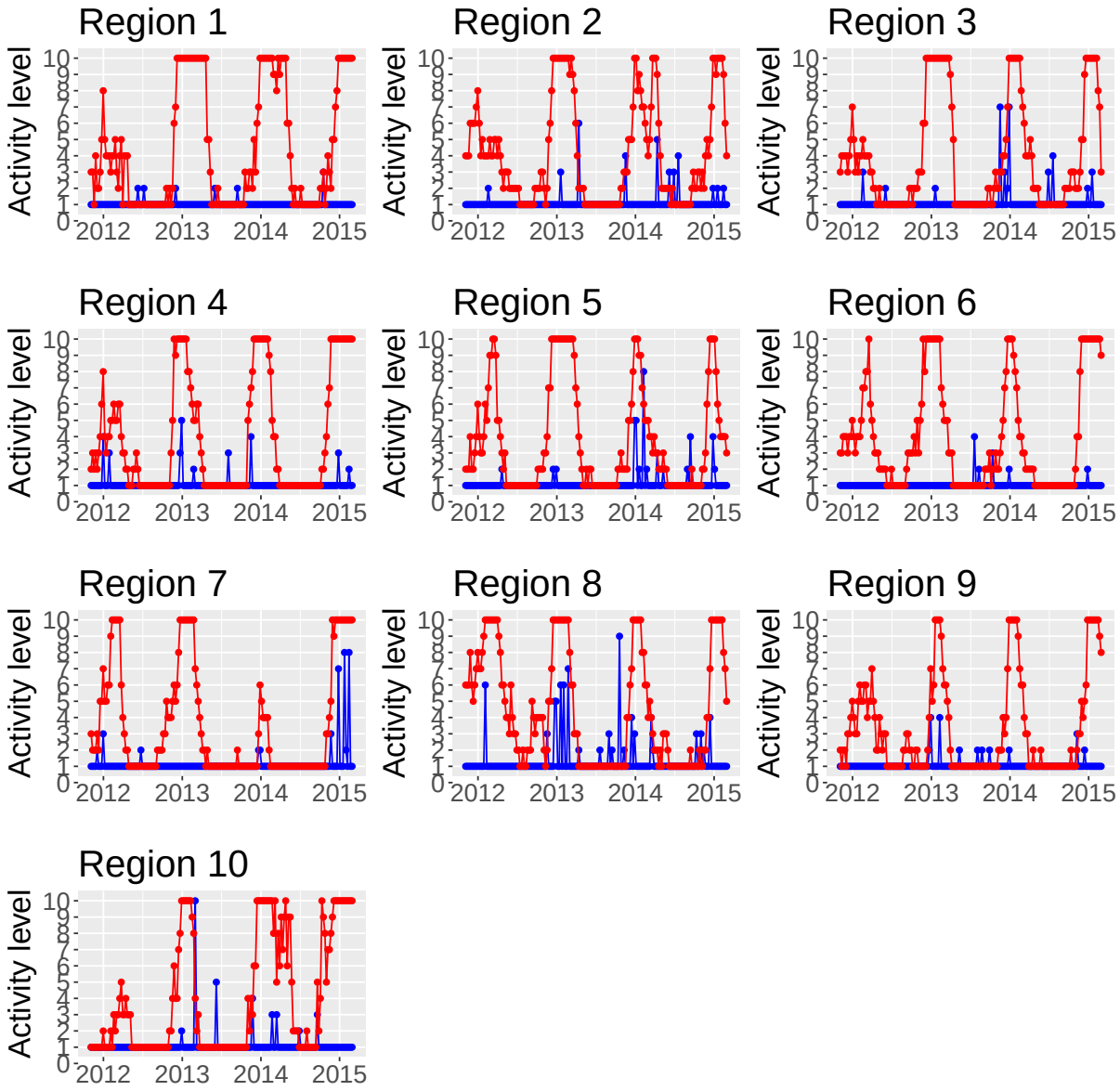


Figure 4.13: Comparison between regional weekly ILI activity levels reported by the CDC and calculated based on relative number of sick tweets per week

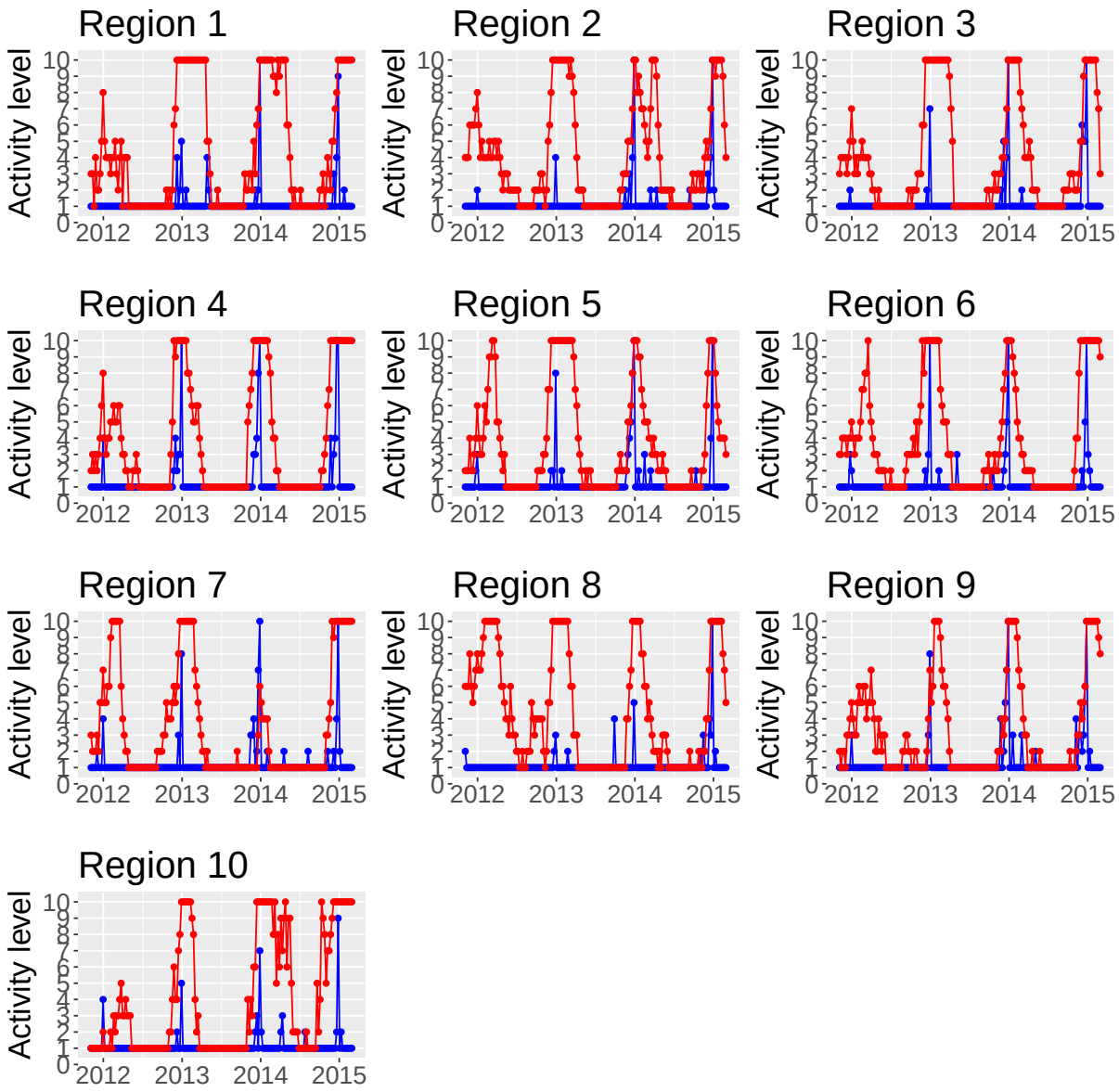
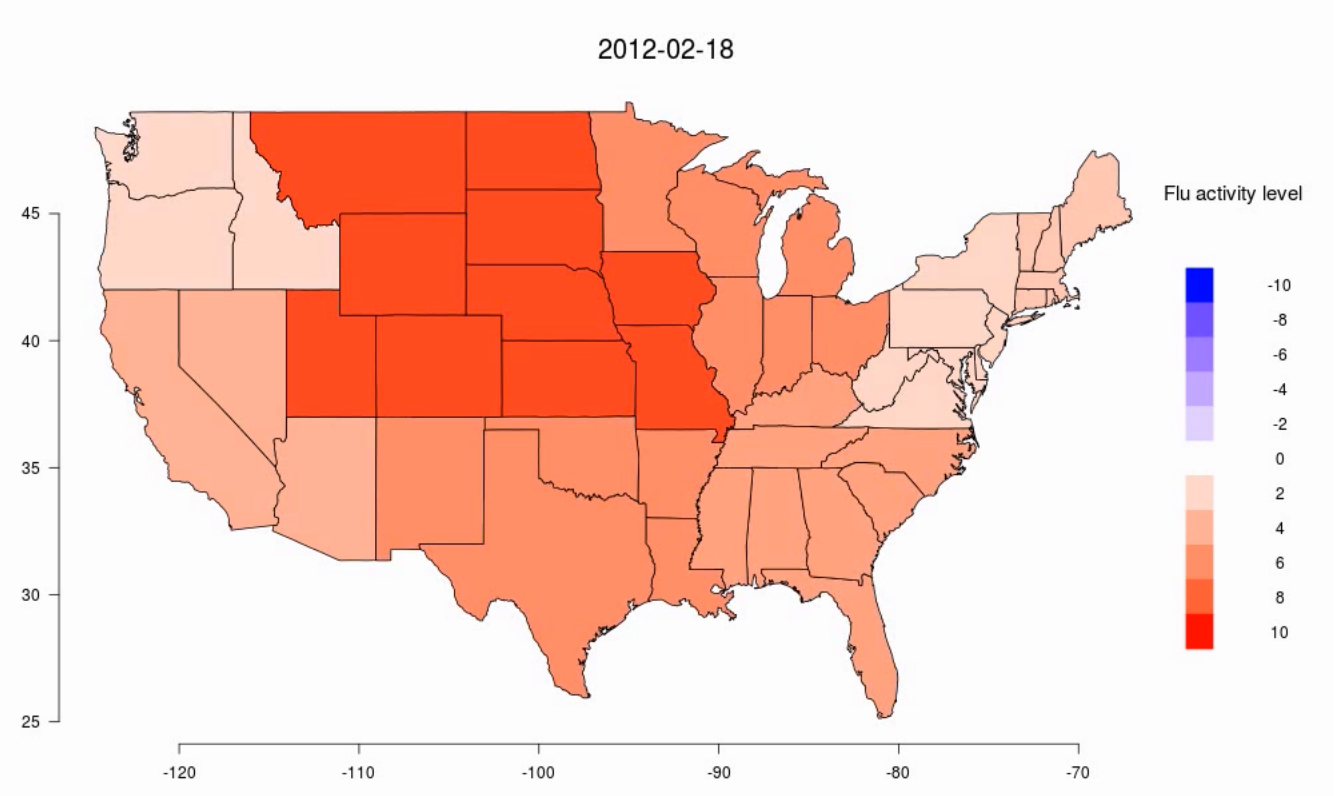
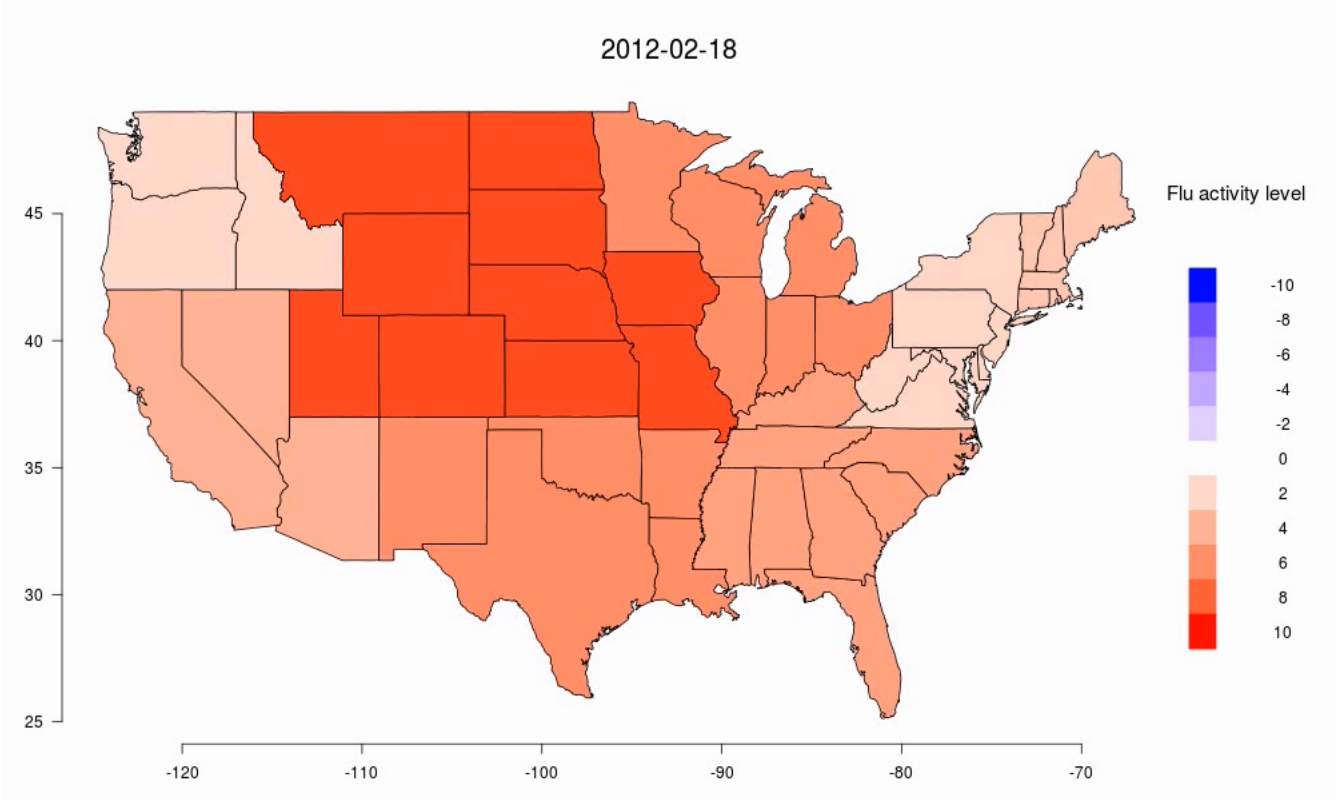


Figure 4.14: Comparison between regional weekly ILI activity levels reported by the CDC and calculated based on the relative number of sick users per week

To get a better understanding of the spatio-temporal pattern of ILI activity levels, I additionally built a 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

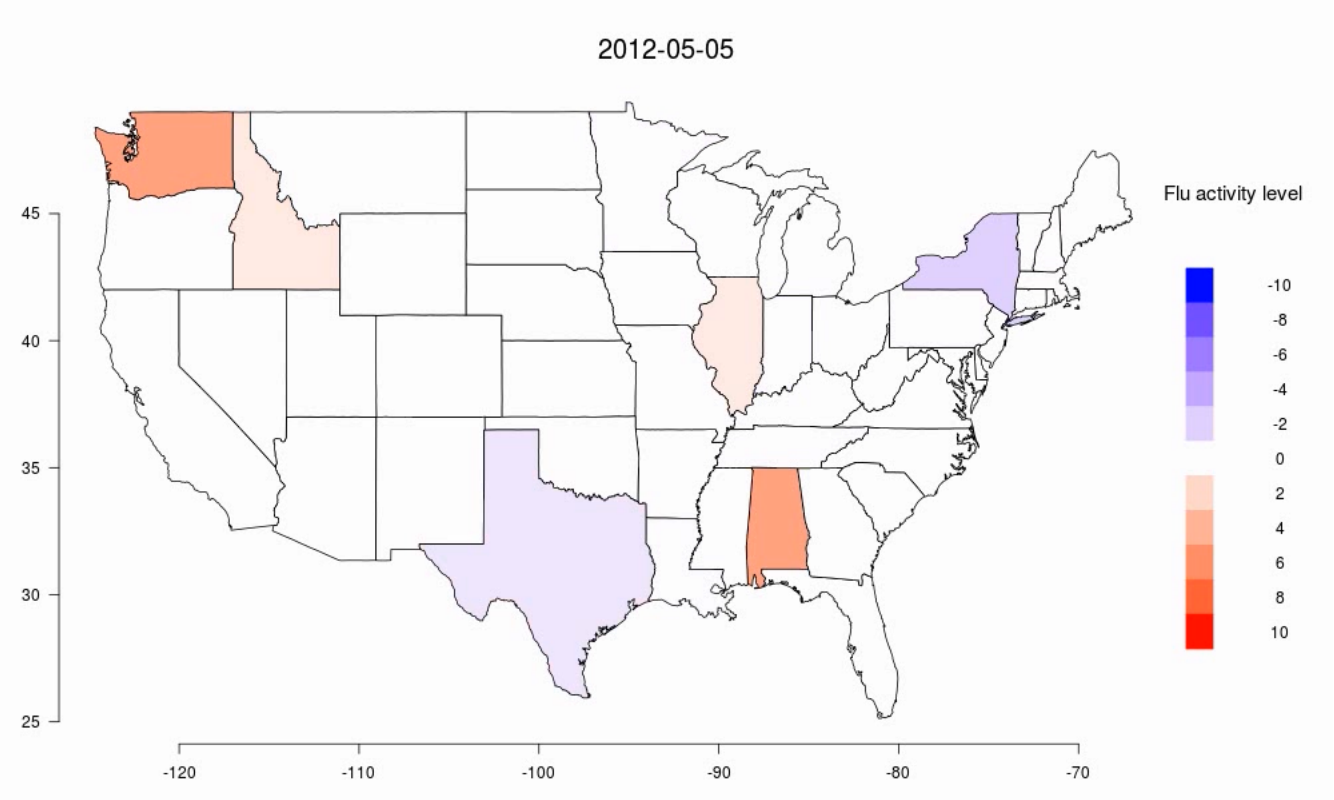
activity levels in a given state than those calculated from the Twitter data set, blue indicates the opposite. As can be seen, the Twitter ILI classifier hardly ever manages to emulate the CDC activity levels and when it does, it mainly happens during off-season weeks.

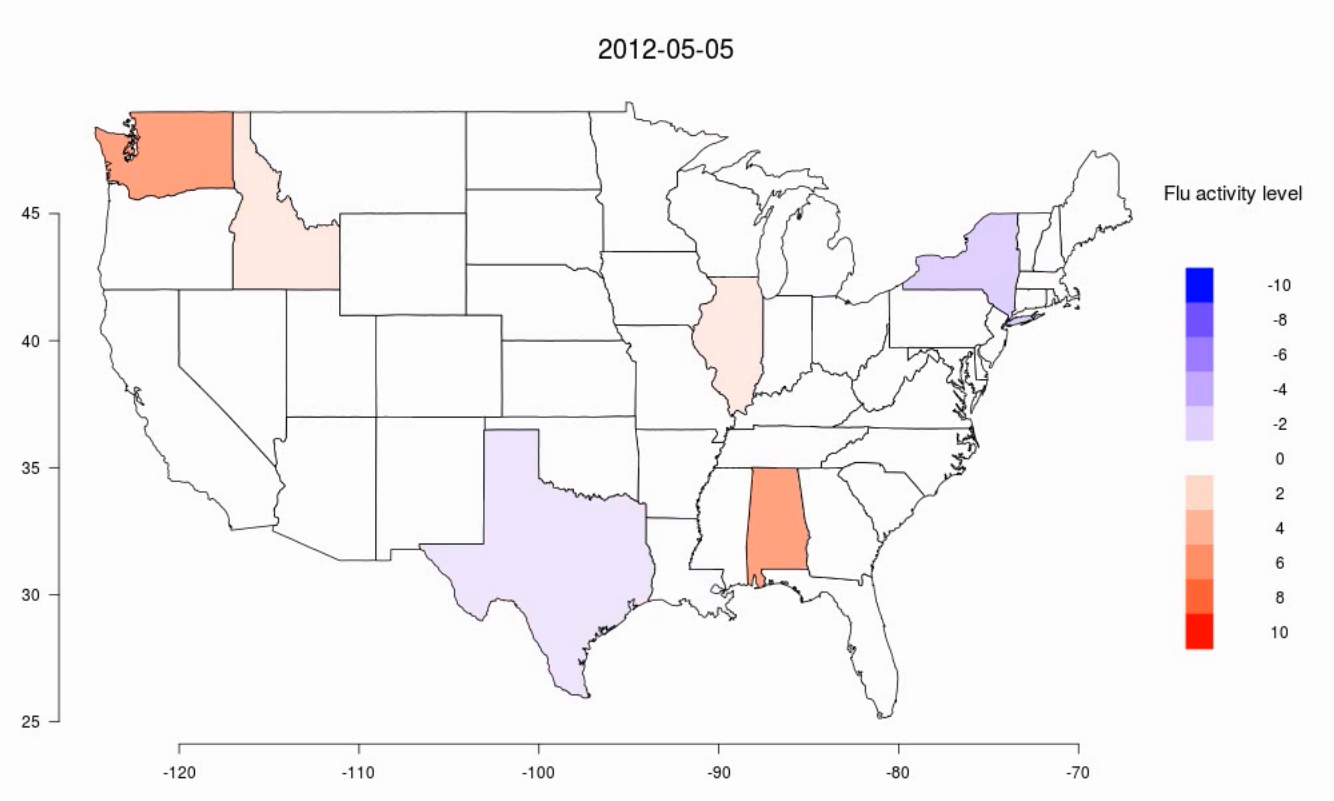




4.3 Comparison on the state level

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.

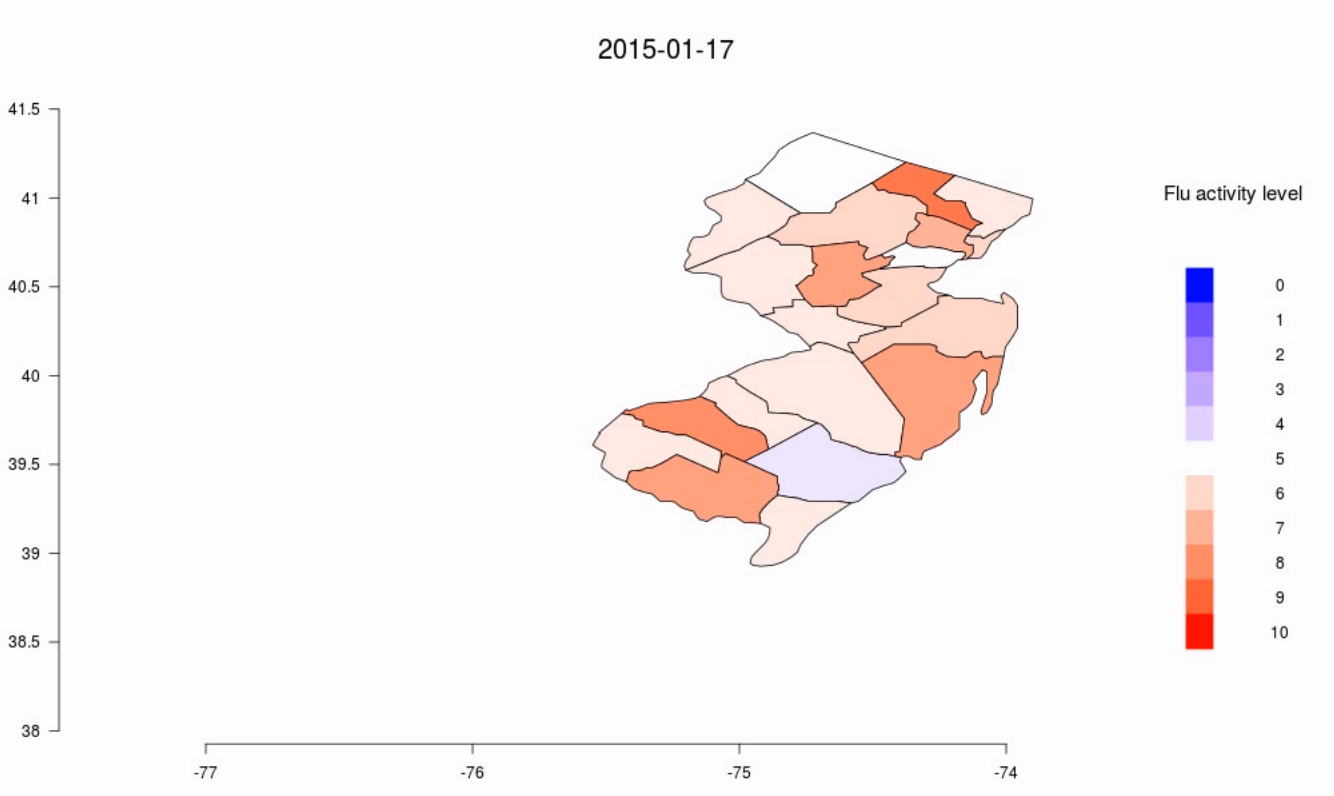

2012-05-05
Flu activity level
-10
-8
-6
-4
-2
0
2
4
6
8
10
45
40
35
30
25
-120
-110
-100
-90
-80
-70

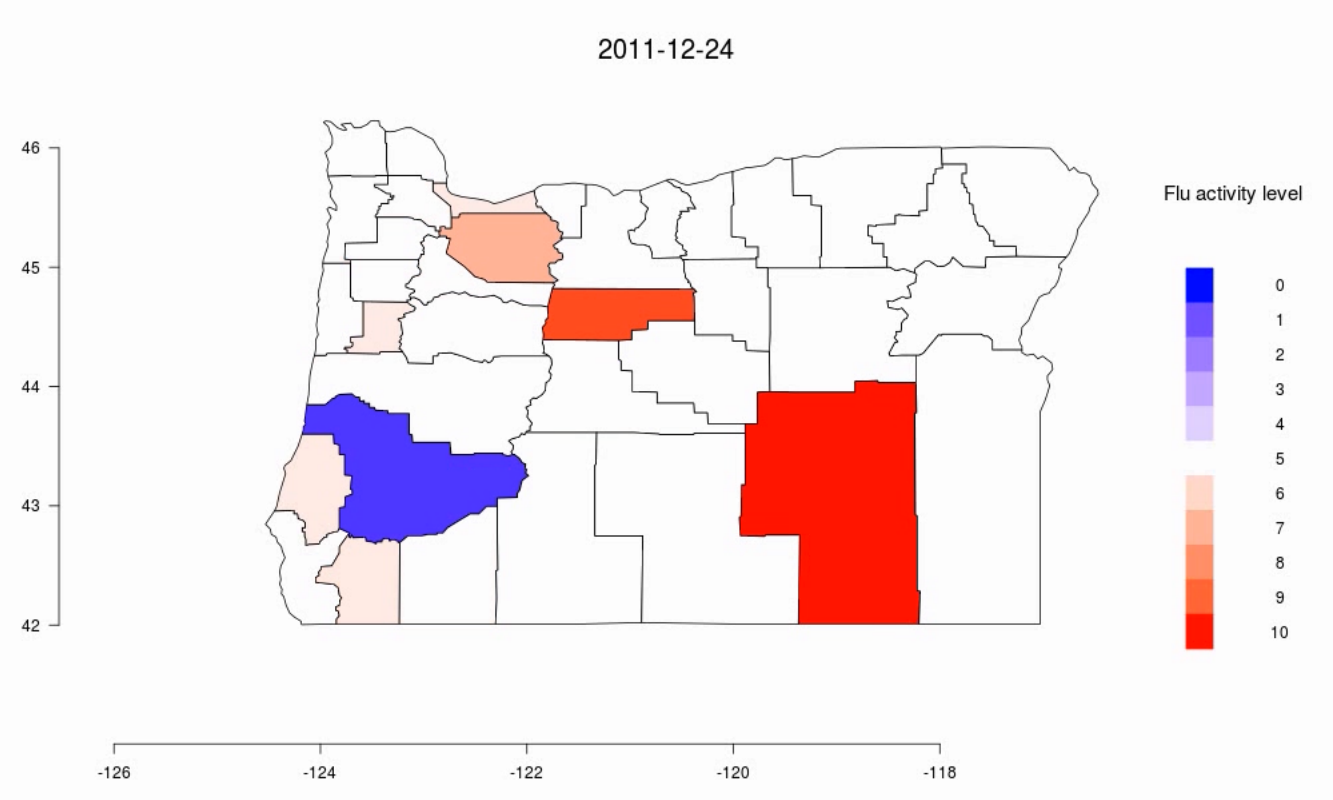

2012-05-05
Flu activity level
-10
-8
-6
-4
-2
0
2
4
6
8
10
45
40
35
30
25
-120
-110
-100
-90
-80
-70

4.4 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).





4.5 Additional attempts to reproduce the findings froms (Bodnar, 2015)

As the results depicted above show, the output from the twitter classifier that served as the basis of my analysis does neither serve as a good approximation of the official CDC data nor does it reflect the results shown in (Bodnar, 2015). There are various potential reasons for this which I will lay out in Chapter 5. Before doing so, however, I will summarise additional attempts of mine to replicate the results.

4.5.1 Attempt to reclassify raw Twitter data

As described in Chapter 3, the data set I was working on a data set that contained the output of the Twitter classifier described in Chapter 1. Since there might be discrepancies between this data set and the one used in (Bodnar, 2015), I wanted to reclassify the geotagged raw Twitter data collected between 2011 and 2015 in order to assess whether these results would diverge from the data set I was working on.

To do so, I retrieved the full Java-based Twitter classifier from Todd Bodnar's Github repository (https://github.com/ToddBodnar/Twitter-Parser). However, some libraries used for building the classifier were missing from the repository, while others (most notably the Amazon Web Services (AWS) SDK for Java: https://aws.amazon.com/de/sdk-for-java/) have in the meantime been superceded by newer vesions which are not backwards compatible with older version and thus incompatible with the Twitter classifier. Even though I managed to retrieve the missing libraries through personal communication with Todd Bodnar and also managed to retrieve the same AWS SDK version that was used for the original version of the Twitter classifier, additional compilation errors remained, so I was unable to compile the classifier.

Finally, I received a compiled version of the Twitter classifier ("TwitterParser.jar") directly from Todd Bodnar, 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. Hence, I still failed to reclassify the raw Twitter data using the Twitter classifier. [2]

4.5.2 Attempt to reproduce the SIR model described in (Bodnar, 2015)

Since I was unable to reproduce the original results of the Twitter classifier due to compilation and runtime errors of said classifier, and hence was unable to assess the validiy of the data set I was working on, I went at it from the opposite direction: I started from the final results described in (Bodnar, 2015) and tried to "reverse engineer" them in order to learn how the excellent fit of the Twitter data with the CDC data came about. Specifically, I focused on chapter 4 of (Bodnar, 2015) and tried to reproduce the findings shown in Figures 2.3 and 2.2 as well as Table 2.1.

In a first step, I asked Todd Bodnar for the data and the code used to create the SIR model as well as the figures. I received an the R-files used to build the SIR model as well csv-files containing the data used to create Figures 2.3 and 2.2 However, I did not receive the R-files or the the model specifications in order to create the data contained in said file from the Twitter data· Hence, I could neither associate the data within these files to the raw results from the Twitter classifier nor to the R-code-files that were provided to me. [3].

[2] All files belonging to the Twitter classifier, including the compiled version of it, can be found on Github (https://github.com/salathegroup/2016_TwitterEpi) in the folder "Twitter_Parser"

[3] I follow-up e-mail regarding this matter is pending response. Data and code files can be found on Github

In order to recreated the figures mentioned above, I used the data available in the file "predictions_r_results.csv", which contained weekly ILI estimates from the CDC as well as additional ILI estimates using different autocorrelation models with or without using the data from the Twitter classifier. The data spanned a four year period, starting on October 3^{rd} 2011 (week 40) and ending on September 28^{th} 2014 (week 39). The 11 columns contained in the data set have the following meaning:[4]

- **cdcoffset** The official ILI data from the CDC
- **predictions_base** The percentage of Twitter users classified as ill based on the predictive base model
- **predictions_autocor** Predictive AR(1) model based on the "cdcoffset" data
- **predictions_autocor2** Predictive AR(2) model based on the "cdcoffset" data
- **predictions_both** Predictive AR(1) model based on the "cdcoffset" data combined with the values from the predictive Twitter base model
- **predictions_both2** Predictive AR(2) model based on the "cdcoffset" data combined with the values from the predictive Twitter base model
- **full_base** The percentage of Twitter users classified as ill based on the retrospective base model
- **full_autocor** Retrospective AR(1) model based on the "cdcoffset" data
- **full_autocor2** Retrospective AR(2) model based on the "cdcoffset" data
- **full_both** Retrospective AR(1) model based on the "cdcoffset" data combined with the values from the retrospective Twitter base model
- **full_both2** Retrospective AR(2) model based on the "cdcoffset" data combined with the values from the retrospective Twitter base model

(https://github.com/salathegroup/2016_TwitterEpi) in the folder "PhDThesisBodnar"

[4]Personal communication by Todd Bodnar; as mentioned above, as of this point I do not know how the predictive or retrospective Twitter model relates to the results from the Twitter classifier. An email requesting additional information is pending response

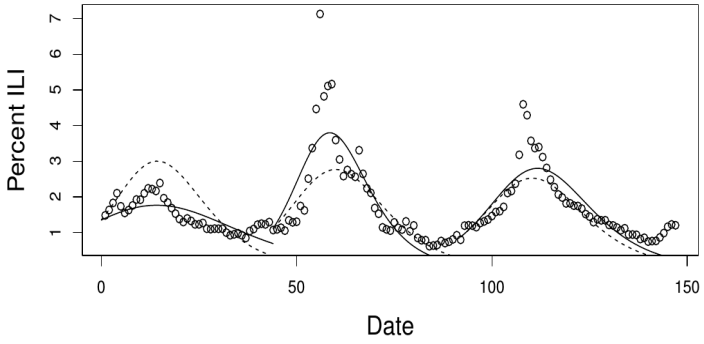
For the predictive models, only values between weeks 0 and $t-1$ were used for model building, for the retrospective models the complete available data was used.

I started with the replication of Figure 2.2 and the corresponding parameters. I did so by using the grid-search method described in (Bodnar, 2015)[5]

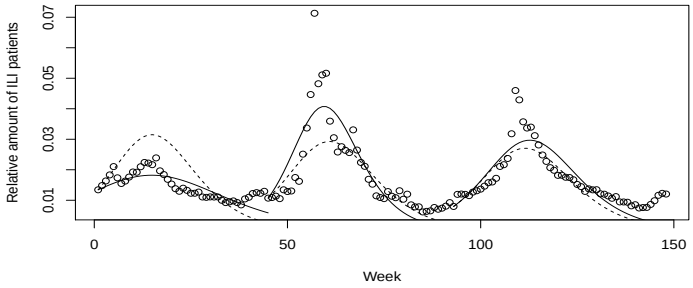
First, I built the SIR model based on the CDC data, since this was the most straightforward way to go. As can be seen from Figure 4.15b, the yearly and combined ILI curves calculated from the model are very similar to the ones depicted in Figure 2.2, however small deviations remain.

Therefore, I went on to fit the very same SIR model that is shown in Figure 2.2. To do so, I needed to know which data the model was built on. I extracted the three starting coordinates of the yearly curves shown in the figure and compared them with the data provided to me by Todd. It turned out that the starting coordinates of the curve only matched the values from the full retrospective Twitter model (i.e. retrospective AR(2) model based on the CDC data combined with the retrospective Twitter base model), so built my second SIR model based on these values. As can be seen from Figure 4.15c, the resulting yearly and combined ILI curves are virtually indistinguishable from the ones shown in (Bodnar, 2015).

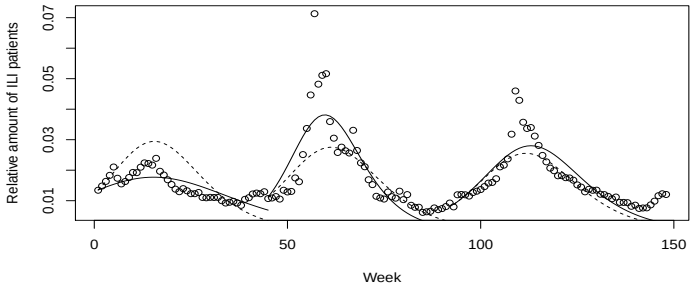
[5] "Specifically, we search through three variables, the two transmission parameters γ,β, and S(0), the initial susceptibility rate which may be less than 1 due to innate immunity or previous vaccination. Next, we generate a logarithmically spaced 25 by 25 by 25 grid of potential values over this range. We then set $I(0)$ to be the same as the first infection value in the data and $R(0)$ (sic!). We then solve an SIR model, with each of the parameter combinations." Note tht in the R-code I received the optimisation only happened for γ and β, not for $S(0)$, as desribed. This makes intuitive sense: Since we know the initial value of $I(0)$ both for the CDC data as well as for the Twitter model, there is no need to algorithmically find the initial value of $S(0)$, since it can simply be calculated as $S(0) = 1 - I(0)$, while $R(0) = 0$. Indeed, this is exactly the way the initial values were set in the algorithm.



(a)



(b)



(c)

Figure 4.15: Comparison between the SIR model depicted in (Bodnar, 2015) (a) and replications based on the official CDC ILI rates (b) and the full retrospective Twitter model (c).

It is very peculiar, however, that the model parameters which I calculated for the two replicated SIR models described above do not match the yearly and combined values for γ and β presented in (Bodnar, 2015) - even though model curves seem to match. As can be seen from Table 4.1, the optimal yearly and combined values of γ and β deviate considerably from the ones depicted in Table 2.1 (i.e. the values reported in (Bodnar, 2015)). This is true for the

SIR model based on the CDC data as well as for the SIR model based on the full retrospective model data. Hence, it is safe to assume that those values do **not** represent the values that were used to calculate the curves shown in Figure 2.2.

Year	γ	β	RSS
2011-2012	0.3962286	0.4446748	4.0608281×10^{-4}
	0.3666311	0.4087479	3.6266689×10^{-4}
2012-2013	0.5121045	0.6777425	0.0029532
	0.5021989	0.6546756	0.0027735
2013-2014	0.451297	0.5699723	0.0014841
	0.4294719	0.535259	0.0013076
Combined	0.4878438	0.6068527	0.008631
	0.4641424	0.5698395	0.0077089

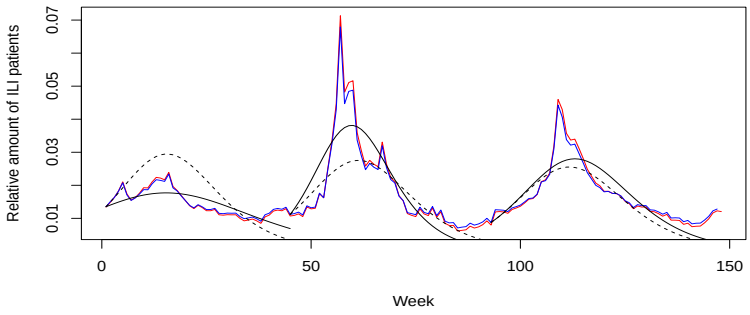
Table 4.1: National best-fit parameters for each year from the CDC's data (white) and the full retrospective model consisting of a retrospective AR(2) model based on the CDC data and the retrospective Twitter base model (grey)

In addition, it is important to note again that the data used to calculate the SIR models is **not** equivalent to the raw results of the Twitter classifier. In fact, the SIR models described above are built on the full retrospective model data which includes the information from the Twitter classifier as well as information from an AR(2) model based on the official CDC data. This stands in contrast to the report in (Bodnar, 2015), where the SIR models are depicted as being based on the information from the Twitter classifier only. However, this cannot be the case. Figure 4.16 clearly shows that the SIR model built solely on the data from the Twitter base model is clearly distinct from the SIR models built on the full retrospective model data.

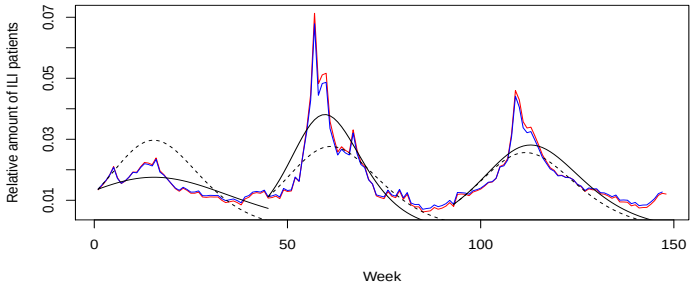
4.5.3 Attempt to reproduce the AR model described in (Bodnar, 2015)

Finally, I tried to reproduce the results depicted in Figure 2.3. The graph is described as the "comparison of Twitter's forecasting (dashed lines) and retroactive measurements (solid lines) to the CDC's reported Influenza rates (circles)", therefory implying that the dashed and solid lines are produces based on the data from the Twitter classifier alone. However, this is clearly not the case. Figure 4.16a shows a reproduction of the original graph depicted in (Bodnar,

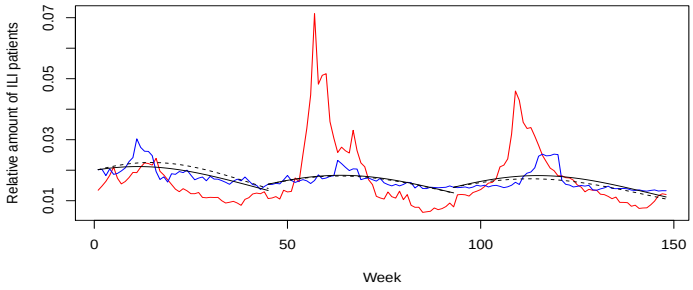
2015), including SIR model fits. However, this reproduction is based on the retrospective **full** model, i.e. a model including information from the Twitter base model as well as from the AR(2) model based on the CDC data. In fact, if we only look at the data from the Twitter base model (Figure 4.16b), we can see that the fit with the official CDC data is much worse. Also, Figure 4.16c shows that a simple AR(2) model based on the official CDC data achieves a model fit that is almost as good as the one from the full model which additional includes information from the Twitter base model.



(a)



(b)



(c)

Figure 4.16: Comparison between the full retrospective Twitter model depicted in (Bodnar, 2015) (a) and replications consisting of a simple retrospective AR(2) model based only on CDC ILI rates (b) and the retrospective Twitter base model (c)

Chapter 5

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

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

5.1 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.

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

As written in the Chapter 3, 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 used to fit to the official CDC data. 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.

5.3 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 it in a second step.

However, in order to assess the quality of and improve said Twitter flu classifier one would not only need access to it, but also get it up and running. Unforunately, I was unable to reclassify the Twitter data in order to confirm my assumptions due to missing / deprecated libraries and runtime errors in the Twitter classifer, respectively. Hence, debugging and updating the Twitter classifier should be a main goal for future replication attempts.

5.4 The findings can 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 source of information to complement already existng estimates and reduce the error. As I have shown in Chapter 4, this also true for the Twitter data used in (Bodnar, 2015): The Twitter base model alone was not able to fit the official CDC data tightly and was clearly outperformed by a simple AR(2) model based on the official CDC data.

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 does certainly improve the fit between the Twitter flu predictions and the official CDC ILI data. However, this then again opens 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 2.2) and in no way comparable to the extraordinary fit shown in Figure 2.3). In the meantime, I received the code used to fit the SIR model as well as the data used to produce Figure 2.3. 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.

In any case, however, the discrepancies between the model fits described in (Bodnar, 2015) and the results presented in Chapter 4 of this thesis are worrying. Whatever the reason may be, it is clear that the information from the Twitter classifier on its on cannot be used to reliably predict CDC ILI rates. Even when disregarding the raw results from the raw Twitter classifier (described in 3) and only focusing on the (processed) data that supposedly served as basis for (Bodnar, 2015), many unpleasant discrepancies are found that. It seems as if the models relying on the results from the Twitter classifier are widely outperformed by a simple AR(2) model based on the CDC data.

5.5 Missing parts

- Formatting and citations
- Cleaning up code & repository (¿ any specific guidelines to follow for this?)

Bibliography

Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., and Fedor, A. (2015). Estimating the reproducibility of psychological science. *Science*, **349**, 1–8.

Abbar, S., Mejova, Y., and Weber, I. (2015). You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3197–3206. ACM.

Achenbach, J. (2015a). Many scientific studies can't be replicated. That's a problem. *The Washington Post*. https://www.washingtonpost.com/news/speaking-of-science/wp/2015/08/27/trouble-in-science-massive-effort-to-reproduce-100-experimental-results-succeeds-only-36-?utm_term=.1a233386e88f. [Online; accessed 13-August-2017].

Achenbach, J. (2015b). No, science's reproducibility problem is not limited to psychology. *The Washington Post*. https://www.washingtonpost.com/news/speaking-of-science/wp/2015/08/28/no-sciences-reproducibility-problem-is-not-limited-to-psychology/?tid=a_inl&utm_term=.d7bf9206f3dc. [Online; accessed 13-August-2017].

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2011). Predicting Flu Trends using Twitter data. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 702–707.

Adrover, C., Bodnar, T., Huang, Z., Telenti, A., and Salathé, M. (2015). Identifying adverse effects of hiv drug treatment and associated sentiments using twitter. *JMIR public health and surveillance*, **1**, .

Anema, A., Kluberg, S., Wilson, K., Hogg, R. S., Khan, K., Hay, S. I., Tatem, A. J., and Brownstein, J. S. (2014). Digital surveillance for enhanced detection and response to outbreaks. *The Lancet. Infectious diseases*, **14**, 1035–1037.

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, 1568–1576. Association for Computational Linguistics.

Asano, E. (2017). How much time do people spend on social media? *Social Media Today*. http://www.socialmediatoday.com/marketing/how-much-time-do-people-spend-social-media-infographic. [Online; accessed 12-August-2017].

Bach, M., Jordan, S., Hartung, S., Santos-Hövener, C., and Wright, M. T. (2017). Participatory epidemiology: the contribution of participatory research to epidemiology. *Emerging Themes in Epidemiology*, **14**, .

Bailey, L., Vardulaki, K., Langham, J., and Chandramohan, D. (2005). *Introduction to Epidemiology*. Open University Press London.

Baker, M. (2016a). Is there a reproducibility crisis? a nature survey lifts the lid on how researchers view the'crisis rocking science and what they think will help. *Nature*, **533**, 452–455.

Baker, M. (2016b). The Reproducibility Crisis Is Good for Science. *Slate*. http://www.slate.com/articles/technology/future_tense/2016/04/the_reproducibility_crisis_is_good_for_science.html. [Online; accessed 13-August-2017].

Baker, M. (2016c). Reproducibility: seek out stronger science. *Nature*, **537**, 703–704.

Bauer, R. (2016). Media (R)evolutions: time spent online continues to rise. *World Bank Blog*. https://blogs.worldbank.org/publicsphere/media-revolutions-time-spent-online-continues-rise. [Online; accessed 12-August-2017].

Begley, C. G. (2013). Six red flags for suspect work. *Nature*, **497**, 431–433.

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science. *Circulation research*, **116**, 116–126.

Belluz, J. (2017). Cancer scientists are having trouble replicating groundbreaking research. *Vox*. https://www.vox.com/science-and-health/2017/1/23/14324326/replication-science-is-hard. [Online; accessed 13-August-2017].

Bodnar, T. (2015). Data science with social media for epidemiology and public health.

Bodnar, T., Barclay, V. C., Ram, N., Tucker, C. S., and Salathé, M. (2014). On the ground validation of online diagnosis with Twitter and medical records. 651–656. ACM Press. bibtex: bodnar_ground_2014.

Bodnar, T. and Salathé, M. (2013). Validating models for disease detection using twitter. *Proceedings of the 22nd international conference on World Wide Web companion*.

Bollen, J., Mao, H., and Pepe, A. (2011a). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, **11**, 450–453.

Bollen, J., Mao, H., and Zeng, X. (2011b). Twitter mood predicts the stock market. *Journal of computational science*, **2**, 1–8.

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., and Hanson, C. L. (2016). Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR mental health*, **3**, .

Broman, K., Cetinkaya-Rundel, M., Nussbaum, A., Paciorek, C., Peng, R., Turek, D., and Wickham, H. (2017). Recommendations to Funding Agencies for Supporting Reproducible Research. *American Statistical Association*. https://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf. [Online; accessed 13-August-2017].

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. (2008). Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Medicine*, **5**, .

Bundesamt für Gesundheit (2016). Saisonale grippe: Antworten auf häufige Fragen. https://www.bag.admin.ch/dam/bag/de/dokumente/mt/infektionskrankheiten/grippe/fach-faq-grippe.pdf.download.pdf/faq-grippe-aw-faq-fach-de.pdf. [Online; accessed 12-August-2017].

Bundesamt für Gesundheit (2017a). Saisonale grippe - Lagebericht Schweiz. https://www.bag.admin.ch/bag/de/home/themen/mensch-gesundheit/uebertragbare-krankheiten/infektionskrankheiten-a-z/grippe.html. [Online; accessed 12-August-2017].

Bundesamt für Gesundheit (2017b). Saisonale grippe (influenza). https://www.bag.admin.ch/bag/de/home/themen/mensch-gesundheit/uebertragbare-krankheiten/infektionskrankheiten-a-z/grippe.html. [Online; accessed 12-August-2017].

Butler, D. (2013). When Google got flu wrong. *Nature*, **494**, 155.

Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., and Dean, H. (2015). Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, **351**, 1433–1436.

Carey, B. (2015a). Many Psychology Findings Not as Strong as Claimed, Study Says. *The New York Times*. https://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html. [Online; accessed 13-August-2017].

Carey, B. (2015b). Science, Now Under Scrutiny Itself. *The New York Times*. https://www.nytimes.com/2015/06/16/science/retractions-coming-out-from-under-science-rug.html. [Online; accessed 13-August-2017].

Casadevall, A. and Fang, F. C. (2010). Reproducible Science. *Infection and Immunity*, **78**, 4972–4975.

Centers for Disease Control and Prevention (2016). Overview of influenza surveillance in the united states. https://www.cdc.gov/flu/weekly/overview.htm. [Online; accessed 24-June-2017].

Centers for Disease Control and Prevention (2017). The flu: What to do if you get sick. https://www.cdc.gov/flu/takingcare.htm. [Online; accessed 12-August-2017].

Chae, B. K. (2015). Insights from hashtag# supplychain and twitter analytics: Considering twitter and twitter data for supply chain practice and research. *International Journal of Production Economics*, **165**, 247–259.

Chen, G. M. (2011). Tweet this: A uses and gratifications perspective on how active twitter use gratifies a need to connect with others. *Computers in Human Behavior*, **27**, 755–762.

Chunara, R., Aman, S., Smolinski, M., and Brownstein, J. S. (2013). Flu near you: an online self-reported influenza surveillance system in the USA. *Online Journal of Public Health Informatics*, **5**, .

Chunara, R., Goldstein, E., Patterson-Lomba, O., and Brownstein, J. S. (2015). Estimating influenza attack rates in the United States using a participatory cohort. *Scientific Reports*, **5**, 9540.

Crotty, D. (2014). Reproducible Research, Just Not Reproducible By You. *The Scholarly Kitchen*. https://scholarlykitchen.sspnet.org/2017/05/24/reproducible-research-just-not-reproducible/. [Online; accessed 13-August-2017].

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*, 115–122. ACM.

Earp, B. D. and Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, **6**, .

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M.,*et al.* (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science*, **26**, 159–169.

Engber, D. (2016). Cancer Research Is Broken. *Slate*. http://www.slate.com/articles/health_and_science/future_tense/2016/04/biomedicine_facing_a_worse_replication_crisis_than_the_one_plaguing_psychology.html. [Online; accessed 13-August-2017].

Feilden, T. (2017). Most scientists 'can't replicate studies by their peers'. *BBC*.

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *PLOS Biology*, **13**, e1002165.

Freifeld, C. C., Chunara, R., Mekaru, S. R., Chan, E. H., Kass-Hout, T., Iacucci, A. A., and Brownstein, J. S. (2010). Participatory epidemiology: use of mobile phones for community-based health reporting. *PLoS Medicine*, **7**, e1000376.

Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S. (2008). HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association : JAMIA*, **15**, 150–157.

Gardy, J., Loman, N. J., and Rambaut, A. (2015). Real-time digital pathogen surveillance — the time is now. *Genome Biology*, **16**, .

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., and Priedhorsky, R. (2014). Global Disease Monitoring and Forecasting with Wikipedia. *PLoS Computational Biology*, **10**, e1003892.

German Research Foundation (2017). Replicability of Research Results. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_en.pdf. [Online; accessed 13-August-2017].

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, **457**, 1012–1014.

Goff, J., Rowe, A., Brownstein, J. S., and Chunara, R. (2015). Surveillance of Acute Respiratory Infections Using Community-Submitted Symptoms and Specimens for Molecular Diagnostic Testing. *PLoS Currents*.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 149–156. IEEE.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, **8**, 341ps12–341ps12.

Google Flu Trends Team (2015). The next chapter for flu trends. https://research.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html. [Online; accessed 20-July-2017].

Hermida, A. (2013). # journalism: Reconfiguring journalism research about twitter, one tweet at a time. *Digital Journalism*, **1**, 295–313.

Himelboim, I., McCreery, S., and Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication*, **18**, 40–60.

Hufnagel, L., Brockmann, D., and Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 15124–15129.

Kenett, R. S. and Shmueli, G. (2015). Clarifying the terminology that describes scientific reproducibility. *Nature methods*, **12**, 699–699.

Khoury, M. J., Lam, T. K., Ioannidis, J. P. A., Hartge, P., Spitz, M. R., Buring, J. E., Chanock, S. J., Croyle, R. T., Goddard, K. A., Ginsburg, G. S., Herceg, Z., Hiatt, R. A., Hoover, R. N., Hunter, D. J., Kramer, B. S., Lauer, M. S., Meyerhardt, J. A., Olopade, O. I., Palmer, J. R., Sellers, T. A., Seminara, D., Ransohoff, D. F., Rebbeck, T. R., Tourassi, G., Winn, D. M., Zauber, A., and Schully, S. D. (2013). Transforming Epidemiology for 21st Century Medicine and Public Health. *Cancer Epidemiology Biomarkers & Prevention*, **22**, 508–516.

Koepsell, T. D. and Weiss, N. S. (2014). *Epidemiologic methods: studying the occurrence of illness*. Oxford University Press (UK).

Kullmann, D. M. (2015). Biomedicine is plagued by failure of replication.

Lamb, A., Paul, M. J., and Dredze, M. (2013). Separating Fact from Fear: Tracking Flu Infections on Twitter. In *HLT-NAACL*, 789–795.

Lampos, V. and Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *2010 2nd International Workshop on Cognitive Information Processing*, 411–416.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, **343**, 1203–1205.

Lee, K., Agrawal, A., and Choudhary, A. (2013). Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining*, 1474–1477. ACM.

Lehrer, J. (2010). The Truth Wears Off. *The New Yorker*. http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off. [Online; accessed 13-August-2017].

Liu, J., Weitzman, E. R., and Chunara, R. (2017). Assessing Behavior Stage Progression From Social Media Data. 1320–1333. ACM Press.

Loken, E. and Gelman, A. (2017). Measurement error and the replication crisis. *Science*, **355**, 584–585.

Love, B., Himelboim, I., Holton, A., and Stewart, K. (2013). Twitter as a source of vaccination information: content drivers and what they are saying. *American journal of infection control*, **41**, 568–570.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S. Y., and others (2013). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2095–2128.

Martcheva, M. (2015). Introduction to Epidemic Modeling. In *An Introduction to Mathematical Epidemiology*, 9–31. Springer.

Marwick, A. E. and Boyd, D. (2011). I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, **13**, 114–133.

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean? *American Psychologist*, **70**, 487–498.

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E. S. (2015). Using twitter for demographic and social science research: tools for data collection and processing. *Sociological Methods & Research*, 0049124115605339.

Meyer, M. (2017). Want to fix sciences's replication crisis? Then replicate. *Wired*. https://www.wired.com/2017/04/want-fix-sciences-replication-crisis-replicate/. [Online; accessed 13-August-2017].

Milinovich, G. J., Magalhães, R. J. S., and Hu, W. (2015). Role of big data in the early detection of Ebola and other emerging infectious diseases. *The Lancet Global Health*, **3**, e20–e21.

Mowery, D., Bryan, C., and Conway, M. (2017). Feature studies to inform the classification of depressive symptoms from twitter data for population health. *arXiv preprint arXiv:1701.08229*.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, **1**, 0021.

Newman, T. P. (2016). Tracking the release of ipcc ar5 on twitter: Users, comments, and sources following the release of the working group i summary for policymakers. *Public Understanding of Science*, 0963662516628477.

Nosek, B. A. and Errington, T. M. (2017). Reproducibility in cancer biology: making sense of replications. *Elife*, **6**, e23383.

Nwosu, A. C., Debattista, M., Rooney, C., and Mason, S. (2014). Social media and palliative medicine: a retrospective 2-year analysis of global twitter data to evaluate the use of technology to communicate about issues at the end of life. *BMJ supportive & palliative care*, bmjspcare–2014.

Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., and Simonsen, L. (2013). Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Computational Biology*, **9**, e1003256.

Paul, M., Dredze, M., Broniatowski, D., and Generous, N. (2015). Worldwide influenza surveillance through twitter. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*.

Paul, M. J. and Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. *ICWSM*, **20**, 265–272.

Paul, M. J., Dredze, M., and Broniatowski, D. (2014). Twitter Improves Influenza Forecasting. *PLoS Currents*.

Peng, R. D. (2009). Reproducible research and Biostatistics. *Biostatistics*, **10**, 405–408.

Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. (2011). Do all birds tweet the same?: characterizing twitter around the world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 1025–1030. ACM.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, **10**, 712–712.

Qiu, L., Lin, H., Ramsay, J., and Yang, F. (2012). You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, **46**, 710–718.

Ray, B., Ghedin, E., and Chunara, R. (2016). Network inference from multimodal data: A review of approaches from infectious disease transmission. *Journal of Biomedical Informatics*, **64**, 44–54.

Rehak, M. (2014). Who made that Twitter bird? *The New York Times*. https://www.nytimes.com/2014/08/10/magazine/who-made-that-twitter-bird.html. [Online; accessed 1-August-2017].

Rolfes, M. A., Foppa, I. M., Garg, S., Flannery, B., Brammer, L., Singleton, J. A., Burns, E., Jernigan, D., Reed, C., Olsen, S. J., and Bresee, J. (2016). Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States. https://www.cdc.gov/flu/about/disease/2015-16.htm. [Online; accessed 13-August-2017].

Rothman, K. J. (2012). *Epidemiology: an introduction*. Oxford university press.

Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L., and others (2012). Digital epidemiology. *PLoS Comput Biol*, **8**, e1002616.

Salathé, M., Vu, D. Q., Khandelwal, S., and Hunter, D. R. (2013). The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science*, **2**, 4.

Schooler, J. W. (2014). Metascience could rescue the'replication crisis'. *Nature*, **515**, 9.

Scott, C. F., Bay-Cheng, L. Y., Prince, M. A., Nochajski, T. H., and Collins, R. L. (2017). Time spent online: Latent profile analyses of emerging adults' social media use. *Computers in Human Behavior*, **75**, 311–319.

Sentinella (2017). Willkommen bei Sentinella. http://sentinella.ch/de/info. [Online; accessed 12-August-2017].

Simonsen, L., Gog, J. R., Olson, D., and Viboud, C. (2016). Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems. *Journal of Infectious Diseases*, **214**, S380–S385.

Sloan, L. and Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE*, **10**, .

Steiger, E., Albuquerque, J. P., and Zipf, A. (2015). An advanced systematic literature review on spatiotemporal analyses of twitter data. *Transactions in GIS*, **19**, 809–834.

Stieglitz, S. and Dang-Xuan, L. (2012). Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, 3500–3509. IEEE.

Stroebe, W. and Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, **9**, 59–71.

Sul, H. K., Dennis, A. R., and Yuan, L. I. (2014). Trading on twitter: The financial information content of emotion in social media. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, 806–815. IEEE.

Swani, K., Brown, B. P., and Milne, G. R. (2014). Should tweets differ for b2b and b2c? an analysis of fortune 500 companies' twitter communications. *Industrial Marketing Management*, **43**, 873–881.

The Economist (2016). Let's just try this again. *The Economist*. https://www.economist.com/news/science-and-technology/21690020-reproducibility-should-be-sciences-heart-it-isnt-may-soon. [Online; accessed 13-August-2017].

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, **10**, 178–185.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2011). Election forecasts with twitter: How 140 characters reflect the political landscape. *Social science computer review*, **29**, 402–418.

Twitter (2013). Annual Report 2013. http://files.shareholder.com/downloads/AMDA-2F526X/4733143221x0x742484/A418947A-E065-4822-8BD4-00FA8EB4E795/Twitter_2013_Annual_Report_-_FINAL.pdf. [Online; accessed 12-August-2017].

Twitter (2017). Annual Report 2017. https://investor.twitterinc.com/common/download/download.cfm?companyid=AMDA-2F526X&fileid=935049&filekey=05E6E71E-D609-4A17-A8BD-B621324A950D&filename=TWTR_2016_Annual_Report.pdf. [Online; accessed 12-August-2017].

Vidal, L., Ares, G., Machín, L., and Jaeger, S. R. (2015). Using twitter data for food-related consumer research: A case study on "what people say when tweeting about different eating situations". *Food Quality and Preference*, **45**, 58–69.

Wang, F., Wang, H., Xu, K., Raymond, R., Chon, J., Fuller, S., and Debruyn, A. (2016). Regional Level Influenza Study with Geo-Tagged Twitter Data. *Journal of Medical Systems*, **40**, .

Widener, M. J. and Li, W. (2014). Using geolocated twitter data to monitor the prevalence of healthy and unhealthy food references across the us. *Applied Geography*, **54**, 189–197.

Wójcik, O. P., Brownstein, J. S., Chunara, R., and Johansson, M. A. (2014). Public health for the people: participatory infectious disease surveillance in the digital age. *Emerging themes in epidemiology*, **11**, 7.

Yong, E. (2016). Psychology's Replication Crisis Can't Be Wished Away. *The Atlantic*. https://www.theatlantic.com/science/archive/2016/03/psychologys-replication-crisis-cant-be-wished-away/472272/. [Online; accessed 13-August-2017].

Zhang, X., Fuehres, H., and Gloor, P. (2012). Predicting asset value through twitter buzz. *Advances in Collective Intelligence 2011*, 23–34.

Zhao, D. and Rosson, M. B. (2009). How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, 243–252. ACM.

Zimmer, M. and Proferes, N. J. (2014). A topology of twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, **66**, 250–261.

R version and packages used to generate this report:

R version: R version 3.4.1 (2017-06-30)

Base packages: stats, graphics, grDevices, utils, datasets, methods, base

Other packages: data.table, knitr

This document was generated on August 14, 2017 at 03:44.