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. Dr. Reinhard Furrer (UZH)
Prof. Dr. Marcel Salathé (EPFL)

Zurich, 2017

Abstract

Digital data have repeatedly proved their value for disease detection and surveillance. Social networks like Twitter promise to be a particularly rich source of epidemiological information, because they give researchers intimate access to a user's health status, habits, and behaviour.

However, most of the resulting epidemiological models use only population-level data and correlate these with traditional disease records. This approach is prone to overfitting and thus needs to be complemented with an alternative approach, in which individual-level disease status can be linked to individual-level social media behaviour. This master thesis assesses the validity of a Twitter flu classifier that was trained on data for which the individual-level disease state of Twitter users was available Bodnar *et al.* (2014). I assessed the performance of this classifier when applied to a large data set of about 2.8 billion geotagged tweets from U.S. mainland by comparing its results with the official flu estimates from the Centers of Disease Control and Prevention (CDC) as well as with results of similar study performed by Bodnar (2015).

In the following, I provide a detailed description of the statistical characteristics of the classified Twitter data set provided to me. I also show that the Twitter classifier identifies the peaks of flu seasons between 2011 and 2015 with some accuracy, but is unable to accurately model the behaviour of the CDC's flu estimates throughout the year. My results also deviate considerably from the ones reported in Bodnar (2015), hinting at discrepancies in the underlying Twitter data sets and the methods used for the respective analyses, or implying missing information and errors in reporting. Attempts to reproduce key figures and tables from Bodnar (2015) failed, too, and uncovered inaccuracies and gaps in reporting.

In conclusion, the results of my analysis of the classified Twitter data set and my partial reproduction of Bodnar (2015) showed that the Twitter classifier on its own does not provide enough information to serve as an accurate estimator of flu trends—neither pro- nor retrospectively. However, it might allow for faster and more accurate influenza surveillance, when used in conjunction with traditional surveillance methods or when fed into more sophisticated disease models. Further research should focus on improving the performance of the classifier's output and testing its reliability in different spatio-temporal settings.

Acknowledements

I thank my two supervisors, Prof. Dr. Reinhard Furrer and Prof. Dr. Marcel Salathé for their academic guidance and critical feedback as well as for the intellectual freedom I was able to enjoy during my thesis. I thank the Digital Epidemiology lab and its members for stimulating conversations, an extraordinary lab retreat, and plenty of free coffee.

In addition, I thank the Swiss Study Foundation, the Canton of Schaffhausen, the Werner Siemens Foundation, and the Ernst Göhner Foundation for their financial and ideational support. It has empowered me and given me the freedom to conduct my studies the way I wanted.

Finally, I thank my mother, my other mother, my other brothers and my friends for reminding me that there is an alternative to books, computer screens, and bad posture. And most importantly, I thank my fiancee for her patience, for laughing about, but mostly with me, and for pretending to care about my rants about malfunctioning code snippets and unsolvable error message conundrums.

Biel, 2017,

Servan Grüninger

"Good statistics is like strong coffee: bitter at times, but always sobering."

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 U.S., 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, 2017b).

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, 2016b) 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 "socially accepted", 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 common, 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 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 combating 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 diseases—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 surprise 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 diseases 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 diseases as well as improve the speed and efficiency 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 (Salathé *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 (Salathé *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 reports 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 heterogeneous 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 lethargically, 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 explicitly 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 platforms 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 application programming interfaces (APIs) provided by the respective 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 unsurprisingly 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 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 U.S. (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.

However, this Twitter flu classifier (see Section 2.4 for details) has only been trained and tested on a very small data set so far. Hence, the first goal of this thesis is to assess the validity of the classifier using 2.8 billion geotagged tweets collected from users on the U.S. mainland over the course of four years (Chapter 4) and to make suggestions on how its performance might be improved (Chapter 5).

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", *i.e.* without collecting new data. 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, experimenters 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 rebuilt by running the corresponding RNW-file

(repeatability), while the findings described therein should be reproducible 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, nowadays, "a wide range of issues [are] 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 distinction given by Goodman *et al.* (2016) 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 exacerbated) effect with the same directionality as the original study.

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

2.2 Ensuring reproducibility: easier said than done

(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? Not much, if we ought to believe the most fervent harbingers of scientific doom.

In fact, the inability to reproduce or replicate many key scientific findings in various fields of research such as drug development (Prinz *et al.*, 2011; Begley and Ellis, 2012), cancer research (Nosek and Errington, 2017), social psychology (Aarts *et al.*, 2015), behavioural economy (Camerer *et al.*, 2016), and biological research (Freedman *et al.*, 2015; Vogt *et al.*, 2016) have spurred many heated scholarly discussions and led to an exponential increase of the number of papers revolving around the issue (Goodman *et al.*, 2016), but it should be noted that the overall body of scientific literature is also increasing exponentially (Bornmann and Mutz, 2015). So, one could say that studies revolving around or involving reproducibility are just trying to keep up with the avalanche of new articles being published each year.

The discussion does not remain within the realms of academia, either, but has already spilled into the pages and websites of mass media outlets, causing a deluge of articles that are further fuelling the discussion about the "reproducibility crisis" (Lehrer, 2010; Carey, 2015a,b; Achenbach, 2015a,b; Yong, 2016; Engber, 2016; Baker, 2016b; The Economist, 2016; Feilden, 2017; Belluz, 2017; Meyer, 2017).

It should be noted, however, that (results) reproducibility is neither a necessary nor a sufficient criterion for good science. There are many areas of science (*e.g.* evolutionary biology, paleontology, geology, climatology astronomy), which make empirical claims about the world based on good scientific practice, but which still fail to be reproducible—simply because it is impossible to replicate dinosaurs marching on earth, volcanoes erupting, or suns exploding, to just name a few examples (German Research Foundation, 2017). Many of these disciplines should strive for methods and inferential reproducibility, but they cannot be blamed for failing the test for results reproducibility.

In experimental research, however, the lack of results reproducibility must give raise to much more concern, even though it should not come as a surprise either. Practices such as multiple

comparisons, data mining, p-hacking, selective outcome reporting, or hypothesising after the results are known (HARKing), can hamper all three forms of reproducibility and regularly occur in daily research even without the researchers exhibiting any nefarious motivations (Goodman *et al.*, 2016). In addition, cognitive and methodical biases such as confirmation, hindsight, framing, or belief bias, and selection, reporting, attrition, recall, or observer bias, respectively, can hamper all three forms of reproducibiliy as well and are commonplace in scientific practice (Munafò *et al.*, 2017).

But even when following every methodological, statistical, and scientific rule to the book in order to boost reproducibility as much as possible, findings can be non-reproducible—at least with regard to results and inferential reproducibility—thanks to the standard null hypothesis testing framework, an unfortunate methodological hybrid between the hypothesis testing concept of R. A. Fisher on one hand and the one of Neyman–Pearson on the other (Amrhein *et al.*, 2017). Within this framework, researchers pit their own hypothesis against the Null hypothesis by conducting a significance test, which reveals the probability of receiving a result that is at least as extreme as the one the researchers have received—under the assumption that the Null hypothesis is true. This probability (the p-value) is than compared with an arbitrarily defined threshold (*e.g.* 0.05), either leading to the verdict that the findings are significantly different from the Null ($p < 0.05$) or not significant enough ($p \geq 0.05$) for the Null to be rejected. Thus, even under perfect condition without no scientific misconduct whatsoever, we would still expect that 5 % of all tests against true Null hypotheses turn out to be (falsely) significant.

This, in combination with publication bias (the preferential publication of statistically significant findings in scientific journals) and the inherent uncertainties of scientific research leads to a strong over-representation of false positive research findings in the scientific literature (Ioannidis, 2005), thereby hampering results reproducibility. And even if the Null hypothesis has been rejected correctly, inferential reproducibility can still be an issue, because scientists often use the rejection of the Null as a sign of support for their alternative hypothesis—even though a significance test is not enough to tell whether this alternative hypothesis is more or less likely to be true than the Null hypothesis. To this, additional information, statistical testing, and inferential reasoning is necessary.

2.3 Putting emphasis on methods reproducibility

As outlined above, ensuring reproducibility is a daunting task that can only be tackled by a multitude of measures and joint efforts from researchers and research institutions.

Nevertheless, individual researchers can ensure to improve the reproducibility of their own

work by following good scientific and statistical practices. Regarding this, I strongly believe that an emphasis on methods reproducibility is the most efficient way to improve overall reproducibility in research, since it provides other researchers with a valuable starting point for their own scientific endeavours.

However, even methods reproducibility remains a challenge that should not be underestimated. As Banks (2011) notes: "Most [research papers] do not come with code or data, and even if they did, I expect a careful check would find discrepancies from the published paper. The reasons for this are innocent: code written by graduate students is continually tweaked and has sketchy documentation. The exact data cleaning procedures are not perfectly remembered when the final version of the paper is written, or may be muddled by miscommunication among multiple authors. And even if a conscientious researcher provided a full description of every cleaning step, every model fitting choice, and all aspects of variable selection, the resulting paper would be so long and tedious that no doubt the foolish editor would demand that it be shortened."

Even in the best of worlds, scientific research cannot follow the textbook of good scientific practice. Still, efforts to improve (methods) reproducibility is paramount to the prolonged reliability of scientific findings. This is especially true for epidemiology in general and digital epidemiology in particular. Epidemiologists aggregate data from different sources such as health agencies, patient reports, or clinical trials, transform them in various ways, and build intricate models based on these aggregations and transformations. In addition, digital epidemiology tries to combine these data with additional sources of information such as billions of Twitter messages, geolocation and sensor data from mobile and tracking devices, or insights from health, fitness, or nutrition apps.

One important aspect of ensuring reproducible research are detailed and complete reports on the experimental and statistical procedures, the study design, potential confounders, techniques, and methods used, study protocols, statistical designs and further important elements that might be necessary to fully reproduce a scientific finding (Kass *et al.*, 2016). Also, it is of paramount importance that researchers publish data and code as complete and accessible as legally and ethically possible and provide clear and concise descriptions of their work flow (Peng, 2006).

Hence, this thesis serves a second and third important goal: It is the attempt to test some findings from Bodnar (2015) for reproducibility as well as to ensure methods reproducibility for this thesis itself. In addition, the first goal, outlined in Chapter 1 and consisting in the validation of the Twitter flu classifier from Bodnar *et al.* (2014), can be regarded as an attempt in reproducibility as well, namely results and inferential reproducibility. Hence, the data, analyses, and discussions presented in the following chapters will focus on the methods, results, and inferential reproducibility of the findings from Bodnar *et al.* (2014) and Bodnar (2015).

2.4 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 on one hand and all other tweets on the other. Out of the 37'599 tweets in the data set, a total of 1609 tweets from 35 users were sent in a month in which their authors were sick.

They then screened the tweets in both categories for the occurrence or absence of a set of seven keywords: {flu, influenza, sick, cough, cold, medicine, fever}. The predictive power of these seven words were then tested by applying five different classification methods: J48 (a Java implementation of the C4.5 algorithm), logistic regression, naive Bayes, random forest, and support vector machine (SVM). All classification methods failed to reliably correlate the keywords with a Twitter user's disease status (see left panel of 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 "healthy" 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 area under the curve (AUC) of 0.8544 when using the top 100 keywords (see right panel of Figure 2.1).

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

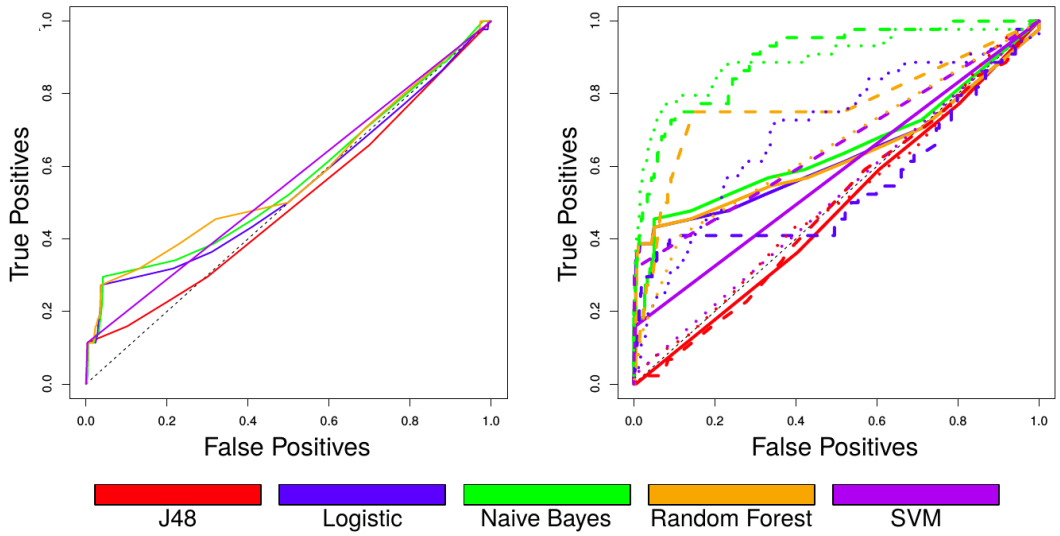


Figure 2.1: The receiver operating characteristic (ROC) of classifiers that use hand chosen key words (left) and algorithmically chosen keywords (right) to determine if an individual is ill. The top 10 (solid line), 100 (dashed line), and 1000 (dotted line) were selected as features. From Bodnar *et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

might be different from that of the average Twitter use. In addition, the exact time of disease of the Twitter users is not known. 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 described 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 U.S. 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 a score 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 as not-sick signal, for example—which, while weaker, are

more prevalent. We chose a step 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."

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 3rd 2011 to March 4th 2015. 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 U.S. mainland were discarded. However, the total number of tweets analysed was not given in any of the documentations available to me.

In a next step, Bodnar created a Kermack–McKendrick SIR model based on the results of the classifier (Martcheva, 2015):

$$\frac{dS}{dt} = -SI\beta, \qquad \frac{dI}{dt} = -SI\beta - I\gamma, \qquad \frac{dR}{dt} = I\gamma. \tag{2.1}$$

Here, S, I, and R represent the relative frequencies of susceptible, exposed, and recovered individuals, respectively, whereas β and γ are the transition probabilities from being susceptible to having 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 influenza-like illness (ILI) rates from the Centers for Disease Control (CDC) in Atlanta using a multi-grid search method. Note that estimates about flu prevalence are usually made using ILI rates as basis, not confirmed flu cases (see Section 4.1 for more details).

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.

Based on the values given in Table 2.1, he could then calculate yearly ILI estimates for the flu seasons of 2011–2012, 2012–2013, and 2013–2014 (see Figure 2.2). In addition, he fitted an autocorrelation model using the results from the Twitter base model and the official CDC data:

$$I_{\text{full}}(t+1) = a \cdot I_{\text{CDC}}(t-1) + b \cdot I_{\text{CDC}}(t) + c \cdot I_{\text{Twitter}}(t) + d. \tag{2.2}$$

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 (Bodnar, personal communication). It is as of yet unclear whether the "Twitter base model" is supposed to consist of the raw results from the Twitter classifier or contains these raw results combined with additional information. However, Bodnar confirmed that the "Twitter base model" is indeed supposed to contain the raw output from the Twitter classifier (personal communication). However, I was not able to retrieve the model parameters a, b, and c, since I only received a file with the model results, but not the parameter specifications.

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

Applying the above-described autocorrelation model, Bodnar was able to achieve a very close fit to the official Twitter data (see Figure 2.3). In the following chapter, I will describe my attempts to reproduce these findings based on the raw results from the Twitter classifier as well as to repeat the tables and figures shown above based on part of the processed data and code sources provided by Bodnar.

In Section 4.1, I will focus on results reproducibility by trying to analyse and compare the results from the Twitter classifier with the official flu data from the CDC. In Section 4.2, on the other hand, I will describe my attempts to reproduce two figures and one table from Bodnar (2015) by using the processed data and code provided to me, *i.e.* it contains my efforts to ensure methods reproducibility. Finally, Chapter 5 contains further remarks on the methods and results reproducibility of the findings in Bodnar (2015) as well as additional comments regarding inferential reproducibility.

Note that neither Section 4.1 nor Section 4.2 contain an exhaustive attempt of achieving

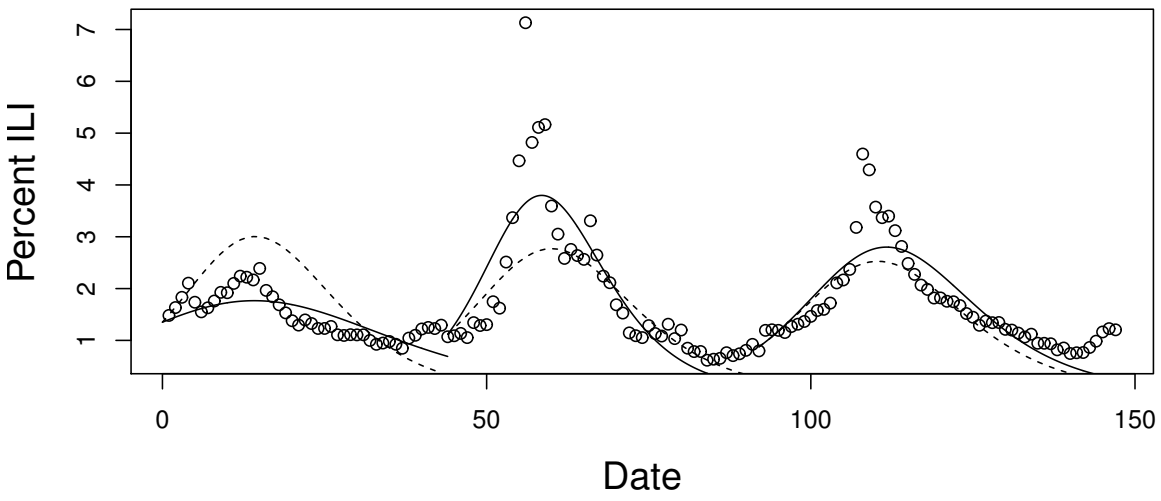


Figure 2.2: The CDC's estimates (circles) of influenza rates for a three year period compared to the best fit SIR models from the Twitter data using combined (dashed line) or yearly (solid line) parameters. The discontinuities stem from recalculating the SIR models for each year with the ILI rates from the Twitter classifier as a starting point for each year. From Bodnar (2015).

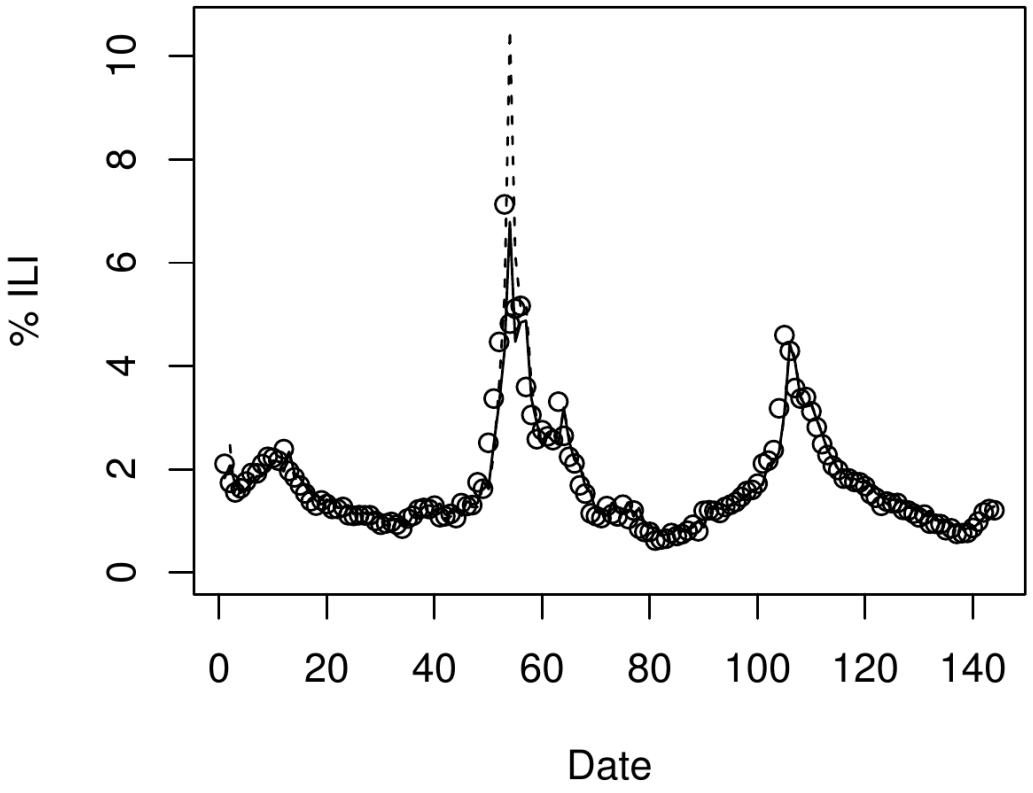


Figure 2.3: Twitter's forecasting (dashed lines) and retroactive measurements (solid lines) compared to the CDC's reported ILI rates (circles) for the whole U.S.. From Bodnar (2015).

results and methods reproducibility, respectively. In addition, one might very well argue that true results reproducibility can only be tested by recollecting a Twitter data set of similar size, rewriting the Twitter classifier, and finally comparing the results from the classified tweets to the official CDC data. However, I argue that using the same classified data set used in Bodnar (2015) still offers the possibility to test results reproducibility, namely by employing different statistical methods to compare the results from the Twitter classifier to the official CDC data. Mere methods reproducibility, on the other hand, would only focus on recreating the exact findings and figures from Bodnar (2015).

Chapter 3

(Almost) big data: How to analyse 2.8 billion tweets

In this chapter, I will describe the basic characteristics of the data set used by me in this thesis. The largest part of the data aggregation, rectification, and analysis was done using statistical programming language "R" (R Core Team, 2017), even though a few aggregation and transformation steps were also done using the Python programming language (Rossum, 1995). In the following, I will cite those packages which I have used most frequently or which were crucial for my data aggregation and analysis. In addition, Appendix A.1 provides the reader with an exhaustive list of all packages I used for this thesis, including the exact version number.

In addition, all packages, data sets, custom-written functions, their output and any additional information that might be important to understand or reproduce this work are made available on Github. See Appendix A.2 for more information about the structure of the Github repository.

3.1 The nature of the data beast

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 (2'847'039'672 tweets);

`one_hundred`: contains the rated tweets of those users who sent at least 100 tweets (42'611'004 tweets);

`sick_users`: contains the rated tweets of all those users who sent at least one sick tweet (4'131'650 tweets).

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`: The geographical longitude of the location in which the tweet was sent (in decimal degrees);

`latitude`: The geographical latitude of the location in which the tweet was sent (in decimal degrees);

`time`: A UNIX timestamp marking the time when tweet was sent;

`sick`: A binary variable indicating whether tweet was labelled as "sick" (=1) or "healthy" (=0) by the Twitter classifier;

`state`: The code for the 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, analysed, and visualised using the packages "data.table" (Dowle and Srinivasan, 2017), "bigmemory" (Kane *et al.*, 2013), "bigtabulate" (Kane and Emerson, 2016), "biganalytics" (Emerson and Kane, 2016), "stats" (R Core Team, 2017), and "ggplot2" (Wickham, 2009).

3.2 Description of the `sick_users` data set

As mentioned above, the `sick_users` 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 pre-processed and filtered the data set in order to remove all those tweets that were sent from outside the U.S. 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 42′860 entries were removed with 4′088′790 entries remaining. As Figure 3.1 shows, most of these tweets were clustered around cities, suburban areas and other population-dense regions of the U.S. You can also see that a considerable amount of the collected tweets was not sent from the United States, but from Mexico or Canada.

In the next step, I ran a custom-written function using a polygon look-up using the "R"-packages "maps" (Becker *et al.*, 2016), "geosphere" (Hijmans, 2016), and "dplyr" (Wickham and Francois, 2016). The function takes the coordinates of each tweet and compares it to the polygons of the U.S. states in order to determine the name of the state from which this tweet was sent or to remove it if it 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 additional 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 analysis. Luckily, this does not seem to be the case. Each state code could clearly be assigned to a specific U.S. 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 3.2).

Most of them either came to lie at the coastline or the Canadian-U.S.-border and the Mexican-U.S.-border, respectively. I removed a total of 180'290 tweets that were sent from either Canada or Mexico (see Figure 3.2). 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 U.S. 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 the majority of their neighbours within a 0.1×0.1 degree window. This way, an additional 211'511 tweets were removed, most of them at the coastline or from the ocean (see Figure 3.2).

After pre-processing, the `sick_users` data set was left with 3'696'989 tweets remaining. These tweets were sent by a total of 213'426 users, meaning that on average, each user sent 17.32 tweets. This is in stark contrast to the mean number of tweets reported in Bodnar (2015) (175.59 tweets per user over the whole study period). A slight decrease in the mean 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 mean tweet number seems suspect

(the time window analysed in Bodnar (2015) was March 3rd 2011 to March 4th 2015, so only slightly longer than in my case). As can be in Figure 3.3 a large part of the Twitter users in the `sick_users` data set only sent one or two tweets over the whole study period. Furthermore, the maximal number of tweets sent per user over the course of the 208 week time window was 86, a very low number given the fact that there are Twitter users out there who send over a hundred tweets *per day*.

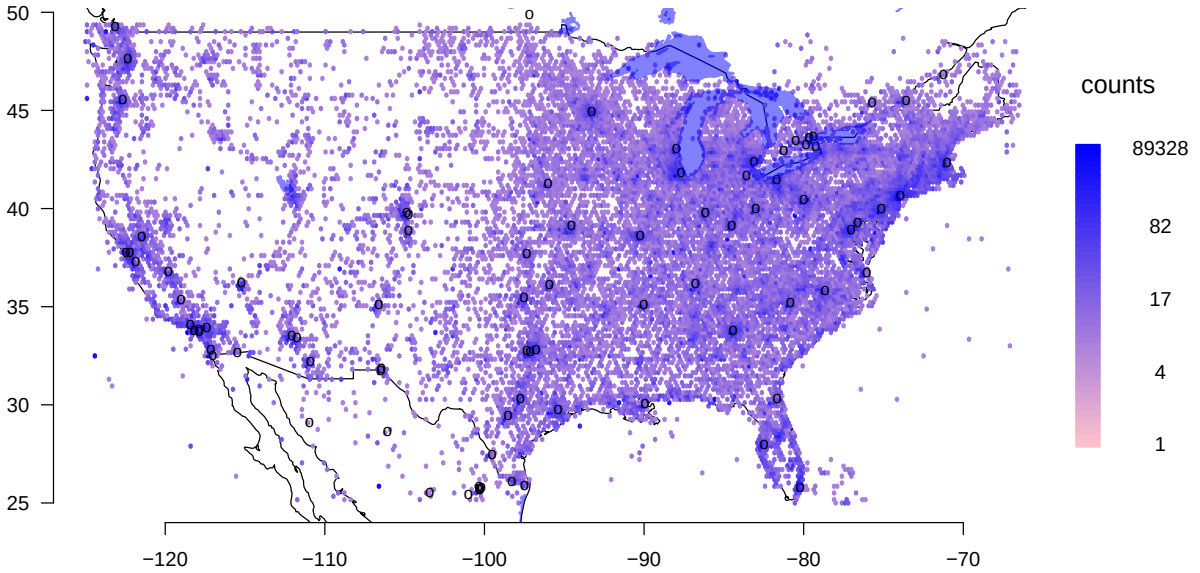
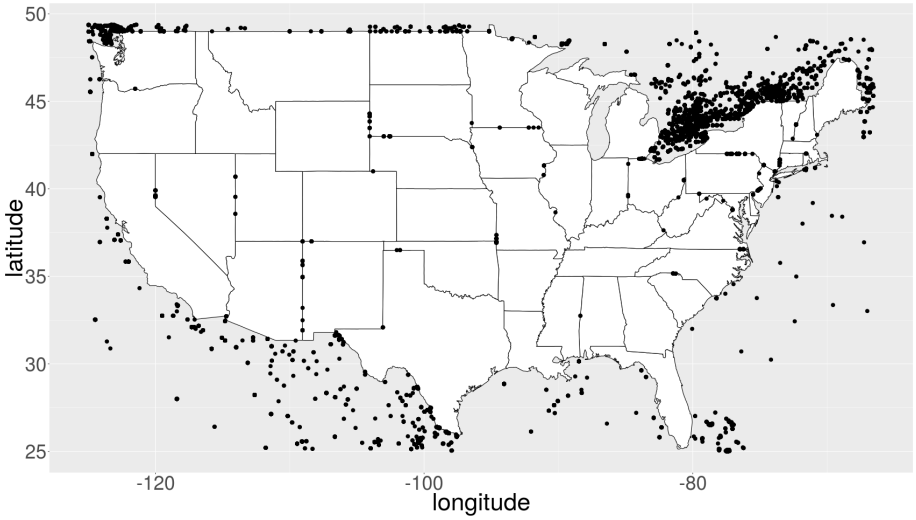


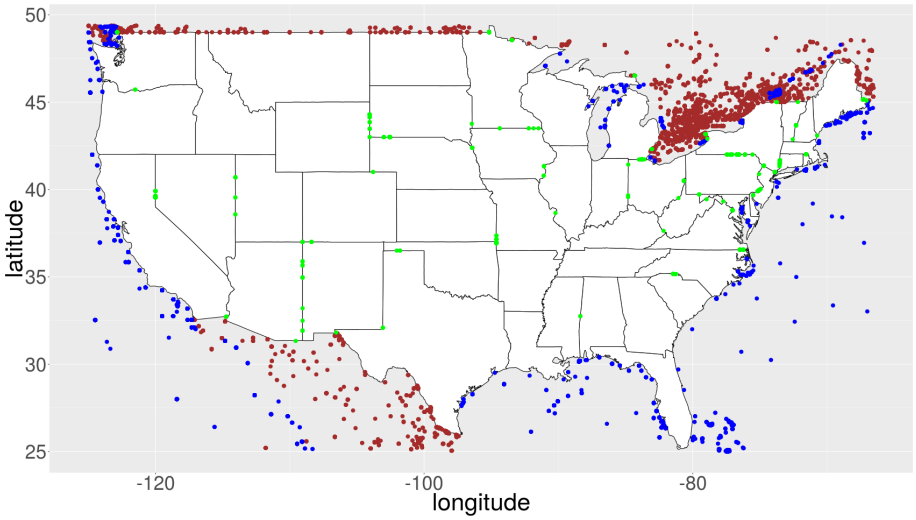
Figure 3.1: Spatial distribution of tweets in the `sick_users` data set, aggregated in 12'405 hexagon cells in grid dimensions 148 vertically by 301 horizontally. The black circles correspond to cities with more than 300'000 inhabitants. Note that the colour label does not follow a linear scale. Due to large differences in the minimal and maximal number of tweets in each bin, I performed a root transformation on the data using the tenth root. The plot was produced using the "R"-packages "hexbin" (Carr *et al.*, 2016), "maps" (Becker *et al.*, 2016), and "grid" (Murrell, 2003, 2007; Zhou and Braun, 2010).

Halfway through 2011, Twitter users sent approximately 200 million tweets per day or 6 billion tweets per month (Twitter, 2011). If we divide the latter number by the 151 million of Twitter users that were active on a monthly basis by the end of June 30th (Twitter, 2013), we get a mean of 39.74 tweets per user per month. Making the conservative assumption that this number remains constant and does not increase, we would expect a mean of 1907.29 tweets per user over the whole four year period between March 3rd 2011 to March 4th 2015—much more than the 86 tweets sent by the most diligent Twitter user in the `sick_users` data set.

One explanation for this discrepancy to the tweet rates found in the `sick_users` data set



(a)



(b)

Figure 3.2: (a) All tweets in the `sick_users` data set whose `state` variable is coded as "56", *i.e.* all tweets that could not be assigned to any specific U.S. state. (b) Tweets whose origins was determined to be in Canada or Mexico (brown) or which could not be assigned to any U.S. state (blue, mainly from the coastline or the ocean) and thus were removed from the `sick_users` data set. The green dots represent the tweets that whose `state` was coded as "56" as a code, but which could be recovered by the polygon look-up 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 a `state` code other than "56", but failed to be assigned to a state by the polygon look-up I performed.

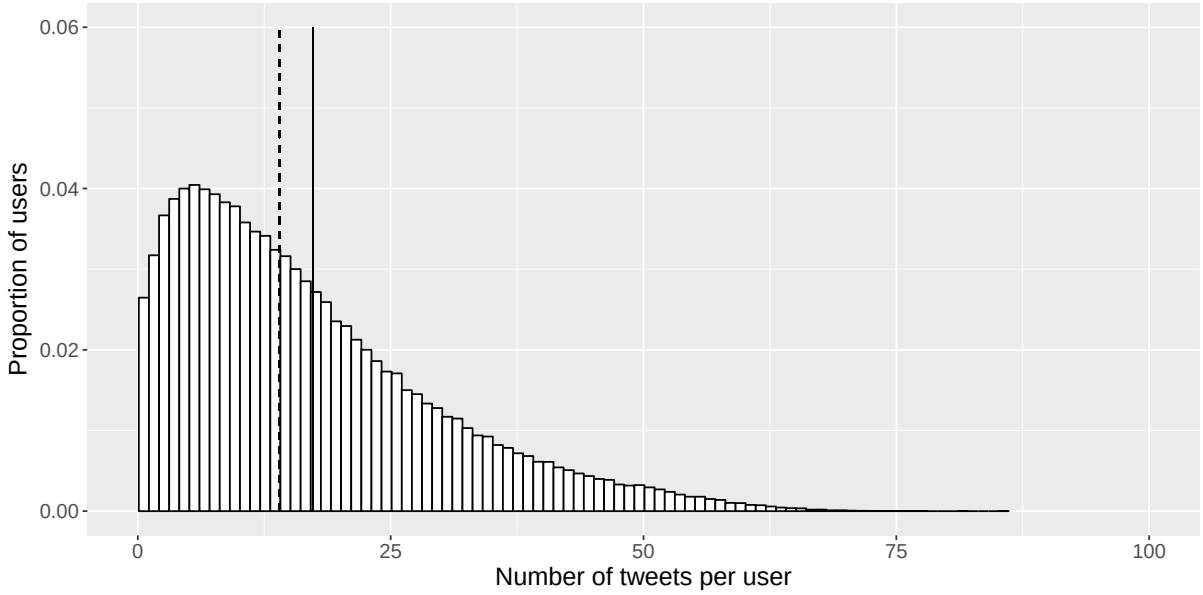


Figure 3.3: Histogram of the number of tweets sent per user in the `sick_users` data set during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1 tweet). Mean = 17.32 (solid line); median = 14 (dashed line).

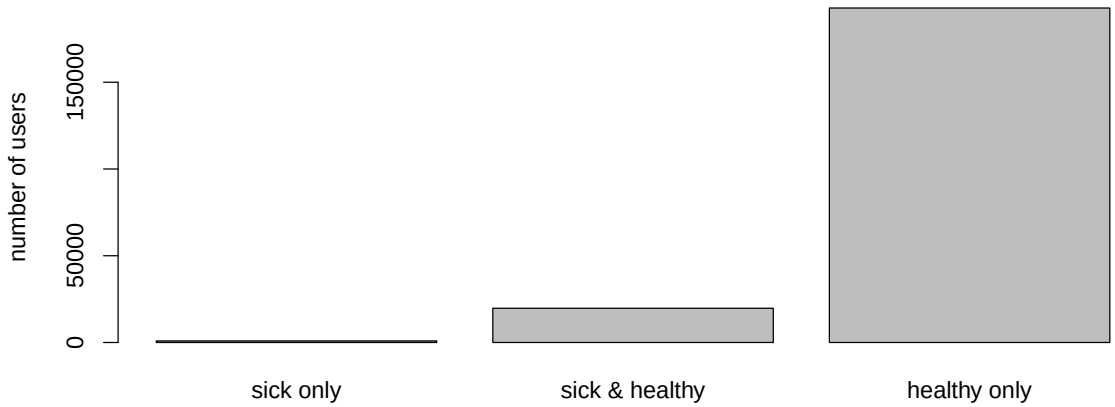


Figure 3.4: The total number of users who only sent tweets labelled as "sick" (919), users who sent at least one tweet labelled as "sick" and "healthy" (19'728), and users who only sent tweets labelled as "healthy" (192'779).

might be found in the fact that the data set only contains geotagged tweets which only represent 0.85 % (Sloan *et al.*, 2013) of all tweets.

If we make the naïve assumption that geotagged tweets represent a subset of all tweets that is uniformly distributed over all Twitter users (in other words: it is assumed that each Twitter user adds geolocation to 0.85% of his tweets) we receive a mean tweet rate of 2037.9 tweets per user in the `sick_users` data set, which is comparable to the mean tweet rate of all Twitter users in the real world (1907.29 tweets per user).

However, Sloan and Morgan (2015) have shown in a recent study that only 58.4 % of all users have location services enabled and only 3.1% of all users ever sent a geotagged tweet. Hence, the complete body of geotagged tweets is clustered heavily around a small subset of Twitter users, so that our assumption of the uniform distribution of geotagged tweets over all Twitters users does not hold.

Nevertheless, the information described above allows us to calculate an estimate of the total amount of tweets and users between March 3rd 2011 to March 4th 2015 using the following formula:

$$\text{tweet rate} = \frac{T_{\text{tot}}}{U_{\text{tot}}} = \frac{T_{\text{geotagged}} \cdot \frac{1}{0.0085}}{U_{\text{geotagged}} \cdot \frac{1}{0.031}}. \tag{3.1}$$

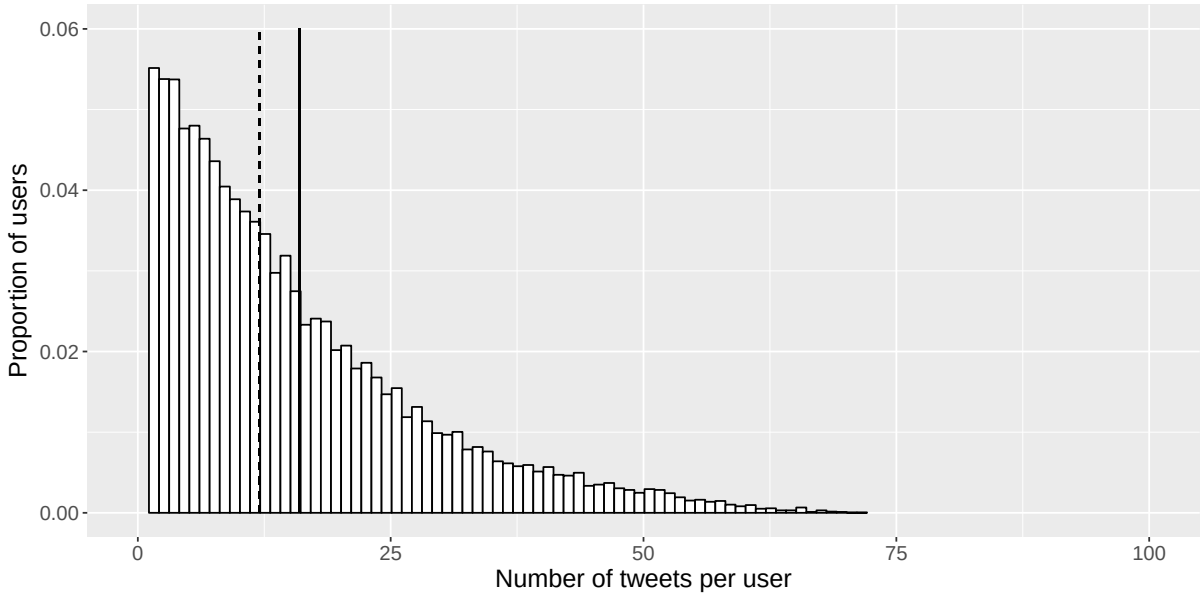
Here, T_{tot} and $T_{\text{geotagged}}$ depict the number of tweets in the real world and the number of tweets in the geotagged subsample, respectively, whereas U_{tot} and $U_{\text{geotagged}}$ denominate the number of Twitter users in the real world and in the geotagged subsample, respectively. Applying this formula yields a mean tweet rate of 63.17 tweets per user in the `sick_users` data set between March 3rd 2011 to March 4th 2015, which is higher then the unadjusted tweet rate per user in the `sick_users` data set, but still remains considerably lower than the conservative estimate of the real-world tweet rate based on the official Twitter data. Hence, I am led to believe that the `sick_users` data set is neither representative of rest of the data set, let alone the total corpus of tweets produced in the real world. This should not come as a surprise, since the `sick_users` data set is supposed to contain the tweets of all those users who sent at least one tweet that was classified as "sick" (Bodnar, personal communication).

However, the large majority of the users within the `sick_users` data set never sent a tweet that was labelled as "sick" by the flu classifier (see Figure 3.4). In fact, only 20'647 out of 213'426 (or 9.67%) ever sent such a tweet.

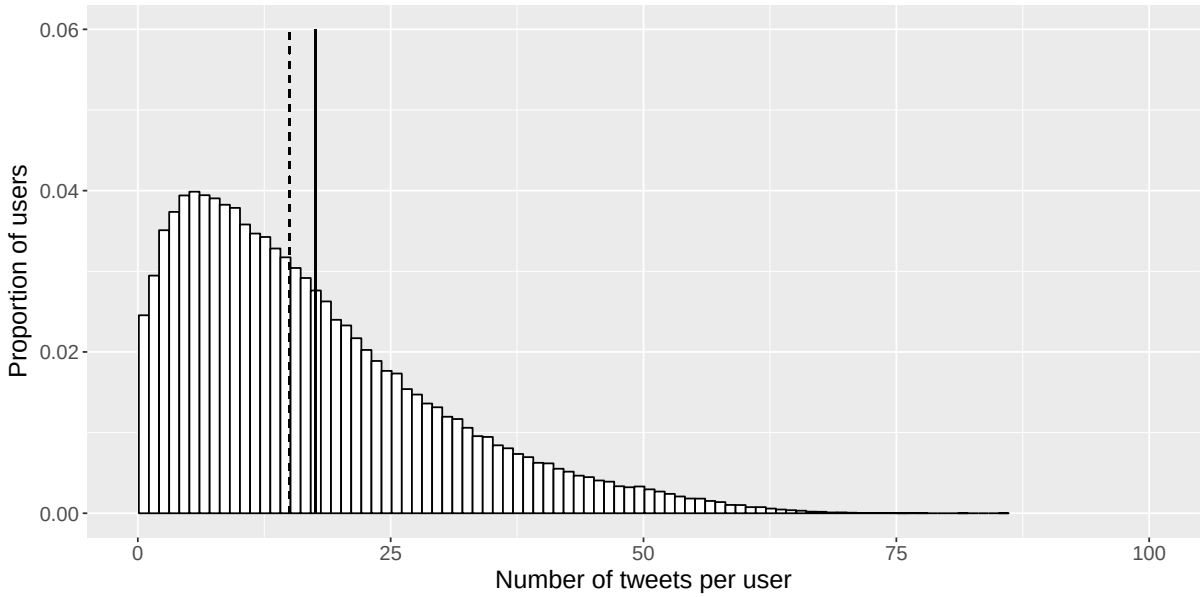
Also, a total of 919 users exclusively sent tweets that were labelled as sick—something that seems rather unlikely to happen. Finally, those 19'728 users who sent both "sick" and "healthy" tweets had a significantly lower average tweet rate than those users who only sent "healthy" tweets (16.01 and 17.53, respectively. $p = 0$ 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 observed in Figure 3.5.

Hence, it is unclear how exactly the `sick_users` 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 "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 `all_tweets` data set.



(a)



(b)

Figure 3.5: Histograms of the number of tweets sent per user during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1 tweet). (a) Users who sent at least one tweet labelled as "sick" and one tweet labelled as "healthy". Mean = 16.01 (solid line); median = 12 (dashed line). (b) Users who never sent a tweet that was labelled as "sick" by the classifier. Mean = 17.53 (solid line); median = 15 (dashed line). 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 "sick" and one tweet labelled as "healthy".

3.3 Description of the `all_tweets` data set

The complete data set consisted of a total of 2'847'039'672 tweets sent by 16'015'981. Hence, it contained more tweets sent by fewer users than the data set reported by Bodnar (2015) which consisted of 2'732'174'105 tweets sent by 15'560'328 users. The difference in the number of tweets might simply stem from the fact that the tweets in my data set were collected until July 2015, while the tweets analysed in Bodnar (2015) were only collected up to March 2015. However, the stark reduction in the number of users is puzzling and hints at the possibility that I was working with a different data set (also see Section 5.2.2 for further discussion). An overview over all differences between the two data sets as well as over key statistical indicators of the `all_users` data set is given in Table 3.1.

Table 3.1: Key differences between the data sets reported in Bodnar (2015) and the one I analysed. Values in white and grey-coloured rows are based on the raw and pruned data set, respectively.

	Bodnar	Grüninger	
# of users	15'560'328	16'015'981	raw data set
# of users, pruned	–	15'229'049	pruned data set
# of tweets	2'732'174'105	2'847'039'672	raw data set
# of tweets	–	2'764'210'962	pruned data set
# of sick users, 2011–2015	182'801	27'052	pruned data set
# of users, 2011	45'086	175'382	pruned data set
mean tweet rate, 2011–2015	175.59	177.76	raw data set
mean tweet rate, 2011–2015	–	181.51	pruned data set
mean tweet rate, weekly	–	31.42	pruned data set
median tweet rate, 2011–2015	10	9	raw data set
median tweet rate, 2011–2015	–	9	pruned data set
median tweet rate, weekly	–	32.53	pruned data set
max. # of tweets per user, 2011–2015	1'119'384	1'413'568	raw data set
max. # of tweets per user, 2011–2015	–	1'413'568	pruned data set
min. # of tweets per user, 2011–2015	1	1	raw data set
min. # of tweets per user, 2011–2015	–	1	pruned data set

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 U.S. mainland (W $-125°$, W $-66°$, N 25° , N 50°) as described above, and aggregated the remaining 2'764'210'962 tweets and 15'229'049 users with

regard to states and weeks in which the tweets where sent, *i.e.* I calculated the number of tweets sent within a given week in a given state.

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 a 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 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 U.S. mainland, 1 label for the District of Columbia and 1 label for the tweets that could not be assigned to any of the other 49 areas), I received a data set with 10'400 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:` The total number of tweets that were labelled as "sick" in the given week and state;

`total:` The total number of tweets sent in the given week and state;

`healthy` The total number of tweets that were not labelled as "sick" in the given week and state;

`sick_per` The percentage of tweets labelled as sick among the total tweets sent in the given week and state.

Next, I added the corresponding date and state name to each week and state index, respectively. In order to assign state names to their respective state labels, I used the label/name relationships established in the `sick_users` 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 3.2), I removed all corresponding state/week pairs from the aggregated data set (223'352'319 tweets), resulting in a data set with 10'192 rows and 16 columns (see below), containing 2'614'711'121 tweets from 15'229'049 users 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 1'189'809 tweets labelled as "sick", a number that is considerably larger than the 20'894 tweets labelled as "sick" in the `sick_users` data set. This further shows that the latter does not contain the full subset of tweets labelled as "sick". Relatively speaking, 0.05% of all tweets in the `all_tweets` data set were labelled as "sick" (as opposed to 9.67% in the `sick_users` data set).

The total amount of users who have sent at least one tweet labelled as sick during the study period was 27'052. Note that this is an upper estimate, since a user could be classified as "sick" more than once during the study period. Since my analysis rests on weekly aggregated data, I was 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 an obstacle 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 over the whole study period is considerably lower than the number reported in Bodnar (2015) (27'052 vs. 182'801 users, respectively), despite the former being an upper estimate of the total number of individual sick users. On the other hand, the number of individual users found in the data set during the first year (2011) for the whole country is larger than the one reported by Bodnar (2015): 175'382 and 45'086, respectively.

Using the information that only 3.1% of all users send geotagged tweets (Sloan *et al.*, 2013), we can estimate the total number of Twitter users that were active in 2011 based on the two mentioned sample estimates, respectively. While the former sample estimate gives us an estimated average of 5'657'476 active users in 2011, the latter estimate based on Bodnar (2015) only amounts to 1'454'387 total active users in 2011. Both numbers are a far cry from the roughly 19–30 million of monthly active Twitter users officially reported by Twitter in 2011 (Twitter,

2013), even though part of this discrepancy might be explained by the first two months of 2011 are missing from the collection of the Twitter data sets Bodnar and I were working on.These points give additional reason to believe that the data reported in Bodnar (2015) are not the same as the data I analysed. In addition, I could observe a peculiar difference in the mean tweet frequency between healthy and sick users. While healthy users had an unadjusted mean tweet rate of 31.42 tweets per week, sick users had an unadjusted mean tweet rate of 45.95 tweets per week (see Figure 3.6), a difference that is highly significant (Mann–Whitney U-Test, p =0).

Also, the unadjusted mean tweet rate per week of all users combined (31.42) is six times smaller than the mean tweet rate over the whole study period (181.51). The same holds true for the unadjusted median tweet rate (32.53 tweets per user per week vs. 9 tweets per user over the whole study period). These differences make intuitive sense, since a user the relative weight of more active Twitter users is more pronounce when aggregating over the whole study period (*i.e.* over four years), instead of only aggregating over the course of a week.

Finally, by applying Equation 3.1, we can estimate the adjusted mean (114.6 tweets per user) and median (118.64 tweets per user) of the Twitter rate of the whole Twitter population in 2011, as well the mean tweet rate for sick (167.6 tweets per user) and healthy (114.58 tweets per user), respectively.

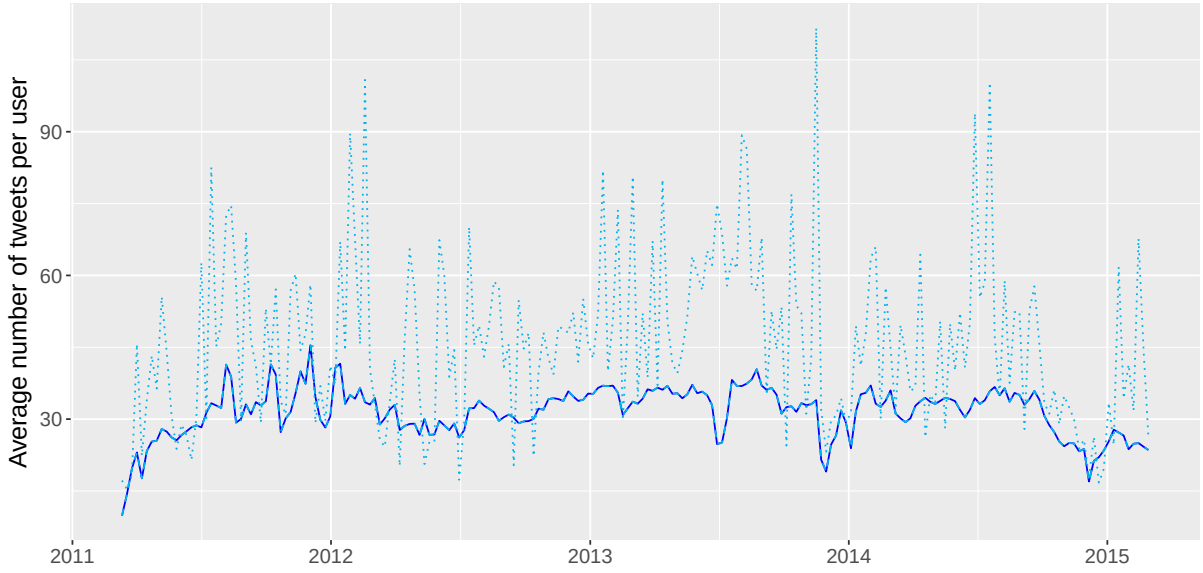


Figure 3.6: The unadjusted mean number of tweets per week sent by sick users (dotted light blue), healthy user (dashed light blue), and total users (solid blue), respectively.

Chapter 4

What does the classified Twitter data set reveal about the flu?

In the first part of this chapter, I will describe the performance of the flu classifier compared to the weekly influenza rate estimates of the CDC (Section 4.1). In addition to this, I will critically assess the Twitter classifiers ability to predict the CDC's ILI reports as described in Bodnar (2015), in order to test which of the described findings are reproducible.

In the second part of this chapter, I will present my attempts to faithfully reproduce two key figures and one table from Bodnar (2015) by using the processed data and the code provided to me. This part will therefore shed light on the question whether the methodological and theoretical descriptions provided in Bodnar (2015) as well as the available data and code basis are sufficient to ensure methods reproducibility.

4.1 Can the Twitter classifier compete for gold?

The CDC publishes weekly ILI reports containing the number of patients with ILI symptoms relative to the total number of patients in the more than 2'800 outpatient healthcare providers that are part of the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). The reports provided by the ILINet are commonly regarded as the gold standard of epidemiological flu data, so comparing the performance of the flu classifier is an important first step to assess its validity.

However, since correctly diagnosing the flu can only be done through microbiological analysis, the CDC uses "influenza-like illnesses (ILI)" as a proxy to assess the prevalence of influenza cases in the United States. ILI are defined as "fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat without a known cause other than influenza" (Centers for Disease

Control and Prevention, 2016b). It is therefore important to note that I am not comparing the Twitter classifier's results to the number of flu cases in the U.S., but only to the number of registered ILI cases, *i.e.* a proxy of flu prevalence. But since the CDC's data are commonly very reliable and since the Twitter classifier itself uses the content of single tweets as a proxy between sick and healthy users, these kinds of comparisons make sense.

Before doing so, however, I will provide additional descriptive statistics of the data set.

4.1.1 Spatio-temporal patterns of Twitter usage

Figure 4.1 shows the total number of tweets sent per week relative to the total number of tweets sent in the study period. No obvious pattern is 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 "healthy" is almost indiscernible from the temporal pattern of the complete data set. When looking at the weekly amount of tweets labelled as "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 "sick" is in fact significantly different from the weekly activity of the tweets labelled as "healthy" ($p = 0.0264$ and $p = 0$ for the normalised and not normalised weekly tweet counts, respectively). As can be seen in Figure 4.2 the tweets labelled as "sick" follow a markedly different temporal pattern than those labelled as "healthy". Also, the distribution of the latter is virtually indistinguishable from the distribution of the total amount of tweets.

Repeating the calculations described above using number of Twitter users instead of number of tweets yields similar results (see Figure 4.3; $p = 0$ and $p = 0$ for the normalised and not normalised weekly user activity, respectively), However, using Twitter users instead of single tweets as basis for the aggregation reduces the weekly fluctuations of the weekly "sick" rates considerably (see Section 4.1.2 for further discussion).

Next, I looked at the total amount of tweets sent in each state. As can be seen in Figures 4.4a, 4.5a, and 4.5b the relative distribution per state largely follows the relative distribution of the state population. Notable exceptions are Maryland, New Jersey, West Virginia, and Delaware, which were the origin of many more tweets than expected, as well as New York, Idaho, Montana, District of Columbia, Wyoming, and Vermont, from where considerably fewer tweets originated than would be expected with regard to its population.

When comparing the relative number of tweets labelled as "healthy" with those labelled as "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 4.7). However, the states with the most pronounced differences (District of Columbia, Montana, South Dakota, 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 0.0801.

Repeating the calculations described above using the number of Twitter users instead of the number of tweets yields similar results (Figures 4.4b, 4.6a, 4.6b, and 4.8; $p = 0.0823$).

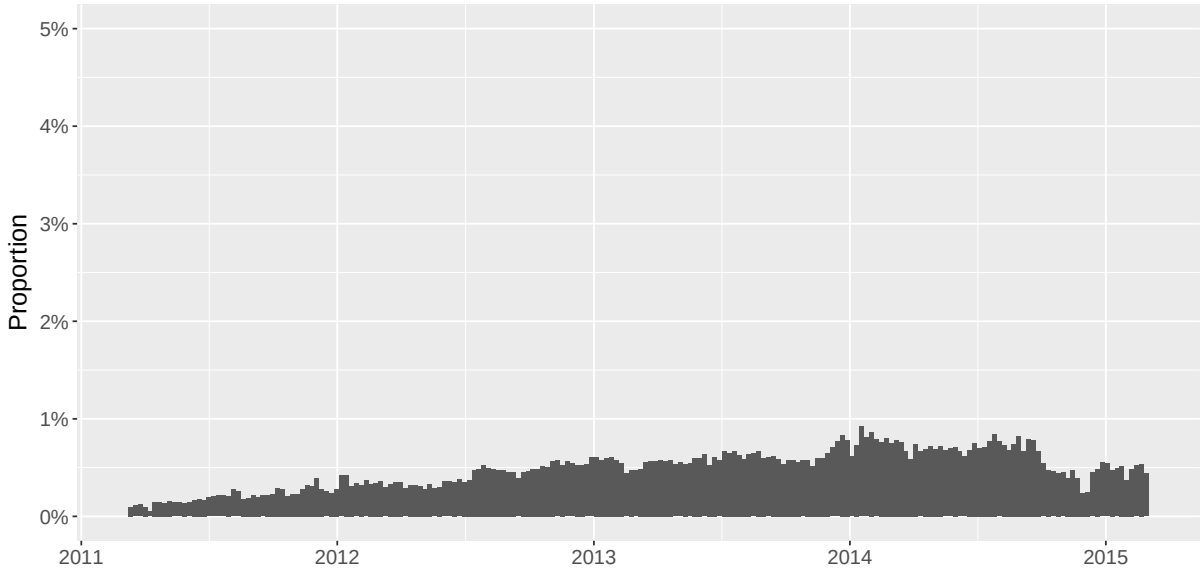


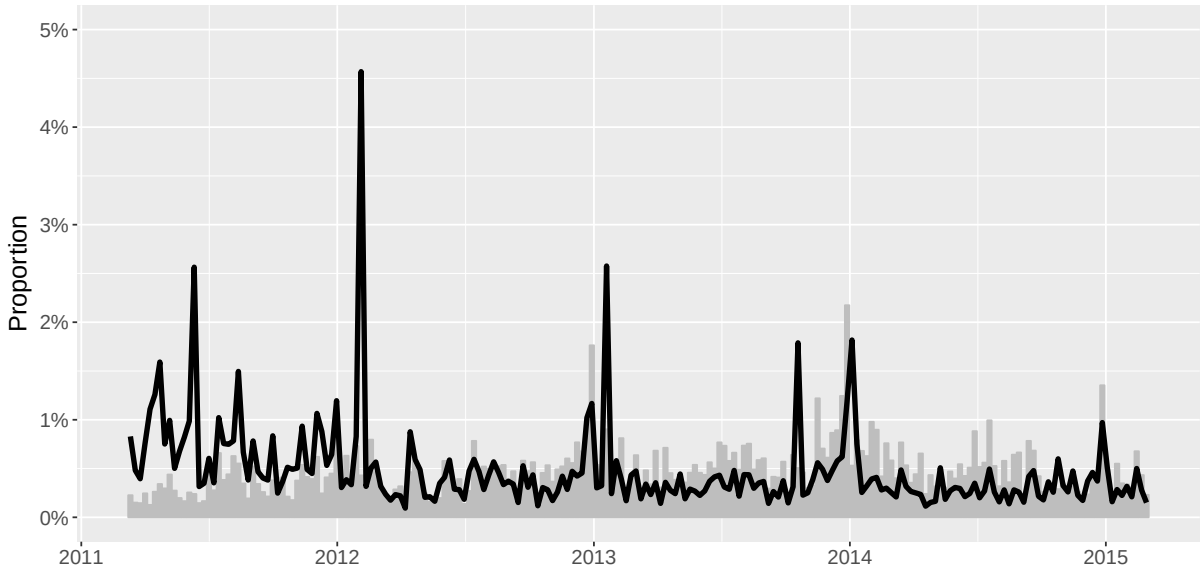
Figure 4.1: Relative number of tweets sent per week in the `all_tweets` data set between 2011-03-05 and 2015-07-11 (bin size = 1 week).

4.1.2 Comparing the flu classifier results with CDC ILI rates

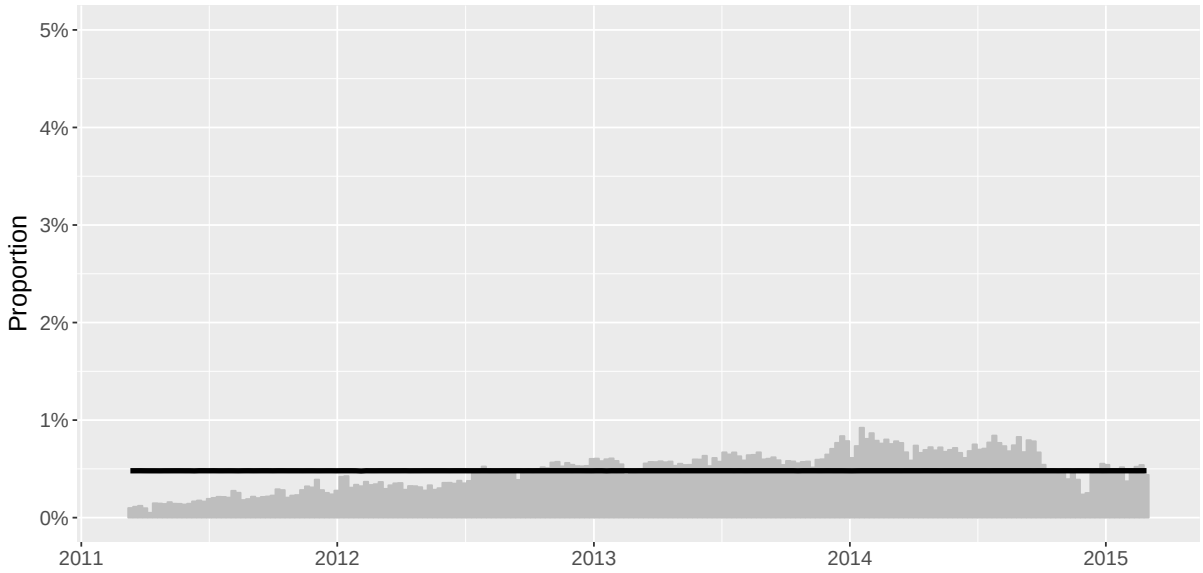
In order to assess the validity of the ILI predictions provided by the flu classifier, I compared the results from the Twitter classifier with the official ILI reports from the CDC on the national, regional and state level, extracted using the "cdcfluview" package (Rudis, 2016).

In a first step, I simply compared the official CDC ILI percentage data on the national level with the relative number of tweets labelled as "sick" per week and the relative number of sick users per week, respectively. As can be seen from Figure 4.9, the relative results from the Twitter classifier are an order of magnitude smaller than the official ILI data from the CDC.

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 4.10 do not represent weekly ILI percentage, but rather the percentual proportion of the relative number of tweets and the ILI percentages, respectively, of

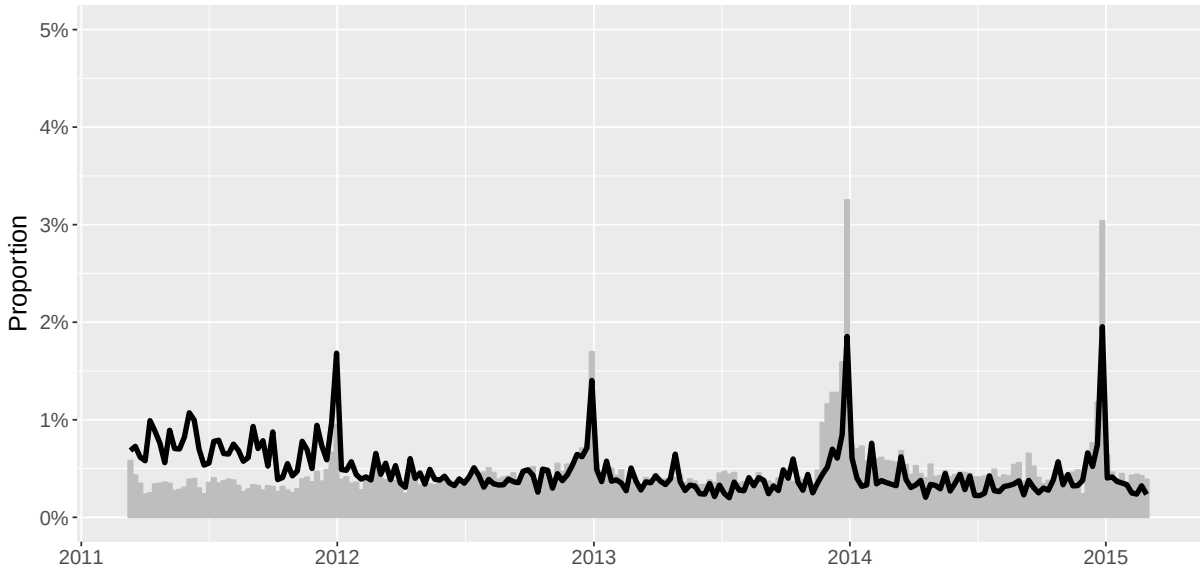


(a)

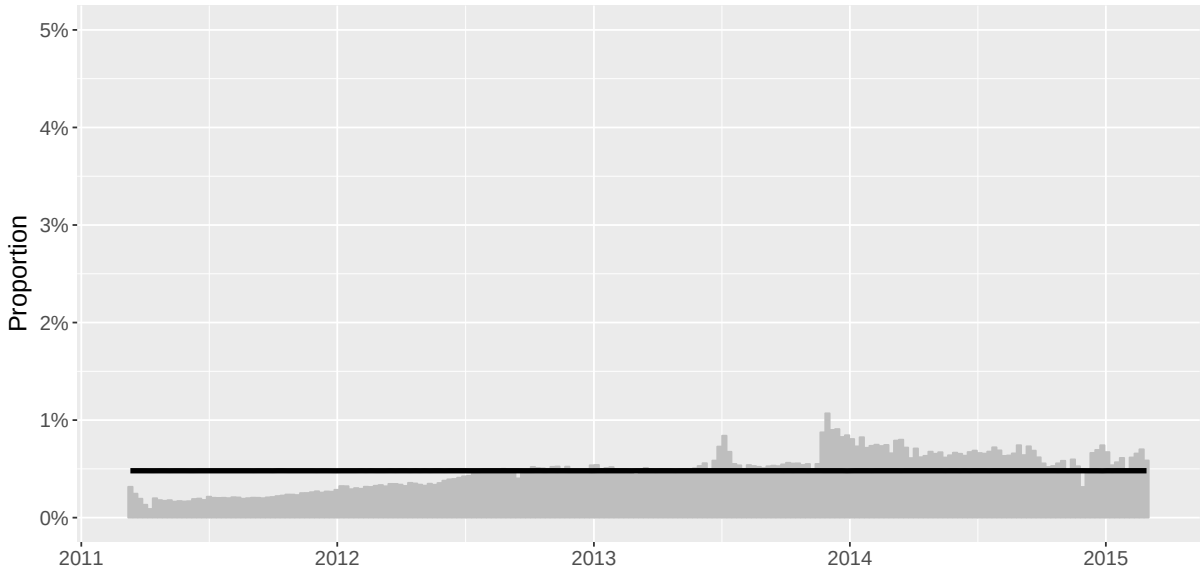


(b)

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) shows only tweets labelled as "sick"; (b) shows only tweets labelled as "healthy" by the classifier. The grey bars indicate the relative number of sick or healthy tweets sent per week without normalisation. The black lines indicate the relative number of sick or healthy tweets sent per week normalised by the total amount of tweets sent in that week.

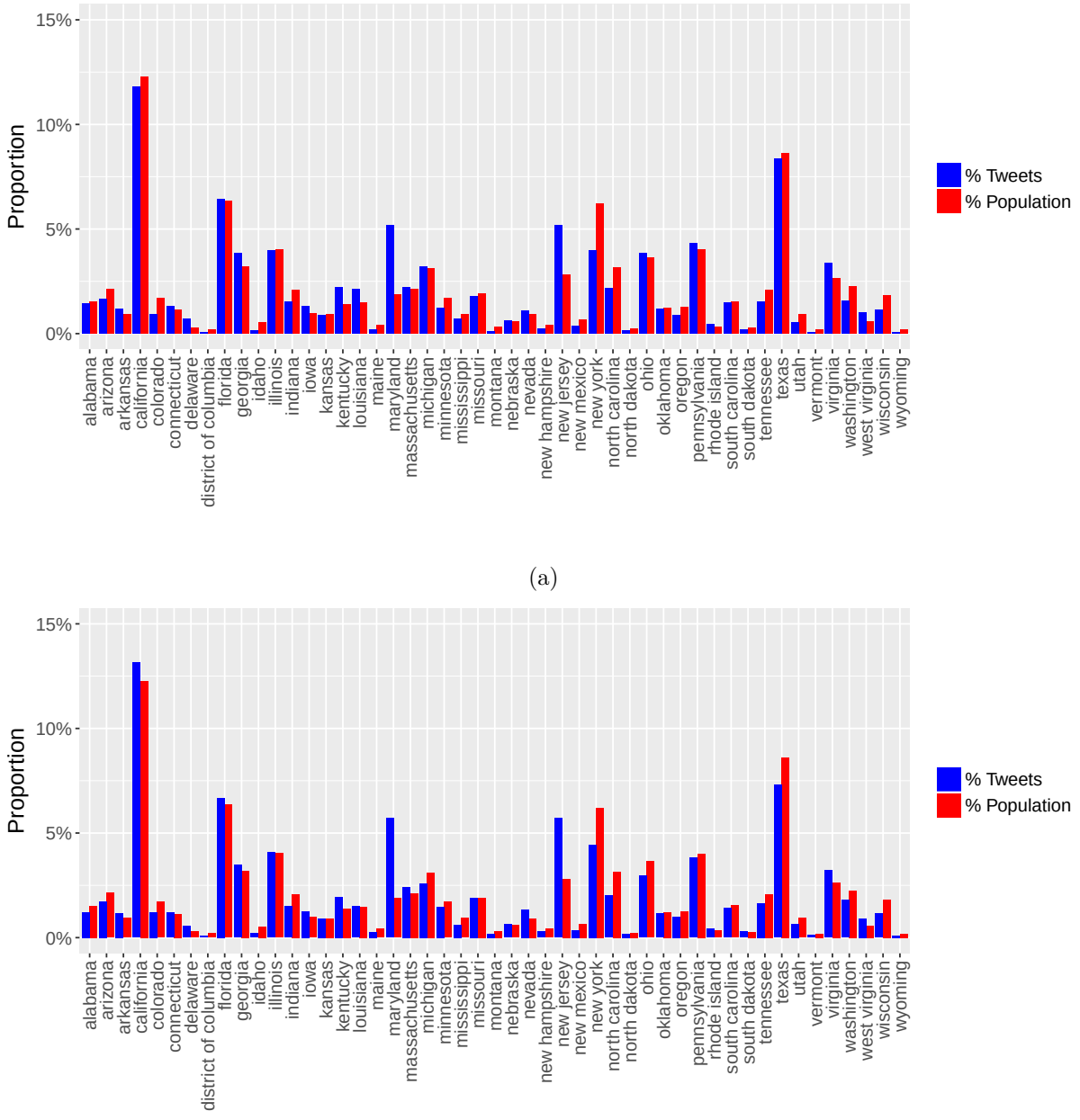


(a)



(b)

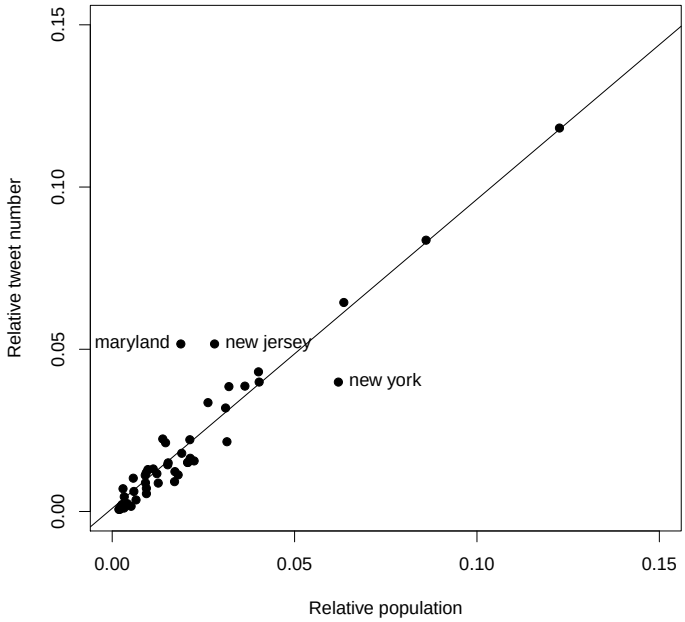
Figure 4.3: Histograms of numbers of the tweets sent per week during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1 week). (a) shows only tweets labelled as "sick"; (b) shows only tweets labelled as "healthy" by the classifier. The grey bars indicate the relative number of sick or healthy tweets sent per week without normalisation. The black lines indicate the relative number of sick or healthy tweets sent per week normalised by the total amount of tweets sent in that week.



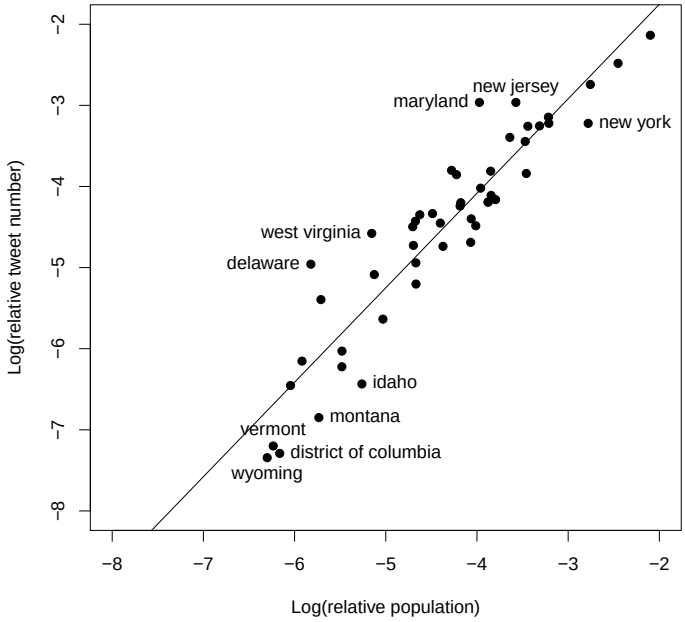
(a)

(b)

Figure 4.4: 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.

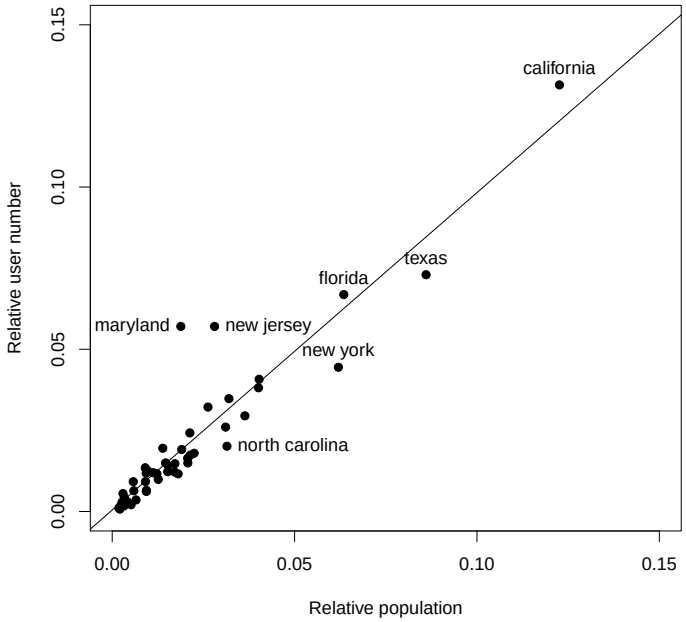


(a)

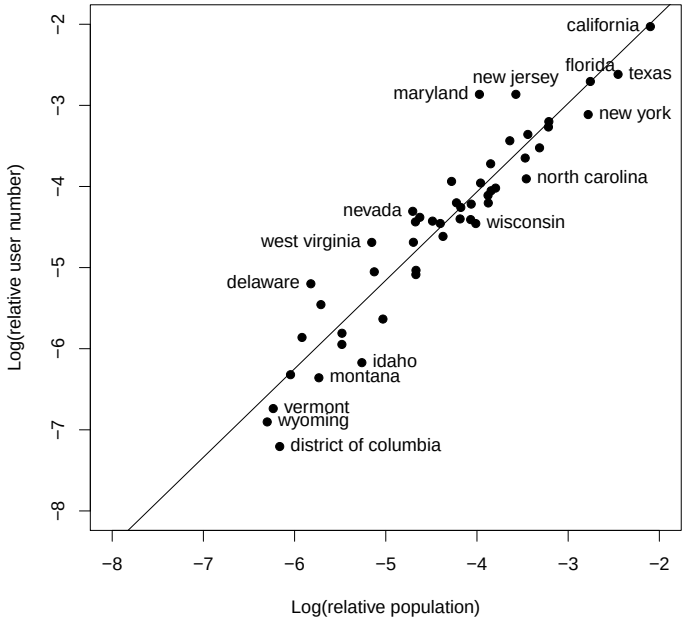


(b)

Figure 4.5: Relative number of tweets sent per state in the `all_tweets` data set between 2011-03-05 and 2015-07-11 plotted against each state's relative population size on a linear (a) and logarithmic scale (b). Some labels were removed to improve readability. The regression lines indicate the best linear fit with intercept = 0 and slope = 0.95 in (a) and intercept = 0.58 and slope = 1.16 in (b), respectively.

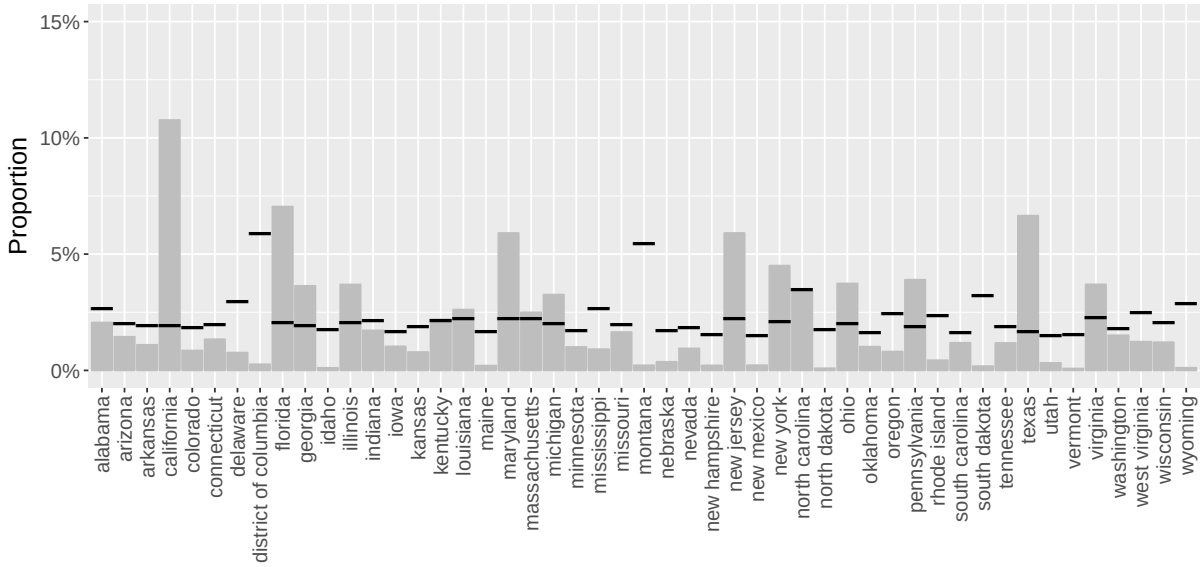


(a)

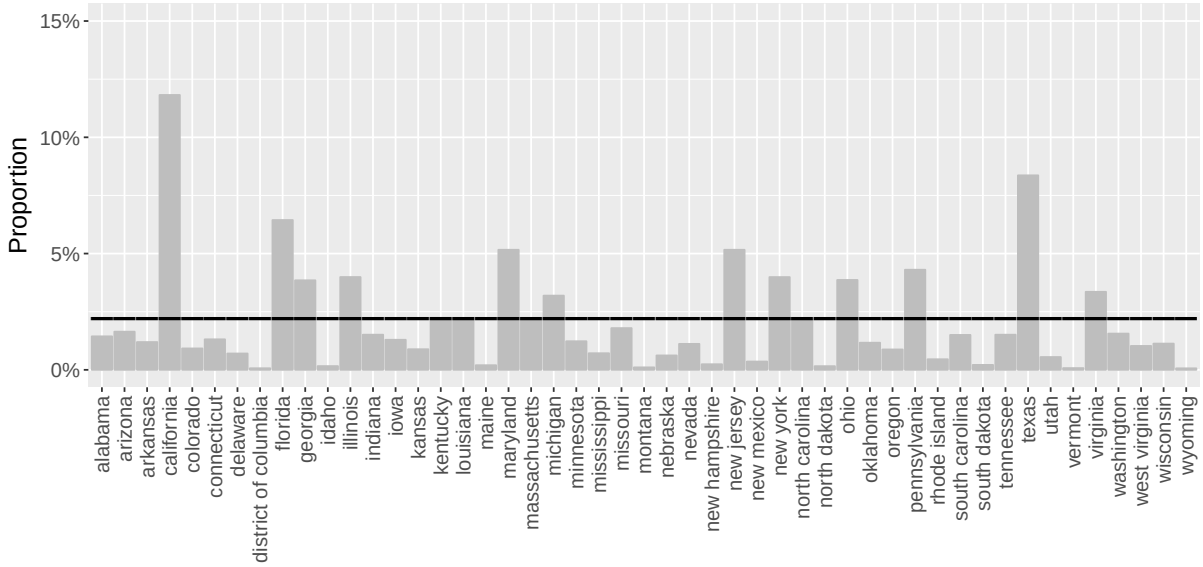


(b)

Figure 4.6: Relative number of Twitter users per state in the `all_tweets` data set between 2011-03-05 and 2015-07-11 plotted against each state's relative population size on a linear (a) and logarithmic scale (b). Some labels were removed to improve readability. The regression lines indicate the best linear fit with intercept = 0 and slope = 0.98 in (a) and intercept = 0.3 and slope = 1.09 in (b), respectively.

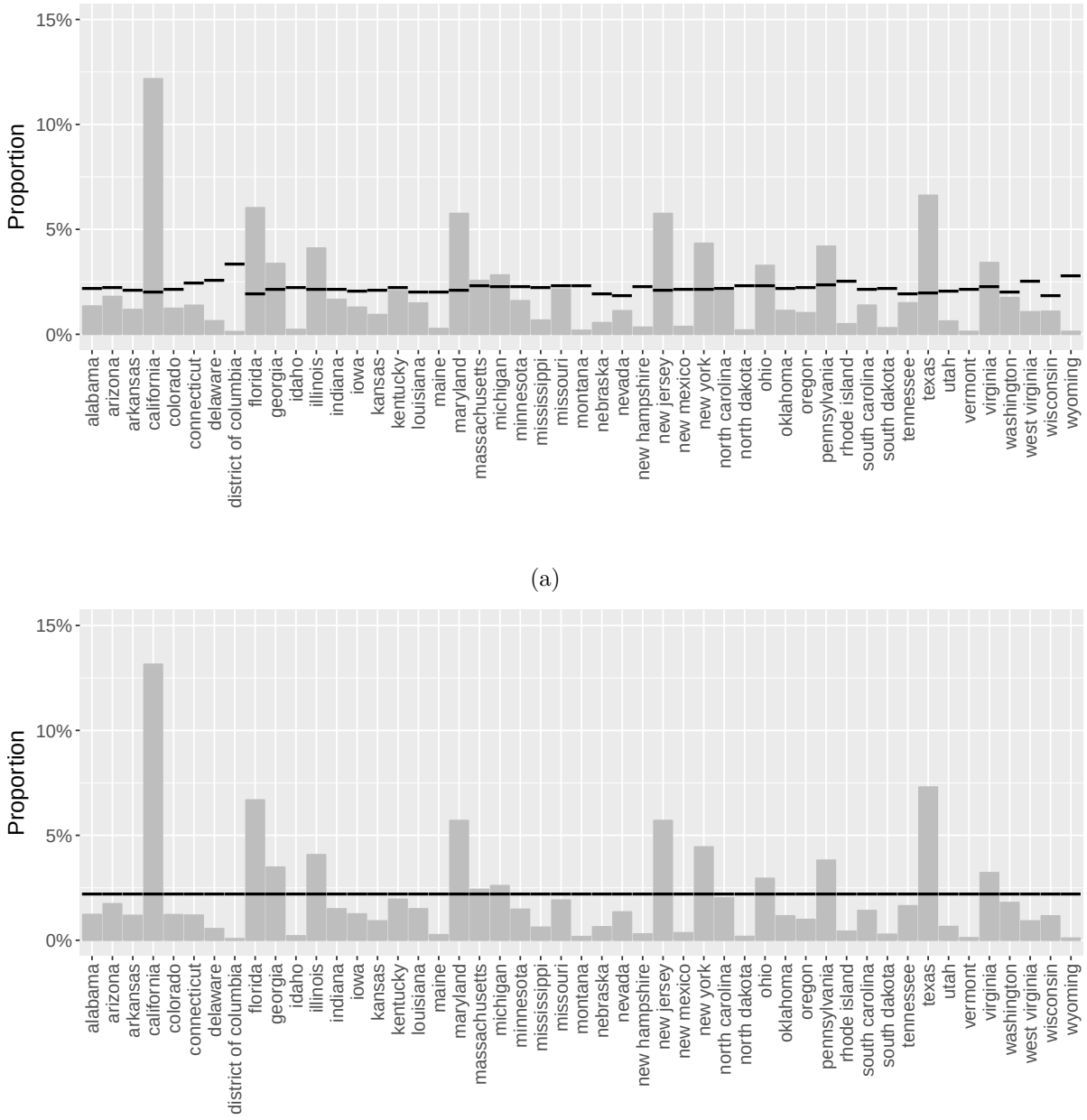


(a)



(b)

Figure 4.7: Histograms of numbers of tweets sent from each state during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1 week). (a) shows only tweets labelled as "sick"; (b) shows only tweets labelled as "healthy" by the classifier. The grey bars indicate the relative number of sick or healthy tweets sent in each state without normalisation. The black lines indicate the relative number of sick or healthy tweets sent in each state normalised by the total amount of tweets sent that state.



(a)

(b)

Figure 4.8: Histograms of numbers of users active in each state during the 208 weeks between 2011-03-05 and 2015-07-11 (bin size = 1 week). (a) shows only users who sent at least one tweet labelled as "sick"; (b) shows only users who were never classified as "sick" during the whole study period. The grey bars indicate the relative number of sick or healthy users active in each state without normalisation. The black lines indicate the relative number of sick or healthy users active in each state normalised by the total amount of tweets sent in that week.

a given week within the whole 208 week study period (In other words: The percentages of each week add up to a 100%, just like the histogram values shown in Figures 4.3–4.1). Since the fluctuations in the Twitter data were very high, I plotted the data again after applying a two-week (Figure 4.11a) and four-week (Figure 4.11b) moving average smoother, using the "forecast" package (Hyndman and Khandakar, 2008). 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 4.12). 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 4.14, 4.15 and 4.16. 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 0.0077 and 0.0077 for tweet- and user-based four-week average curves, respectively).

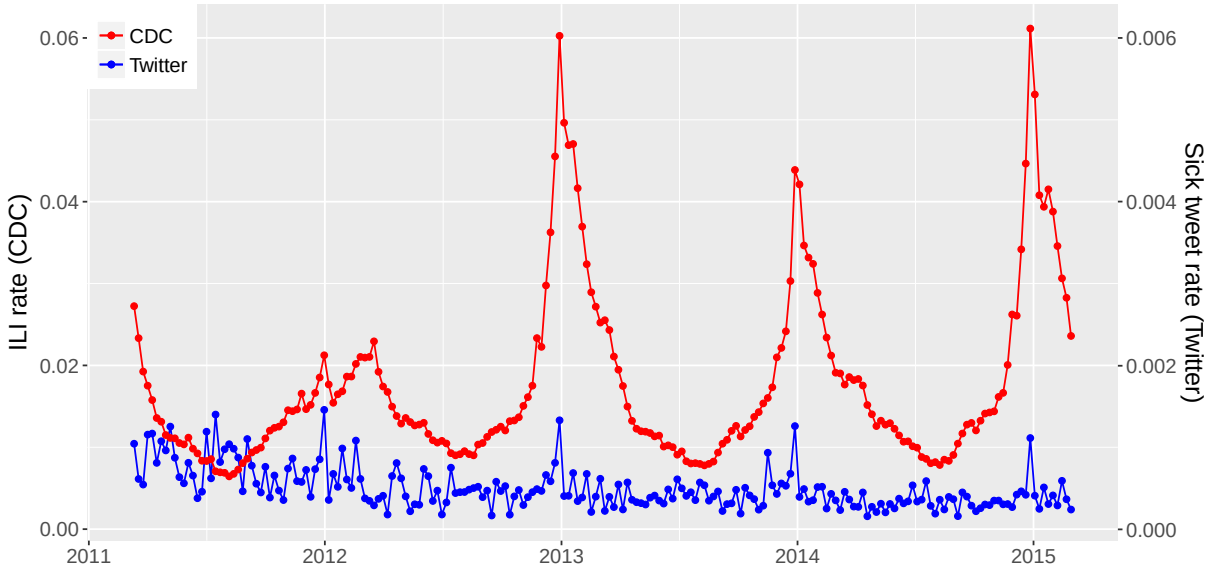
4.1.3 Comparing classifier results with CDC activity levels

Next, I attempted to reduce the fluctuations 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.

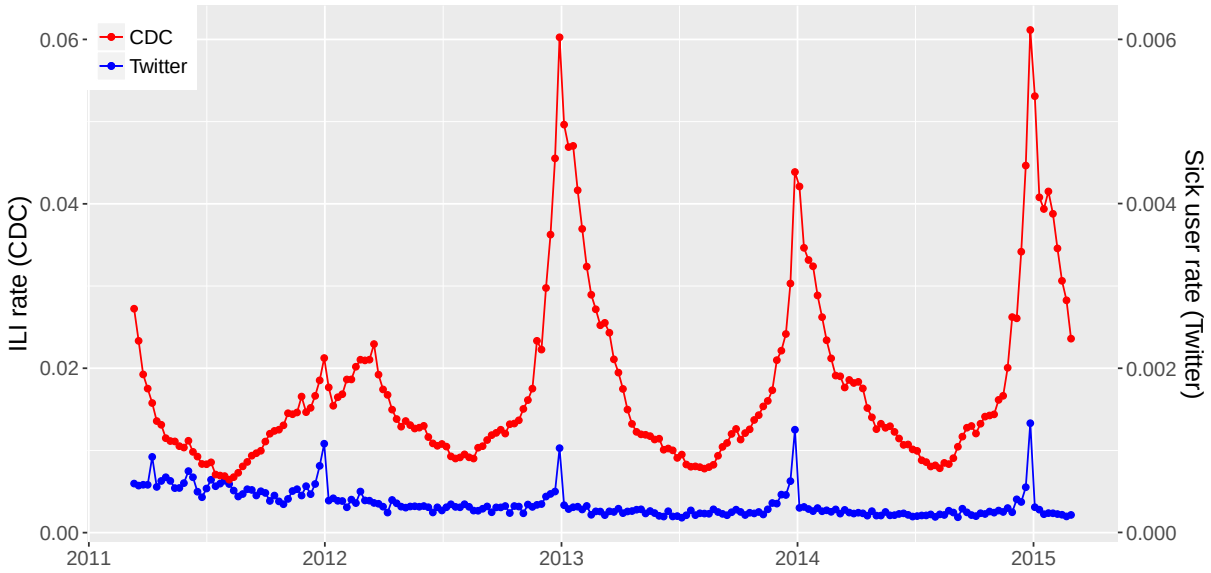
The CDC differentiates between ten different ILI activity levels which represent the deviation relative to the ILI baseline values. The activity levels compare the mean reported percent of visits due to ILI for the current week to the ILI baseline based on the number of reported ILI cases during non-influenza 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, 2016b).

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, 2016b).

Since a similar threshold does not exist for the Twitter data, I simply used the relative number of tweets labelled as "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 4.17a and Figure 4.18 shows the comparison on the national and regional level, respectively.



(a)



(b)

Figure 4.9: 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. Not that the Twitter rates are given in order of per mille of all tweets and users, respectively (right y-axis), while CDC ILI rates are given in order of percent of all patients visiting ILINet outpatient clinics (left y-axis).

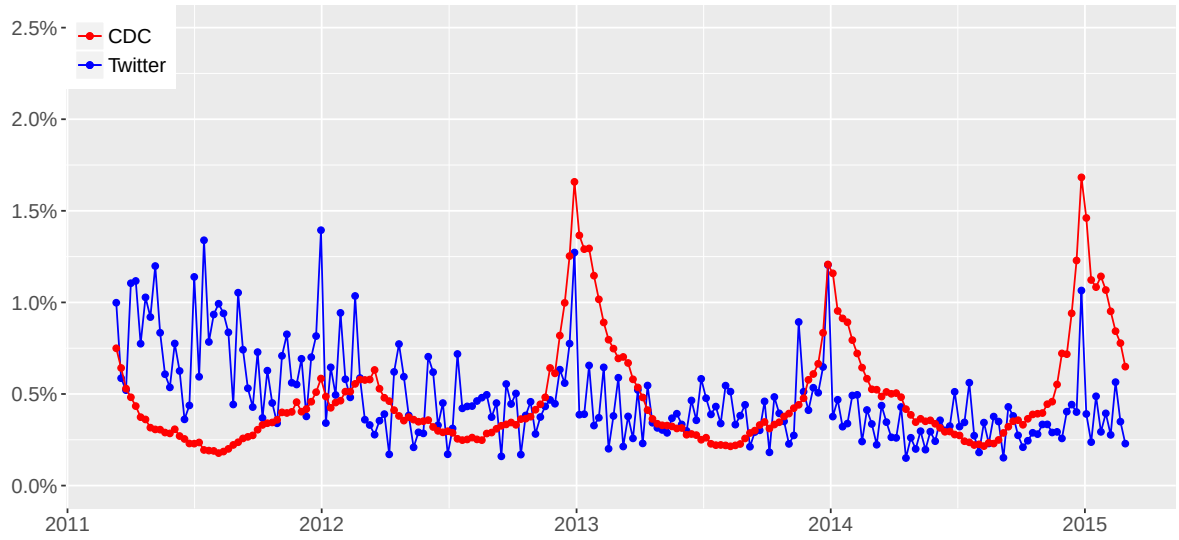
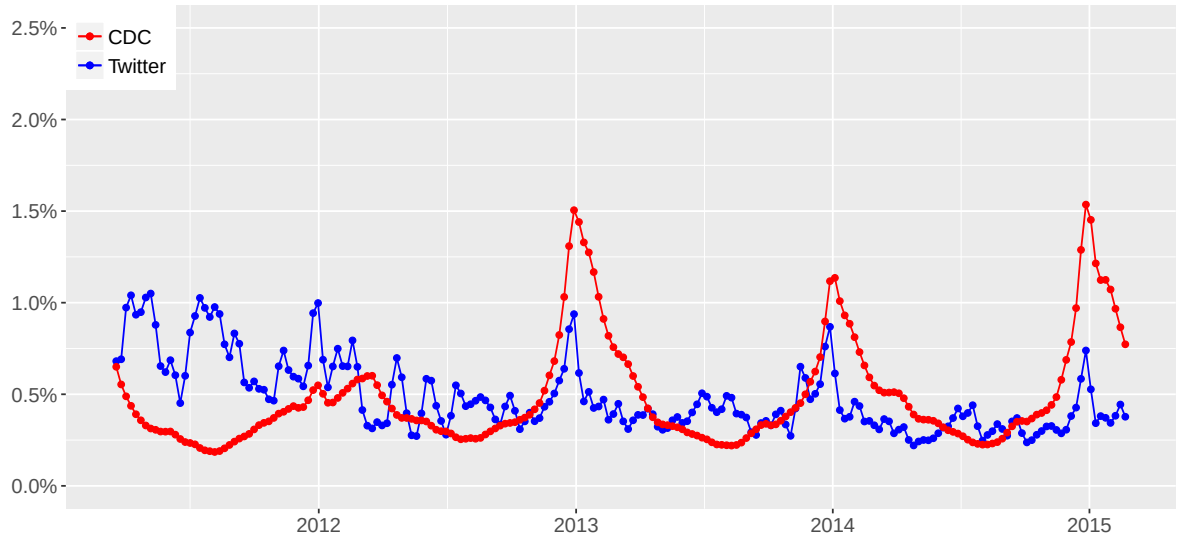
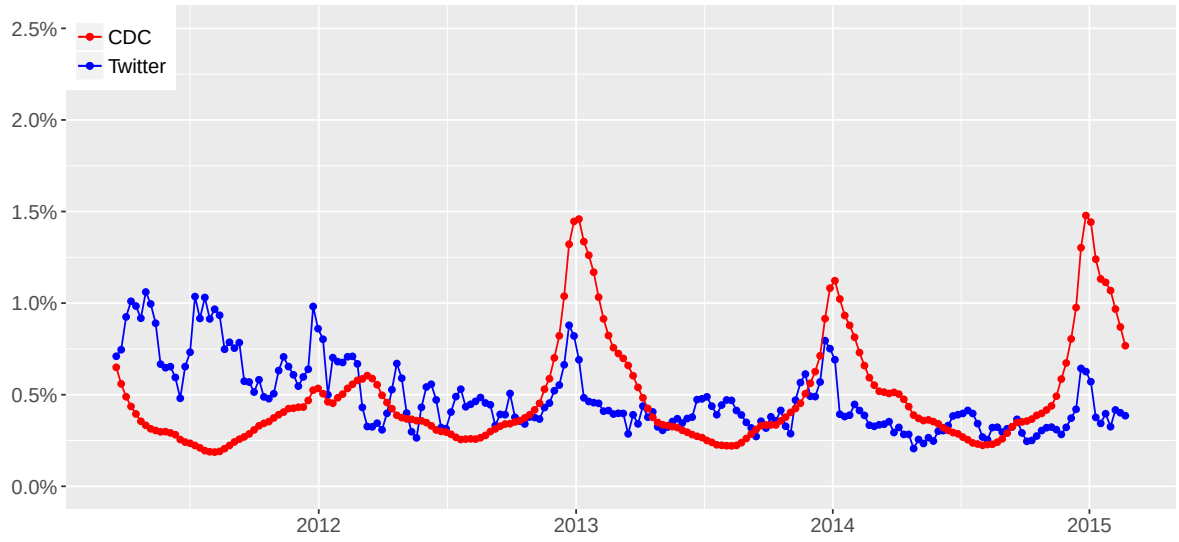


Figure 4.10: Comparison between weekly CDC ILI rates (red) and the relative amount of tweets labelled as "sick" from the Twitter flu classifier (blue). The data has been normalised in order to make them comparable, *i.e.* the percentages do not represent weekly ILI percentages, but instead sum up to a 100% over the whole time period.



(a)



(b)

Figure 4.11: Comparison between weekly CDC ILI rates (red) and the relative amount of tweets labelled as "sick" from the Twitter flu classifier (blue) after applying a two-week moving average smoother (a) and after applying a four-week moving average smoother (b), respectively. 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.

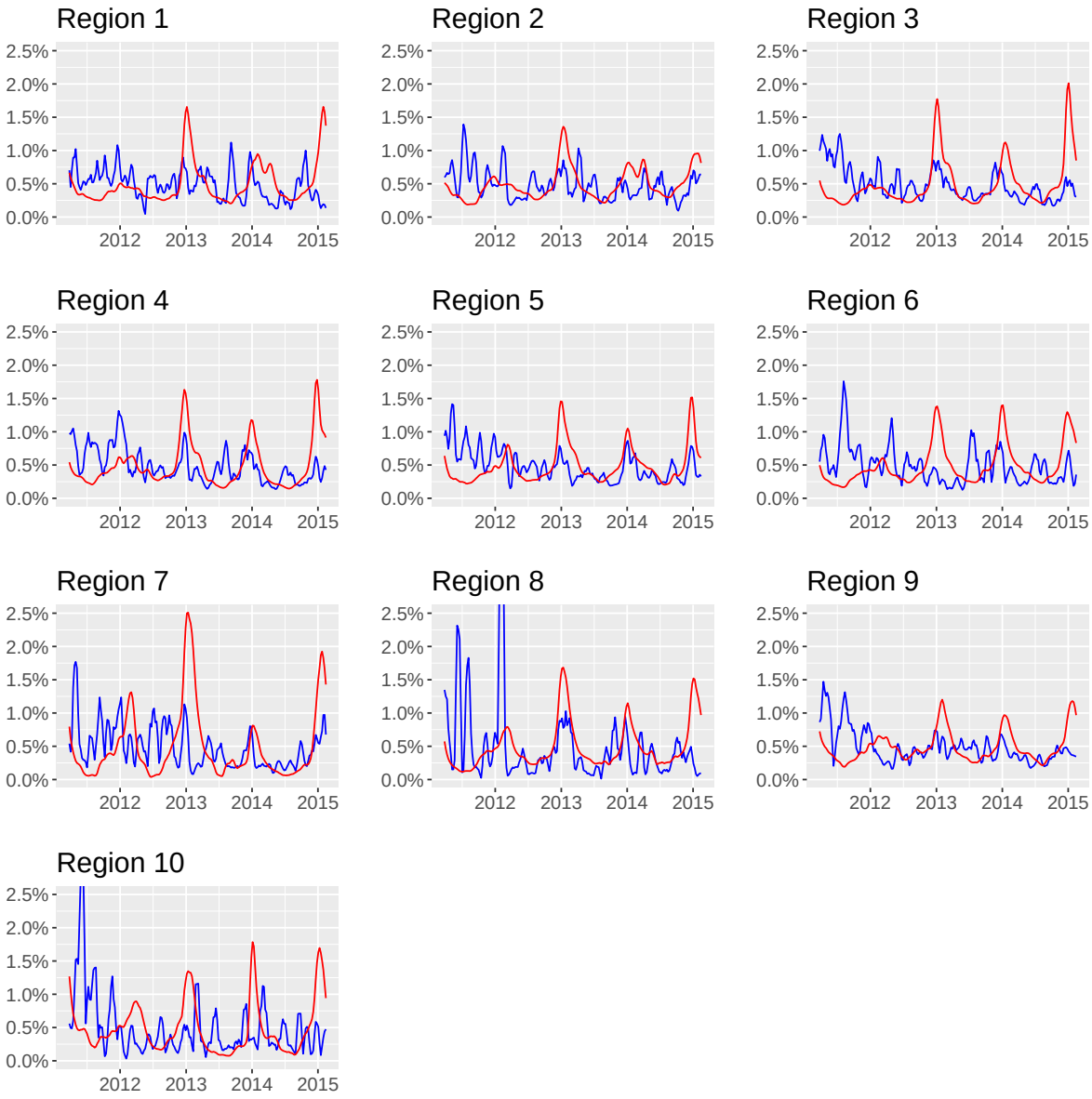


Figure 4.12: Relative number of tweets sent within each CDC ILI surveillance region per week (blue) compared with weekly ILI percentages in those regions (red). Data has been normalised and processed with a four-week moving average smoother. Note that Region 2 contains "Puerto Rico" and "Virgin Islands", Region 9 contains "Hawaii" and Region 10 contains "Alaska", all of which are missing from the Twitter data set.

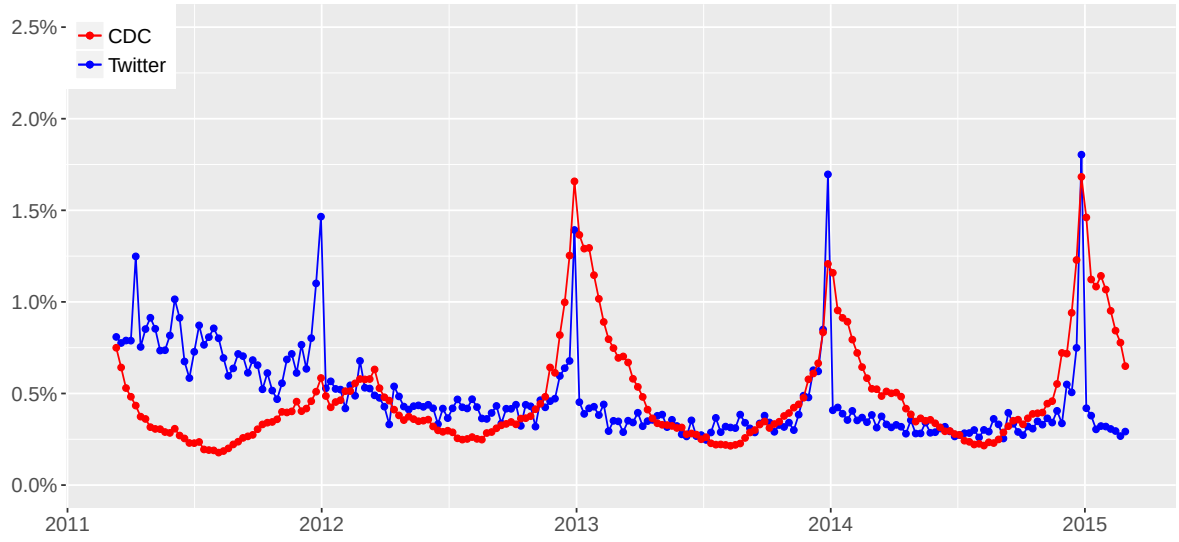
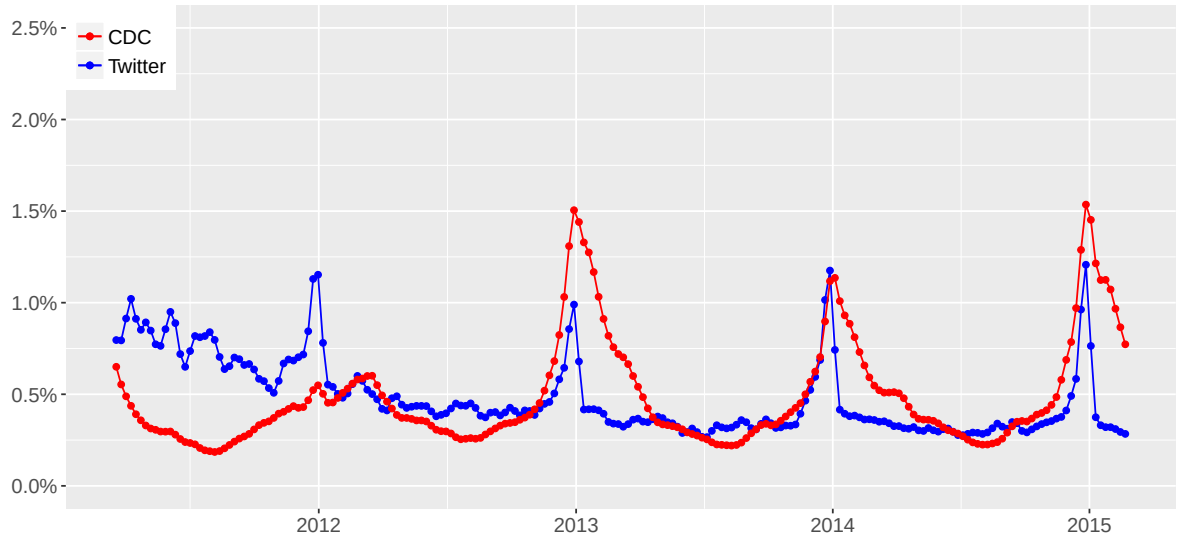
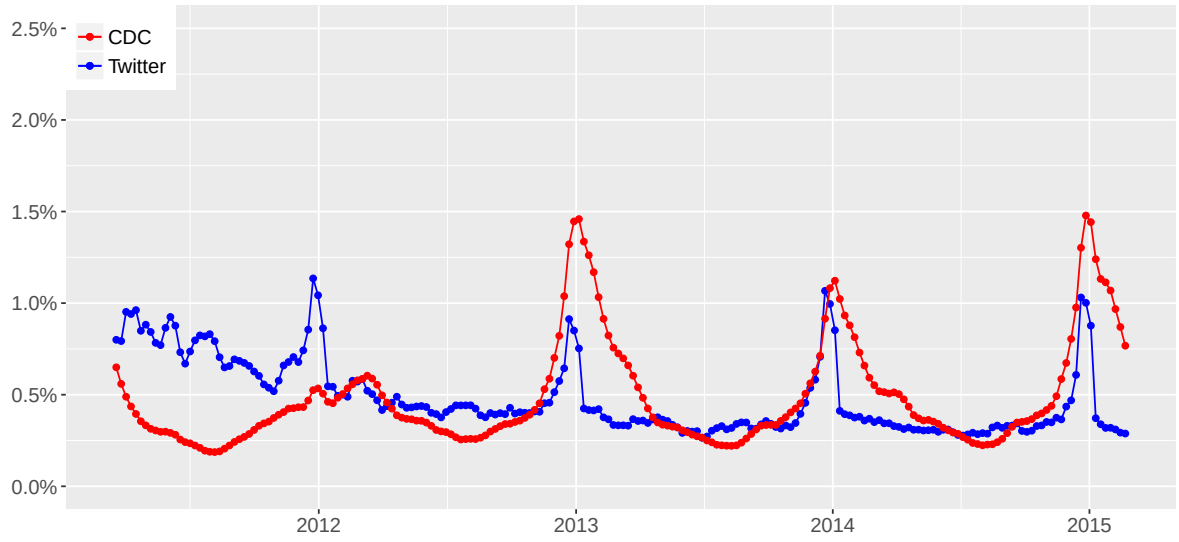


Figure 4.13

Figure 4.14: Comparison between weekly CDC ILI rates (red) and the relative amount of users who sent at least one tweet classified as "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)



(b)

Figure 4.15: Comparison between weekly CDC ILI rates (red) and the relative amount of users who sent at least one tweet classified as "sick" from the Twitter flu classifier (blue) during a specific week (a) and after applying a four-week moving average smoother (b), respectively. 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.

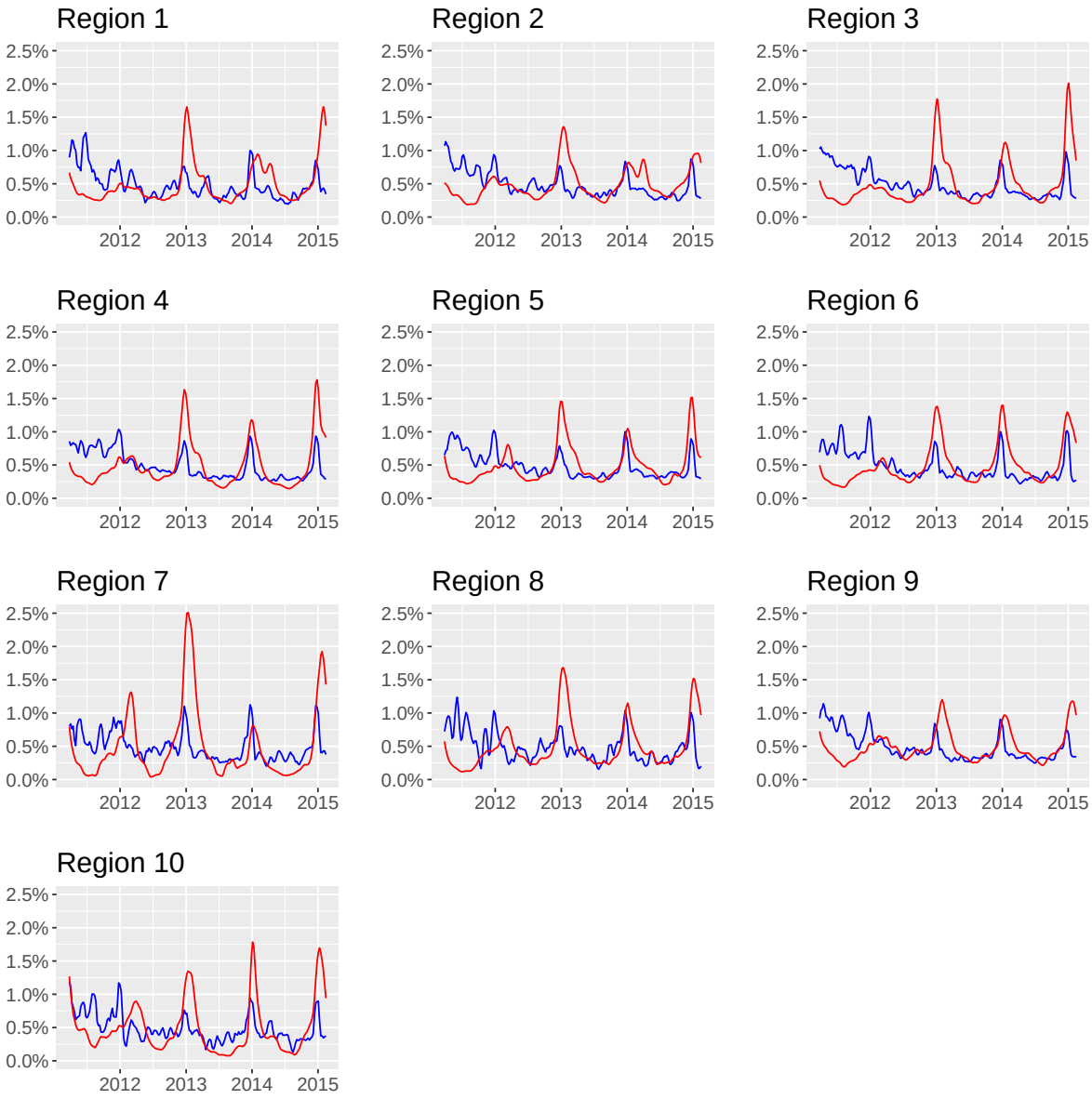


Figure 4.16: Relative number of users who sent at least one tweet labelled as "sick" within each CDC ILI surveillance region per week (blue) compared with weekly ILI percentages in those regions (red). Data has been normalised and processed with a four-week moving average smoother. Note that Region 2 contains "Puerto Rico" and "Virgin Islands", Region 9 contains "Hawaii" and Region 10 contains "Alaska", all of which are missing from the Twitter data set.

The regions are selected according to the ten regional offices of the U.S. Department of Health and Human Services (United States Department of Health and Human Services, 2014), for which the CDC provides separate ILI rate estimates:

Region 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode island, Vermont;

Region 2: New Jersey, New York, Puerto Rico, the U.S. Virgin Islands;

Region 3: Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, West Virginia;

Region 4: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee;

Region 5: Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin;

Region 6: Arkansas, Louisiana, New Mexico, Oklahoma, Texas;

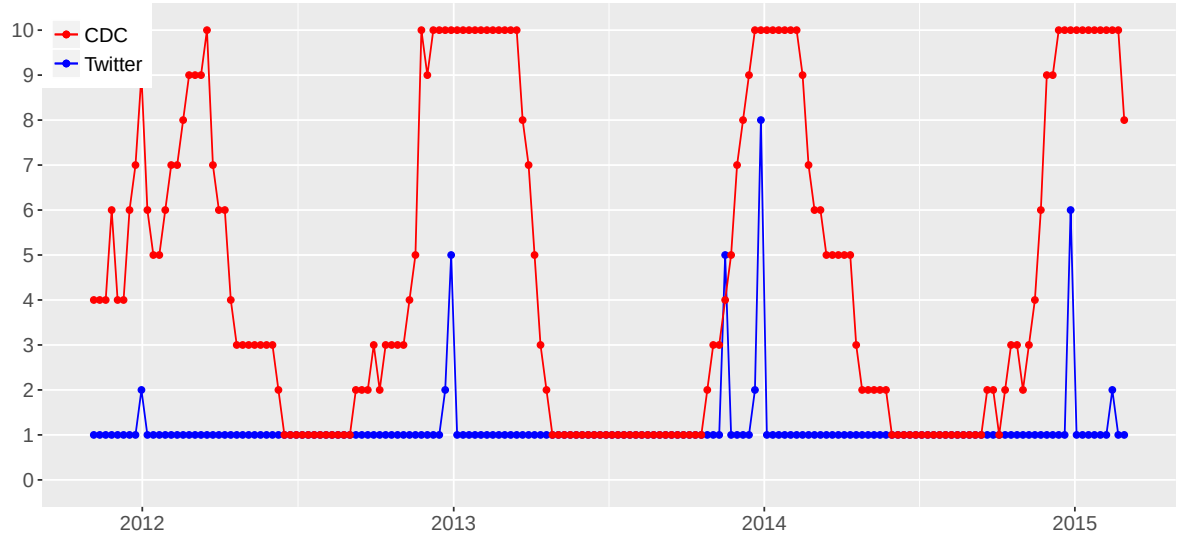
Region 7: Iowa, Kansas, Missouri, and Nebraska;

Region 8: Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming;

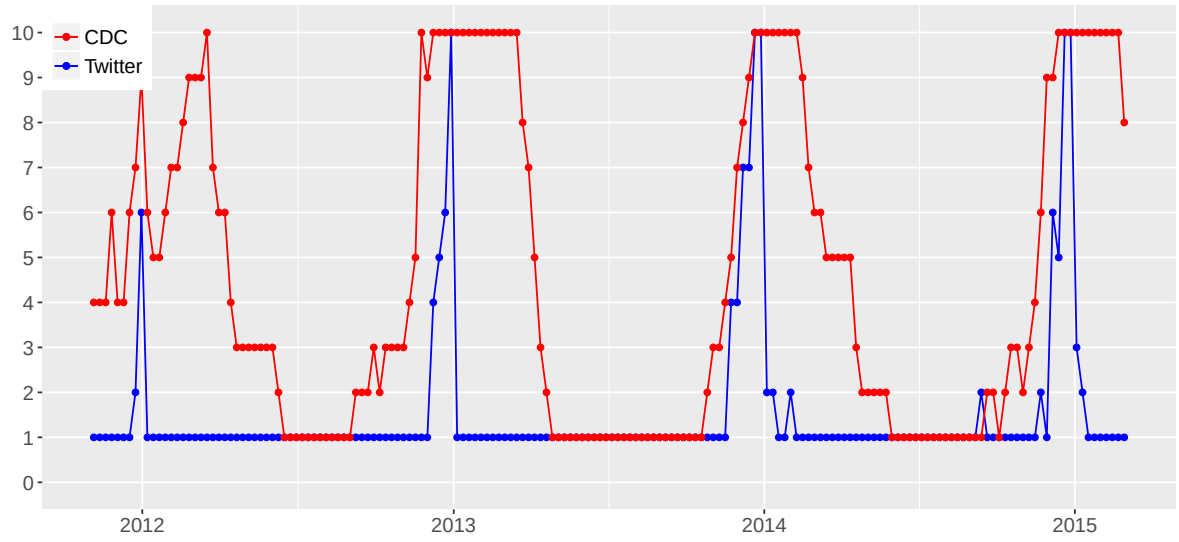
Region 9: Arizona, California, Hawaii, and Nevada;

Region 10: Alaska, Idaho, Oregon, and Washington.

I then did the same using the relative number of sick users (as opposed to sick tweets) instead (Figure 4.17b and Figures 4.19). Note that the Twitter data available to me only spanned the time period between 2011 and 2015. In order not too lose too many years worth of data to the baseline calculation, 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 4.17: Comparison between the weekly ILI activity levels reported by the CDC and the activity levels calculated based on the Twitter data set. (a) Twitter activity levels based on the relative number of sick tweets (b) Twitter activity levels calculated based on the relative number of sick users.

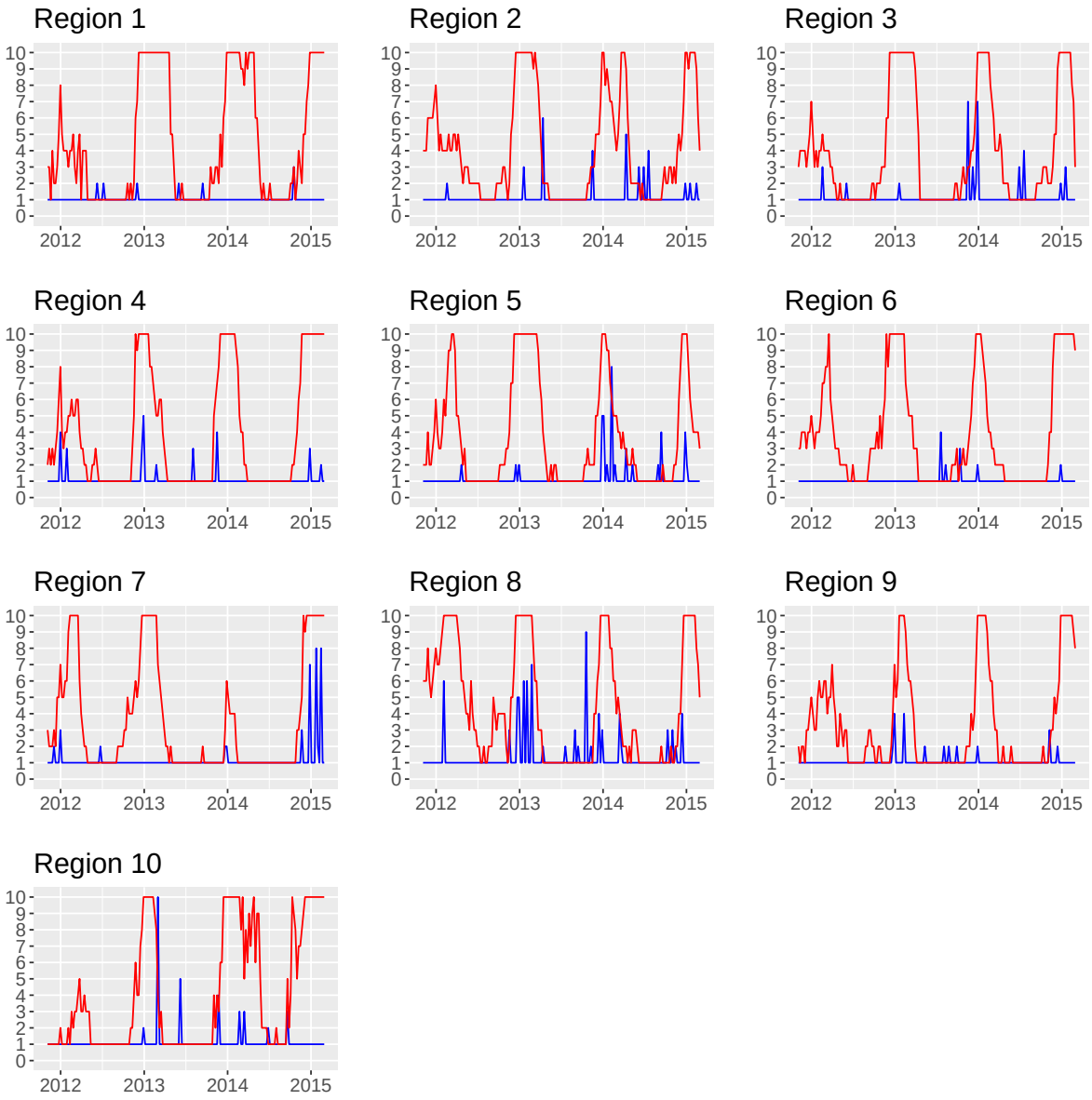


Figure 4.18: Comparison between the regional weekly ILI activity levels reported by the CDC and the activity levels calculated based on the relative number of sick tweets per week.

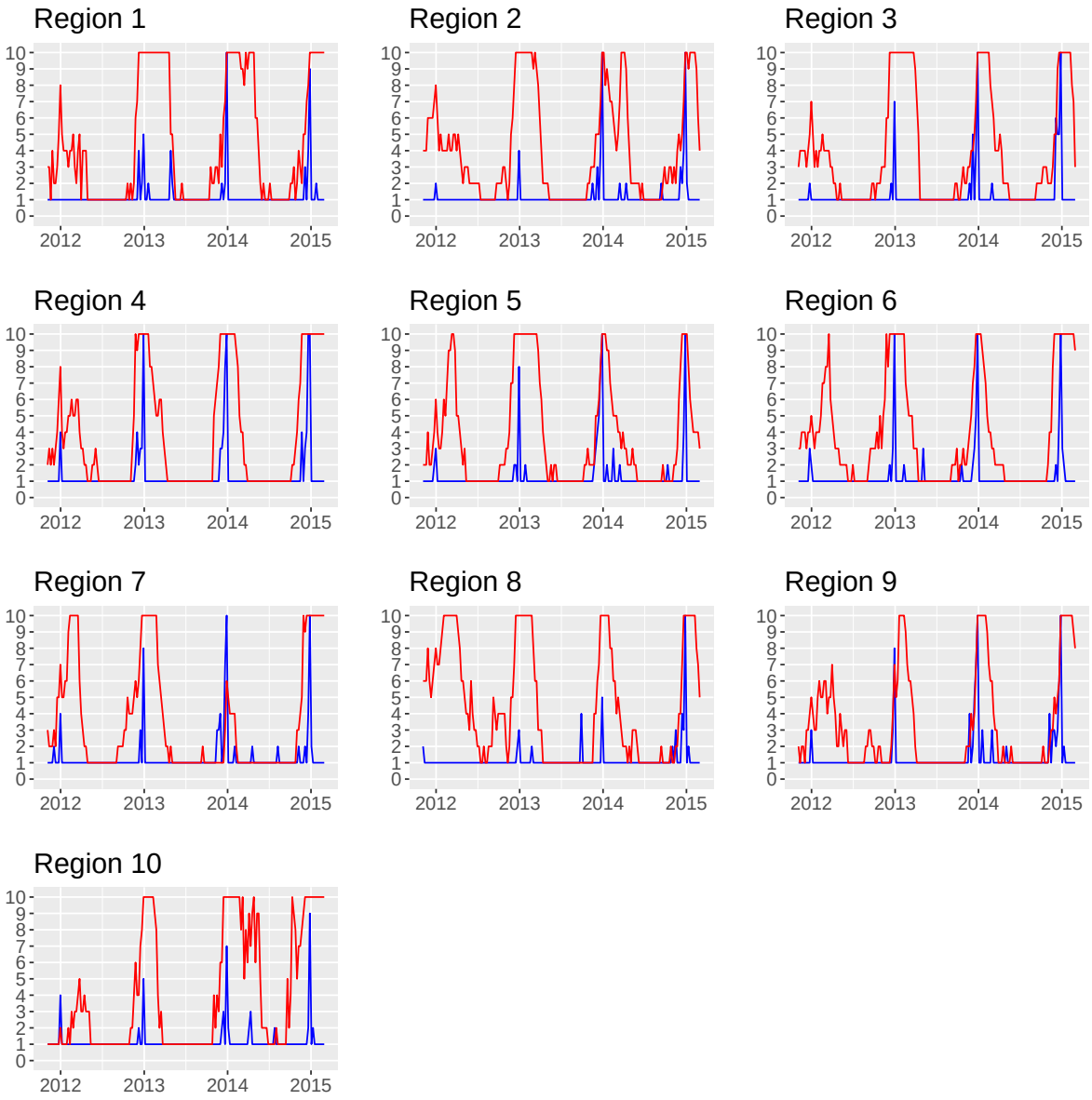
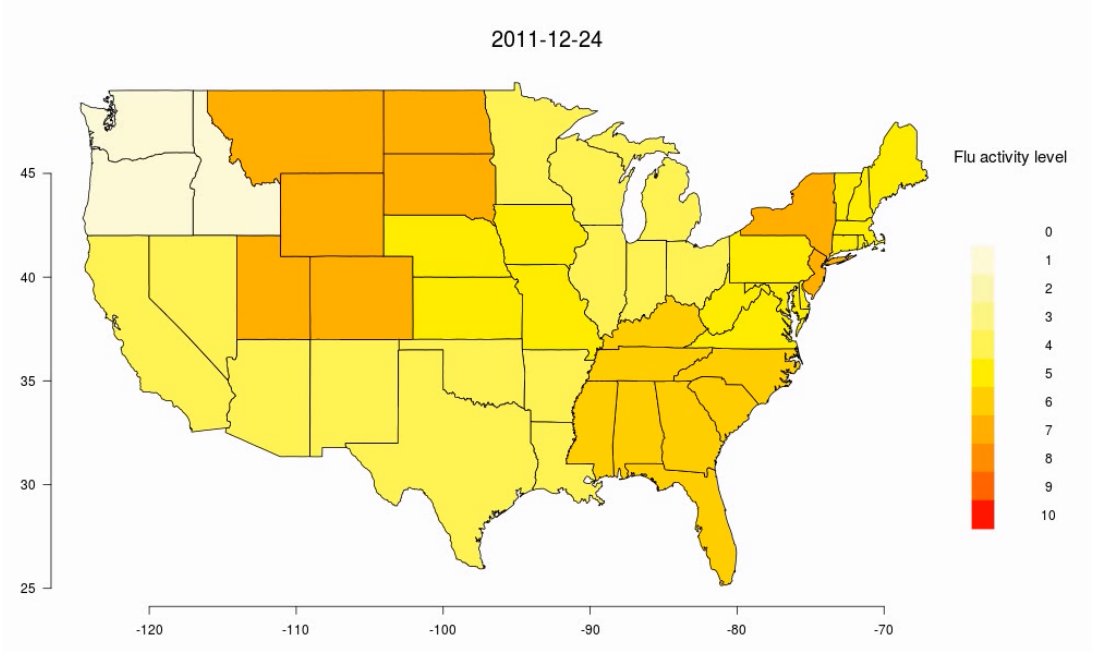


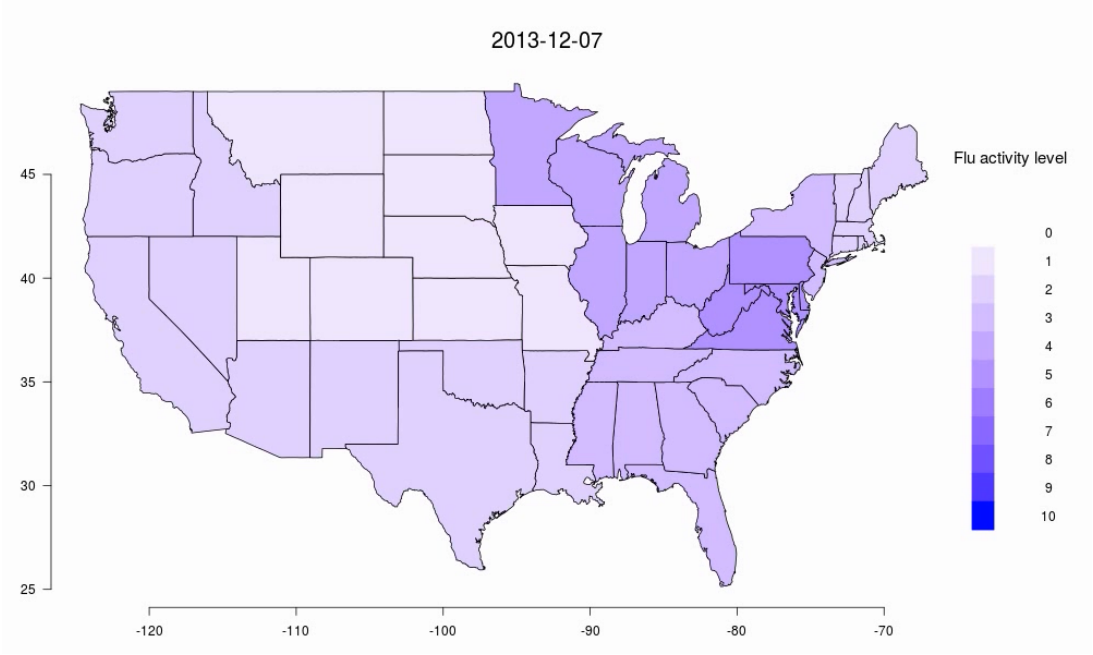
Figure 4.19: Comparison between the regional weekly ILI activity levels reported by the CDC and the activity levels calculated based on the relative number of sick users per week.

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 function was built using the "R"-packages "maps" (Becker *et al.*, 2016), "mgcv" (Wood, 2006, 2016), "animation" (Xie *et al.*, 2015), and "grid" (Murrell, 2003, 2007; Zhou and Braun, 2010).

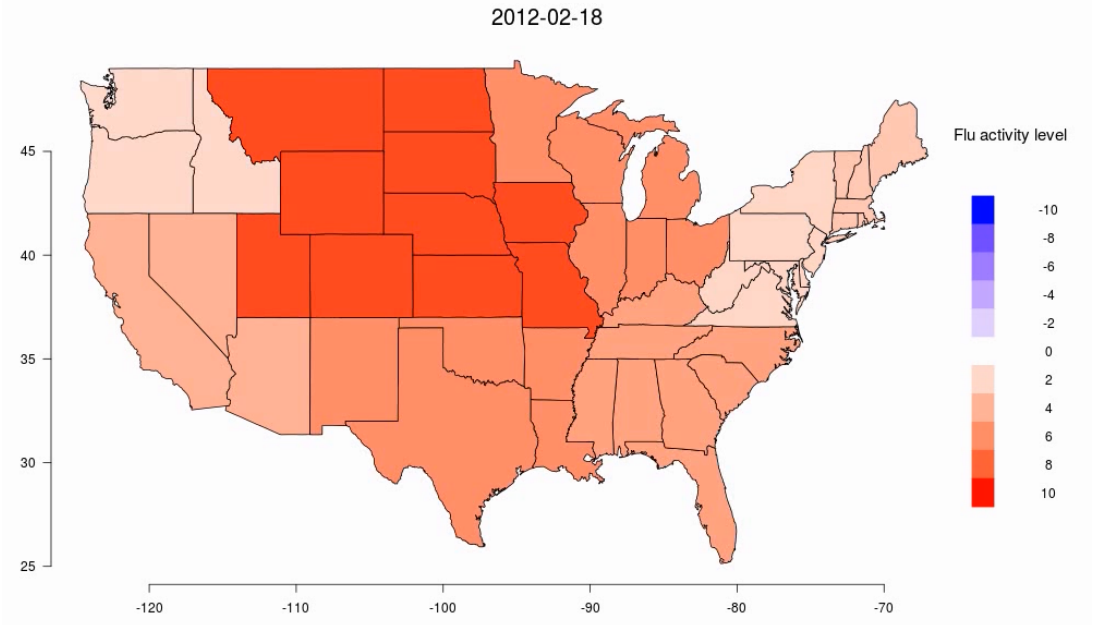
Video 4.1 contains a visualisation of the regional ILI activity levels provided by the CDC. It serves as proof of principle for the validity of the time-lapsed flu map, allowing for a direct visualisation of the yearly flu epidemics in the USA. Video 4.2 also shows regional activity levels, but with the results from the Twitter classifier.



Video 4.1: ILI activity levels as provided by the CDC for the 10 surveillance regions. Red corresponds to activity level 10, while light yellow corresponds to activity level 1, which indicate the highest and lowest ILI activity, respectively. The PDF-version of this thesis contains a playable video.



Video 4.2: Flu activity levels as calculated based on the relative number of Twitter users classified as "sick" in each week aggregated within each of the 10 surveillance regions. Dark blue corresponds to activity level 10, while light blue corresponds to activity level 1, *i.e.* the highest and l deo.

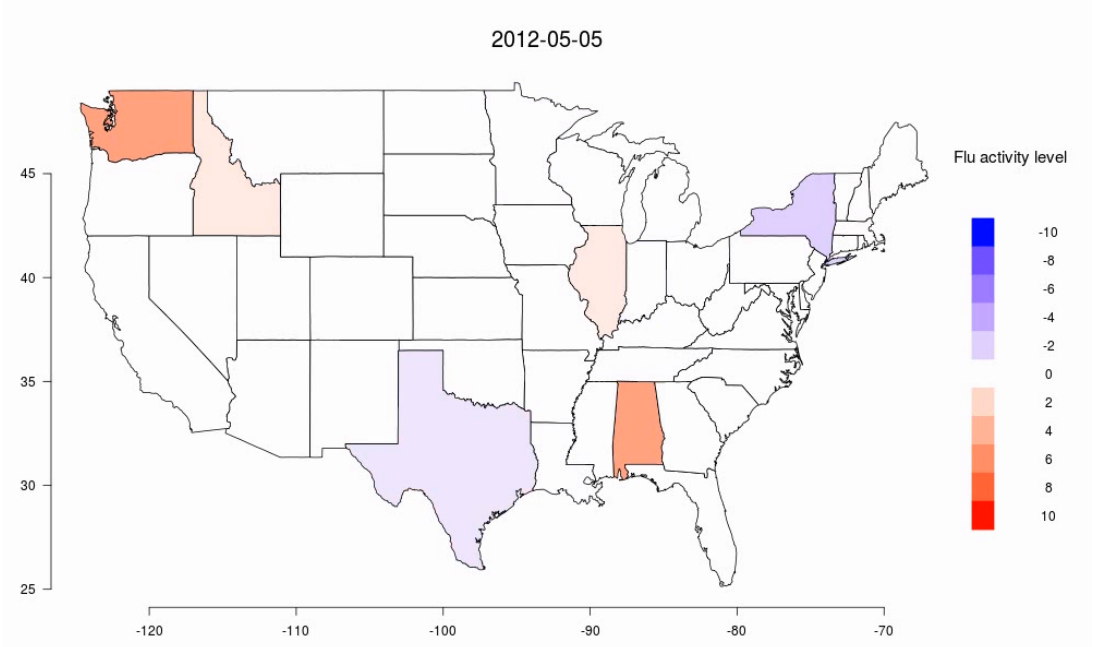


Video 4.3: Difference between the CDC ILI activity levels and those calculated based on the relative amount of sick Twitter users in each corresponding region and week. White means that CDC and Twitter activity levels are exactly the same, red means that the CDC reported higher ILI activity levels in a given state and week than those calculated from the Twitter data set, blue indicates the opposite. The PDF-version of this thesis contains a playable video.

In order to directly compare the CDC and Twitter flu activity levels, I subtracted the Twitter activity levels in each week and region from each corresponding CDC activity level. I then remapped the values onto the USA over the whole 208 week interval for which I had data. Video 4.3 shows performance of the Twitter classifier results over space and time. 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.

4.1.4 Comparing classifier results and ILI rates on the state level

In a next step, I assessed the performance of the Twitter flu classifier by looking at state-level data. I again calculated activity levels for each state and week based on the relative amount of tweets labelled as "sick" and the relative number of users classified as "sick", respectively. 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 thesis, but only a video showing the spatio-temporal differences in ILI activity levels (all additional videos and figures can be found on Github, however). Video 4.4 contains the comparison with activity levels based on the relative number of users classified as "sick" during a specific week and within a given state.

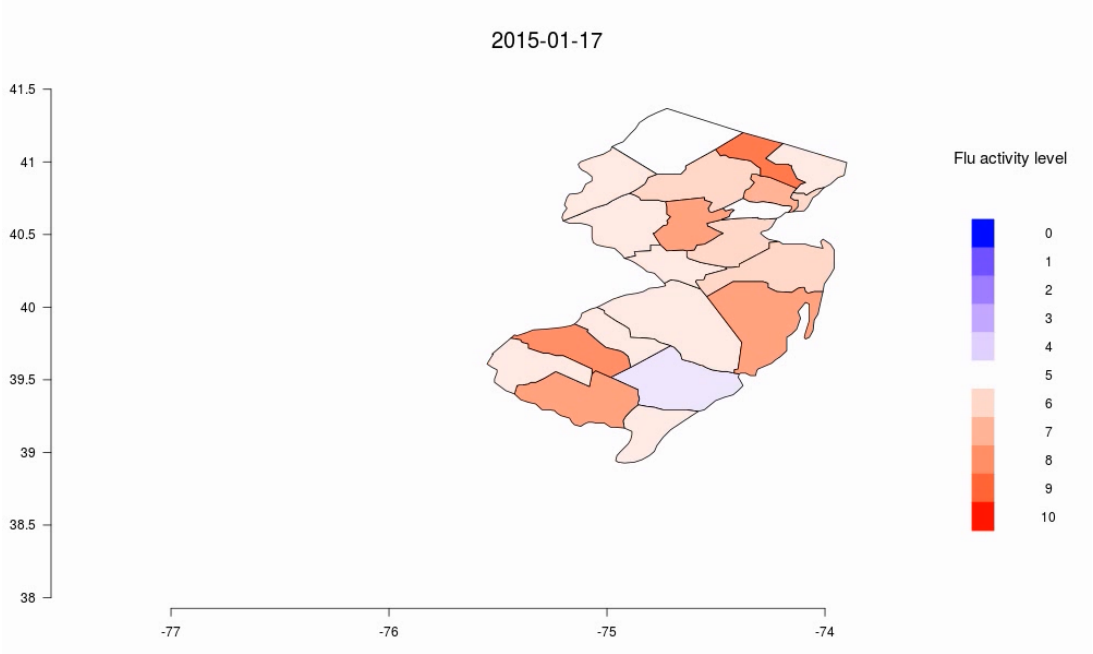


Video 4.4: Difference between the CDC ILI activity levels and those calculated based on the relative amount of sick Twitter users in each corresponding state and week. White means that CDC and Twitter activity levels are exactly the same, red means that the CDC reported higher ILI activity levels in a given state and week than those calculated from the Twitter data set, blue indicates the opposite. The PDF-version of this thesis contains a playable video.

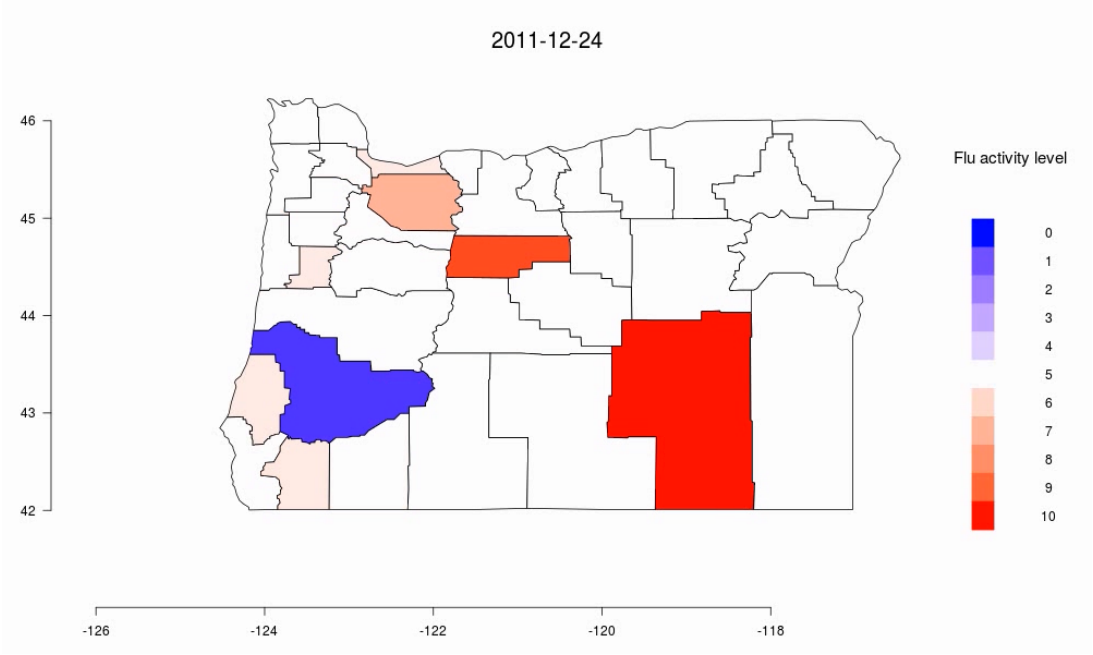
4.1.5 Comparing classifier results and ILI rates on the county level

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. 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 (Video 4.5) and Oregon (Video 4.6). Note that Oregon did not provide ILI estimates for all counties (most of the northeastern counties are missing, for example).



Video 4.5: Difference between the official ILI activity levels provided by the state of New Jersey and those calculated based on the relative amount of sick Twitter users in each corresponding week and county of New Jersey. White means that the official and Twitter activity levels are exactly the same, red means that the state of New Jersey reported higher ILI activity levels in a given county and week than those calculated from the Twitter data set, blue indicates the opposite. The PDF-version of this thesis contains a playable video.



Video 4.6: Difference between the official ILI activity levels provided by the state of New Jersey and those calculated based on the relative amount of sick Twitter users in each corresponding week and county of Oregon. White means that the official and Twitter activity levels are exactly the same, red means that the state of Oregon reported higher ILI activity levels in a given county and week than those calculated from the Twitter data set, blue indicates the opposite. The PDF-version of this thesis contains a playable video.

4.2 Attempts to reproduce key figures and findings from 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 very 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 reproduce some key results reported in Bodnar (2015).

4.2.1 Reclassification of the raw Twitter data

As described in Chapter 3, the data set I was working with a data set that contained the output of the Twitter classifier described in Chapter 1. Since there might very well 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. 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) have in the meantime been replaced by newer versions 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 errors 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. All files belonging to the Twitter classifier, including the compiled version of it, can be found on Github in the folder "Twitter_Parser".

4.2.2 Reproduction of 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 validity 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 four 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 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 model specifications in order to create the data contained in said file from the Twitter data set. 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. A follow-up e-mail regarding this matter is pending response. Data and code files can be found on Github in the folder "PhDThesisBodnar".

In order to recreate 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 3rd 2011 (week 40) and ending on September 28th 2014 (week 39). The 11 columns contained in the data set have the following meaning (Bodnar, personal communication):

`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

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. According to Bodnar `full_base` column corresponds to the weekly aggregated results of the Twitter classifier (personal communication).

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):

"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 that in the R-code I received the optimisation only happened for γ and β, not for $S(0)$, as described. 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.

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 4.20, 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 I built my second SIR model based on these values. As can be seen from Figure 4.21b, the resulting yearly and combined ILI curves are virtually indistinguishable from the ones shown in Bodnar (2015).

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 the 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. At the same time, however, the curves that result when creating an SIR model based on the values in Table 4.1 and the `full_both2` data provided to me are virtually indistinguishable from the curves depicted in Bodnar (2015), *i.e.* in Figure 2.2. This implies that the β and γ values reported in Bodnar (2015) are not the ones that were used to create Figure 2.2.

In order to corroborate this assumption, I recreated SIR model curves based on the values reported in Table 2.1 and compared them to the SIR model curves resulting from the values in Table 4.1, which are depicted in cyan and blue, respectively, in Figure 4.22. The comparison reveals a marked difference between the two model curves and makes clear that the model curves based on the recreated values in Table 4.1 are much more similar to the original curves depicted in in Figure 2.2 than the model curves based on the values from Bodnar (2015) depicted in Table 2.1.

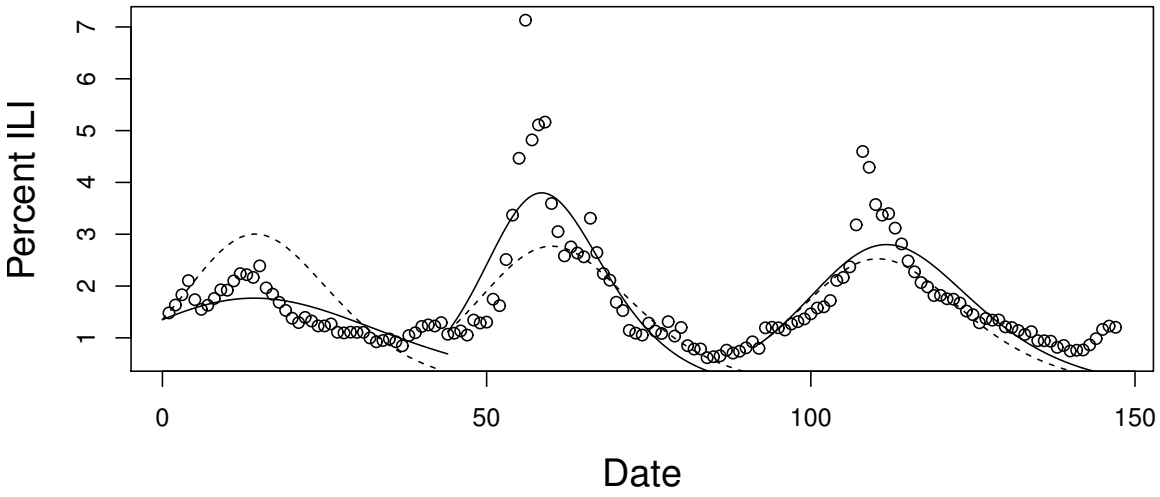
Table 4.1: Combined and yearly best-fit parameters for the SIR model I recalculated based on the CDC's data (white) and the Twitter base model data (grey) provided to me by Bodnar (personal communication).

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

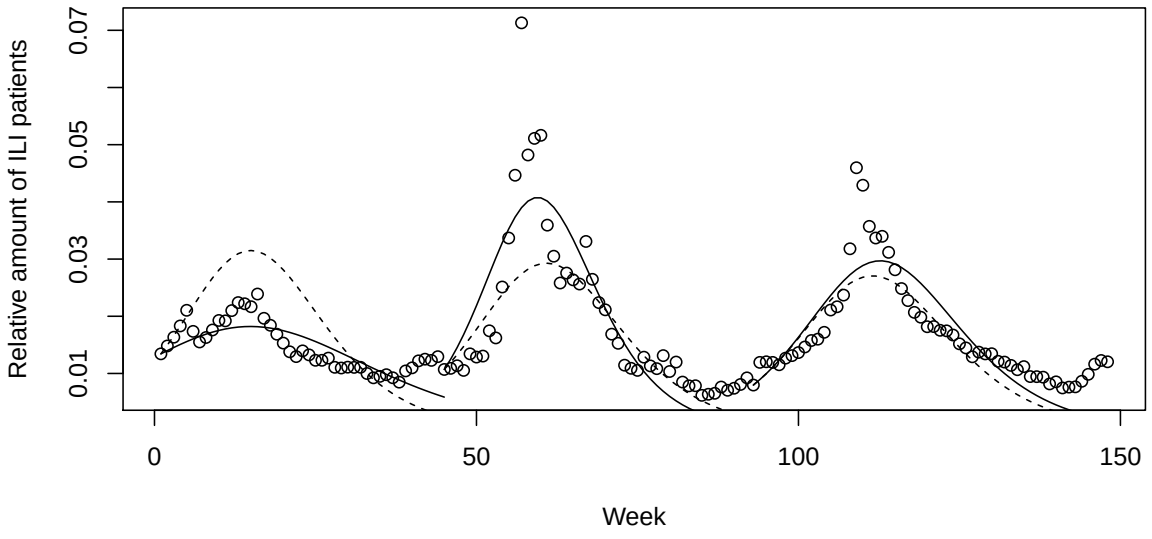
4.2.3 Attempt to reproduce the AR model described in Bodnar (2015)

It is important to repeat that the data used to calculate the SIR models in Bodnar (2015) are *not* equivalent the `full_base` data, which are supposed to represent the raw output from the Twitter classifier according to Bodnar (personal communication). 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, as I have shown in the previous section, this cannot be the case and it is neither the case for the results shown in Figure 2.3.

However, this is clearly not the case. Figure 4.23a 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.* the model corresponding to the `full_both2` data and including information from the Twitter base model as well as from the AR(2) model based on the CDC data. If we only look at the data from the Twitter base model (Figure 4.24), we can see that the fit with the official CDC data is worse and that the corresponding SIR model is also decidedly different from the SIR model based on the `full_both2` values. Also, Figure 4.23 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 includes additional information from the Twitter base model.

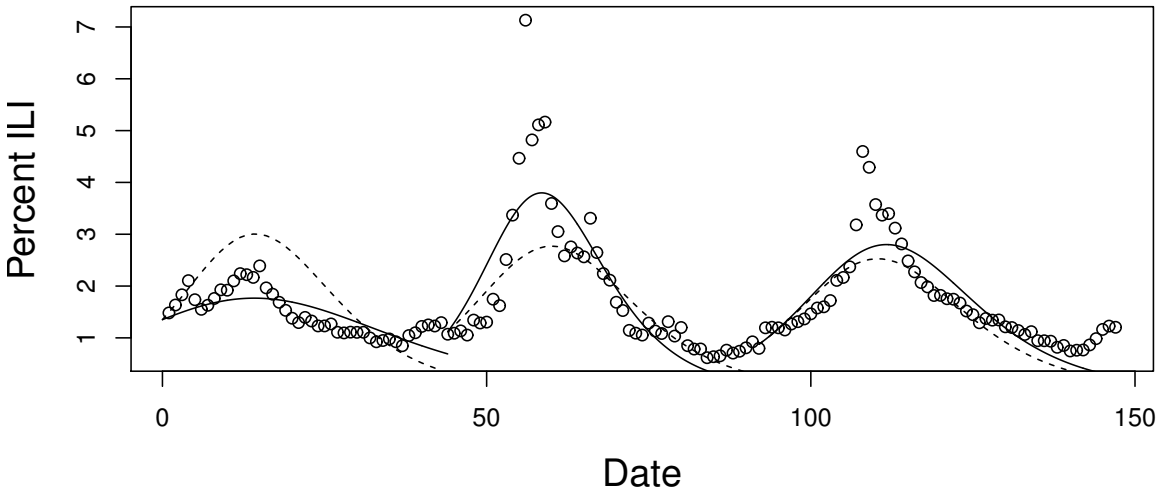


(a)

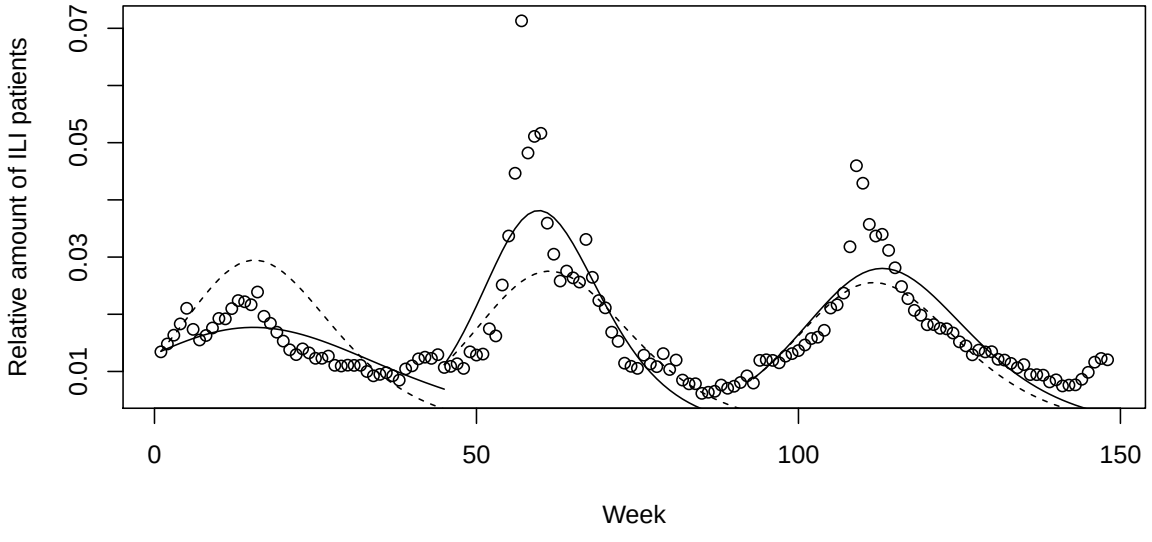


(b)

Figure 4.20: Comparison between the SIR model depicted in Bodnar (2015) (a) and a replication based on the official CDC ILI rates (b).

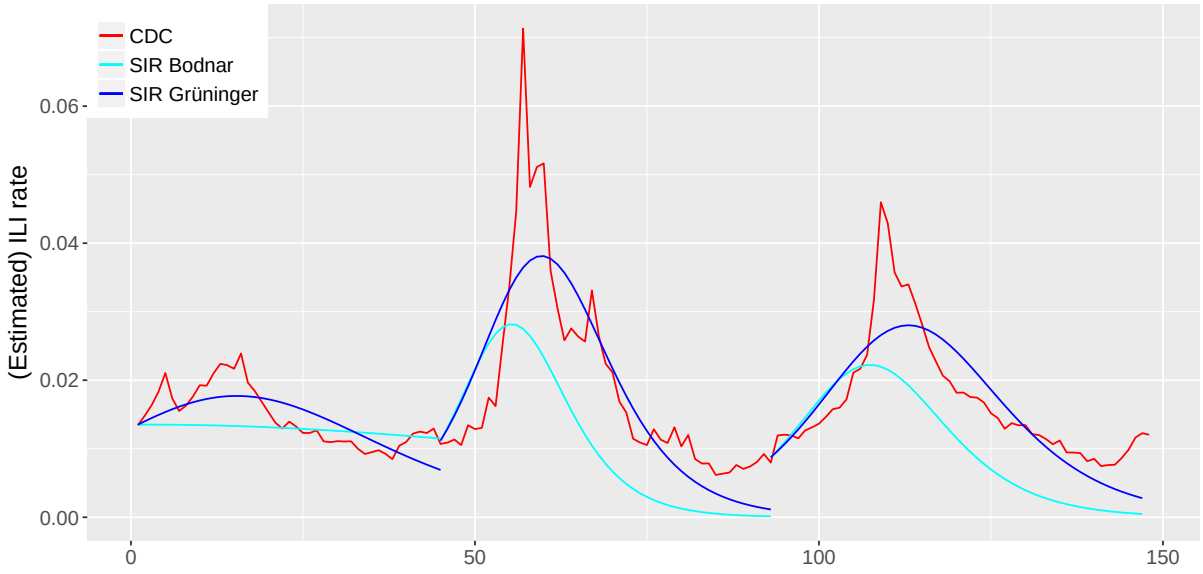


(a)

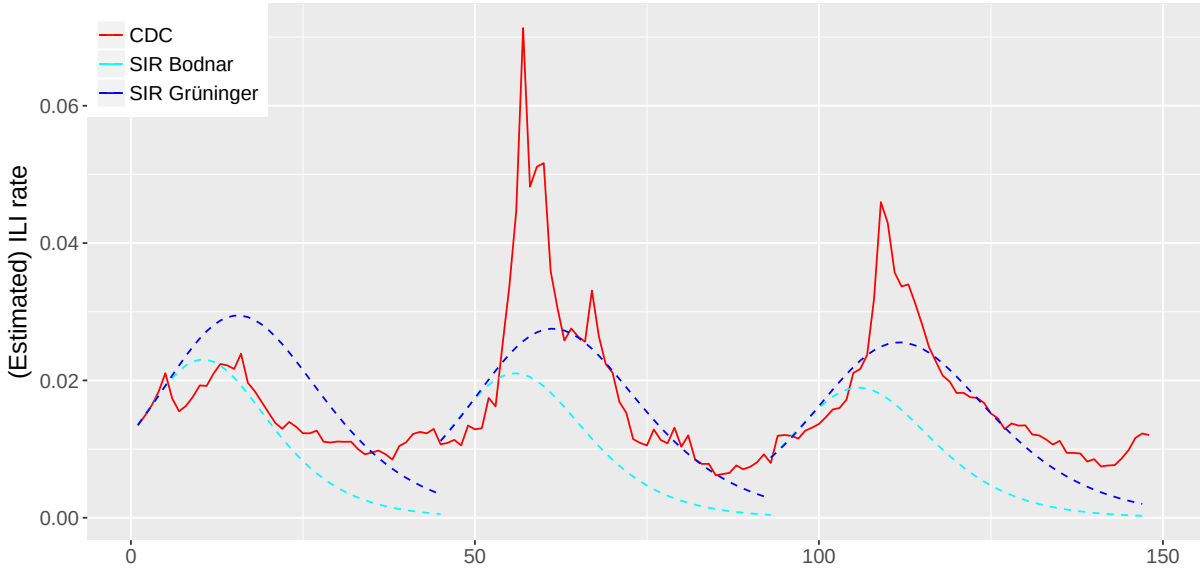


(b)

Figure 4.21: Comparison between the SIR model depicted in Bodnar (2015) (a) and a replication based on the full retrospective Twitter model (c).

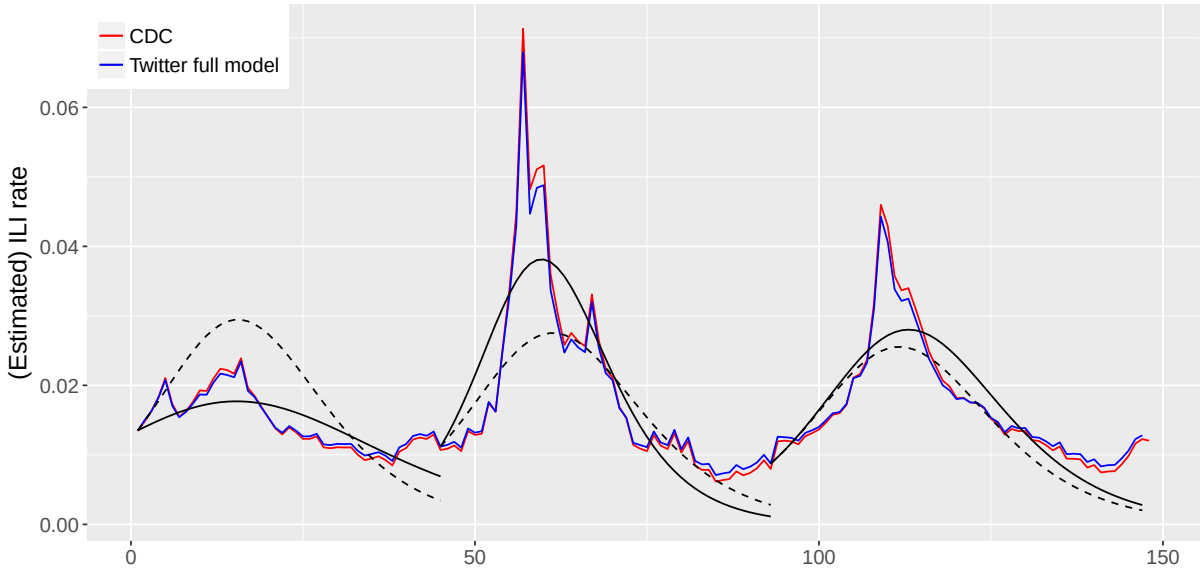


(a)

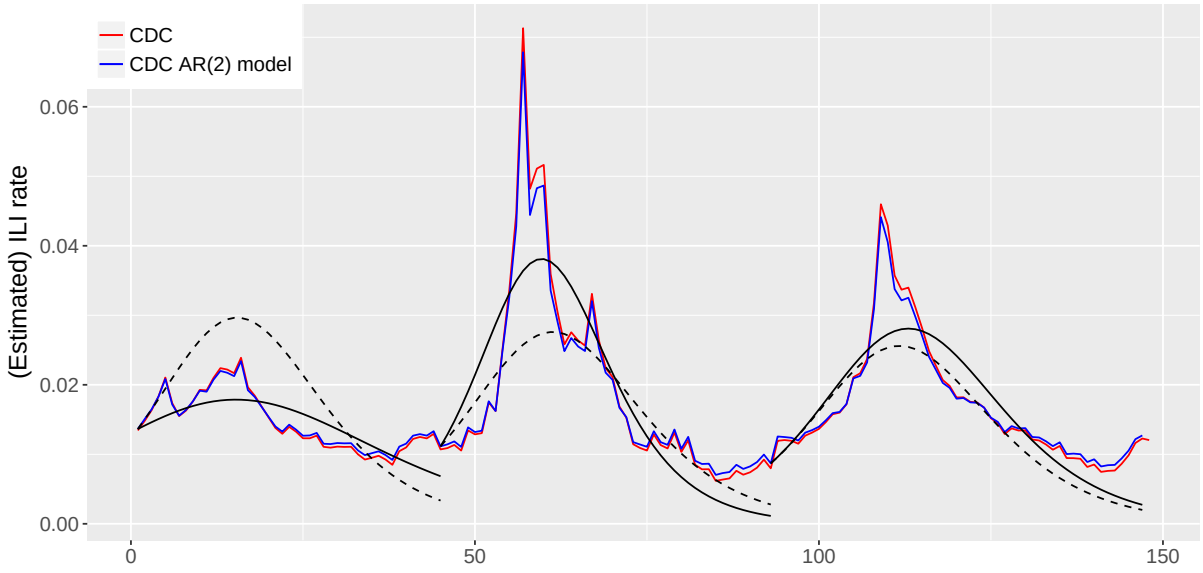


(b)

Figure 4.22: SIR model curves using yearly (a) and combined (b) parameters to model the ILI prevalence throughout one year. Red indicates the official ILI rates from the CDC, while cyan and blue show the SIR model curves based on the parameters reported in Bodnar (2015) and the ones recalculated by me, respectively, based on the very same data (*i.e.* the `full_both2` values).

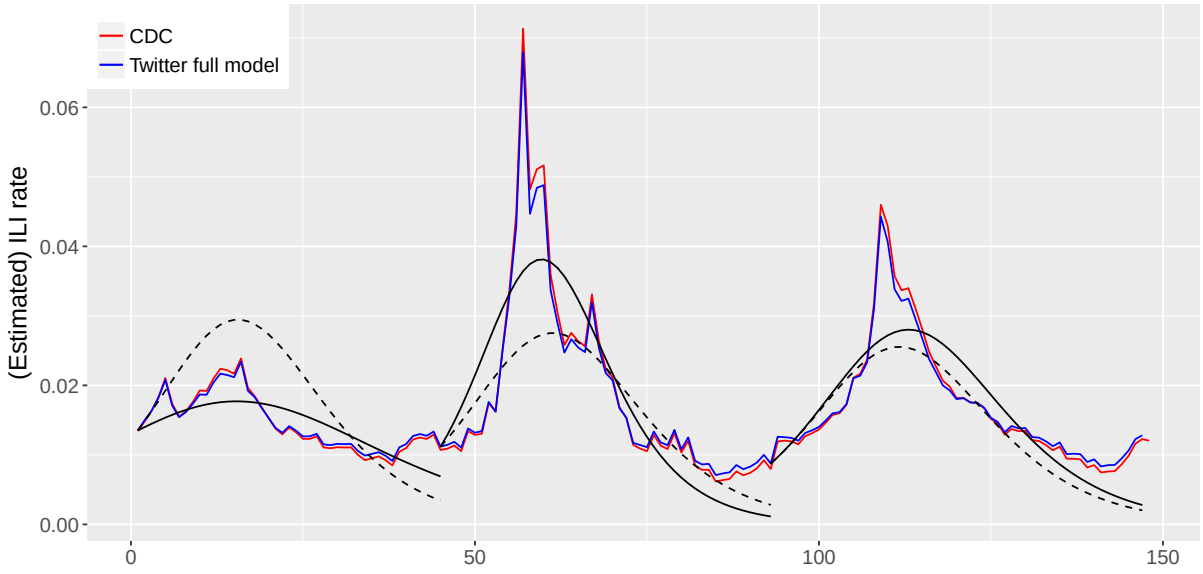


(a)

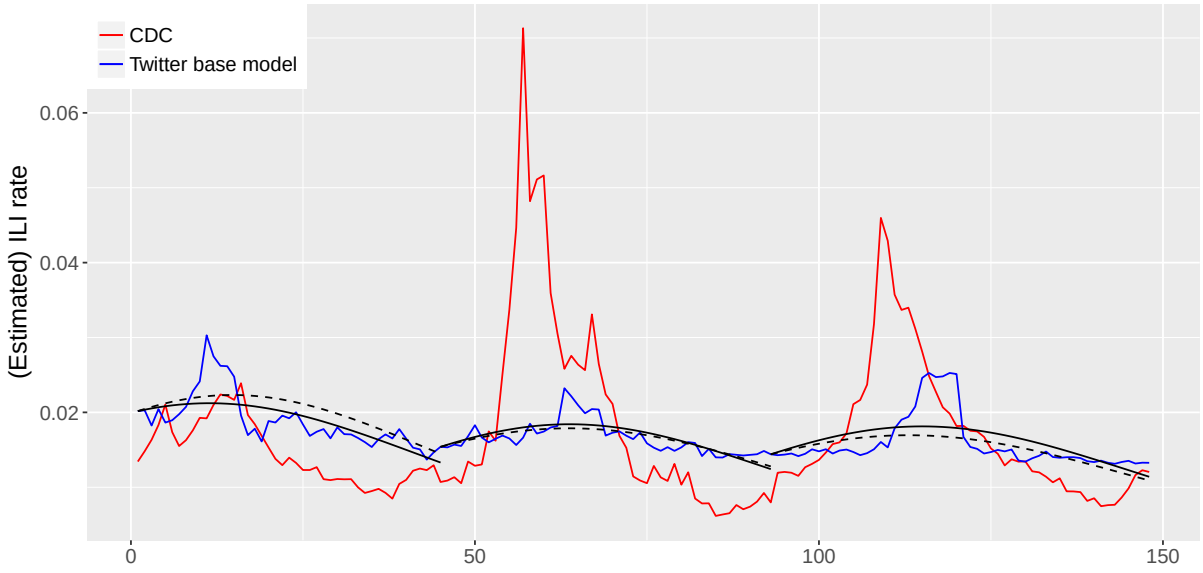


(b)

Figure 4.23: Comparison between the full retrospective Twitter model depicted in Bodnar (2015) (a) and a replication consisting of a simple retrospective AR(2) model based only on CDC ILI rates (b). The black lines indicate the results of the corresponding SIR models using yearly (solid) and combined (yearly) parameters.



(a)



(b)

Figure 4.24: Comparison between the full retrospective Twitter model depicted in Bodnar (2015) (a) and a replication consisting of the retrospective Twitter base model (b). The black lines indicate the results of the corresponding best-fit SIR models using yearly (solid) and combined (yearly) parameters.

Chapter 5

Discussion and outlook

In this last chapter, I will first discuss the findings summarised in Chapters 3 and 4 with a special emphasis on potential future applications of the Twitter classifier or related methods for flu surveillance.

In a second part, I will then put emphasis on the discussion of the different issues of reproducibility that I have outlined in Chapter 2 and that have arisen during the course of this work. I will also make suggestions about how to overcome such issues in the future.

5.1 Has "Larry the Bird" deserved our trust?

The results of my comparison of the results of the Twitter classifier with the official ILI data from the CDC are mixed. The following sections will discuss the reliability of the classifier on different levels and make suggestions on how to make improvements in the future.

5.1.1 The Twitter classifier can detect seasonal flu peaks on a national level

The classifier has shown that it can faithfully predict the Influenza peaks recurring on a national level every year between the months of December and February (see Figures 4.9, 4.11, and 4.15).

However, one can also see from these figures that the peaks are very short and sharp. This might be explained by the way the Twitter classifier labels a certain tweet—and thereby its author—as "sick". The classifier does so by aggregating the daily signals within a 4-week-window in order to classify the tweets of the first day either as "sick" or "healthy" (see Section 2.4 for details).

But this way of assigning disease status has one problematic implication, namely that information from future tweets is incorporated while information from past tweets is ignored. This is insofar counter-intuitive as a Twitter user who tweeted about being sick on day 1, will most certainly still be sick on day 2—regardless of what he tweets about. However, the algorithm will

not take this into account, 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 healthy again.

In addition, the algorithm is bound to classify tweets too early as being sick. For example, if particularly strong "flu" signal turns appears, 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.

Of course, one can argue that a Twitter user does not necessarily tweet about her flu symptoms on the first day of the onset of the symptoms so that the exact start of the disease is difficult to assess in any case. This is true, but as long as we do not have any additional information about the probability distribution of when the user tweets about her sickness, it would be prudent to assume a uniform distribution and treat each day of the disease as having equal probability of being a day during which the user tweets about her disease.

In any case, the fact that the classifier assigns disease status only on the first day of the 4-week-sliding window and only by taking into account prospective information, might artificially shorten the duration in which a user is classified as having the flu. On a population level, this might lead to the sharp and short peaks we can observe in the figures referenced above. Future revisions of the Twitter classifier should therefore also take into consideration retrospective information from a Twitter user's timeline and not classify the tweets of the first day of the 4-week-sliding window, but the tweets of the 14th or 15th day, *i.e.* a day in the middle of said window.

5.1.2 The Twitter classifier might be able to detect ILI symptoms in summer

Another peculiar finding from the comparison of CDC and Twitter classifier data is the overestimation of ILI symptoms in summer by the classifier. This can clearly be seen in Figures 4.9, 4.11, and 4.15 as well as in Figure **??** and is most prominent in the summer of 2011.

Since the classifier was built using confirmed medical records that could be correlated on an individual basis with the output of specific Twitter users, it is very unlikely that the peaks in summer are due to overfitting, especially because the set of users on which the classifier was validated was very small.

However, as outlined in Section 2.4, the small sample size that served to build the classifier in combination with its low temporal resolution could have introduced other biases which might

jeopardise the generalisability of the classifier up to the point at which it is useless outside the very specific environment in which it was built (the Pennsylvania State University and part of its student body in the year 2011–2012). However, if that were the case, we would expect the classifier to fail to detect the flu peaks in winter, for example. Since this is not the case, however, we can safely assume that rules which the classifier uses to detect flu-related tweets can be applied on national level data as well.

Hence, the most likely explanation for the peaks in summer might also be the most promising: If we can assume that the classifier is able to pick up ILI symptoms with some reliability, then we should expect the classifier to detect these signals regardless of whether the Twitter user has the flu or another disease that symptomatically resembles the flu. This is true for common colds, which lead to symptoms such as sore throat, runny nose, coughing, sneezing, headaches, and body aches, and whose occurrence coincide with that of the flu, but which are most often caused by rhinoviruses (Heikkinen and Järvinen, 2003; Centers for Disease Control and Prevention, 2017a).

However, there are also many cold viruses that predominantly spread during the summer, the non-polio enteroviruses being the most notorious representative (Pons-Salort *et al.*, 2015). An infected person can exhibit symptoms akin to those of the flu, such as fever, runny nose, sneezing, coughing, body and muscle aches (Pons-Salort *et al.*, 2015; Centers for Disease Control and Prevention, 2016a).

Hence, the off-season peaks of the Twitter classifier in 2011, 2012, 2013, and 2014 could be caused by other infectious diseases that exhibit similar symptoms to that of the flu. Further assessments of the validity of the flu classifier should take this into account, for example by comparing its performance with the surveillance data of other infectious disease capable of causing ILI symptoms.

5.1.3 Autoregression trumps Twitter data

As Paul *et al.* (2014) noted, autoregressive models are very strong baseline models to forecast ILI prevalence—better than only using Twitter for this task. This can also be observed in Figures 4.24 and 4.23, which show that a simple retrospective AR(2) correlates much better with the official CDC ILI rates than the Twitter base model. As already explained in Section 2.4, the Twitter base model is not equivalent with the raw results from the Twitter classifier, but most likely consists of these results combined with additional sources information.

However, Twitter data can still be of use, especially when it comes to now-cast ILI rates, *i.e.* to assess the current prevalence of influenza-like symptoms based on the preliminary CDC data that has been published (Paul *et al.*, 2015). This can also be seen in Figures 4.23 and 4.24,

which show that the the CDC AR(2) combined with the Twitter baseline model provide slightly better results than the AR(2) model alone, respectively.

It should be noted that most disease models, including the ones described in this thesis, are using *revised* CDC data which are already corrected for mistakes by the CDC (Paul *et al.*, 2014). However, Twitter data might be much more useful when used with the unrevised data instead. As Aramaki *et al.* (2011) report, Twitter data is most useful for the early detection of influenza epidemics, increasing the speed of flu surveillance and the reliability of the uncorrected ILI data from the CDC.

Hence, future assessments of the performance of the Twitter classifier should focus on comparisons with uncorrected CDC data in order to find out whether the classifier results can help to increase the accuracy of these results. In addition, the results from the Twitter classifier might be combined with an autoregressive model of the official ILI data, even though this would partially defeat the purpose of building a classifier which is *independent* of population-level ILI data. After all, the underlying reason to built a Twitter flu classifier that can detect flu occurrences independent of correlations with population-level flu data was to circumvent the many problems that might arise with regard to overfitting.

5.1.4 Low signal to noise ratio on regional, state, and county level

Using geotagged tweets offers the intriguing prospect of accurately tracking transmission dynamics on a national, regional, state-level, and county-level scale. In addition, it might allow for more fine-grained spatio-temporal surveillance of the flu due the fact that flu occurrences can be tracked for every day and up to street resolution.

However, as shown in Sections 4.1.5, 4.1.4, and 4.1.3, this approach does not yet seem to yield feasible results. The higher the spatial resolution become (moving from the national level with the lowest to the county level with the highest resolution), the more pronounced the influence of statistical noise becomes. One likely explanation for this is the fact that the number of available tweets is reduced the smaller the geographical region of interest becomes. In return, this increase the impact of individual tweets on the ratio of tweets labelled as "sick" to tweets labelled as "healthy". Even though this is to be expected, the extent of it is still surprising.

After all, a state like New Jersey has a comparably large and active Twitter population (see Figure 4.4). Hence, one would assume that the sheer amount of tweets sent from within these states would be big enough to make them somewhat impervious to statistical noise and large outliers. However, the flu rates are calculated on a weekly basis, therefore heavily reducing the number of tweets that are available to calculate the ratio of sick to healthy users in a given state or week. As I have shown with Figures 4.11 and 4.15, these fluctuations can be reduced

by applying a simple moving average over two or four weeks.

However, if one needs to aggregate the data on a temporal (from days over weeks to months) or spatial (from streets over counties and states to regions and the whole continent) scale in order to retrieve meaningful information from the results of the Twitter classifier, one abandons the two most important advantages for using geotagged Twitter data for flu surveillance: Increased speed and spatial acuity.

Using Twitter data for epidemiological purposes has the potential to provide faster and more spatially accurate information about the distribution of diseases such as the flu. But by aggregating the available data too heavily, this advantage is lost. Therefore, future revisions of the Twitter classifier should focus on making it more reliable with smaller data sets or—if this is not possible—on classifying a higher number of tweets in order to increase the content of the data source. For example, one could try to classify *all* tweets from a small, but densely populate state like New Jersey that also offers access to official CDC data in order to find out whether an increase in the Twitter data source can help to offset the increase in statistical noise that happens when aiming for high spatio-temporal resolution.

5.2 What the failure to reproduce can teach us

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

Faulty data handling: Errors in the aggregation and processing of the Twitter data led to different results.

Differences in the data set: The findings from Bodnar (2015) are based on Twitter data that are different from the data I analysed.

Classification errors: There are errors present in the Twitter classifier code that led to wrongly labelled tweets.

Additional modelling needed: The output from the Twitter classifier alone is not sufficient to make ILI predictions.

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

5.2.1 Faulty data handling

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 calculations 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.

Nevertheless, for which I am fairly confident that the results reported in this thesis do not contain any fundamental errors based on faulty code. For almost every step in the description, aggregation and analysis of the data I have usually chosen at least two different approaches (not all of them are reported in this thesis, but 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, in order to make this thesis as reproducible as possible and in order to facilitate any follow-up analysis, I will further clean up the code and the database structure.

5.2.2 Differences in the 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 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 correspondence with Bodnar confirmed that the data set described in his thesis and the one stored in the data base dumps of the Salathé research group supposedly have to be the same. In this case, the only remaining explanation would be that the data set was inadvertently changed at some point after the end of Bodnar's thesis.

Any additional explanation might be that I was working with a processed version of the results of the Twitter classifier. As Bodnar (2015) noted, he removed both Spam-bots with very high tweet rates as well as infrequent tweeters with less then 10 tweets over the whole study period from the data set. However, this explanation is contradicted by the fact that the data

set I used still contained Twitter users who only sent one tweet over the whole study period as well as very spammy accounts with almost 1.5 million tweets sent between the beginning of 2011 and the beginning of 2015.

5.2.3 Classification errors

This idea was in some regards the 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. Unfortunately, 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 classifier, respectively. Hence, debugging and updating the Twitter classifier should be a main goal for future replication attempts.

5.2.4 Additional modelling needed

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 existing estimates and reduce the error. As I have shown in Chapter 4, this is 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.

In fact, Bodnar (2015) fitted two kinds of models based on the results from the Twitter flu classifier in order: A linear regression model, including autoregressive information from the CDC ILI data (see Equation 2.2), and an SIR model (see Equations 2.1). However, the SIR model reported in Bodnar (2015) and depicted in Figure 2.2 was not based on the output of the Twitter classifier, but rather on the results from the linear regression model described in Equation 2.2, partially defeating the purpose of having a classifier independent of population level data. In addition, the results I was able to receive from the Twitter classifier results were an order of magnitude smaller than both the official CDC data and the Twitter `base_full` model data provided to my by Bodnar (personal communication). It is therefore unclear how exactly the Twitter classifier data reported in Bodnar (2015) came about.

In any case, the discrepancies between the model fits described in Bodnar (2015) and the results presented in Chapter 4 of this thesis are problematic. 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 the figures and tables in Bodnar (2015), many discrepancies are found. Also, 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.

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-times/`. [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/`. [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**, .

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017). The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ*, **5**, e3544.

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.

Banks, D. (2011). Reproducible research: A range of response. *Statistics, Politics, and Policy*, **2**, 4–1–0.

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].

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., and Deckmyn, A. (2016). *maps: Draw Geographical Maps*. R package version 3.1.1.

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.

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.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, **66**, 2215–2222.

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**, e151.

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].

Carr, D., ported by Nicholas Lewin-Koh, Maechler, M., and contains copies of lattice functions written by Deepayan Sarkar (2016). *hexbin: Hexagonal Binning Routines*. R package version 1.27.1.

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

Centers for Disease Control and Prevention (2016a). Non-Polio Enterovirus. https://www.cdc.gov/non-polio-enterovirus/about/symptoms.html. [Online; accessed 14-August-2017].

Centers for Disease Control and Prevention (2016b). 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 (2017a). Common Colds: Protect Yourself and Others. https://www.cdc.gov/features/rhinoviruses/index.html. [Online; accessed 14-August-2017].

Centers for Disease Control and Prevention (2017b). 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**, e133.

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.

Dowle, M. and Srinivasan, A. (2017). *data.table: Extension of 'data.frame'*. R package version 1.10.4.

Earp, B. D. and Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, **6**, 1–11.

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.

Emerson, J. W. and Kane, M. J. (2016). *biganalytics: Utilities for 'big.matrix' Objects from Package 'bigmemory'*. R package version 1.1.14.

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].

Heikkinen, T. and Järvinen, A. (2003). The common cold. *The Lancet*, **361**, 51–59.

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

Hijmans, R. J. (2016). *geosphere: Spherical Trigonometry*. R package version 1.5-5.

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.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, **26**, 1–22.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, **2**, e124.

Kane, M. J., Emerson, J., and Weston, S. (2013). Scalable Strategies for Computing with Massive Data. *Journal of Statistical Software*, **55**, 1–19.

Kane, M. J. and Emerson, J. W. (2016). *bigtabulate: Table, Apply, and Split Functionality for Matrix and 'big.matrix' Objects*. R package version 1.1.5.

Kass, R. E., Caffo, B. S., Davidian, M., Meng, X.-L., Yu, B., and Reid, N. (2016). Ten Simple Rules for Effective Statistical Practice. *PLOS Computational Biology*, **12**, e1004961.

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: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**, 1–9.

Murrell, P. (2003). Integrating grid graphics output with base graphics output. *R News*, **3**, 7–12.

Murrell, P. (2007). grid Graphics.

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*, 6–11.

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. (2006). Reproducible Epidemiologic Research. *American Journal of Epidemiology*, **163**, 783–789.

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.

Pons-Salort, M., Parker, E. P., and Grassly, N. C. (2015). The epidemiology of non-polio enteroviruses: recent advances and outstanding questions. *Current opinion in infectious diseases*, **28**, 479–487.

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.

R Core Team (2017). R: A Language and Environment for Statistical Computing. Technical report, R Foundation for Statistical Computing, Vienna, Austria.

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].

Rossum, G. (1995). Python Reference Manual. Technical report, CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands.

Rothman, K. J. (2012). *Epidemiology: An Introduction*. Oxford university press.

Rudis, B. (2016). *cdcfluview: Retrieve U.S. Flu Season Data from the CDC FluView Portal*. R package version 0.5.1.

Salathé, 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 Computational Biology*, **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**, e0142209.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., and Rana, O. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological research online*, **18**, 7.

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 (2011). 200 million Tweets per day. *Twitter Blog*. `https://blog.twitter.com/official/en_us/a/2011/200-million-tweets-per-day.html`. [Online; accessed 14-August-2017].

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].

United States Department of Health and Human Services (2014). Regional offices. `https://www.hhs.gov/about/agencies/iea/regional-offices/index.html`. [Online; accessed 13-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.

Vogt, L., Reichlin, T. S., Nathues, C., and Würbel, H. (2016). Authorization of Animal Experiments Is Based on Confidence Rather than Evidence of Scientific Rigor. *PLOS Biology*, **14**, e2000598.

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**, 1–8.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. and Francois, R. (2016). *dplyr: A Grammar of Data Manipulation*. R package version 0.5.0.

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.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, S. N. (2016). Just Another Gibbs Additive Modeler: Interfacing JAGS and mgcv. *Journal of Statistical Software*, **75**, 1–15.

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.

Xie, Y., Mueller, C., Yu, L., and Zhu, W. (2015). *animation: A Gallery of Animations in Statistics and Utilities to Create Animations*. R package version 2.4.

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.

Zhou, L. and Braun, W. J. (2010). Fun with the r grid package. *Journal of Statistics Education*, **18**, 1–35.

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

Appendix A

Appendices

A.1 R-Packages

```
print( sessionInfo(), locale=FALSE)

## R version 3.4.1 (2017-06-30)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.3 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## attached base packages:
##  [1] parallel  grid      compiler  stats     graphics  grDevices utils
##  [8] datasets  methods   base
##
## other attached packages:
##  [1] usmap_0.2.0            tseries_0.10-37        sfsmisc_1.1-1
##  [4] openWAR_0.2.3.9001     mosaic_0.14.4          Matrix_1.2-11
##  [7] mosaicData_0.14.0      MASS_7.3-45            maps_3.1.1
## [10] geosphere_1.5-5        forecast_7.3           timeDate_3012.100
## [13] zoo_1.7-14             dplyr_0.5.0            doParallel_1.0.10
## [16] iterators_1.0.8        deSolve_1.14           XML_3.98-1.5
## [19] vcd_1.4-3              timezone_0.1           proj4_1.0-8
```

```
## [22] tabulizer_0.1.24       stringr_1.1.0          scales_0.4.1
## [25] rworldxtra_1.01        rworldmap_1.3-6        rvest_0.3.2
## [28] xml2_1.1.1             R.utils_2.5.0          R.oo_1.21.0
## [31] R.methodsS3_1.7.1      rgeos_0.3-22           RCurl_1.95-4.8
## [34] bitops_1.0-6           qdap_2.2.5             RColorBrewer_1.1-2
## [37] qdapTools_1.3.1        qdapRegex_0.7.2        qdapDictionaries_1.0.6
## [40] profvis_0.3.3          png_0.1-7              plotrix_3.6-4
## [43] pdftools_1.2           pageviews_0.3.0        mgcv_1.8-19
## [46] nlme_3.1-131           maptools_0.8-41        sp_1.2-4
## [49] lubridate_1.6.0        lattice_0.20-35        hexbin_1.27.1
## [52] gtable_0.2.0           gridExtra_2.2.1        gridBase_0.4-7
## [55] ggplot2_2.2.1          ggdendro_0.1-20        gganimate_0.1
## [58] feather_0.3.1          devtools_1.12.0        chron_2.3-49
## [61] cdcfluview_0.5.1       bit64_0.9-5            bit_1.1-12
## [64] bigtabulate_1.1.5      biganalytics_1.1.14    biglm_0.9-1
## [67] DBI_0.5-1              foreach_1.4.3          bigalgebra_0.8.4
## [70] bigmemory_4.5.19       bigmemory.sri_0.1.3    animation_2.4
## [73] data.table_1.10.4      knitr_1.15.1
##
## loaded via a namespace (and not attached):
##  [1] backports_1.0.5     spam_1.4-0          Hmisc_4.0-2
##  [4] plyr_1.8.4          igraph_1.0.1        lazyeval_0.2.0
##  [7] splines_3.4.1       openNLP_0.2-6       digest_0.6.12
## [10] htmltools_0.3.5     gender_0.5.1        gdata_2.17.0
## [13] magrittr_1.5        checkmate_1.8.2     memoise_1.0.0
## [16] xlsx_0.5.7          tm_0.6-2            cluster_2.0.6
## [19] readr_1.0.0         wordcloud_2.5       Sxslt_0.91-4
## [22] colorspace_1.3-2    jsonlite_1.2        survival_2.40-1
## [25] V8_1.2              Rcpp_0.12.12        htmlTable_1.9
## [28] foreign_0.8-69      Formula_1.2-1       htmlwidgets_0.8
## [31] httr_1.2.1          acepack_1.4.1       rJava_0.9-8
## [34] openNLPdata_1.5.3-2 nnet_7.3-12         venneuler_1.1-0
## [37] reshape2_1.4.2      munsell_0.4.3       tools_3.4.1
## [40] evaluate_0.10       purrr_0.2.2         slam_0.1-40
```

```
## [43] curl_2.3            tibble_1.2          stringi_1.1.2
## [46] highr_0.6           fields_8.10         tabulizerjars_0.9.2
## [49] lmtest_0.9-34       R6_2.2.0            latticeExtra_0.6-28
## [52] KernSmooth_2.23-15  codetools_0.2-15    reports_0.1.4
## [55] gtools_3.5.0        assertthat_0.1      xlsxjars_0.6.1
## [58] withr_1.0.2         fracdiff_1.4-2      hms_0.3
## [61] quadprog_1.5-5      rpart_4.1-10        tidyr_0.6.1
## [64] NLP_0.1-9           base64enc_0.1-3
```

A.2 Github

In order to provide the reader with all the information needed to reproduce the findings reported in this thesis and to build upon it, all code files, processed data sets, and figures are publicly available on Github (`https://github.com/salathegroup/2016_TwitterEpi`). The repository contains the following main folders:

DataCleansing: Contains all code files needed for aggregating, transforming and pruning the raw output from the Twitter classifier. It also contains the processed data sets.

StatisticalAnalysis: Contains all code files needed to perform the statistical analyses of the data sets and the comparisons with the CDC data. It also contains the corresponding figures and output summaries.

PhDThesisBodnar: Contains all information revolving around Todd Bodnar's PhD thesis, including a summary of his communication to me. it also includes the code files and corresponding results of my reproduction attempts.

TwitterParse: Contains all information provided to me by Todd Bodnar with regard to the Java-based Twitter classifier, as well as my own additions.

TestingCode: Contains a collection of code files which just served for "quick-and-dirty" testing and will most likely not work without some tweaking. They are just included for completeness' sake. and to satisfy the especially curious ones. None of the output is directly used for this report, yet some of the code elements have found their way into the final code version provided in the other folder.

Report: Contains all necessary files to reproduce this Master thesis.

However, due to limited storage space and size limitations on Github, the raw files are not available. People interested in working with the original data can get in touch with EPFL's Digital Epidemiology Lab.