Britt Dijkstra

The effects of weather circumstances on the onset of influenza-like illness using the Grote Griepmeting data

Master thesis, defended on January 09, 2015

Thesis supervisor:
Prof. Dr. Jaqueline J. Meulman

Thesis advisors:
Prof. Dr. Theo Stijnen
Rosa Meijer, MSc

Specialization: Statistical Science

Leiden University Mathematics and Psychology Depts.

Leiden University Medical Centre

VU University Amsterdam Mathematics Dept

VU Medical Centre Amsterdam

Wageningen University and Research Centre

Abstract

As the gathering of data becomes easier, by for instance using computers and the internet, datasets keep becoming larger. This makes it more difficult to find an appropriate way to analyze the data. In this thesis I have analyzed the data of de Grote Griepmeting (influenzanet), which is a dataset that contains over 300,000 measurements. The goal of this thesis is to find out if weather circumstances have an effect on the incidence of influenza. To do this the data of de Grote Griepmeting are combined with weather variables gathered from the KNMI. The data of the Grote Griepmeting contains repeated measurements, multicollinearity, the covariates are likely to be nonlinear to the response variable and because influenza is a contagious disease most likely there will be dependence between the subjects. These issues all need to be accounted for in the analysis. In this thesis several possibilities to analyze the data are considered; the Cox proportional hazards model, a logistic regression model, generalized estimating equations and the generalized linear mixed model. Finally, it was decided to use a logistic regression, with a lasso penalty to account for multicollinearity and B-splines for nonlinearity. For using the B-splines a lot of extra variables need to be created, so the data expand even more. Computations become bothersome, and a trick from the medical field that is used in case-control studies is introduced to reduce computational time.

Contents

1. De Grote Griepmeting

1.1 Influenza

Influenza is an acute viral infection that spreads easily from person to person[1]. This happens through people coughing, causing infected droplets in the air, which others breath in. Influenza can also spread by for instance touching an infected hand. It is a serious public health problem that causes severe illness. It can affect anybody in every age group, but it is especially dangerous for high risk populations, such as children under 2 years old, the elderly above 65 and people with severe medical conditions, in which influenza can cause death. Worldwide the incidence of influenza is about 3 to 5 million cases of severe illness and about 250.000 to 500.000 deaths per year. Most deaths associated with influenza in industrialized countries occur among people of age 65 or older. Influenza also takes an economic toll through lost workforce productivity and overwhelmed hospitals during influenza epidemics.

The symptoms of influenza are a sudden onset of high fever, a dry cough, sore throat, runny nose, headache, muscle and joint pain and severe malaise. Most people recover of the fever in a week, but the cough can last for over 2 weeks. Medical attention is not needed in most cases. The incubation time for influenza is around 2 days[2].

This thesis will focus on influenza in The Netherlands and its potential to be influenced by weather circumstances. The Netherlands has a temperate climate, which means that the changes in weather between summer and winter are moderate. In regions with a temperate climate, like The Netherlands, the influenza epidemics will peak during winter[1]. This might indicate a relation between weather circumstances and influenza.

The incidence of influenza in The Netherlands is around 40 per 1000 persons each year [3]. During an epidemic this can run up to about 50 to 200 per 1000 persons. The incidence of influenza is greatest in children from 0-5 years old. In the Netherlands less than 1 per 100.000 persons die each year from influenza. During an epidemic about 15.000 to 30.000 people are hospitalized.

Articles have been written on the influence of the weather on the mortality of influenza[4][5] and on the relationship between respiratory illnesses and the weather[6][7]. There are also articles on the relationship between the incidence of influenza and weather variables [7][8]. However, all studies on the relation between influenza and the weather were performed in Northern America, the East of Asia and Australia. It seems that similar studies have not yet been conducted in Europe. In all current articles on this topic, the subjects were retrieved from clinics.

In this thesis the data from the Grote Griepmeting[9] will be used. The Grote Griepmeting is an internet questionnaire, used in The Netherlands, in which everybody can participate, so not just the subjects that report to clinics are used. Because of the large peaks in influenza-like illness during autumn and winter time in temperate climates, as can be seen in The Netherlands as well, the question rises whether weather circumstances play a part in inducing the onset of influenza. In this thesis I do not want to see if there is just a relation between the incidence of influenza and the weather, but if it is possible to predict the incidence of influenza using weather variables, which will be retrieved from the Koninklijk Nederlands Meteorologisch Instituut (KNMI)[10]. This thesis will try to answer the following question: Can the incidence of influenza-like illness be predicted using weather variables?

1.2 Introduction to the Grote Griepmeting

Until recently, the only insight in influenza in the Netherlands came from information obtained by general practitioners (GP's). Not everyone with symptoms of influenza will check in with their GP, so this information will not be completely representative.

In 2003 the Grote Griepmeting kicked off. The Grote Griepmeting tries to get an insight in the distributions of colds and influenza in the Netherlands by using an internet questionnaire asking "plain" people about their symptoms. Participants are asked to check boxes indicating the symptoms they experience. It is not possible to know if the participant has influenza, because there is no simple home test or test the general practitioner can perform to find out if the symptoms experienced are caused by influenza. This is why participants are not asked if they suffer from influenza. Experts have established criteria regarding which combinations of symptoms are likely to reflect influenza. If a participants symptoms fit the criteria we say the participant has “influenza-like illness” (ILI), since we cannot be positive whether it really is influenza. For this thesis I have been using the raw data and established my own definition of ILI.

The Grote Griepmeting was founded by Carl Koppeschaar[11], an astronomer, physicist and mathematician. The team of the Grote Griepmeting also contains biologists, virologists, medical doctors, computer scientists and web hosts.

The Grote Griepmeting started in the season of 2003/2004 on the 3rd of October and was the first in the world to collect data on colds and influenza through the internet[9]. The goal of the Grote Griepmeting is to collect as much detailed data as possible on cold- and influenza epidemics. The purpose for this data is for researchers to be able to create models on influenza or create simulations on for instance the spread of influenza. A lot of topics could be researched with these data, for instance the effectiveness of the flu shot or the typical geographical distribution of influenza like illness. The distribution of ILI can consecutively be compared with for instance transportation methods, like public transportation. These are just a few examples.

In the season of 2004/2005 the project was expanded with additional research on the relation between stress and influenza. In the season of 2005/2006 the project became international when Portugal joined and in 2008 Italy followed. In collaboration with the Portuguese and Italian colleagues the European influenza measurement was founded. This was a central part of the 4-year lasting European research program EPIWORK. This project was funded from 2009 until 2013 by the European Commission. In the past year also Great-Britain, Sweden, France and Spain joined the project. The project went worldwide when also Australia and the United States joined.

Participants of de Grote Griepmeting are recruited in several ways[12]. Articles have been published on a regular basis in prominent (and less prominent) newspapers like Sp!ts and Trouw, magazines like the Libelle, but also on websites like nu.nl, ziekenhuis.nl and gezondheidsplein.nl. Also several television and radio interviews were conducted. Scientific articles have also been published based on the Grote Griepmeting, for instance articles of Marquet et al. (2006)[13], Friesema et al. (2009)[14] and van Noort (2012)[15]. There are also publications based on the influenzanet, the international version of the Grote Griepmeting.

1.3 Data obtained from the Grote Griepmeting

I had access to the data of 8 seasons from the Grote Griepmeting, ranging from season 2003/2004 until season 2010/2011. The seasons contain over 300.000 measurements each. Therefore it was decided to only use season 2010/2011 to start with.

The Grote Griepmeting is an internet questionnaire consisting of two parts; the intake and the weekly questionnaire. In the intake questionnaire, questions are asked to obtain demographic information about the subject, which is not of much interest for this study, except for the postal code. Via the postal code, the subject’s hometown can be linked to the nearest KNMI weather station. We have chosen this approach, since we expect substantial differences in values for the weather variables for places landward and on the coast.

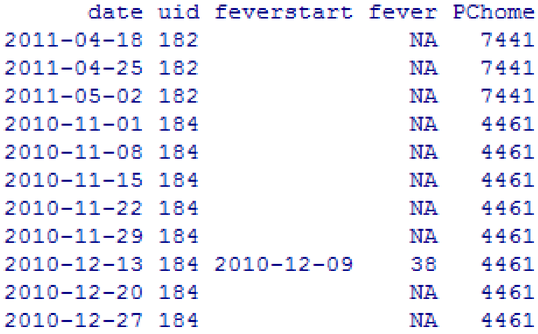
For this study, the primary interest is in the weekly questionnaire. The questionnaire is given in Appendix A. The data obtained from the weekly questionnaire contains a maximum of 17 variables. Not all questions from the questionnaire are filled out when the subject reports no symptoms.

The following variables were selected from the weekly symptoms’ questionnaire (and the postal code is added from the intake questionnaire);

- Date : The calendar date at which the measurement is taken

- Uid : a unique number indicating a subject
- Feverstart: The date the fever started
- PChome : Postal code corresponding to the subjects home address
- Fever : Variable indicating the body temperature of the subject at the feverstart date

Since the date of the measurement and the date on which an episode of fever for a subject started can be different, the variable feverstart was added to the data. In Table 1 a short extract of the data is presented.



```
      date uid feverstart fever PChome
2011-04-18 182                NA   7441
2011-04-25 182                NA   7441
2011-05-02 182                NA   7441
2010-11-01 184                NA   4461
2010-11-08 184                NA   4461
2010-11-15 184                NA   4461
2010-11-22 184                NA   4461
2010-11-29 184                NA   4461
2010-12-13 184 2010-12-09     38   4461
2010-12-20 184                NA   4461
2010-12-27 184                NA   4461
```

Table 1. A short extract of the data

The first column “date” is the calendar date at which the measurement is recorded. As can be seen from this column the questionnaire was filled out every week. Subjects get a reminder per e-mail each week to fill in the questionnaire. Of course participating is free of obligation, so it is possible that participants refrain from filling out the questionnaire every week. This can be seen in Table 1 between the 8th and 9th row, where the measurements for subject 184 are 2 weeks apart. In the second column of Table 1, the variable ‘uid’ indicates that the data of subject 182 and 184 are shown.

The variable “feverstart” shows that subject 184 experienced an episode of ILI, which started at 2010-12-09, which the subject filled out on the calendar date 2010-12-13. Here, the following situation is recognized; the variable “feverstart” indicates when the episode of ILI started, but there is no variable indicating when the episode ends. What is known is that subject 184 reports to be healthy at the date of the next measurement at 2010-12-20. The ‘fever’ variable from line 9 indicates the measured body temperature in case of illness. For example we know that the body temperature of subject 184 was 38.0 degrees Celsius.

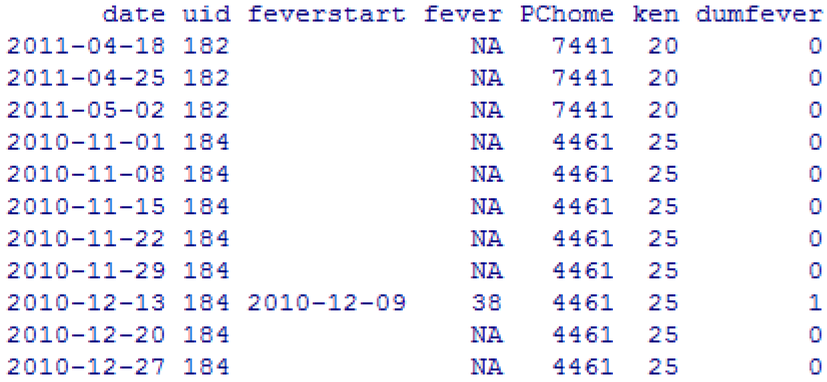
1.4 Data preparation: creating variables

The definition of ILI in this thesis will be based on body temperature. A subject is affected by ILI if the body temperature is equal or higher than 38.0 degrees. This definition is chosen because it is hard to define ILI from the symptoms listed by the subjects, especially when one does not know the severity of the symptoms and whether or not there are other ways of explaining them. Therefore the body temperature (commonly known as fever) seems the most clear-cut way to define ILI. To enable statistical analysis, a dichotomous (0/1) variable for the outcome measure fever needs to be created. The variable “feverstart” cannot simply be turned into an indicator variable when a date occurs, since it contains dates for which the temperature is lower than 38.0 degrees. Therefore, the variable “dumfever” is created, using information from the ‘fever’ variable, which will be a dichotomous variable with value equal to ‘0’ indicating no signs of ILI and value equal to ‘1’ indicating a body temperature equal or over 38.0 degrees, hence an episode of ILI.

The original survey data of season 2010/2011 consists of 17 columns and 307,321 rows. The rows being the 307,321 measurements obtained from 17,871 subjects. This means there are on average

17 measurements per subject. From these 17,871 subjects, 13,496 never experienced an episode of influenza-like illness, whereas 4,375 people did experience at least one episode of ILI.

The weather stations are located in different locations in The Netherlands. To connect the subject's home postal code to the nearest weather station, a variable "ken" is created. The nearest weather station will be assigned the same "ken" number, so the "Grote Griepmeting" data and the "KNMI" data can be merged by the "ken" variable. An example of the extended dataset in which both the variable "dumfever" and "ken" are added is given in Table 2.



```
      date uid feverstart fever PChome ken dumfever
2011-04-18 182               NA   7441  20        0
2011-04-25 182               NA   7441  20        0
2011-05-02 182               NA   7441  20        0
2010-11-01 184               NA   4461  25        0
2010-11-08 184               NA   4461  25        0
2010-11-15 184               NA   4461  25        0
2010-11-22 184               NA   4461  25        0
2010-11-29 184               NA   4461  25        0
2010-12-13 184 2010-12-09    38   4461  25        1
2010-12-20 184               NA   4461  25        0
2010-12-27 184               NA   4461  25        0
```

Table 2. Two additional variables: dumfever and ken

1.5 Data preparation: cleaning up the data

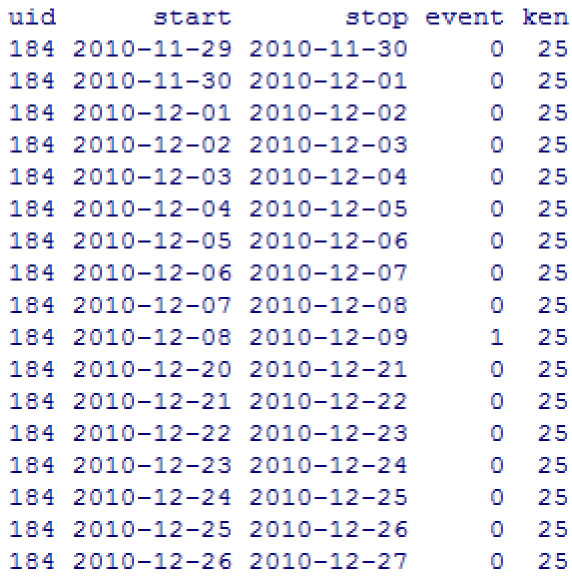
The data is checked for large gaps in between the measurements. If subjects have large gaps in between their measurements, this could indicate that they do not fill out the questionnaire consistently and that they might not be reliable subjects. It is likely that subjects loose interest in the Grote Griepmeting when they do not experience ILI and only fill out the questionnaire when they do experience symptoms. This would induce bias. Hence, the subjects with gaps between measurement larger than 40 days, which would equal missing 5 measurements, are excluded from the study. For the smaller gaps it is assumed that no episode of influenza-like illness occurred. Now, 888 subjects were excluded (5%), which accounted for a total of 8,620 measurements (2%).

The number of measurements per subjects is checked in the data. This is done because subjects with only a few measurements have low reliability, because there is no telling whether they fill out the questionnaire consistently, which could also induce bias. Decided is to exclude subjects with less than 5 measurements. Another 3,199 subjects were excluded from the study, which accounted for 6,066 measurements. In total 4,087 subjects were excluded from the study (23%) which represented 14,686 measurements (5%).

The prepared data now contains 13,784 subjects and 292,639 measurements.

1.6 Data preparation: expanding the data

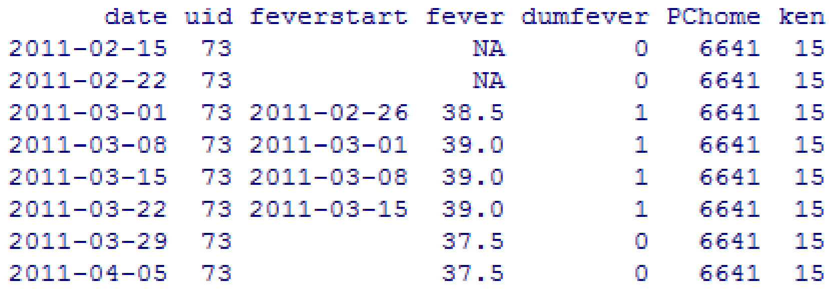
In this thesis we are interested whether weather variables are related to the onset of ILI. In order to answer this question, the weather variables need to be matched to the days the subject is at risk of experiencing the event. As can be seen from Table 2, the data at this point contains intervals of a week or more. We want to create intervals of one day, so weather variables can be connected to each day. What is known is that, if in two consecutive measurements no fever is reported and there is no reported 'feverstart' date, there will be no incidence of ILI in between these measurements. For example, subject 182 reports no fever at 2010-04-18 and 2010-04-25. For this period we can create 7 intervals of one day, where no fever or feverstartdate is reported and so the dumfever variable is zero. If one day intervals are to be created for periods that contain an episode of ILI, it gets more complicated, since it is unknown at what date the episode ends. Intervals at which the subjects still experience ILI after the onset date should not be included, since the subject is not at risk to be infected with another episode of ILI at this time. Since there is no way of knowing the exact date of recovery from the episode of ILI, it was decided to exclude the period after the date of the feverstart date. The first interval that will be reported after an episode of ILI, will be the first calendar date at which the subject reports healthy after an episode of ILI. For example, for subject 184 (see Table 2), for the period between calendar dates 2010-11-29 and 2010-12-20 the following 1 day intervals will be created. From 2010-11-29 until 2010-12-08, 1-day intervals will be created with dumfever = 0. The 1-day interval 2010-12-08 until 2010-12-09 will have dumfever =1. The intervals from 2010-12-10 until 2010-12-19 will be missing, since it is unknown in this period whether the subject is healthy or still experiences ILI, so whether or not the subject is at risk. Table 3 presents an example of what the data will look like at this point. In this table it can be seen that at calendar date interval 2010-11-08 until 2010-11-09 an episode of ILI starts. The next interval is 2011-11-20 until 2011-11-21, so 11 intervals are missing, because it is unknown in these intervals if the subject is experiencing ILI.



```
uid      start       stop event ken
184 2010-11-29 2010-11-30     0  25
184 2010-11-30 2010-12-01     0  25
184 2010-12-01 2010-12-02     0  25
184 2010-12-02 2010-12-03     0  25
184 2010-12-03 2010-12-04     0  25
184 2010-12-04 2010-12-05     0  25
184 2010-12-05 2010-12-06     0  25
184 2010-12-06 2010-12-07     0  25
184 2010-12-07 2010-12-08     0  25
184 2010-12-08 2010-12-09     1  25
184 2010-12-20 2010-12-21     0  25
184 2010-12-21 2010-12-22     0  25
184 2010-12-22 2010-12-23     0  25
184 2010-12-23 2010-12-24     0  25
184 2010-12-24 2010-12-25     0  25
184 2010-12-25 2010-12-26     0  25
184 2010-12-26 2010-12-27     0  25
```

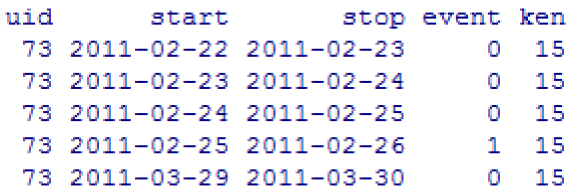
Table 3. Creating 1-day intervals

In Table 4 and 5 an example is shown of how intervals are created when the subject reports consecutive episodes of ILI. Probably the subject did not recover from the episode it reported at the first measurement and this can be seen as one long episode of ILI. As can be seen in Table 5 all intervals after the first interval in which ILI was reported are excluded. The next interval starts when the subject report healthy again. In Table 4 it can be seen that only a body temperature of 38.0 degrees Celcius or more is denoted as an episode of ILI. When subject 73 experienced a body temperature of 37.5 the dumfever variable was set to zero.



```
      date uid feverstart fever dumfever PChome ken
2011-02-15  73               NA        0   6641  15
2011-02-22  73               NA        0   6641  15
2011-03-01  73 2011-02-26  38.5        1   6641  15
2011-03-08  73 2011-03-01  39.0        1   6641  15
2011-03-15  73 2011-03-08  39.0        1   6641  15
2011-03-22  73 2011-03-15  39.0        1   6641  15
2011-03-29  73             37.5        0   6641  15
2011-04-05  73             37.5        0   6641  15
```

Table 4. Consecutive ILI reported



```
uid      start       stop event  ken
 73 2011-02-22 2011-02-23     0   15
 73 2011-02-23 2011-02-24     0   15
 73 2011-02-24 2011-02-25     0   15
 73 2011-02-25 2011-02-26     1   15
 73 2011-03-29 2011-03-30     0   15
```

Table 5. Intervals created for the period of consecutive ILI reported

2. KNMI data

2.1 Koninklijk Nederlands Meteorologisch Instituut (KNMI)

The data containing the weather variables are obtained from the "Koninklijk Nederlands Meteorologisch Instituut" (KNMI)[10]. The obtained data consists of 39 weather variables and a variable containing the date of the measurement. In appendix B the complete list of variables obtained from the KNMI is demonstrated. Based on an article of Steven Zhixiang ZHOU (2009)[16] on seasonal influenza all around the world several variables are selected.

Motivation for variables:

- Average temperature:
  1. Cold air inhaled can reduce the respiratory defences such as mucociliary clearance and phagocytic activity of leukocytes.
  2. Cooling of the body surface can cause vaso-constriction in the nose, resulting in reduced blood flow and leukocyte supply and increased susceptibility to infection.
  3. Cold temperature influences host behaviours by driving people to stay indoors, increasing the proximity between susceptible individuals and infected hosts.
- Change in temperature:
  1. Abrupt change in temperature causes people to develop more influenza-like symptoms.
- Humidity:
  1. Viruses remain longer viable when humidity is low.
  2. In a dry atmosphere, virus carrying particles expelled from infected hosts can remain suspended in the air for a longer time, increasing the likelihood for transmission. When the air is dry, large drops partially evaporate, creating smaller lighter drops that are more likely to remain airborne for extended periods of time.
  3. Inhalation of air with low humidity could dry the mucus, impairing the subject's defences against infection.
- Dew point:
  1. A strong correlation between influenza activity and low dew point temperature in temperate and arctic regions was found.
  2. Dew point is a comprehensive estimator of temperature and humidity with less chance for misinterpretation than the effect of pure vapour amount in the air.
- Solar radiation:
  1. Studies have indicated that the influenza virus is sensitive to the electric magnetic spectrum near 254-nm and can be inactivated by solar radiation.
- Sun hours:
  1. Solar radiation exposure and length of daily photoperiod affect a subject's vitamin D level and emotion. Vitamin D modulates the effectiveness of macrophages and induces antimicrobial peptide gene expression. Emotions can affect the susceptibility to the common cold.
- Precipitation:

1. Precipitation is usually highly correlated to weak solar radiation, reducing germicidal effect and Vitamin D level.
2. It can affect subject behaviours, mainly the frequency of contact with other hosts.
3. The grey outside world may induce negative feelings, which can significantly reduce natural killer cytotoxicity in the immune system and increase the susceptibility to the common cold.

Not all the above mentioned variables are already available from the obtained data from the KNMI. The unavailable variables will be created from the KNMI data. This will be described in the data preparation.

2.2 Data preperation: creating variables

From the KNMI data we can already extract the following variables (names used in the tables are in brackets):

- Average daily wind speed in 0.1 m/s (wind)
- Average daily temperature in 0.1 degrees Celsius (temp)
- Minimum daily temperature in 0.1 degrees Celsius (mintemp)
- Maximum daily temperature in 0.1 degrees Celsius (maxtemp)
- Percentage of maximum potential sunshine duration (zonduur)
- Global radiation in J/cm2 (straling)
- Precipitation duration in 0.1 hour (neerslagduur)
- Daily precipitation amount in 0.1 mm (-1 for <0.05 mm) (neerslagsom)
- Daily average relative atmospheric humidity in percents (vochtigheid)
- Potential evapotranspiration (Makkink) in 0.1 mm (ref)

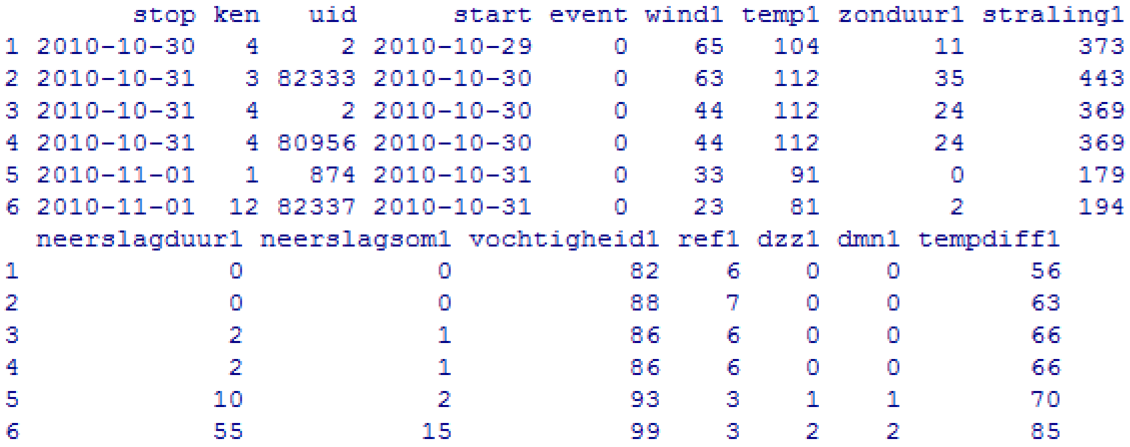
The “neerslagsom” variable turns -1 when the daily precipitation amount is less than 0.05 mm. Since the precipitation amount cannot be less than zero, the value -1 is turned into zero.

The following variables are created from the KNMI-data:

- Consecutive days with less than an hour of sun in days (dzz)
- Consecutive days with precipitation in days (dmn)
- Maximum difference in temperature per day in 0.1 degrees Celsius (tempdiff)

3. Merging of the data

In general, the incubation period for influenza is estimated to range from 1 to 4 days with an average of 2 days[2]. It has to be taken into consideration when the Grote Griepmeting data is merged with the weather variables. To do this 1 (2, 3 or 4) day(s) is added to the dates of the weather variables. In this way the weather variables of the day before the measurement are matched to the measurement. Table 6 presents the data when just the weather variables of 1 day before the measurement are added in addition to the measurements of one day before. For the complete dataset, weather variables belonging to respectively 2,3,4 days before the measurement are added. The days are indicated by the number behind the variable. The total dataset now contains 57 columns and 2,137,304 rows.



```
        stop ken   uid      start event wind1 temp1 zonduur1 straling1
1 2010-10-30   4     2 2010-10-29     0    65   104       11       373
2 2010-10-31   3 82333 2010-10-30     0    63   112       35       443
3 2010-10-31   4     2 2010-10-30     0    44   112       24       369
4 2010-10-31   4 80956 2010-10-30     0    44   112       24       369
5 2010-11-01   1   874 2010-10-31     0    33    91        0       179
6 2010-11-01  12 82337 2010-10-31     0    23    81        2       194
  neerslagduur1 neerslagsom1 vochtigheid1 ref1 dzz1 dmn1 tempdiff1
1             0            0           82    6    0    0        56
2             0            0           88    7    0    0        63
3             2            1           86    6    0    0        66
4             2            1           86    6    0    0        66
5            10            2           93    3    1    1        70
6            55           15           99    3    2    2        85
```

Table 6. Piece of Grote Griepmeting and weather variables 1 day before measurement.

The data of the Grote Griepmeting and the KNMI are combined now. With this dataset it will be tried to answer the following question: Can the incidence of influenza-like illness be predicted using weather variables?

4. Points of attention in the data

4.1 Repeated measurements

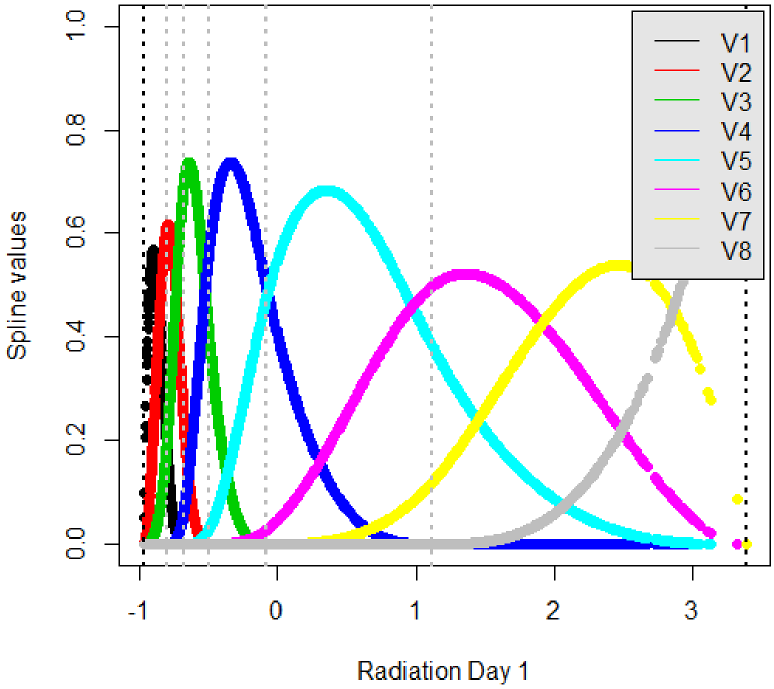
The data of the Grote Griepmeting contains repeated measurements on a single individual on different occasions, since each individual fills out the questionnaire more than once[17]. The repeated measurements for the subjects in the Grote Griepmeting are ordered in time.
Measurements within a cluster will typically exhibit a positive correlation, which should be accounted for in the analysis. The positive correlation violates the assumptions of independence, which is crucial in most standard statistical analysis techniques. So it is necessary to find a technique that takes into account this correlation between measurements.

4.2 Nonlinearity

It is expected that the relation between the explanatory variables (i.e. the weather variables) and the response variable (incidence of ILI) will not be perfectly linear[18]. There are different spline methods, but in this thesis I have chosen to use B-splines[19]. First the range of a weather variable X is divided into intervals. The endpoints of these intervals are called the knots. The minimum and maximum of the range of X are called the boundary knots, and the other points in between the boundary knots, are called the inner knots (K). Thus, if for example 4 intervals are chosen, we have K=3 inner knots, and two boundary knots. The B-splines are of a chosen degree (q) and this degree defines the degree the polynomials the B-splines consist of have. Thus, on each interval the effect of X is modeled by a polynomial of the degree q. The polynomials should be smoothly connected, which is done by forcing them to be connected in the knots with the same first up to the (q-1)th derivative, which depends on the degree chosen for the B-splines. Thus, splines consist of several polynomial pieces, smoothly joint in so-called spline knots. The knots are inserted at chosen values of the weather variable.
In this thesis it was chosen to use cubic B-splines (so degree, q = 3), because a cubic spline gives a curve that appears smooth to the human eye and higher degrees will probably not improve the model much. For a cubic spline the first and second derivative of the polynomials should be equal in the knots. If there are no inner knots, 4 regression parameters are needed to model a cubic relationship including an intercept. If one inner knot is added, another 4 parameters are added, but only one of them is effectively free, because the two third degree polynomials are forced to be

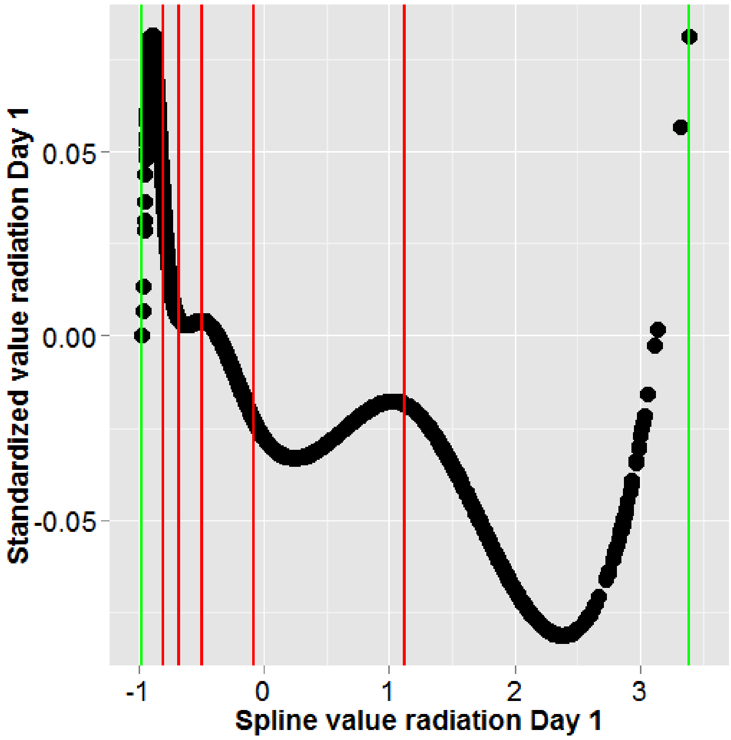
connected and to have the same first and second derivative. Thus, the B-spline has 5 regression parameters if the model includes an intercept and 1 inner knot, and for each extra inner knot one extra parameter is added. In general the number of regression coefficients is q+1+K including the intercept, or q+K excluding the intercept. The corresponding q+K covariates are called the spline basis functions.

In this thesis I will mostly use cubic B-splines with 5 inner knots, so 8 covariates have to be calculated. There are several methods to calculate these covariates, among them the method based on the recurrence relation of Carl de Boor[20].
In Graph 1 an example of 8 B-spline basis functions is shown for the Radiation Day 2 weather variable. The grey dotted lines show the positions of the inner knots, the boundary knots are indicated by the black dotted lines. As can be seen, 8 B-splines are created, each ranging over 5 adjacent knots and consisting of 4 intervals, being described by third degree polynomials.
The R-package 'splines' was used to create the B-splines[21]. It creates extra 8 columns for the B-splines in the dataset per weather variable.



Graph 1. Example of B-spline

If a linear combination is made, by multiplying the 8 columns of values for the splines with the corresponding value of the weather variable and then add up the results of the multiplication a smooth curve appears. This curve can be seen in Graph 2.



Graph 2. B-spline curve

4.3 Multicollinearity

Another issue that needs to be addressed is multicollinearity[22]. Multicollinearity refers to a situation in which one or more covariates in the model are highly correlated. If multicollinearity occurs, regression coefficients can be determined, but the standard errors will be large, which means that the coefficients cannot be estimated with great accuracy. Collinearity can cause unstable estimates of covariates, which can mask or amplify the underlying true effect. Also, some covariates may be dropped from the model, although they are important in the research. Collinearity can be explored by examining the correlation matrix or the eigenvalues.

The dataset in this thesis contains over 40 variables, mainly weather variables, consisting of a set of 11 variables, repeatedly measured over 4 consecutive days. These 11 variables, will be highly correlated over the different days. Also between different variables on the same day a high correlation can be expected, because some weather variables influence each other. For example, when there is a great amount of precipitation, the amount of sun hours will be low.

Bridge regression[23] is a special family of penalized regression methods that deals with collinearity. The bridge regression minimizes the negative log-likelihood subject to constraint $\Sigma|\beta_j|^{\Upsilon} \leq t$ with $\Upsilon>0$, where the β_j's are the regression coefficients corresponding to the weather variables in some chosen statistical model. In this way the estimates are shrunken towards zero. By shrinking the estimates a little more bias is introduced in order to reduce the variance of the estimates and reduce the mean squared error. The regression model can have better predictor accuracy this way.

A special case of bridge regression is ridge regression[24], where $\Upsilon=2$. Ridge regression shrinks variables towards zero, but keeps all the variables in the model. Another special case is the “least absolute shrinkage and selection operator” (lasso)[25], where $\Upsilon=1$. This method will shrink some variables towards zero and set others to zero, in this way performing variable selection. A disadvantage of the lasso penalty is that it possibly will select different variables when the dataset is slightly changed[26]. Also standard errors cannot be calculated in the conventional way[25] for ridge and lasso regression,, so an alternative method like the bootstrap should be used to calculate a measure for variability. In this thesis there are a great amount of variables. The lasso selects variables, which makes the model more workable and easier to interpret and because of this the lasso method was chosen for this analysis. The lasso solves the problem of finding the estimates for β by minimizing the "penalized"negative log-likelihood in the following way[27]:

$$\hat{\mu}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda) = arg\min_{\mu,\beta}(-\textstyle\sum_{i=1}^{n}\log(p_{\mu,\boldsymbol{\beta}}\,(\boldsymbol{y_i}|X_{\boldsymbol{i}}) + \lambda||\boldsymbol{\beta}||_1),$$

where $p_{\mu,\beta}(\boldsymbol{y_i}|X_i)$ is the probability density function through which $\boldsymbol{y_i}$ depends on $X_{i,}$ $\boldsymbol{y_i}$ is the observed outcome variable, X_i is a vector containing the covariates, $\boldsymbol{\beta}$ is a vector of unknown regression parameters and λ is a tuning parameter. Usually the intercept term is not penalized and this will also not be done in this thesis. The tuning parameter controls the amount of shrinkage that is applied to the estimate. The higher the tuning parameters, the more shrinkage is introduced and the more estimates will be set to zero. Determination of the tuning parameter can be done by cross validation. The lasso regression can for instance be performed with the R-package 'penalized'[28].

The splines that will be used in the analysis consist of 8 different columns of values, which are added to the design matrix. Since the lasso penalty can set variables, or in this case spline parts of the variables, to zero, incomplete spline variables can enter in the model, which was not intended. We would want to select all 8 parts the spline consists off or set all of them to zero. The accomplish this the group lasso[29] will be used. The group lasso function was originally developed to perform the lasso penalty on categorical variables, but it seems suitable for the spline variables too. For the spline variables it will select all parts of the spline or it will set the complete spline to zero. The group lasso can for instance be performed by the R-package 'grppenalty'[30].

4.4 Independence of subjects

In the analysis it is assumed that the subjects in the study are independent from each other. Influenza-like illness is a contagious disease, so the subjects will probably not be independent. If one of the subjects is infected with influenza-like illness, the disease will spread and more people will get infected. As long as there is no subject who is infected the disease will not spread.
This relationship is not modelled in the analysis, since we do not know an appropriate way to model the correlation between subjects.

5. Proposals for analysis and why they were rejected

5.1 Cox proportional hazards model

The first analysis that probably comes to mind for this kind of data will be survival analysis, since the data contains periods of time until ILI occurs. In this thesis we are not so much interested in the time until the episode of ILI occurs, but into the effects the weather circumstances have on the incidence of ILI. The Cox proportional hazards model can be used to estimate the hazard rate[31]. The hazard rate is the rate with which the subjects at risk of ILI, so the subjects that were healthy the day before, are experiencing an episode of ILI at the chosen time point. The time points in this thesis being the one-day intervals created earlier. The hazard λ rate is defined in the following way:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t},$$

where $\lambda(t)$ is the function giving the hazard rate at time t and T is the time at which the episode of ILI starts.

With the Cox proportional hazards model one can estimate the hazard rate using the weather variables as explanatory variables. The weather variables will change over days, so the model should contain time-dependent covariates (covariates that change over time).

The Cox proportional hazards model describes the hazard ratio as dependent on the weather variables in the following way:

$$\lambda(t|\boldsymbol{x_{it}}) = \lambda_0(t)e^{\boldsymbol{x_{it}\beta}},$$

where $\lambda(t|\boldsymbol{x_{it}})$ is a function giving the hazard rate at time t (a one-day interval) given the weather variables $\boldsymbol{x_{it}}$, λ_0 is a function containing the baseline hazard at time t, $\boldsymbol{\beta}$ is a vector of unknown regression parameters and $\boldsymbol{x_{it}}$ is a vector containing the values of the weather variables at 1,2,3 and 4 days previous to time t, but in the design matrix these values are corresponding to time t. By taking the values of the covariates at time t, the covariates are made time-dependent. If $\boldsymbol{x_{it}}$ was replaced by $\boldsymbol{x_i}$, $\boldsymbol{x_i}$ would not depend on time and the variable would be stationary. Stationary and time-dependent covariates can simultaneously be included in the model.
The estimation of the $\boldsymbol{\beta}$-coefficients can be done using the maximum likelihood approach. By maximizing the likelihood the values for the coefficients are found that are most “likely” for the data. This can be done with for instance R-software package ‘coxph’[32].
Subjects participating in the Grote Griepmeting can experience multiple episodes of ILI. The correlation introduced by these recurrent events need to be accounted for in the Cox model. This can be done in two ways, namely with the Andersen-Gill method (A-G method)[33] or with a frailty variable[31].
The Andersen-Gill method adds a covariate to the Cox model, which counts the number of events, in this case episodes of ILI, that the subject has yet experienced.
Another way to account for the recurrent events in a Cox model is to add a frailty variable. A frailty variable is a random effect that is added to the model to describe the frailty of the subjects in the study, so each subject has its own value for the frailty. The higher the value of the frailty for a subject, the higher the chance this subject will be infected with influenza-like illness. In this analysis the frailty variable is not used to see which subject is most frail, but to adjust for overdispersion, or said otherwise, for the correlation between the outcomes of the same subject. The frailty adjusts for effects that are not measured by the other covariates in the model, but which can be important for the model. If this is ignored the hazard rate estimates may be biased.
The difference between the two methods is that the A-G method is a population-averaged method, which means it predicts the mean response of the population and the frailty method is subject-specific, which means that the estimated coefficients describe the changes in the mean response of a specific subject.

The Cox proportional hazards model is less appropriate for my goal for the analysis, because of the baseline hazard. The baseline hazard changes for each time t. Because of the different baseline hazards for each time t, the baseline hazard rate contains the effect for the different days. The β-coefficients will now only contain the effect caused by the minor differences in weather circumstances on a specific day, for instance May 5th 2011 , but at different places in The Netherlands. The effects of the much larger weather differences between the days are not picked up by the regression coefficients, but disappear in the baseline hazard. What we are aiming for in this thesis is to find the effects of the different weather circumstances on all different days of the year, since we want to know if subjects would experience ILI more often in for instance low temperatures in January instead of high temperatures in June.

5.2 Logistic regression

The next idea that came to mind is the logistic regression model[17]. In logistic regression the observed outcome variable, referred to as y_{it} here, is binary, which means the variable only contains 0/1 outcomes. In this thesis the observed outcome variable is whether the subject experiences the start of an episode of ILI $(y_{it} = 1)$ or not $(y_{it} = 0)$ on a specific day. Since there are two possible outcomes for y_{it}, the probability distribution for y_{it} is Bernouilli, with $Pr(y_{it} = 1) = p_{it}$ and $Pr(y_{it} = 0) = 1 - p_{it}$. The logistic regression describes the effect the covariates have, in this case the weather variables measured 1,2,3 and 4 days previously to a specific time day, on the chance that $(y_{it} = 1)$, so a case of ILI. In a linear predictor $\boldsymbol{x_{it}\beta}$ the predicted outcomes can range from minus infinity to infinity, but the observed outcomes range from 0 to 1. A transformation is needed to confine the predicted outcome to the range of 0 to 1. In case of the logistic regression this is done by a logit link. The mean of each response $E(y_{it}|\boldsymbol{x_{it}}) = p_{it}$ depends on the covariates in the following way:

$$logit(p_{it}(\boldsymbol{x_{it}})) = \log\left(\frac{p_{it}(\boldsymbol{x_{it}})}{1 - p_{it}(\boldsymbol{x_{it}})}\right) = \boldsymbol{x_{it}\beta}\ ,$$

where $\boldsymbol{\beta}$ is a vector of unknown regression coefficients and $\boldsymbol{x_{it}}$ is a vector containing the covariates. If p_{it} is the probability of occurrence of ILI or in general an event, then $p_{it}/(1 - p_{it})$ is the odds of an event occurring, so $\log(p_{it}/(1 - p_{it}))$ is the log odds of an event occurring.

The logistic model can be expressed in terms of probability of an event in the following way:

$$p_{it}(\boldsymbol{x_{it}}) = \frac{e^{x_{it}\boldsymbol{\beta}}}{(1 + e^{x_{it}\boldsymbol{\beta}})},$$

The estimation of the $\boldsymbol{\beta}$-coefficients is done using the maximum likelihood approach. If outcomes of the same subject would be independent, the likelihood for the logistic regression is as following:

$$L(\boldsymbol{\beta}, y) = \prod_{i,t} p_{it}^{y_{it}}\,(1 - p_{it})^{(1-y_{it})}$$

From which the following log-likelihood can be derived:

$$l(\boldsymbol{\beta}, y) = \sum_{i,t}(y_{it}log(p_{it}) + (1 - y_{it})\log(1 - p_{it}))$$

The logistic regression model and the Cox proportional hazards model have a lot of similarities. It is actually possible to convert the Cox model into a logistic regression model.

The formula's for the models are already described above, but will be written again in a way so the resemblance becomes more clear.

Cox proportional hazards model: $\lambda(t|x_{it}) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_p x_p}$,

where $\lambda(t|x_{it})$ is a function giving the hazard rate at time t (a one-day interval) given the covariates x_{it}, λ_o is a function containing the baseline hazard at time t, $(\beta_1 \dots \beta_p)$ is a vector of unknown regression parameters and $(x_1 \dots x_p)$ is a vector of the covariates.

Logistic regression model: $\log\left(\frac{p_{it}}{1-p_{it}}\right) = \beta_0(t) + \beta_1 x_1 + \cdots + \beta_p x_p$ -> $\frac{p_{it}}{1-p_{it}} = e^{\beta_0(t) + \beta_1 x_1 + \cdots + \beta_p x_p}$, where $p_{it}/(1-p_{it})$ is the odds of experiencing ILI, $\beta_0(t)$ is the intercept which depends on time t, which in case of logistic regression are one-day intervals, $\beta_1 \dots \beta_p$ is a vector of unknown regression parameters and $x_1 \dots x_{pn}$ is a vector of the covariates. The main difference between the two models, is that the Cox model predicts a probability (time is discrete, so the hazard reduces to a conditional probability) and the logistic model predicts odds. Influenza-like illness is an event that occurs rarely in the Grote Griepmeting. If the event occurs rarely, we have $\frac{p_{it}}{1-p_{it}} \approx p_{it,}$ since p_{it} will be close to zero. The odds that is predicted in the logistic model is now very close to the chance predicted in the Cox model, which makes that the models resemble each other very much.

The disadvantage of the logistic regression is that the correlation between repeated measurements in an individual is not modelled here.

5.2.1 Generalized Estimating Equations

The next idea that came to my mind is the marginal modeling approach[17]. The marginal modeling approach is a way to extend generalized linear models to handle the repeated measurements, as are present in the Grote Griepmeting data. In marginal modeling no assumptions about the distribution of the observations are required in general. The specification of a marginal model contains three steps. The first step is specifying a model for the mean response. The mean response μ_{it} of the outcome y_{it} is modeled in a linear way by choosing a known link function and assuming that $g(\mu_{it}) = \boldsymbol{x_{it}\beta}$. Since the outcome data of this thesis are binary, the model used here will be the logistic model and the link function used will be the logit link, as were both described in section 5.2. The second step is to specify a model for the conditional variance of y_{it} given the covariates, which will be done in the form of: $Var(y_{it}) = \phi v(\mu_{it})$, where $v(\mu_{it})$ is a known variance function and ϕ is a scale parameter that is known or needs to be estimated. As y_{it} is Bernouilli, the variance function is:

$$Var(y_{it}) = \mu_{it}(1-\mu_{it})$$

As can be seen, the first two steps of the marginal model are nearly equal to those of the generalized linear model. The difference being that no distributional assumptions need to be made and the variance function is not conceived as the true variance for the marginal model, but as the so called 'working variance'.

In the third step the within-subject association among the repeated measurements of the same individual is specified. In the marginal model it is only assumed that the relation between the mean response and the covariates is modelled correctly. The above mentioned model for the variance and the model for the correlation matrix R ($Corr(y_{it})$) are not assumed to be necessarily correct. That is why they are called the working variance model and the working correlation model. Together these two form the working covariance matrix: $V_i = A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}$, where A is a diagonal matrix with $Var(y_{it}) = \mu_{it}(1-\mu_{it})$ along the diagonal and R is the correlation matrix of y_{it}. It is possible that the working correlation depends on unknown parameters. In this case there can be chosen from several popular correlation structures, for instance the independence, exchangeable, AR(1) or unstructured. I have chosen here to use an exchangeable working correlation structure. The third step is what distinguishes the marginal model for repeated measurements from the ordinary generalized linear model.

Maximum likelihood requires full specification of the joint distribution of $y_{it.}$ Since this is not done for the marginal model an alternative approach is needed to find estimates for the regression coefficients for the data. This approach will be the generalized estimating equations (GEE). The GEE have the following form:

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T V_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) = 0,$$

where V_i is the working covariance matrix, $\partial\mu_i/\partial_{\boldsymbol{\beta}}$ is the derivative matrix containing the derivative of $\boldsymbol{\mu_i}$ with respect to the components of $\boldsymbol{\beta}$, $\boldsymbol{\mu_i}$ is the vector of mean responses for a subject and $\boldsymbol{y_i}$ is the vector of observed responses for a subject. These estimating equations are the direct analogue of the estimating equations for the ordinary generalized linear model. One can prove that $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, which means that $\widehat{\boldsymbol{\beta}}$ is with high probability close to the population regression parameters $\boldsymbol{\beta}$ in large samples. For the marginal model it is only required that the mean response is correctly specified and $\widehat{\boldsymbol{\beta}}$ is a consistent estimator whether or not the within-subject association is specified correctly. The sandwich estimator can be used to correctly estimate the standard errors. It is useful to choose the within-subject association structure as close as possible to the real structure, because this yields more efficiency.

As any regression model, this approach is not resistant to high collinearity due to the high correlation between the covariates in the model. Hence, the model can have inaccurate estimation and prediction due to collinearity. Therefore a penalty should be added to the model, for instance a lasso penalty. The difficulty for adding a penalty to the GEE model is that the GEE model lacks a joint likelihood, which is required for a classical approach to penalty models. Through penalized generalized estimating equations[34] the penalty term will be added to the GEE. The next function will add the bridge estimator to the GEE function:

$$-\sum_{i=1}^{n} D_i^T V_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) + \boldsymbol{\lambda} d(\boldsymbol{\beta}, \gamma) = 0,$$

where $\boldsymbol{\lambda}$ is the tuning parameter and $d(\boldsymbol{\beta},\gamma) = \gamma|\boldsymbol{\beta}|^{\gamma-1} sign(\boldsymbol{\beta})$, is the partial derivative of the penalty function with respect to $\boldsymbol{\beta}$.

The bridge penalty is reduced to the lasso estimator when $\gamma = 1$. $\boldsymbol{\lambda}$ is the tuning parameter of the lasso penalty. Due to the lack of joint likelihood in the GEE model, it is impossible to use generalized cross validation for the estimation of the tuning parameter. A quasi generalized cross validation method was developed for selecting the tuning parameter.
Unfortunately, to my knowledge no software is yet developed in R, to use the penalized estimating equations, so in this case it is not possible to use this method for analysis. Taken a closer look to this approach I realized that it might not have been as advantageous as I initially thought. The main attraction of the GEE approach is that it could make a population-averaged estimate of the coefficients and model the within-subject association separately, so standard errors could be properly estimated. Since the data contains covariates which are highly correlated a penalty (in this case the lasso) is required, because otherwise the $\boldsymbol{\beta}$-coefficients will not be properly estimated. Because of the shrinkage that is introduced by the lasso penalty, the standard errors estimated by the GEE will not be valid for these estimates. So the advantage of the GEE of estimating proper standard errors disappears when using the lasso penalty.

5.2.2 Generalized linear mixed models

The generalized linear mixed model (GLMM)[17] is, contrary to the generalized estimating equations, a subject-specific model. This means that the coefficients of the covariates are the changes in the mean response of a specific subject. Since the goal of this thesis was to make a population-averaged estimate, the GLMM was my choice after the GEE approach was dismissed. There is a way to transform the subject-specific coefficients into population-averaged estimates. First more about the generalized linear mixed models.

The GLMM is an extension of the generalized linear models to longitudinal data, by using variables that can vary between individuals. As was mentioned in the section 5.1, the data are binary, so the distribution of y_{it} is Bernouilli. The logistic regression model is again used for the GLMM. A known link function, in case of the logistic model, the logit link is used to transform the predicted outcomes from the unrestricted linear predictor scale into a range between 0 and 1. The mean y_{it} depends upon the fixed and random covariates in the following way:

$$g\{E(y_{it}|\boldsymbol{b_i})\} = \eta_{it} = \boldsymbol{x_{it}\beta} + \boldsymbol{z_{it}b_i},$$

where $\boldsymbol{x_{it}}$ is a vector containing the fixed covariates, $\boldsymbol{z_{it}}$ is a vector containing q covariates, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients for the fixed covariates and $\boldsymbol{b_i}$ is a subject specific vector of random regression parameters, called the random effects. Given $\boldsymbol{b_i}$**,** the y_{it}'s are assumed to be independent within a subject.
For the random effects some probability distribution is assumed. This can be any kind of multivariate distribution, but for computational purposes it is common to use a multivariate normal distribution, with zero mean and a $q * q$ covariance matrix G.
The variance function for the logistic model given the random effects is specified as $Var(y_{it}|\boldsymbol{b_i}) = E(y_{it}|\boldsymbol{b_i})\{1 - E(y_{it}|\boldsymbol{b_i})\}$.

In generalized linear mixed effect models the joint distributions are fully specified (contrary to the GEE) so the maximum likelihood approach can be used. The following likelihood equation should be maximized to find the estimates for the regression coefficients:

$$L(\boldsymbol{\beta}, G) = \prod_{i=1}^{n} \int f(\boldsymbol{y_i}|\boldsymbol{b_i}) f(\boldsymbol{b_i}) d\boldsymbol{b_i},$$

The GLMM also has trouble with the collinearity caused by the highly correlated covariates. The lasso penalty looks like the best solution in this model. Again the difficulty in this model appears when the penalty is added for the model.
GLMM with a lasso penalty is a fairly new way of analyzing data, so it was quite hard to find an article in which this was done. The article that was found on the subject was by Groll and Tutz (2014)[35]. The analysis in the article is meant for high-dimensional data, which the data of the Grote Griepmeting are not, but since the problem of collinearity is similar we tried to apply the method to the data. The method in the article of Groll and Tutz (2014) was based on an article of Goeman (2010)[36]. This method will be described first. Goeman developed an algorithm using a combination of gradient ascent optimization and the Newton-Raphson algorithm. The gradient ascent optimization method has great computational simplicity, but needs a lot of steps until convergence, which makes that it takes great computation time. The Newton-Raphson algorithm converges faster compared to the gradient ascent optimization method. Unfortunately the Newton-Raphson algorithm can only be applied when the target function is concave and twice differentiable, which is not the case, because of the lasso penalty. The gradient ascent method will be used for starters and when this comes near the optimum of the function, the target function will be concave and twice differentiable and the Newton-Raphson algorithm will take over.
Goeman developed his method for generalized linear models, but did not make an application for generalized linear mixed models. For this reason the method of Groll and Tutz was used. Groll and Tutz based their method on that of Goeman, but used a combination of gradient ascent optimization and the Fisher scoring algorithm. The algorithm is implemented in the R-package *glmmLasso*[37]. The first analysis that was tried with this package was done on a small dataset, containing 5000

observations and 7 variables (instead of 2,137,304 observations and 44 variables). The application has run for 4 days and still did not converge. It was then decided to stop the computations. This method already takes a great deal of computation time for a small part of the data set, so will probably take great computation time for the whole dataset.

6. Analysis

For the final analysis it was decided to use a logistic regression model. B-splines will be used to account for the possible non-linearity of the effects of the covariates in the model. The B-splines will contain 5 inner knots and 2 boundary knots and no intercept will be included. The positions of the inner knots will be chosen in a way that each interval contains an equal amount of observations. Because there is high collinearity between the weather variables in the dataset, it will be necessary to use a penalty, for which in this case the lasso was chosen . For each weather variable, 8 columns of B-spline values were created in the design matrix, which are added to the model. We would want to include the complete weather variable, so all 8 B-spline parts, or none of the B-spline parts at all. To obtain this the group lasso[29] will be used. The group lasso will have the following form:

$$\hat{\beta}(\lambda) = -l(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} s(df_g) || \boldsymbol{\beta_g} ||_2$$

where $l(\boldsymbol{\beta})$ is the log-likelihood function:

$$l(\boldsymbol{\beta}, y) = \sum_{i,t} y_{it} log(p_{it}) + (1 - y_{it}) \log(1 - p_{it}),$$

Here $\boldsymbol{\lambda}$ is the tuning parameter that controls the amount of penalization, $\boldsymbol{\beta}$ is a vector containing the unknown regression coefficients and $s(df_g)$ is a function used to rescale the penalty with respect to the dimensionality of the parameter vector $\boldsymbol{\beta_g}$.

The amount of data for the analysis is very large, especially since the splines introduce 352 (44*8) covariates to the dataset, instead of the 44 original weather variables available. The number of records in the dataset is 2,137,304. Combined with the 352 covariates, the size of the dataset becomes so big that computation time becomes bothersome. However, the number of events is 3,751, which is very small relative to the number of records. An old trick to reduce the computation time is found in the medical field of case-control studies. Instead of taking all the records without an event, we take a sample of them. It has been proven that the parameters estimated by logistic regression on the reduced dataset are the same as for the original dataset, except for the value of the intercept[38]. The value of the intercept can be adjusted by a simple formula, shown later, depending on the ratio cases/controls and the prevalence of the event in the original dataset. In medical statistics a rule of thumb is that a ratio of 4 controls to 1 case is used. The power of the study will not greatly improve if more controls are added. In the medical field it takes time to gather control subjects. This is not the case in this thesis, since we have a great amount of data. The trouble of this study is that we have too much data. We could have chosen any amount of controls, but we decided to use the medical field as a guideline and use the ratio of 4 controls to 1 case, although more controls are available.

It might have been noticed, that in the final analysis the correlation introduced by repeated measurements has not been accounted for. If in the above likelihood all records after the day of the first ILI episode of a subject were ignored, the likelihood would be correct. The data contains 13,784 individuals after cleaning up the data. 2,480 of the subjects experienced influenza like illness once and 563 subjects experienced ILI more than once, so a total of 3,043 subject experienced ILI. So 18.5% of the 3,043 subject that experienced ILI, have experienced ILI more than once. Since this is only a small percentage it was decided to keep the multiple episodes of ILI in the analysis, since I thought the dependence could be safely ignored. Moreover, the correlation would only effect the standard errors of the estimations, which would not be interpretable in any case in the analysis due to the shrinkage introduced by the lasso penalty.

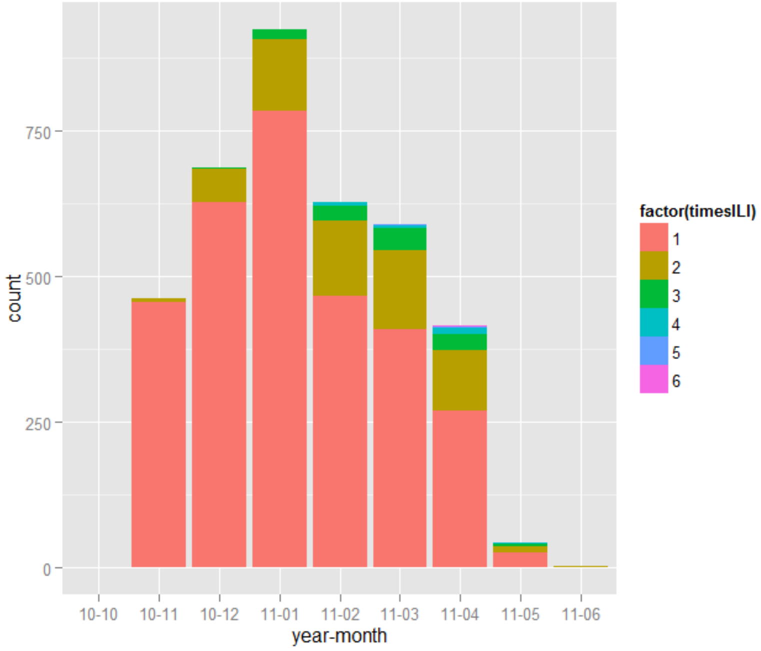
I could have introduced a covariate counting the ILI periods like the Andersen-Gill approach, but I did not do that.

The distribution of the episodes of ILI is shown in Graph 3. The exact amount of episodes of ILI can be found in Table 7.

As expected, it can be seen from the Graph that at the start of the season, people experience the first episode of ILI, while as the season progresses more and more of the ILI infections are a second episode of ILI or even more. The greatest amount of ILI is experienced in January. In Table 7 it is shown that the maximum number of times that ILI is experienced by the same individual equals 6 times, which happens twice in the dataset.

Table 7. Frequency of times ILI experienced

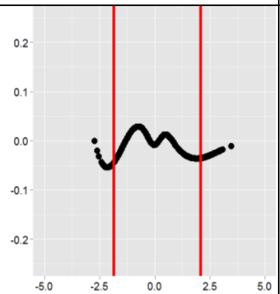
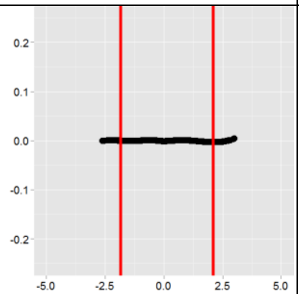
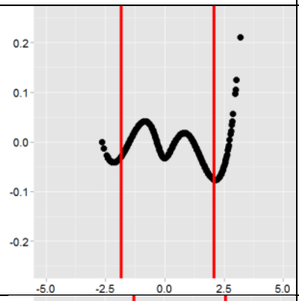
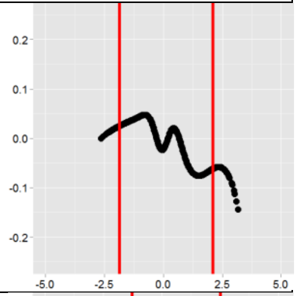
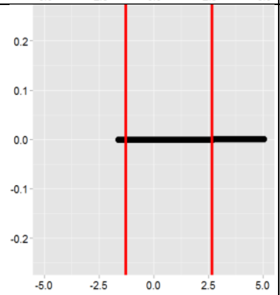
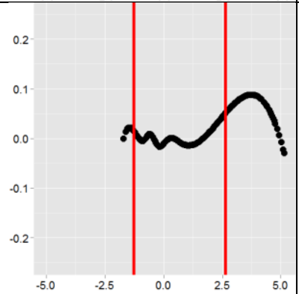
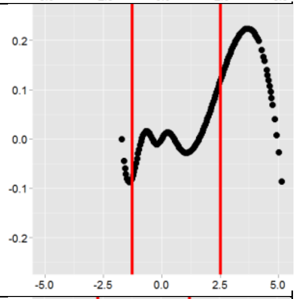
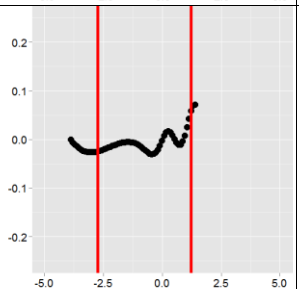
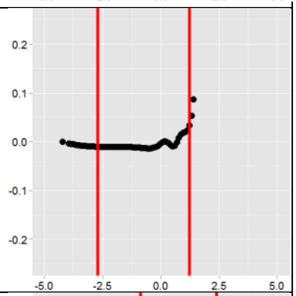
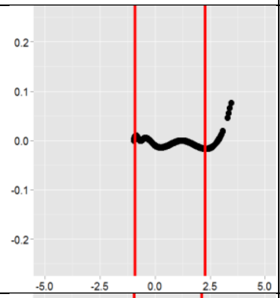
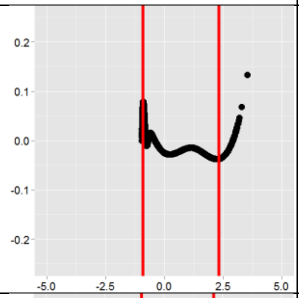
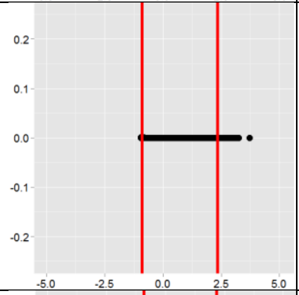
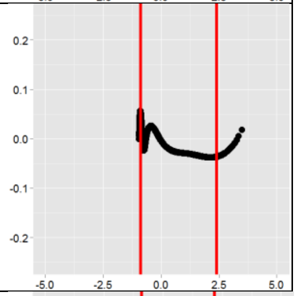
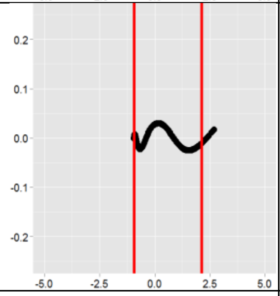
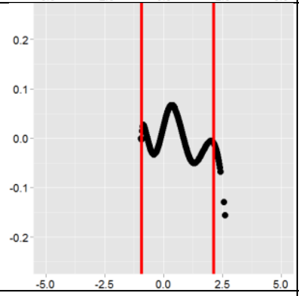
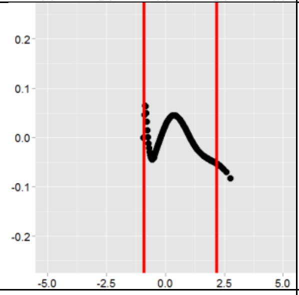
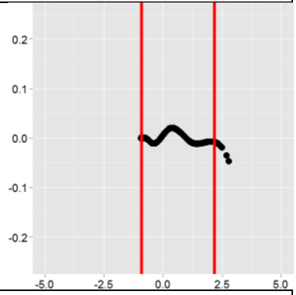
1	**2**	**3**	**4**	**5**	**6**
2480	449	90	19	3	2

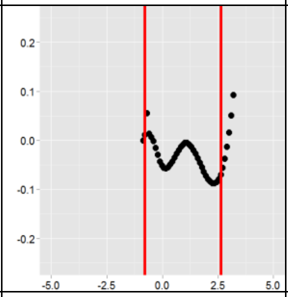
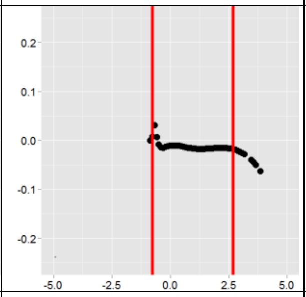
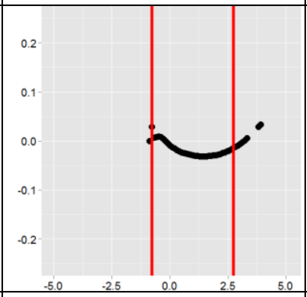
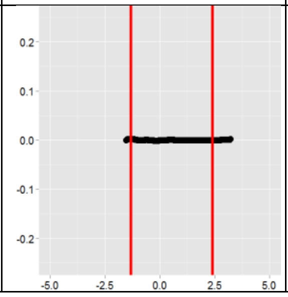
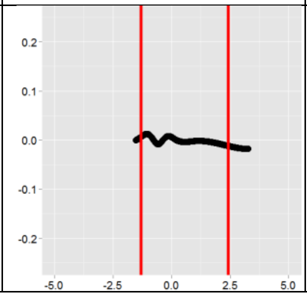
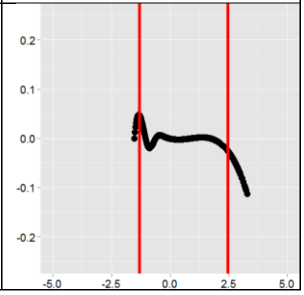


Graph 3. Frequency of ILI in different months.

In Table 8 the effects of the weather variables as modeled through the splines are plotted against the associated standardized weather variables.

Table 8. Plotting variables vs. b-spline variables

	Day 1	*Day 2*	*Day 3*	*Day 4*
Temperature				
Wind				
Humidity				
Radiation				
Sunhours				
Precipitationhours				

Sum of precip				
Evapotranspiration				
Days without sun				
Days with rain				
Tempdiff				

The rows of Table8 show the different weather variables. The columns of the Table display the different days before the day of measurement. As can be seen, not all entrances of the Table are filled. The group lasso penalty has selected variables. The blanc entrances are variables that are not selected by the lasso penalty and so the coefficients returned by the analysis for these variables are zero. If a plot would be created for these variables, it would give a straight line through zero on the x-axis. These variables are left out of the table to give a clearer representation of the variables that are selected by the group lasso.

The group lasso selected 23 from the 44 variables. 7 different weather variables were selected; temperature, wind, humidity, radiation, sun hours, evapotranspiration and temperature difference. For all these variables several different days before measurement were selected. The variables precipitation hours, sum of precipitation, days without sun and days with rain are never selected. Note that, the model on which this table is based, was made on a subset of the whole dataset.

The values of the spline displayed on the Y-axis variables are calculated in the following way:

SplineTemperatureDay1:B-splinevalueTemperatureDay1[column1]*coefficientTemperatureDay1[1] +
B-splinevalueTemperatureDay1[column2]*coefficientTemperatureDay1[2] +
B-splinevalueTemperatureDay1[column3]*coefficientTemperatureDay1[3] +
B-splinevalueTemperatureDay1[column4]*coefficientTemperatureDay1[4] +
B-splinevalueTemperatureDay1[column5]*coefficientTemperatureDay1[5] +
B-splinevalueTemperatureDay1[column6]*coefficientTemperatureDay1[6] +
B-splinevalueTemperatureDay1[column7]*coefficientTemperatureDay1[7] +
B-splinevalueTemperatureDay1[column8]*coefficientTemperatureDay1[8]

Since the coefficients are still on the log odds scale, the spline values are also on the log odds scale. In the plots in Table 8 on the x-axis the standardized weather variable is displayed. The x-axis is held constant from -5 to 5.2. On the y-axis the B-spline variable is displayed (calculation shown above). The length of the y-axis is held constant between -0.25 and 0.25, so it is possible to compare the effects the different weather variables have of experiencing ILI on the log odds scale. In each figure two vertical, red lines are displayed. Between these red lines lies 95% of the data. To the left and the right of the lines a little over 500 points are displayed, so a total of a little over 1000 points are outside the red lines. Inside the red lines over 19500 points are presented.
As can be seen from Table 8 no really large effects are shown, especially the effects between the red lines are small. In fact, all effects are smaller than 0.25, except for that of evapotranspiration variable on Day 1. For the evapotranspiration variable on Day 1, 4 data points are excluded from the graph, because otherwise the scale of all the graph should be changed on the y-axis to -0.5 to 1.5, which would make it a lot more difficult to see the effects of the other variables. Also the effects of the evapotranspiration Day 1 that were large were corresponding to extreme weather circumstances. No large effects were found for weather circumstances that were not as extreme.

The chance for a person to experience influenza-like illness under the given weather circumstances has to be calculated. The just described calculations for the variables selected by the group lasso are used. The intercept is also added to the formula, but the intercept from the analysis cannot directly be used. For the analysis it was decided to use a sample of the whole dataset. This sample contained all cases from the whole dataset and a subsample from the controls (ratio: 1 case to 4 controls). Because this way of sampling was chosen, the amount of cases will grossly be overrepresented in the subsample. This overrepresentation of the cases will affect only the coefficient of the intercept in the analysis, which will be overestimated. To find the intercept for the whole sample data, the intercept from the subsample needs to be adjusted in the following way:

$$\beta_{0(whole\ sample)} = \beta_{0(subsample)} - \log\left(\frac{n_1}{n_0}\right) + \log\left(\frac{\pi}{1-\pi}\right),$$

Where $n_1$is the size of the subsample of measurements in which an episode of ILI started, n_0 is the size of the subsample of measurements in which no episode of ILI started and π is an estimate of the prevalence of ILI, in this thesis from the whole dataset.
This gives the following formula for calculating the log odds on ILI:

Out = $\beta_{0(whole\ sample)}$ + SplineTemperatureDay1 + SplineWindDay1 + SplineRadiationDay1 + SplineSunhoursDay1 + SplineEvapotranspirationDay1 + SplineTemperatureDifferenceDay1 + SplineTemperatureDay2 + SplineRadiationDay2 + SplineSunhoursDay2 + SplineEvapotranspirationDay2 + SplineTemperatureDay3 + SplineWindDay3 + SplineHumidityDay3 + SplineRadiationDay3 + SplineSunhoursDay3 + SplineEvapotranspirationDay3 + SplineTemperatureDifferenceDay3 + SplineTempDay4 + SplineWindDay4 + SplineHumidityDay4 + SplineRadiationDay4 + SplineSunhoursDay4 + SplineTemperatureDifferenceDay4

The out-variable calculated above is still on the log odds scale and will have to be transformed back, so it ranges between 0 and 1 again and is an estimate of the chance a subject experiences ILI given the weather circumstances. This will be done in the following way:

$$P(event = 1) = \frac{1}{1 + e^{-out}}$$

Roc-curve

The Receiver Operating Characteristic (ROC) curve is used to assess the discriminatory performance of the model[39]. The discriminatory performance of the model is good if the covariates discriminate between cases and controls.

From the logistic regression analysis predicted chances are calculated as above describing the chance of the measurement being a case. The cut-off point for the predicted values has to be chosen in such a way, that the trade-off between sensitivity and specificity is as high as possible. Sensitivity being the true positive rate (TPR = positives/cases) and specificity being the true negative rate (TNR = negatives/noncases).

In the ROC-curve the trade-off between sensitivity and 1-specificity is plotted for a range of cut-off points. 1 – specificity gives the false positive rate, which reflects the noncases that are falsely predicted to be cases. We would like 1-specificity to be close to zero and the sensitivity to be close to 1. The larger the area under the ROC-curve (AUC)[39], the better the discriminatory performance of the model is.

The AUC can be found by calculating the proportion of all possible case/noncase pairs, where the predicted value of the case is higher than the predicted value of the noncase. Because there will be ties for the predicted values in the pairs, a so called fair equation is developed, called the c-statistic. In this fair equation one half of the ties is added to the proportion where the cases have a higher prediction and the other half to the proportion where the noncase has a higher prediction than the case. The c-statistic[38] is given by the following formula:

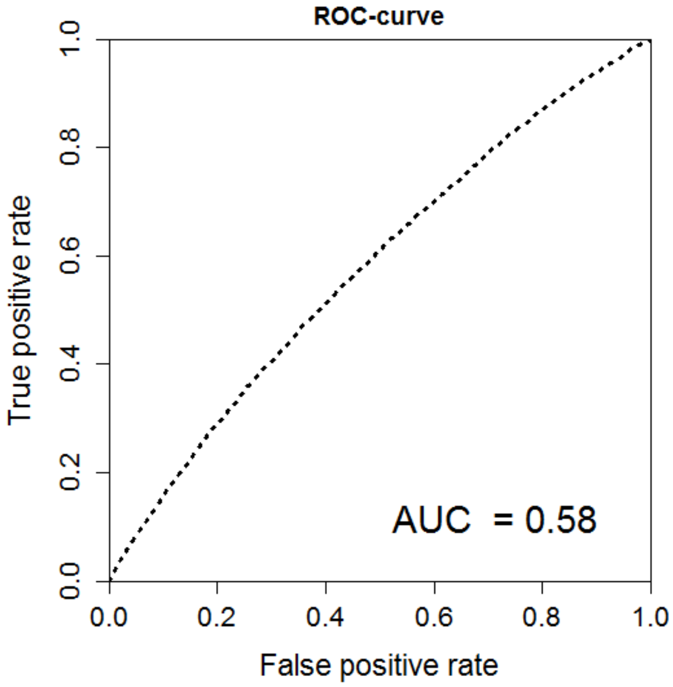
$$c = \frac{w + 1/2z}{n_p} = AUC$$

where n_p is the number of all possible pairs, w is the number of pairs where the prediction of the case is higher than the prediction for the noncase, and z is the number of pairs where the prediction is a tie. Below it is shown how the values of AUC are graded[39].

Grading AUC-values

- *0.90 – 1.00 = excellent discrimination*
- *0.80 – 0.90 = good discrimination*
- *0.70 – 0.80 = fair discrimination*
- *0.60 – 0.70 = poor discrimination*
- *0.50 – 0.60 = failed discrimination*

In Graph 4 the ROC-curve with the AUC-value of the analysis is shown. From the graph we can see the ROC-curve is almost a diagonal. The AUC of the model is 0.58, which indicates that the model fails to discriminate cases from noncases.



Graph 4. ROC-curve with calculated AUC.

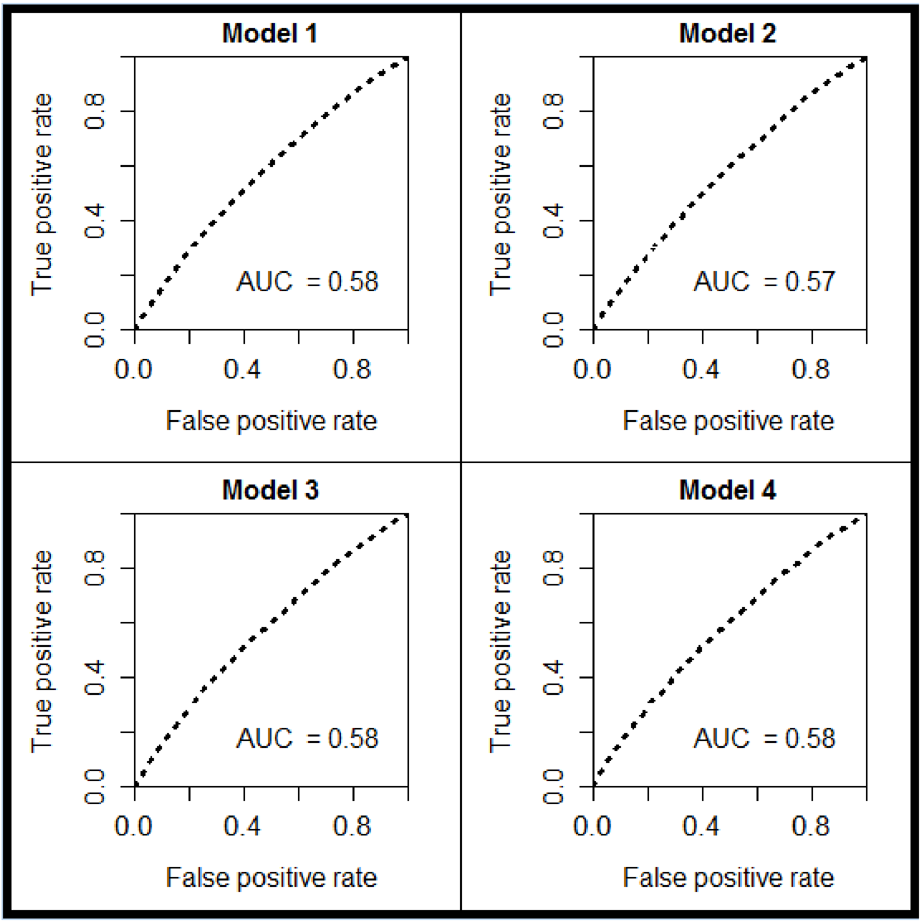
Since it is known that a pitfall of the lasso method is, that it can select quite different variables when the dataset is slightly changed, I chose to create 3 control groups, so it is possible to see if the same weather variables are selected and this is not the case, how much the selections differ.
The 3 control groups are created in the same way the first sample was created. Table 9 shows the variables selected in the first sample, which is named 'Model 1' and is the one that was used for the analysis shown above. The 3 control samples from the data are named 'Model 2', 'Model 3' and 'Model 4'. As can be seen in the table, Model 1 is the most elaborate model, in which 23 variables were selected. In Model 2,3 and 4 the same 7 variables that were mentioned before were selected, although for some variables different days previously to the measurement were selected. So the models selected are not exactly the same, but they are pretty similar, since the same weather variables were selected.

Table 9: Variables selected for control groups

Model1	***Model2***	***Model3***	***Model4***
Temperature Day1	Temperature Day1	Temperature Day1	Temperature Day1
Wind Day1			
		Humidity Day1	
Radiation Day1			
Sunhours Day1	Sunhours Day1	Sunhours Day1	Sunhours Day1
Evaporation Day1	Evaporation Day1	Evaporation Day1	Evaporation Day1
Tempdiff Day1			
Temperature Day2	Temperature Day2	Temperature Day2	
Radiation Day2	Radiation Day2	Radiation Day2	Radiation Day2
Sunhours Day2		Sunhours Day2	Sunhours Day2
Evaporation Day2	Evaporation Day2		Evaporation Day2
Temperature Day3	Temperature Day3	Temperature Day3	Temperature Day3
Wind Day3		Wind Day3	Wind Day3
Humidity Day3	Humidity Day3	Humidity Day3	Humidity Day3
Radiation Day3			Radiation Day3
Sunhours Day3	Sunhours Day3	Sunhours Day3	Sunhours Day3

Evaporation Day3	Evaporation Day3	Evaporation Day3	
Tempdiff Day3			Tempdiff Day3
Temperature Day4	Temperature Day4	Temperature Day4	Temperature Day4
Wind Day4		Wind Day4	Wind Day4
Humidity Day4	Humidity Day4	Humidity Day4	Humidity Day4
Radiation Day4	Radiation Day4	Radiation Day4	Radiation Day4
Sunhours Day4		Sunhours Day4	Sunhours Day4
		Evaporation Day4	Evaporation Day4
Tempdiff Day4		Tempdiff Day4	Tempdiff Day4

In Graph 5 the ROC-curves are shown with the calculated AUC's. As can be seen there is barely any difference in the AUC for the different models. None of the models has a sufficient AUC to make it a satisfactory model. So although the lasso-method selects different variables for different samples of the data, the models are evenly efficient.



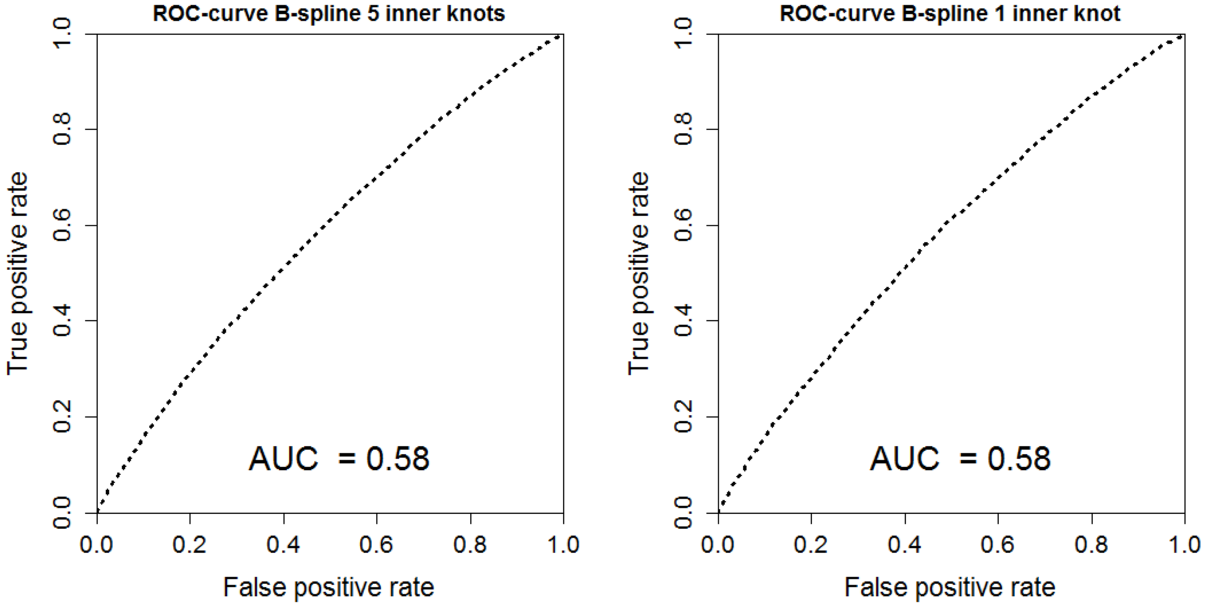
Graph 5. ROC-curves and AUC for control groups

Using B-splines with 1 inner knot:
From Table 8, it seems that overfitting occurs in most of the variables for the model with 5 knots for the B-splines. This can be concluded from the excessive amount of peaks in the graphs. For the B-splines with 5 inners knots 8 parameters are used. We wanted to reduce the amount of parameters to 4 to decrease the overfitting, so B-splines with 1 inner knot were used. For the model with 1 inner knot the lasso method selected 39 variables. Table 10 displays the variables selected for the 5 and 1 inner knot models.

Table 10. Comparing models	
5 inner knot model	**1 inner knot model**
Temperature Day 1	Temperature Day 1
Wind Day 1	Wind Day 1
	Humidity Day 1
Radiation Day 1	Radiation Day 1
Sunhours Day 1	Sunhours Day 1
	Precipitationhours Day 1
	Sum of precipitation Day 1
Evaporation transpiration Day 1	Evaporation transpiration Day 1
	Days without sun Day 1
	Days with precipitation Day 1
Temperature difference Day 1	Temperature difference Day 1
Temperature Day 2	Temperature Day 2
	Humidity Day 2
Radiation Day 2	
Sunhours Day 2	Sunhours Day 2
	Sum of precipitation Day 2
Evaporation transpiration Day 2	Evaporation transpiration Day 2
	Days without sun Day 2
	Day with precipitation Day 2
	Temperature difference Day 2
Temperature Day 3	
Wind Day 3	Wind Day 3
Humidity Day 3	Humidity Day 3
Radiation Day 3	
Sunhours Day 3	Sunhours Day 3
	Precipitationhours Day 3
	Sum of precipitation Day 3
Evaporation transpiration Day 3	Evaporation transpiration Day 3
	Days without sun Day 3
	Day with precipitation Day 3
Temperature difference Day 3	Temperature difference Day 3
Temperature Day 4	Temperature Day 4
Wind Day 4	Wind Day 4
Humidity Day 4	Humidity Day 4
Radiation Day 4	Radiation Day 4
Sunhours Day 4	Sunhours Day 4
	Precipitationhours Day 4
	Sum of precipitation Day 4
	Evaporation transpiration Day 4
	Days without sun Day 4
	Day with precipitation Day 4
Temperature difference Day 4	Temperature difference Day 4

A Table comparing the B-splines with 1 and 5 inner knots can be found in Appendix C.
In Table 10 can be seen that the lasso penalty has not selected the same variables for the model with B-splines with 5 inner knots and the model with 1 inner knot. For all variables that can be compared for the two models, the variables from the 1 inner knot model seem to be less affected by overfitting.

In Graph 6 the ROC-curve with the AUC calculation for the model with 5 inner knots and 1 inner knot are shown. The AUC values for the two models are equal, which would not have been expected, since the 5 inner knot model is more elaborate, so it would be expected it would discriminate better between cases and noncases. The 1 inner knot model has selected 16 more variables (23 compared to 39), but even though it has 16 more variables the 5 inner knot model uses more parameters, since the 5 inner knots model uses 23 x 8 = 184 parameters and the 1 knot model uses 39*4 = 156 parameters. The 156 parameters used are still a lot of parameters to use in the model, but it was chosen to not further simplify the model.
Next the model will be validated on another dataset. If overfitting has occurred the 1 inner knot model will perform better than the 5 inner knot model on new data.

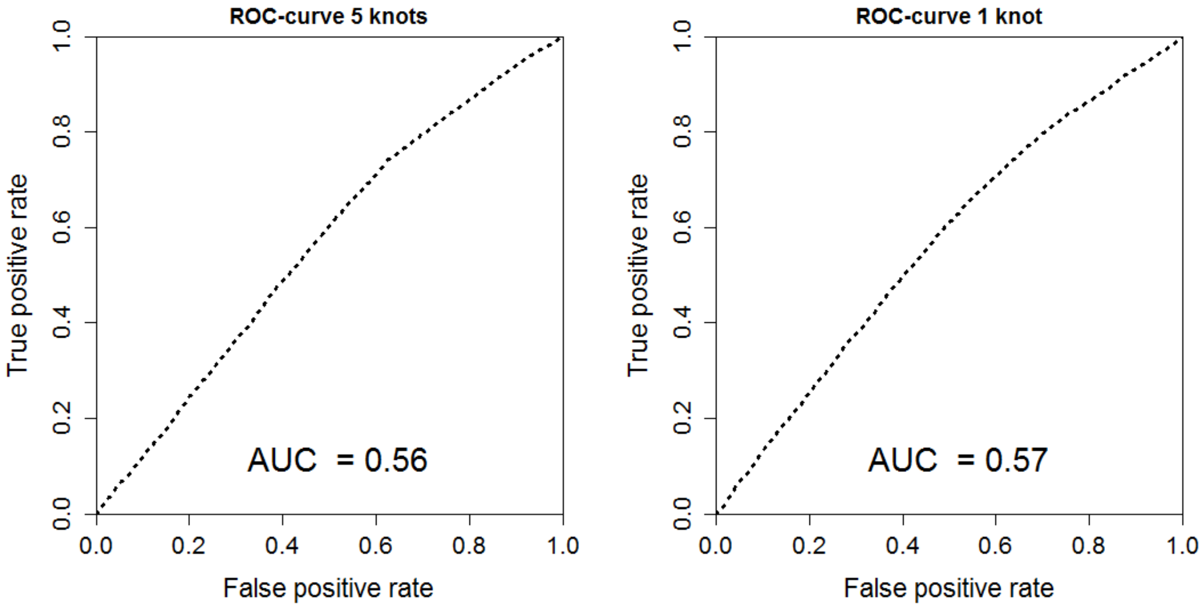


Graph 6. ROC curve and AUC for 5 and 1 inner knots B-spline model

Season 2009-2010

For validating the model several methods could have been used, like creating a training and test data set out of the original data, or using cross-validation or bootstrapping. In this thesis we will validate the model by using a new dataset. As mentioned before I had access to 8 seasons of data from the Grote Griepmeting, ranging from season 2003/2004 until season 2010/2011. The model will be validated using data of the Grote Griepmeting from season 2009/2010. The data of season 2009/2010 will be constructed in the exact same manner as the data of season 2010/2011. The data of season 2009/2010 consists of 545,969 measurements and was filled out by 20,521 subjects. This is an average of 26 measurements per subject. After cleaning up the data, it contained 508,870 measurements and 16,151 subjects. After creating the one day intervals, the data from season 2009/2010 contain 5,671 cases and 3,886,086 control measurements. The weather variables for this season can be collected from the KNMI site.
The B-splines are reconstructed in exactly the same way, so the knots are at the same values for the weather variables. The coefficients that were obtained from the analysis on season 2010/2011 are used to fit the data. This was done for the 5 inner knots model and the 1 inner knot model. From this the following ROC curves and AUC- values in Graph 7 are created.
It was expected that the 1 inner knot model would perform better than the 5 inner knot model, since some overfitting was expected for the 5 inner knot model. Graph 7 shows that the 1 inner knot performs slightly better than the 5 inner knot model with an AUC of 0.57 compared to 0.56. The AUC-values for both the 1 inner knot model and the 5 inner knots model for the validation dataset also barely differ from the models for the original dataset. The discriminatory power of the model for

the validation dataset is also low, but this was expected, since the discriminatory power of the original model was low and the model for the validation dataset will not outperform that of the model for the original data.



Graph 7. ROC curve for season 2009-2010.

7. Results and discussion

From the analysis it can be concluded that it was not possible to predict the onset of influenza-like illness using the weather variables. Both the model with 5 inner knots and 1 inner knot have poor discriminatory performance. Even though low discriminatory performance was found in the premier analysis, it was decided to use the data of season 2009/2010 to validate the model. The discriminatory performance of the model for the validation dataset is only slightly worse than the performance of the model for the original dataset.

From the results it is clear that there are no variables found that have a great influence on ILI, since the values on the Y-axis of Table 8 are very small, so no weather circumstance can be pointed out that causes the start of an episode of ILI.

There could be different reasons why the model fits poorly. The main reason is probably the fact that ILI is a contagious disease and this is not taken into account in the model. If a first case of ILI presents itself, the amount of cases will probably increase rapidly, but there has to be one subject that infects other subjects. The wave-like form this will cause cannot be modelled in the analysis as was mentioned in section 4.4, since we do not know how to model this.

In the analysis no interaction effects have been added. Several interactions would have been possible, interactions between variables on the same day, but also interactions between variables on different days might be possible. If all the possible interactions would have been added to the model, the number of variables would have been a total of more than 20,000 interactions. This would be impossible to model. It could have been decided to use only interactions of variables on the same day of the measurement or choose the interactions that would be expected to have an effect. It was decided to focus on the main effects in this thesis, but possible effects of interactions are not ruled out by this thesis.

A lot of other variables probably have great influence on the incidence of ILI. For instance the amount of contact subjects have with other people will greatly influence the chance of infection. If weather variables do have an effect on the onset of ILI, it would also depend on how long subjects

are exposed to weather circumstances, but also how well they are prepared for this weather. For instance if the weather changes for the worse, do subjects have their warm coat with them?
The model used here for the analysis was not perfect for the data. The model was unable to model the correlation for the repeated measurements. Also the computation time would be very large if the complete dataset would be used, instead of 4 controls for each case. Another question raised is how reliable the model is using the lasso method, since it possibly chooses different variables every time the model is fit on a different data sample.
Though the analysis may not be perfect, I think it would be reasonable to conclude it is not possible to find a model that can predict the incidence of influenza-like illness using weather circumstances measured previously to the day of onset, from the data of the Grote Griepmeting.
There are probably a lot more and stronger influences in play that are not taken into account in the analysis.
It is possible that the way influenza-like illness is determined here is not accurate and because of this I was unable to predict the incidence of ILI from the data. The definition chosen here was the definition we thought was most straight forward, but other definitions would have been possible.
A way to possibly improve the model would be by adding a variable indicating time to the model. This does could be different indicators of time, for instance the season or the month the measurement was taken in.

8. References

[1] Media Center, The World Health Organization (WHO). Influenza (seasonal) factsheets [EB/OL]. [2009-04-12]. http://www.who.int/mediacentre/factsheets/fs211/en/

[2] Centers for Disease Control and Prevention (CDC). http://www.cdc.gov/h1n1flu/guidelines_infection_control.htm .

[3] Stichting Nationaal Programma Grieppreventie. http://www.snpg.nl/medische_informatie/influenza.htm#4

[4] Kunst, A.E., Looman, C.W.N. and Mackenbach, J.P., (1993). Outdoor air temperature and mortality in The Netherlands: A time series analysis. American journal of epidemiology 137(3): 331-341.

[5] Reichert, T.A., Simonsen, L., Sharma, A., Pardo, S.A., Fedson, D.S. and Miller, M.A., (2004). Influenza and the winter increase in mortality in the United States, 1959-1999. American journal of epidemiology 160(5): 492–502.

[6] Costilla-Esquivel, A., Corona-Villavicencio, F., Velasco-Castanõn, J.G., Medina-de la Garza, C.E., Martínez-Villareal, R.T., Cortes-Hernández, D.E., Ramirez-Lopez, L.E., and González-Farias, G. (2014). A relationship between acute repiratory illnesses and weather. Epidemiology and infection 142(7): 1375-1383.

[7] Tang, J.W., Lai, F.Y.L., Nymadawa, P., Deng, Y.M., Ratnamohan, M., Petric, M., Loh, T.P., Tee, N.W.S., Dwyer, D.E., Barr, I.G. and Wong, F.Y.W., (2010). Comparison of the incidence of influenza in relation to climate factor during 2000-2007 in five countries. Journal of medical Virology 82(11): 1958-1965.

[8] Urashima, M., Shindo, N., Okabe, N., (2003). A seasonal model to simulate influenza oscillation in Tokyo. Japanese journal of infectious diseases 56(2): 43-47.

[9] De Grote Griepmeting. De Grote Griepmeting > het project. https://www.degrotegriepmeting.nl/nl/project/

[10] Koninklijk Nederlands Meteorologisch Instituut (KNMI). Home > klimatologie > daggegevens. http://www.knmi.nl/klimatologie/daggegevens/download.html

[11] De Grote Griepmeting. De Grote Griepmeting > het project>Het grote griepmeting team. https://www.degrotegriepmeting.nl/nl/project/ggm-team/

[12] De Grote Griepmeting. De Grote Griepmeting > het project. [4] De Grote Griepmeting. De Grote Griepmeting > het project>De GGM in de media. https://www.degrotegriepmeting.nl/nl/project/de-grote-griepmeting-de-media

[13] Marquet, R.L., Bartelds, A.I.M., Van Noort, S.P., Koppeschaar, C.E., Paget, J., Schellevis, F.G. and Van der Zee, J. (2006). Internet-based monitoring of influenza-like illness (ILI) in the general population of the Netherlands during the 2003–2004 influenza season. BMC Public Health 6: 242-249.

[14] Friesema, I.H.M., Koppeschaar, C.E., Donker, G.A., Dijkstra, F., Van Noort, S.P., Smallenburg, R., Van der Hoek, W. and Van der Sande, M.A.B. (2009). Internet-based monitoring of influenza-like illness in the general population: Experience of five influenza seasons in the Netherlands. Vaccine 27: 6353–6357.

[15] Van Noort, S.P., Águas, R., Ballesteros, S and Gomes, M.G.M. (2012). The role of weather on the relation between influenza and influenza-like illness. Journal of Theoretical Biology. 298: 131–137.

[16] Zhixiang Zhou, S., (2009). A seasonal influenza theory and mathematical model incorporating meteorological and socio-behavioral factors. Journal of Tropical Meteorology 15(1).

[17] Fitzmaurice, G.M., Laird, N.M., Ware, J.H., (2004). Applied Longitudinal Analysis. Wiley.

[18] Harrell, F.E., (2002). Regression modeling strategies. Springer series in statistics.

[19] Eilers, P., Ellers, H.C and Marx, B.D., (1996). Flexible smoothing with B-splines and penalties. Statistical Science 11(2): 89-121.

[20] De Boor, C., (1978). A practical guide to splines. New York, Springer-Verlag.

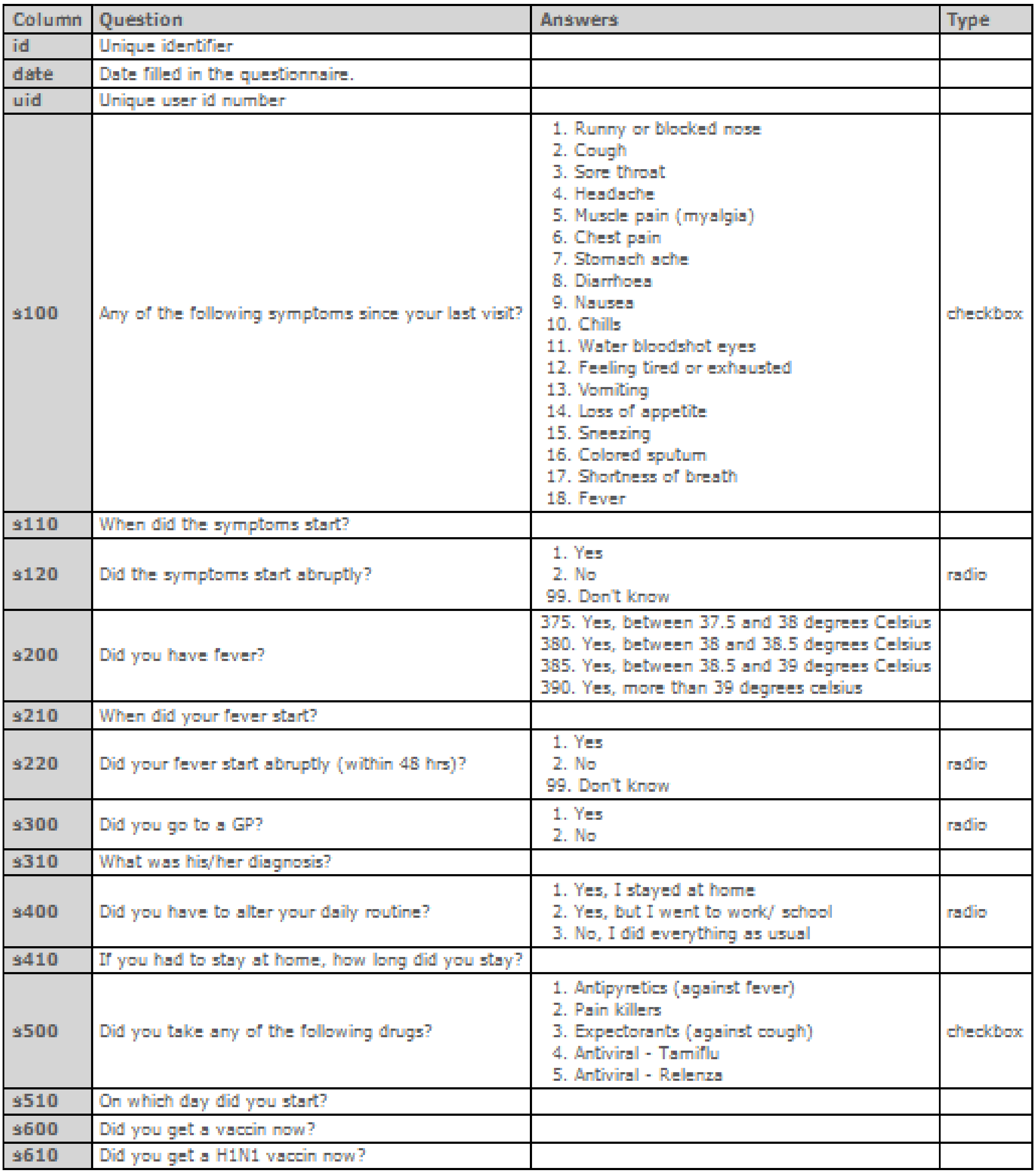
[21] Bates, D. and Venables, B., (1998). Package: Splines.

[22] Xiaonan Xue , Kim, M.Y. and Shore, R.E. , (2007). Cox regression analysis in presence of collinearity: An application to assessment of health risks associated with occupational radiation exposure. Lifetime Data Analysis 13: 333–350.

[23] Fu, W.J. (1998). Penalized regressions: The bridge versus the lasso. Journal of computational and Graphical statistics, 7(3): 397-416.
[24] Hoerl, A.E. and Kennard R.W. (1970). Ridge regression: Biased estmation for non-orthogonal problems. Technometrics, 12: 55-67.
[25] Tibshirani, R., (1995). Regression Shrinkage and Selection via the Lasso. J. R. Statist. Soc. 58(1): 267-288.
[26] Leng, C., Lin, Y., and Wahba, G., (2004). A note on the lasso and related procedures in model selection. Statistica Sinica, 16(4): 1273-1284.
[27]Bühlmann, P., and Van de Geer, S., (2011). Statistics for high-dimensional data. Springer.
[28] Goeman, J. J., (2010). L1 penalized estimation in the Cox proportional hazards model. Biometrical Journal 52(1), 70-84.
[29] Meier, L., van de Geer, S. and Bühlmann, P., (2008). The group lasso for logistic regression. J.R. Statist. Soc. B. 70(1): 53-71.
[30] Jiang, D., (2013). Concave 1-norm and 2-norm group penalty in linear and logistic regression.
[31] Klein, J.P. and Moeschberger, M.L., (2005). Survival analysis: Techniques for censored and truncated data, second edition. Springer.
[32] Therneau, T.M. (2009). Survival Analysis. http://r-forge.r-project.org.
[33] Andersen, P.K. and Gill, R.D., (1982). Cox's regression model for counting processes: a large sample study. The annals of statistics 10(4):1100-1120.
[34] Fu, W.J., (2003). Penalized Estimating Equations. Biometrics 59(1): 126-132.
[35] Groll, A. and Tutz, G. ,(2014). Variable selection for generalized linear mixed models by L_1-penalized estimation. Stat Comput 24: 137-154.
[36] Goeman, J.J., (2010). L1-penalized estimation in the Cox proportional hazards model. Biom. J. 52: 70-84.
[37] Groll, A., (2014). Variable selection for generalized linear mixed models by L1-penalized estimation. *Package 'glmmLasso'*.
[38] Prentice, RL and Pike, R., (1979). Logistic disease incidence models and case-controls studies. Biometrika 66(3): 403-411.
[39] Kleinbaum, D.G. and Klein, M., (2002). Logistic regression: A self-learning text. Third edition, Springer.

Appendix A: Weekly symptoms' questionnaire

Every participant is reminded weekly to complete a symptoms questionnaire. The final questions are only asked if the participant reported any symptoms. Not all questions were present in every country or season.

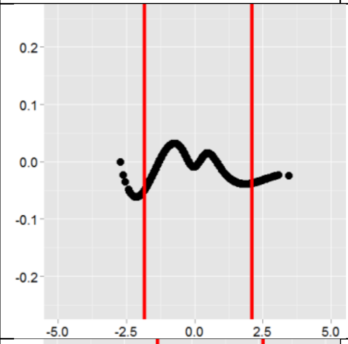
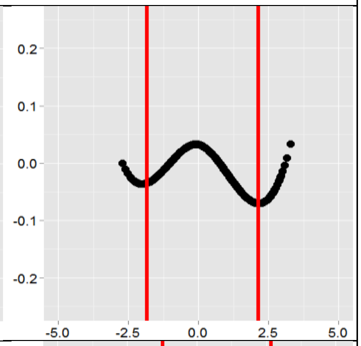
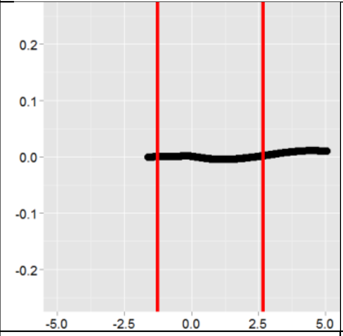
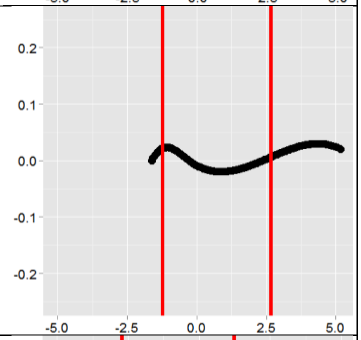
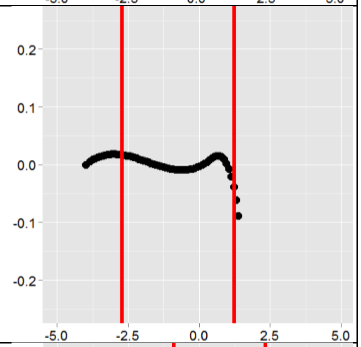
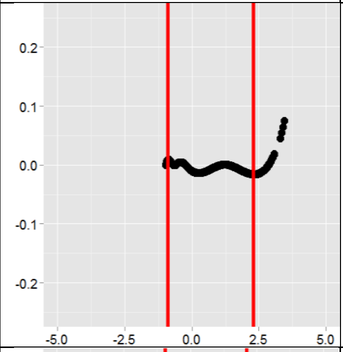
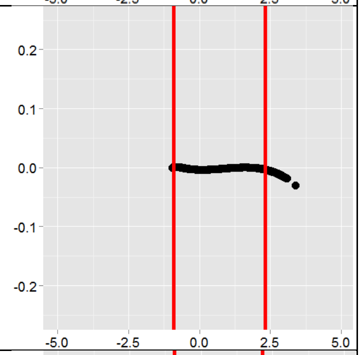
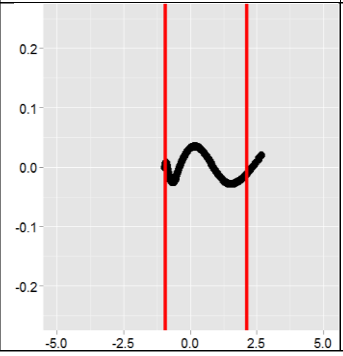
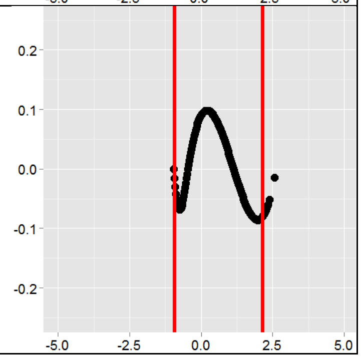


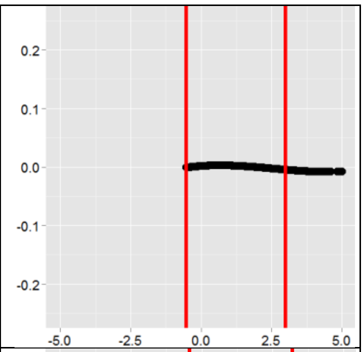
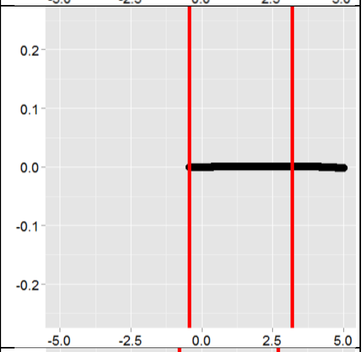
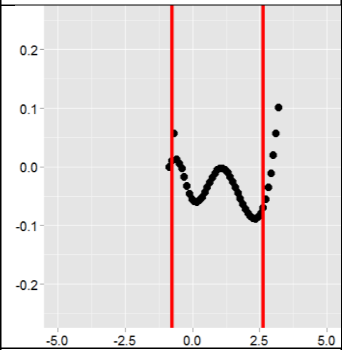
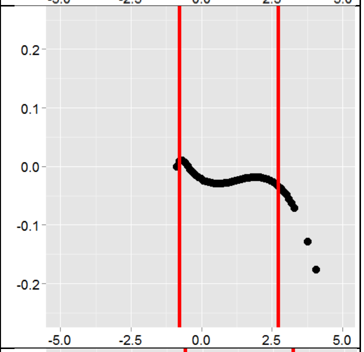
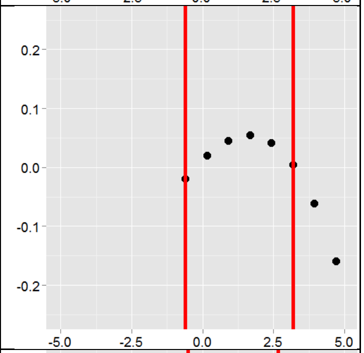
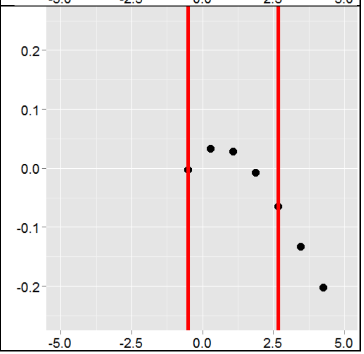
Column	Question	Answers	Type
id	Unique identifier		
date	Date filled in the questionnaire.		
uid	Unique user id number		
s100	Any of the following symptoms since your last visit?	1. Runny or blocked nose 2. Cough 3. Sore throat 4. Headache 5. Muscle pain (myalgia) 6. Chest pain 7. Stomach ache 8. Diarrhoea 9. Nausea 10. Chills 11. Water bloodshot eyes 12. Feeling tired or exhausted 13. Vomiting 14. Loss of appetite 15. Sneezing 16. Colored sputum 17. Shortness of breath 18. Fever	checkbox
s110	When did the symptoms start?		
s120	Did the symptoms start abruptly?	1. Yes 2. No 99. Don't know	radio
s200	Did you have fever?	375. Yes, between 37.5 and 38 degrees Celsius 380. Yes, between 38 and 38.5 degrees Celsius 385. Yes, between 38.5 and 39 degrees Celsius 390. Yes, more than 39 degrees celsius	
s210	When did your fever start?		
s220	Did your fever start abruptly (within 48 hrs)?	1. Yes 2. No 99. Don't know	radio
s300	Did you go to a GP?	1. Yes 2. No	radio
s310	What was his/her diagnosis?		
s400	Did you have to alter your daily routine?	1. Yes, I stayed at home 2. Yes, but I went to work/ school 3. No, I did everything as usual	radio
s410	If you had to stay at home, how long did you stay?		
s500	Did you take any of the following drugs?	1. Antipyretics (against fever) 2. Pain killers 3. Expectorants (against cough) 4. Antiviral - Tamiflu 5. Antiviral - Relenza	checkbox
s510	On which day did you start?		
s600	Did you get a vaccin now?		
s610	Did you get a H1N1 vaccin now?		

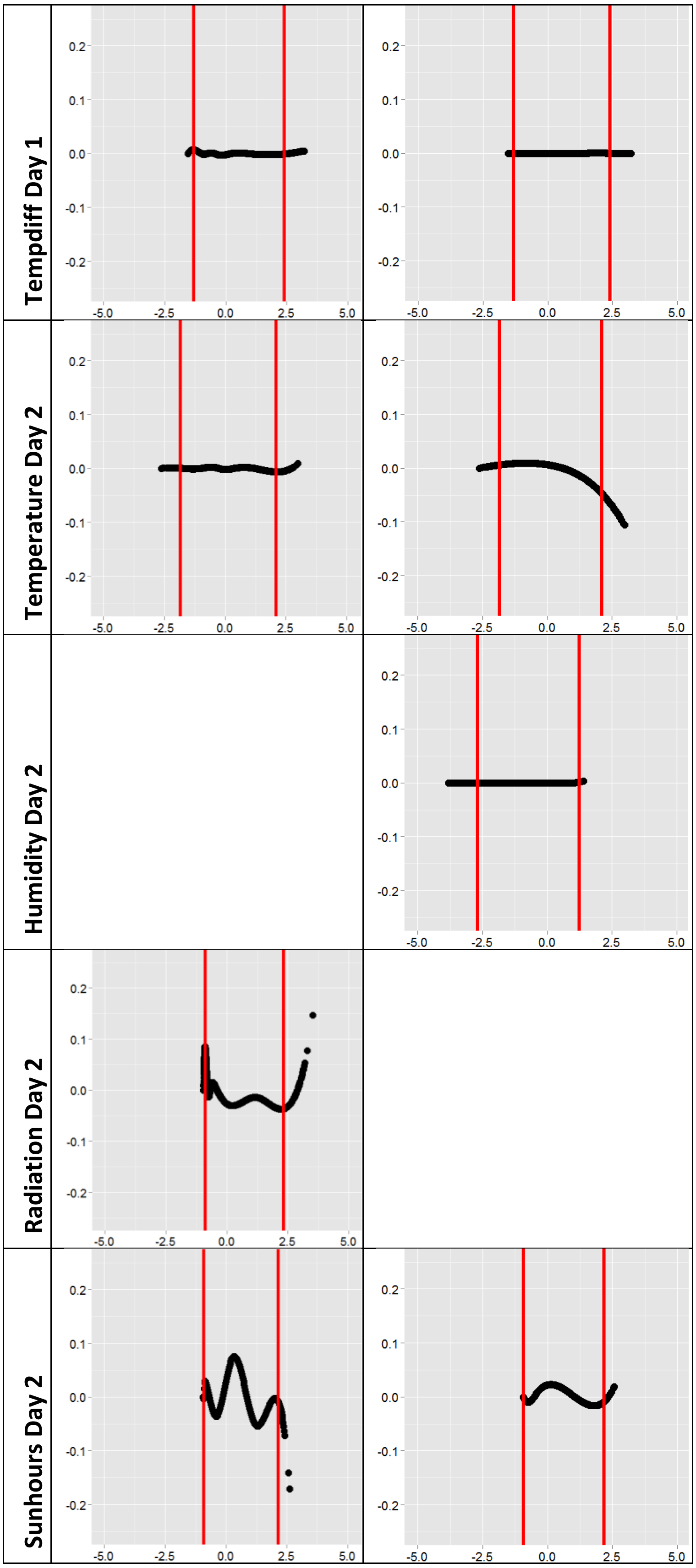
Appendix B: Weather variables

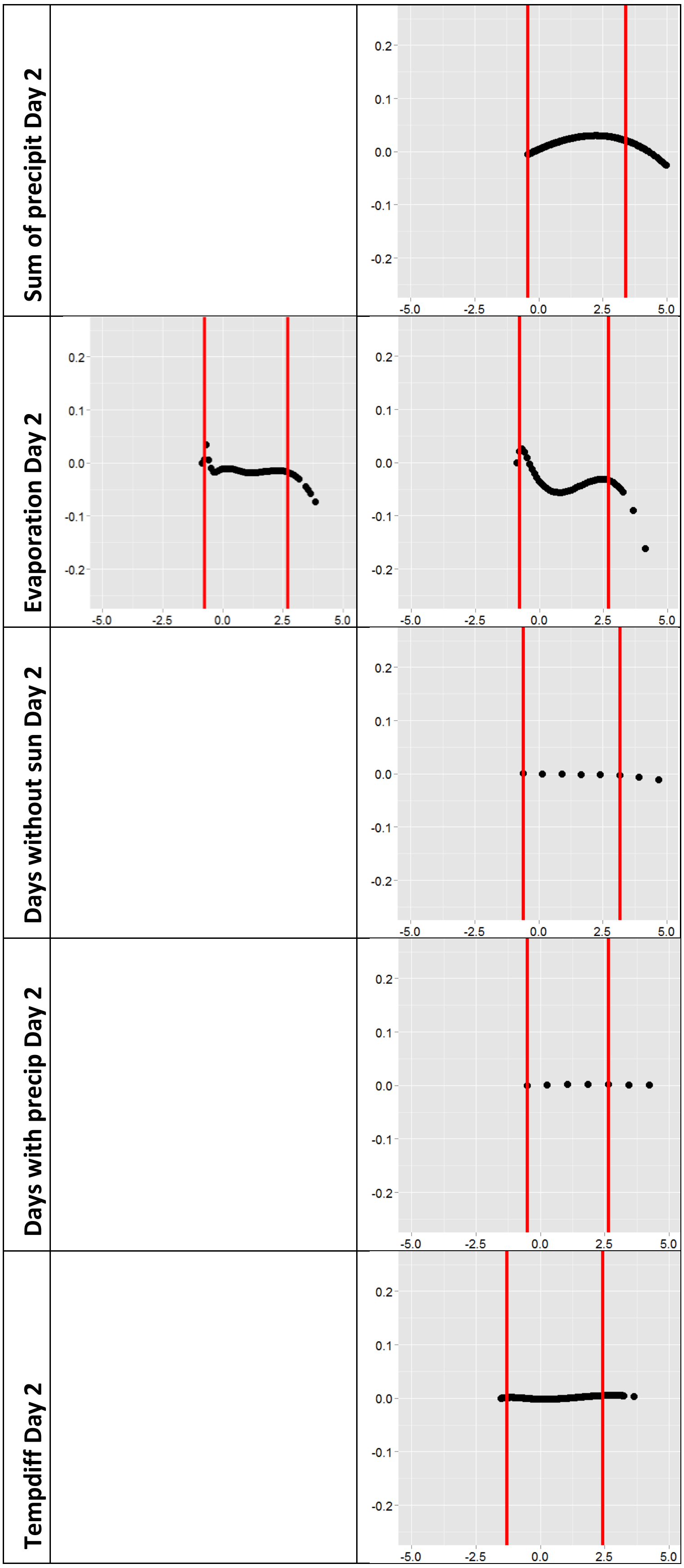
DDVEC = Vector mean wind direction in degrees (360=north, 90=east, 180=south, 270=west, 0=calm/variable)
FHVEC = Vector mean windspeed (in 0.1 m/s)
FG = Daily mean windspeed (in 0.1 m/s)
FHX = Maximum hourly mean windspeed (in 0.1 m/s)
FHXH = Hourly division in which FHX was measured
FHN = Minimum hourly mean windspeed (in 0.1 m/s)
FHNH = Hourly division in which FHN was measured
FXX = Maximum wind gust (in 0.1 m/s)
FXXH = Hourly division in which FXX was measured
TG = Daily mean temperature in (0.1 degrees Celsius)
TN = Minimum temperature (in 0.1 degrees Celsius)
TN = Hourly division in which TN was measured
TX = Maximum temperature (in 0.1 degrees Celsius)
TXH = Hourly division in which TX was measured
T10N = Minimum temperature at 10 cm above surface (in 0.1 degrees Celsius)
T10NH = 6-hourly division in which T10N was measured; 6=0-6 UT, 12=6-12 UT, 18=12-18 UT, 24=18-24 UT
SQ = Sunshine duration (in 0.1 hour) calculated from global radiation (-1 for <0.05 hour)
SP = Percentage of maximum potential sunshine duration
Q = Global radiation (in J/cm2)
DR = Precipitation duration (in 0.1 hour)
RH = Daily precipitation amount (in 0.1 mm) (-1 for <0.05 mm)
RHX = Maximum hourly precipitation amount (in 0.1 mm) (-1 for <0.05 mm)
RHXH = Hourly division in which RHX was measured
PG = Daily mean sea level pressure (in 0.1 hPa) calculated from 24 hourly values
PX = Maximum hourly sea level pressure (in 0.1 hPa)
PXH = Hourly division in which PX was measured
PN = Minimum hourly sea level pressure (in 0.1 hPa)
PNH = Hourly division in which PN was measured
VVN = Minimum visibility; 0: <100 m, 1:100-200 m, 2:200-300 m,..., 49:4900-5000 m, 50:5-6 km, 56:6-7 km, 57:7-8 km,..., 79:29-30 km, 80:30-35 km, 81:35-40 km,..., 89: >70 km)
VVNH = Hourly division in which VVN was measured
VVX = Maximum visibility; 0: <100 m, 1:100-200 m, 2:200-300 m,..., 49:4900-5000 m, 50:5-6 km, 56:6-7 km, 57:7-8 km,..., 79:29-30 km, 80:30-35 km, 81:35-40 km,..., 89: >70 km)
VVXH = Hourly division in which VVX was measured
NG = Mean daily cloud cover (in octants, 9=sky invisible)
UG = Daily mean relative atmospheric humidity (in percents)
UX = Maximum relative atmospheric humidity (in percents)
UXH = Hourly division in which UX was measured
UN = Minimum relative atmospheric humidity (in percents)
UNH = Hourly division in which UN was measured
EV24 = Potential evapotranspiration (Makkink) (in 0.1 mm)

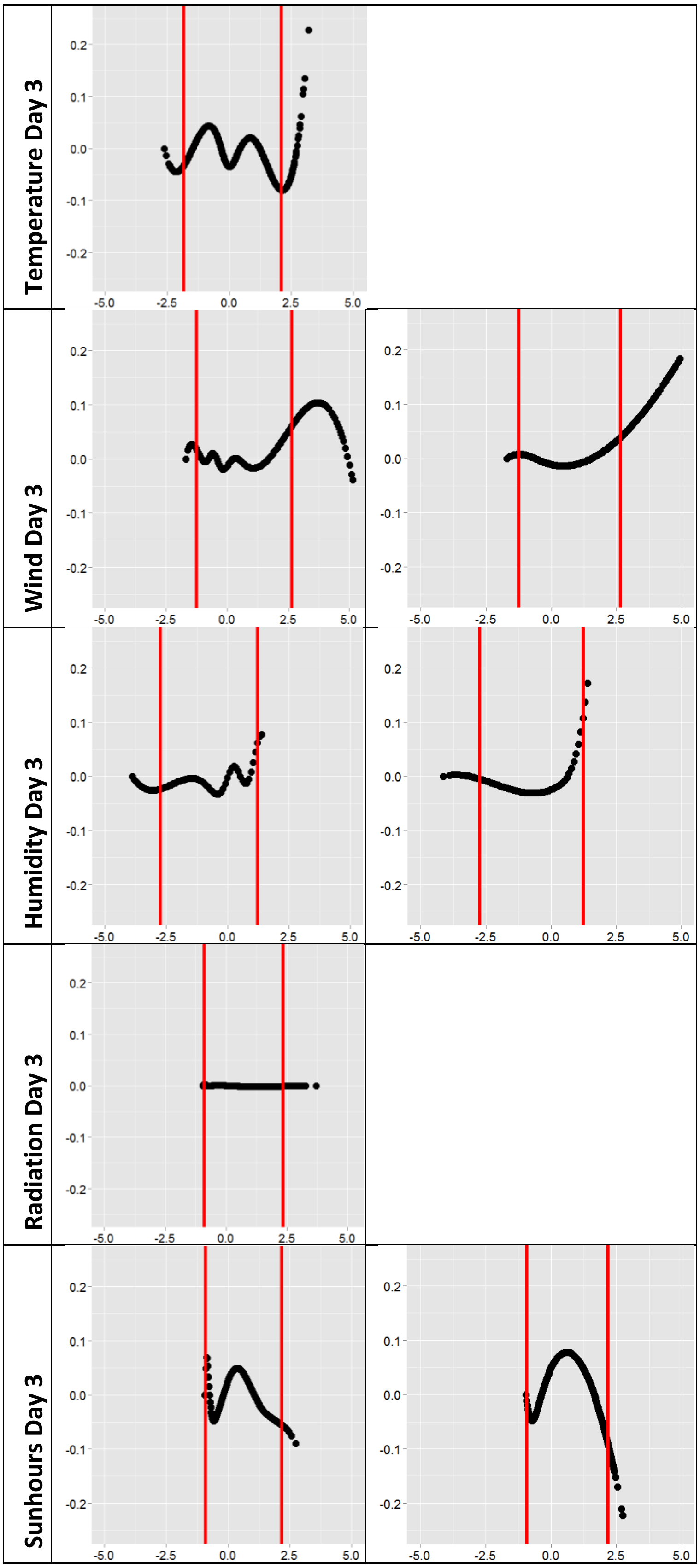
Appendix C. Comparing models

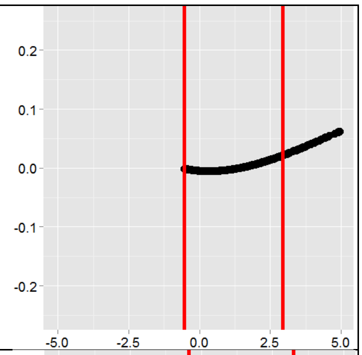
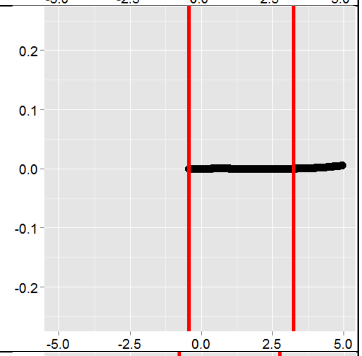
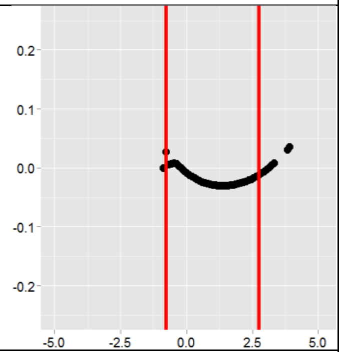
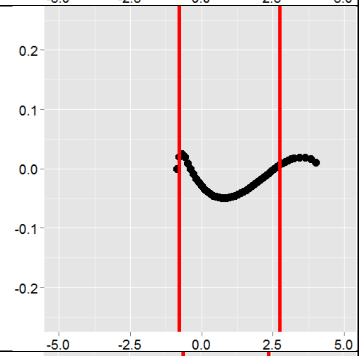
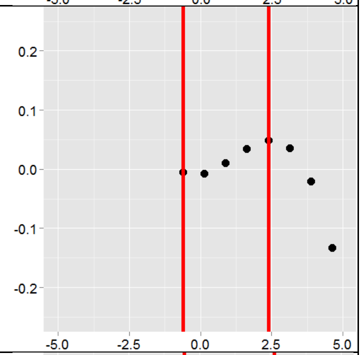
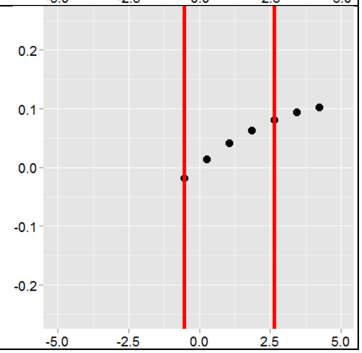
	5 inner knots	*1 inner knot*
Temperature Day 1	 	
Wind Day 1	 	
Humidity Day 1		
Radiation Day 1	 	
Sunhours Day 1	 	

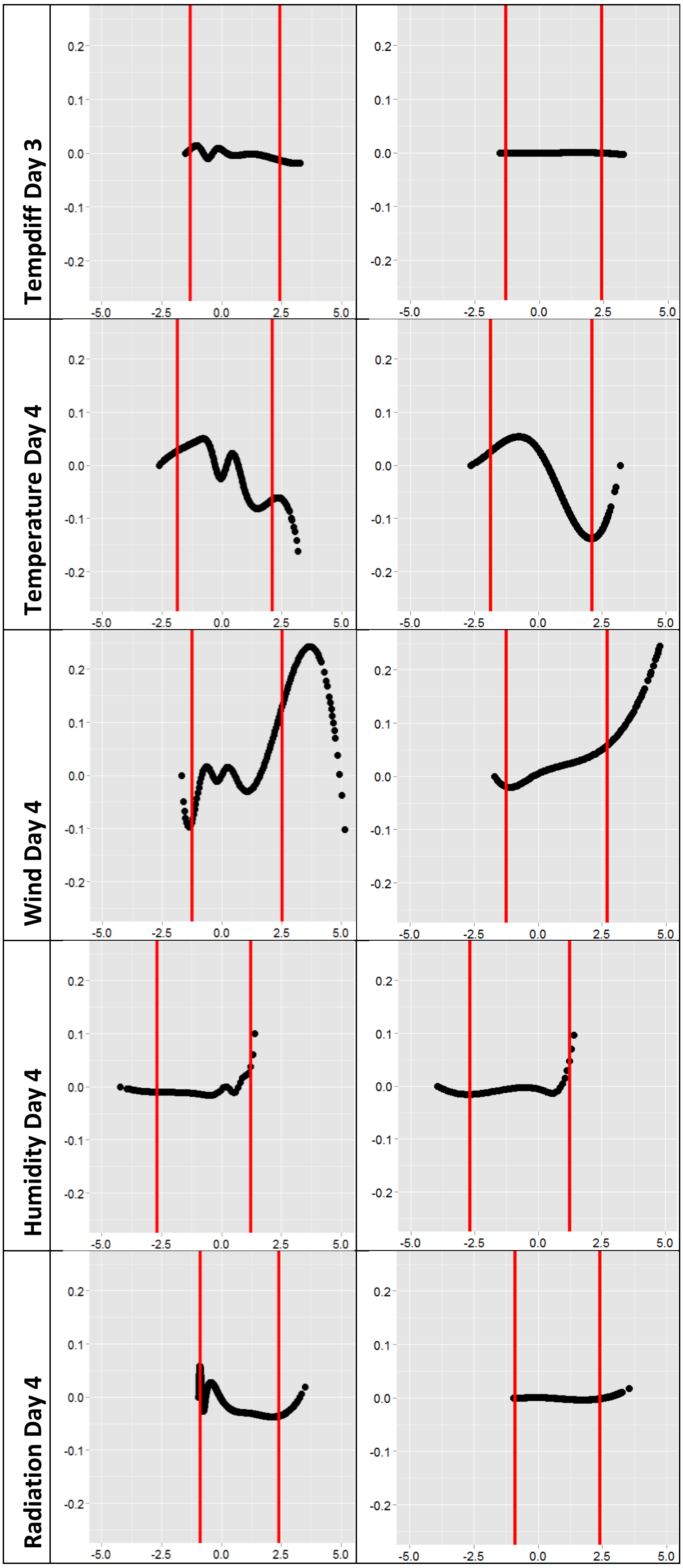
Precipithours Day 1		
Sum of precipit Day 1		
Evaporation Day 1		
Days without sun Day 1		
Days with precip Day 1		

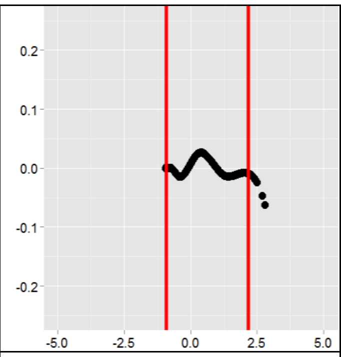
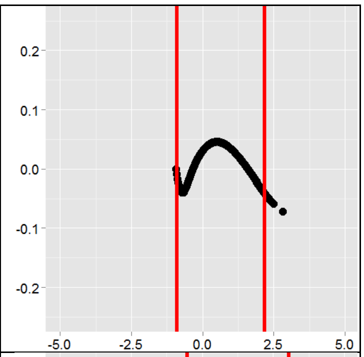
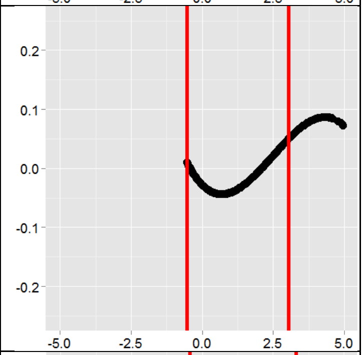
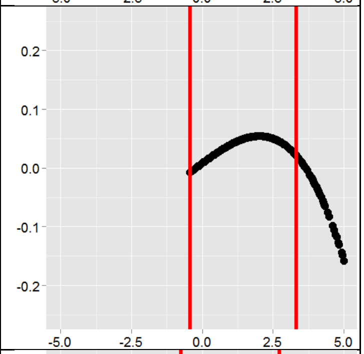
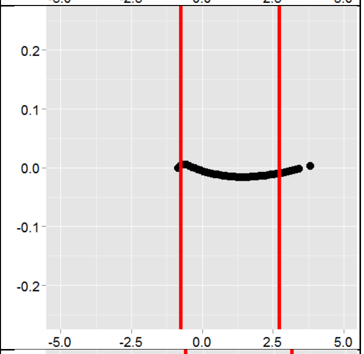
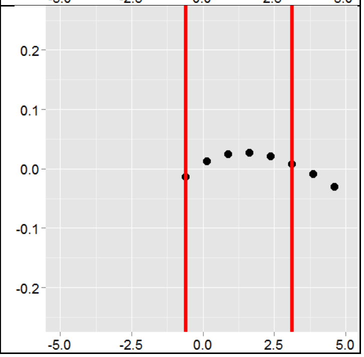

Tempdiff Day 1
Temperature Day 2
Humidity Day 2
Radiation Day 2
Sunhours Day 2
0.2
0.1
0.0
-0.1
-0.2
-5.0
-2.5
0.0
2.5
5.0

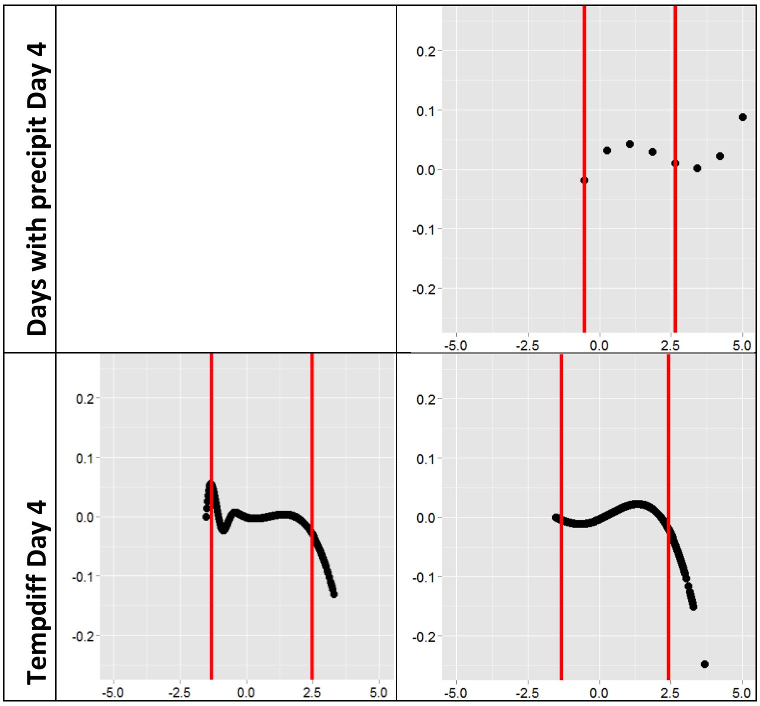

Sum of precipit Day 2
Evaporation Day 2
Days without sun Day 2
Days with precip Day 2
Tempdiff Day 2
0.2
0.1
0.0
-0.1
-0.2
-5.0
-2.5
0.0
2.5
5.0


Temperature Day 3
Wind Day 3
Humidity Day 3
Radiation Day 3
Sunhours Day 3
0.2
0.1
0.0
-0.1
-0.2
-5.0
-2.5
0.0
2.5
5.0

Precipithours day 3		
Sum of precipit Day 3		
Evaporation Day 3		
Days without sun Day 3		
Days with precipit Day 3		


Tempdiff Day 3
Temperature Day 4
Wind Day 4
Humidity Day 4
Radiation Day 4
0.2
0.1
0.0
-0.1
-0.2
-5.0
-2.5
0.0
2.5
5.0

Sunhours Day 4		
Precipithours Day 4		
Sum of precipit Day 4		
Evaporation Day 4		
Days without sun Day 4		


Days with precipit Day 4
Tempdiff Day 4
0.2
0.1
0.0
-0.1
-0.2
-5.0
-2.5
0.0
2.5
5.0