Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Epidemiology & Public Health Methodology

Masterproef

Estimating trends of influenza-like illness based on an observational study

Promotor :
Prof. dr. Christel FAES

Yannick Vandendijck

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Epidemiology & Public Health Methodology

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Epidemiology & Public Health Methodology

Masterproef

Estimating trends of influenza-like illness based on an observational study

Promotor :
Prof. dr. Christel FAES

Yannick Vandendijck

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Epidemiology & Public Health Methodology

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Acknowledgment

First and foremost I would like to thank my promoter, Prof. Dr. Christel Faes. Thanks for the assistance and aid during this thesis. You always had new insights, interesting thoughts and helpful comments during our meetings. I am looking forward to the cooperation in the next four years, in which I hopefully can succeed in my Ph.D. thesis. I also want to thank all professors at Censtat that aid me in the last two years. Special thanks go to Sander van Noort, Carl Koppeschaar and Ronald Smallenburg, from *De Grote Griepmeting* in The Netherlands, to provide the data and helpful extra information.

Many thanks to Steven and Stijn for the five enjoying years we had at Hasselt University, especially during the many projects during the Master of Statistics. Hopefully, we can stay friends for a much longer period. Finally, many thanks to Leen, for all the love and support you gave me in the last four years.

Yannick Vandendijck
Diepenbeek, 11 September 2011

Summary

The interest in this thesis is to estimate influenza-like-illness (ILI) trends in Flanders and Brussels in Belgium for the influenza season 2010/2011 based on observational data coming from the great influenza survey (GIS). The GIS is a surveillance system in the Netherlands and Belgium that is based on participation of volunteers that respond weekly to a symptoms questionnaire.

In total 4634 volunteers participated during the 2010/2011 winter season. 89 % of the participants responded more than three times. Comparing the demographics of the GIS population with the Flanders/Brussels population revealed that especially the younger and elderly are underrepresented, whereas participants aged 40-69 years are over-represented. Analysis is performed without and with correcting for this selection bias. A good association is observed between the ILI trend estimated from the GIS and the trend of a traditional surveillance system (EISN). The best association is observed when correcting for the selection bias and including all participants into the analysis, with a Pearson correlation of 0.866 (95 % CI: 0.724 - 0.938). It is also found that the trend of the GIS lagged that one of EISN by one to two weeks.

ILI trend are estimated using semiparametric logistic models. The time variable in the models is modelled using penalized splines, based on the truncated power basis of degree one and the O'Sullivan basis. The results from the latter basis are found to be more numerically stable and smooth. For fitting the models, the close connection of penalized splines with mixed models is used. Again it is observed that the estimated trend are very similar as the one obtained from the EISN surveillance system, Pearson correlation equal to 0.937 (95 % CI: 0.867 - 0.971). Using semiparametric models it is straightforward to incorporate different risk factors into the model. In a univariate analysis, correcting for the selection bias by post-stratification weighting, it is found that the following risk factors increased the risk on ILI incidence during the 2010/2011 season: not being vaccinated for seasonal influenza, having asthma and/or diabetes, having one or more allergies (hay fever, dust mite allergy and allergy for pets), living with at least one child and smoking. It is found that young children have the highest risk to obtain ILI. A second increase in ILI incidence risk was found for adults aged 25-35 years and a last, but smaller, increase for elderly aged 50-70 years. Based on a multiple semiparametric logistic model, only the risk factors allergies, smoking and age are identified.

From the presented work, it can be concluded that the GIS is a reliable surveillance system to monitor influenza-like-illness in Flanders and Brussels. Based on semiparametric models using penalized splines for the time, smooth estimated trends can be constructed together with 95% confidence bands. The GIS has the important advantage of yielding individual data, which can be used to identify risk factors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Surveillance of Influenza-Like Illness

Almost every winter in Belgium, thousands of individuals experience influenza, although, it often is not more than a serious cold. Infection with the influenza virus is a serious underestimated disease and influenza epidemic causes illness with associated increase in medical consumption and excess deaths. Simonsen *et al.* (1997) estimated that in the US during the influenza seasons from 1972 until 1992 their were on average 5600 yearly excess deaths associated with influenza and pneumonia. Jansen *et al.* (2007) found that excess mortality in individuals ≥ 50 year-old is significantly related to the influenza virus active periods. Influenza-associated hospitalisation was highest and about equal for $0-1$ year-olds and the elderly, and also significant for low-risk adults. Vaccination for influenza is not mandatory in Belgium, but strongly recommended for risk groups (e.g. ≥ 65 year-olds) and health care workers. Because for many people, infection with the influenza virus is often hard to distinguish from a serious cold, the concept of influenza-like illness (ILI) is used. ILI is defined an an illness that has the same symptoms as infection with the influenza virus, although it is not clinically tested for it.

Policy makers need rapid reliable data on influenza-like illness in the population to be able to make decisions and respond to both seasonal and pandemic influenza (Flahault, 2006). Surveillance of influenza aims to mitigate the burden of influenza and control the disease (ECDC, 2009). Traditional surveillance of influenza is based on virological and clinical data, coming from ILI consultations at physicians, mostly general practioners (GPs). However, Tilston *et al.* (2010) argue that this has some potential drawbacks as they require individuals to attend physicians when they are ill. In Belgium this surveillance system is organised by the Scientific Institute of Public Health and is based upon weekly ILI consultations from general practitioners participating in the Belgium Sentinel Network. Weekly ILI consultations rates are provided by the physicians and a sample of the consulting ILI patients are tested for influenza virus infection by nose and throat swabs, and when positive further cultured for type of strain. The Belgium Sentinel Network provides its information to surveillance schemes such as WHO/Europe Influenza Surveillance (EuroFlu) and the European Influenza Surveillance Network (EISN) organised by the European Centre for Disease Prevention and Control (ECDC).

Besides the traditional surveillance systems, many systems based on voluntary participation of individuals have been set up to track ILI in the population. In Denmark a year-round telephone reporting system was established in 2008 in collaboration with the Danish medical on-call service. This system was proved to be useful and timelier than traditional surveillance systems (Harder *et al.*, 2011). Thanks to the ever increasing access to internet new tools to surveillance are available. Influenzanet is based on the participation of volunteers in the population who weekly respond to an internet questionnaire about influenza symptoms. Every individual can join the survey at any time by completing an online registration form and are then weekly asked to report on their ILI symptoms experienced since their previous visit. Participation is stimulated by email newsletters, online educational materials, presentations and other educational activities (Marquet *et al.*, 2006). The Netherlands and Belgium launched the first influenzanet, *De Grote Griepmeting* or the The Great Influenza Survey (GIS) [www.degrotegriepmeting.nl/], during the 2003/2004 influenza season. Other countries such as Portugal, Italy, Australia, Brasil, Mexico and the United Kingdom followed in the recent years. A major concern with web-based surveys is the non-representation of the population and the self-selection of participants, therefore it is likely to give biased results (Marquet *et al.*, 2006). This is due to the fact that internet is not penetrated equally in all age groups and individuals who respond are almost certain to be different from those that do not respond. An additional internet based surveillance tool to track ILI is based on influenza related queries entered at online search engines. The most popular one is Google Flu Trends, see Ginsberg *et al.* (2009).

1.2 Objectives

The purpose of this thesis is to investigate ILI trends in Belgium during the influenza season 2010/2011 using data coming from the great influenza survey (GIS). A first research objective is to investigate the demographics of the GIS population and compare it with the Belgium population. Influenza-like illness trends will be constructed based on different assumptions and criteria. A comparison of these trends with trends based on the traditional surveillance system, obtained via the EISN, is made. Based on semiparametric logistic regression models using penalized splines, smooth ILI trends are investigated for different subgroups of the GIS population. Based on these models it is possible to identify risk factors for ILI and to provide timeframes in which ILI incidences are different for certain subgroups.

1.3 Structure of the Thesis

In chapter 2, the design of the GIS is further explained and the data is introduced. A first analysis of the GIS data is done in chapter 3. The demographics of the GIS are investigated and incidence trends of ILI are constructed and compared with the Belgium sentinel of general practioners. In chapter 4, the semiparametric modeling technique is explained. A penalized spline will be used to model the ILI trend. The close connection of penalized splines and mixed models will be highlighted and used to fit semiparametric logistic models to the GIS data. Finally, in model 5 the most important conclusions are summarized and a brief discussion to further research is provided.

Chapter 2

The Great Influenza Survey (GIS)

2.1 Design of the GIS

The Great Influenza Survey or *De Grote Griepmeting* was launched in the Netherlands and the Dutch-speaking part of Belgium (Flanders) in 2003 (Marquet *et al.*, 2006; and Friesema *et al.*, 2009). Since then, every winter during the influenza season, GIS monitors the trend of ILI based on an internet system. At the beginning of each season, volunteers are encouraged to participate by press releases and direct mailings to schools and universities. Participants of the previous season receive an email for participation in the new season. Also during the season participants are mainstained to be encouraged by press releases, daily updates of ILI news on the website, competitions, presentations and other activities. Participation starts by completing an online intake questionnaire, involving baseline characteristics such as demographical, medical and lifestyle questions. In table A1 in appendix A the intake questionnaire of the season 2010/2011 is presented. Participants of the previous season need to fill in this questionnaire again at the start of each influenza season, because mostly new questions are taken up.

Weekly, participants receive an email with a link to a questionnaire about ILI symptoms they might have experienced since their last survey. In table A2 in appendix A this symptoms questionnaire is presented for the 2010/2011 influenza season. A list of 12 symptoms can be chosen from and more than one symptom can be selected. Information on body temperature is asked next. If symptoms or/and fever (defined as a body temperature of 38 degrees of higher) are reported, participants are asked to the date of onset, this date can maximally go 14 days back. The volunteers are questioned whether they consulted a GP for these symptoms/fever and if so, the outcome of the consultation. Lastly, it is asked whether these symptoms/fever led to change of daily behaviour and if so, how many days he/she stayed at home. After the questionnaire participants are given an indication on their health status.

Strict symptomatic criteria are formulated to distinguish ILI from a common cold. The criteria of ILI used by the GIS are an acute onset of fever ($\geq$38°C), accompanied with muscle pain and at least one respiratory symptom (running nose/ cough/ sore throat/ chest pain).

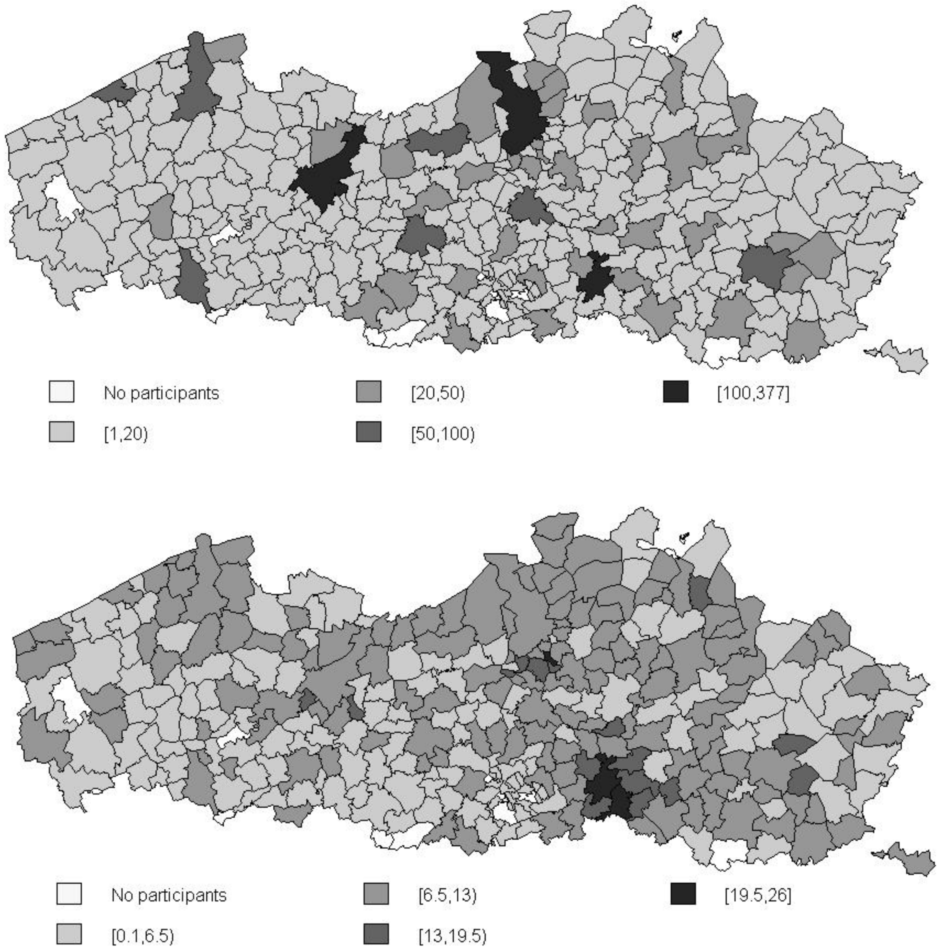

Figure 2.1: *Number of GIS participants (upper) and per 10000 inhabitants (lower) in each city in the Flanders and Brussels region for the influenza season 2010/ 2011.*

2.2 Data

The data used in this thesis is the GIS in the Netherlands and Belgium for the season 2010/2011. It was provided by C. Koppeschaar, R. Smallenburg and S.P. van Noort. The data collected between week 43 in 2010 and week 19 in 2011 is used for analysis, corresponding to the period of 25/10/2010 until 15/05/2011. Only data from the northern part of Belgium, Flanders and Brussels, are used. This is done because the GIS is in Dutch (mainly spoken in these two regions), therefore only few data is available on the southern part, the Walloon region. In total 4634 volunteers participated at least once in the 2010/2011 season. These participants yielded a total of 85092 symptoms questionnaires. In figure 2.1 it can be observed that especially in the major towns many individuals participate, with a maximum of 377 participants in Antwerp. In most towns less than 13 individuals on 10000 participated. The most dense region is around the student city of Leuven. In almost every town in Flanders or Brussels there were volunteers.

Chapter 3

Analysis of the Great Influenza Survey (GIS)

In this chapter, the Belgium GIS data is analysed using similar methods as performed in literature for the GIS in the Netherlands and the Flusurvey in the UK. Results for the Netherlands can be found in Marquet *et al.* (2006), Friesema *et al.* (2009), van Noort *et al.* (2009) and van Noort *et al.* (2011). In the latter two, results for the Belgium GIS are briefly discussed. Tilston *et al.* (2010) present the results for the UK. The first aim is to investigate the demographics of the GIS population. The second aim is to estimate ILI trends based on the GIS data and compare it with the traditional surveillance system of general practioners in Belgium.

3.1 Methods

3.1.1 Sample Used in the Analysis

In order to reduce the effect of participants that only participate occasionally and those who only participate as a response to their current symptoms, most analysis in literature was done by restricting the dataset to participants fulfilling certain criteria. Marquet *et al.* (2006) restrict their analysis using those participants who responded at least five times and data from visitors who joined GIS when ILI was already epidemic were excluded. Friesema *et al.* (2009) and Tilston *et al.* (2010) use both a dataset by restricting to participants' second and subsequent reports. The latter authors also used participants who participated more than once in an additional analysis. In both publications of van Noort *et al.* (2007) and van Noort *et al.* (2011) participants had to participate at least three times to be taken up in the analysis. To minimize selection bias in recruiting sick volunteers, they exclude the first symptoms questionnaire of a participant concerning the week(s) before registration. The real-time analysis on the web site of the GIS follows the same criteria for inclusion and exclusion of participants in the analysis. However, it is unclear which criteria for inclusion or exclusion leads to the most trustworthy results. Friesema *et al.* (2009) discusses that the criterion of at least five times was only necessary in the first GIS season (2003/2004), therefore this criterion seems to be too strict in our case. Table 3.1 presents the two datasets that will be used in this thesis.

Table 3.1: *The two datasets that will be used for the analysis based on criteria for inclusion and exclusion of participants, together with the number of participants and reports.*

Dataset	*Criteria*	*Participants*	*Reports*
Complete	All participants	4634	85092
Restricted	Include all participants that participated at least three times and exclude the first symptom questionnaires concerning the week(s) before the registration date of a participant	4151	80282

3.1.2 Demographics of the GIS Population

An important concern with internet based surveys is the non-representativeness of the internet population (Gosling *et al.*, 2004). To investigate the demographics of the GIS population the complete dataset is used. Age, gender and spatial distributions of the GIS population are compared with the Flanders/Brussels population. Data on population sizes per age group per town and per gender are used of the year 2005. Age distributions are considered into age groups of 10 years. Following Marquet *et al.* (2006) prevalence on asthma and diabetes are also compared, because these are important health factors. Prevalence numbers for these chronic diseases are obtained from the Health Interview Survey in Belgium of the year 2008 (Van der Heyden *et al.*, 2010). The vaccination rates for influenza are compared between the GIS and the Flanders/Brussels population, both for the total population as for the 65+ population. The Flanders/Brussels vaccination rates for influenza were also obtained from the Health Interview Survey in Belgium of the year 2008 (Gisle *et al.*, 2010). For both chronic diseases and vaccination rates of influenza the prevalence of Flanders and Brussels were weighted according to their population sizes, to obtain one estimate for the Flanders/Brussels region. To obtain the vaccination rates of the 65+ population the vaccination rates of the 65-84 year olds and the $\geq$85 year olds were accordingly weighted to their population sizes.

3.1.3 Estimation of ILI Incidence Rates

Estimation of ILI incidence rates is concerned with the estimation of the number of new cases of ILI in a certain time period divided by the size of the population initially at risk. Estimating based on the GIS suffers from a correct definition to do this. In literature many approaches are undertaken to define both the nominator and denominator for the estimation of ILI incidence rates. Most attention is put on the construction of a denominator. Estimation of ILI incidence is based both on the complete and the restricted dataset.

To determinate the nominator, only one approach is considered. Following Friesema *et al.* (2010), the nominator in week i (Monday-Sunday) is measured as the number of ILI participants with ILI onset in week i. It is important to note that the nominator is constructed based on the onset of ILI and not on the week the participant filled in the

Table 3.2: *Definitions that are used to calculate the nominator and denominator to estimate the ILI incidence rate.*

Method	*Definition*
Nominator	
	The nominator in a certain week is measured as the number of ILI participants with ILI onset in that same week
Denominator	
D1	The denominator for each week is the number of participants that completed the symptoms questionnaire that same week
D2	The denominator includes all GIS participants
D3	The denominator is the number of all active participants in a certain week

symptom questionnaire. If a participant responded more than one time in the same week, he/she is only counted once.

As different numbers of participants report to the GIS each week, it would be misleading simply to report the nominator week by week. Instead, an appropriate denominator is needed, to allow for the estimation of ILI incidence rates. Three approaches to obtain a denominator for the incidence rate are considered. The most straightforward approach is to take for each week the number of participants that completed the symptoms questionnaire that same week. Tilston *et al.* (2010) also argue that the possibility that the denominator should include all GIS participants, whether or not they completed the symptoms questionnaire that week. A third manner to define a denominator is to consider all active participants in a certain week (van Noort *et al.*, 2007). A participant is considered to be active between the day of the completion of his/her symptoms questionnaires and the day of the last completed symptoms questionnaire. Table 3.2 summarizes the definitions that are used to calculate both the nominator and denominators in this chapter.

To visualize the trends of the ILI incidence rates, the moving average approach of time series plots is used (Diggle, 1990). This is done to enhance a smoother curve that highlights trends amid the variation (Zeger *et al.*, 2006).

3.1.4 Weighting

Because of the demographic bias contained within the GIS population (see section 3.2.2) in terms of age and gender distribution, post-stratification weighting (Kalton and Flores-Cervantes, 2003) of the sample is considered in the analysis. The sample is split in age groups of 0-19, 20-29, 30-39, 40-49, 50-59, 60-69 and 70+ years for each gender and weighted to match the Flanders/Brussels age and gender distribution. The post-stratification weight for each participant is assigned according to the formula:

$$w_i = \frac{p_i^{F/B}}{p_i^{GIS}}$$

where, w_i is the weight of participant i, $p_i^{F/B}$ is the proportion of the Flanders/Brussels population in the same age and gender category as participant i and p_i^{GIS} is the proportion of the GIS population in the same age and gender category as participant i. It is important to mention that the sum of the weights must again be equal to the sample size. Note that the weights for each participant is time unvarying. This is based on the assumption that age and gender distribution remains the same throughout the influenza season. That this assumption is reasonable can be seen from the figures in appendix C.

3.1.5 Comparison Between ILI Incidence Rates in GIS and EISN

The estimated ILI incidence rates based on the GIS are compared with the incidence rates of the EISN surveillance system. The ILI incidence is determined by the number of ILI patients visiting the GP, divided by the total number of individuals belonging to the practices of the participating GPs. The weekly incidence rates for many of the countries, including Belgium, are weekly reported in the Weekly Influenza Surveillance Overview (WISO). The data, for comparison, was obtained from the WISO reports from week 40 in 2010 until week 20 in 2011. In these reports ILI cases per 100,000 inhabitants are reported. The weekly incidence, covering the period from Monday to Sunday, is published in the WISO the following Wednesday or Thursday. The case definition of ILI as used by EISN is the same as used by the GIS.

An important question between the two surveillance systems is to see whether the rising and declining trend of the GIS and the EISN network are comparable (Truyers *et al.*, 2010). Increasing/decreasing trends in one network should coincide with increases/decreases in the other network at the same time. A similar trend will indicate good validity to detect rapid changes. For comparison, Pearson correlation coefficients are calculated between the ILI incidence rates of the GIS and the EISN. It is also examined whether there is a better association between the two networks when various time lags are introduced. This is possible by means of cross-correlation coefficients (Diggle, 1990), because the trends actually represent two time series. An estimator of the cross-correlation coefficient for two time series $\{(x_t, y_t) : t = 1, ..., n\}$ at lag time k is defined as

$$r_{xy}(k) = \frac{g_{xy}(k)}{\sqrt{g_{xx}(0)g_{yy}(0)}}$$

where

$$g_{xx}(0) = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2, \qquad g_{yy}(0) = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2$$

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n}x_t, \qquad \bar{y} = \frac{1}{n}\sum_{t=1}^{n}y_t$$

and

$$\mathrm{g}_{xy}(k) = \begin{cases} \frac{1}{n}\sum_{t=k+1}^{n}(x_t - \bar{x})(y_t - \bar{y}), & \mathrm{k} \geq 0 \\ \frac{1}{n}\sum_{t=1}^{n+k}(x_t - \bar{x})(y_t - \bar{y}), & \mathrm{k} < 0. \end{cases}$$

A cross correlation on the original incidence rates is investigated to indicate how many weeks the EISN network needs to be shifted to match the GIS network.

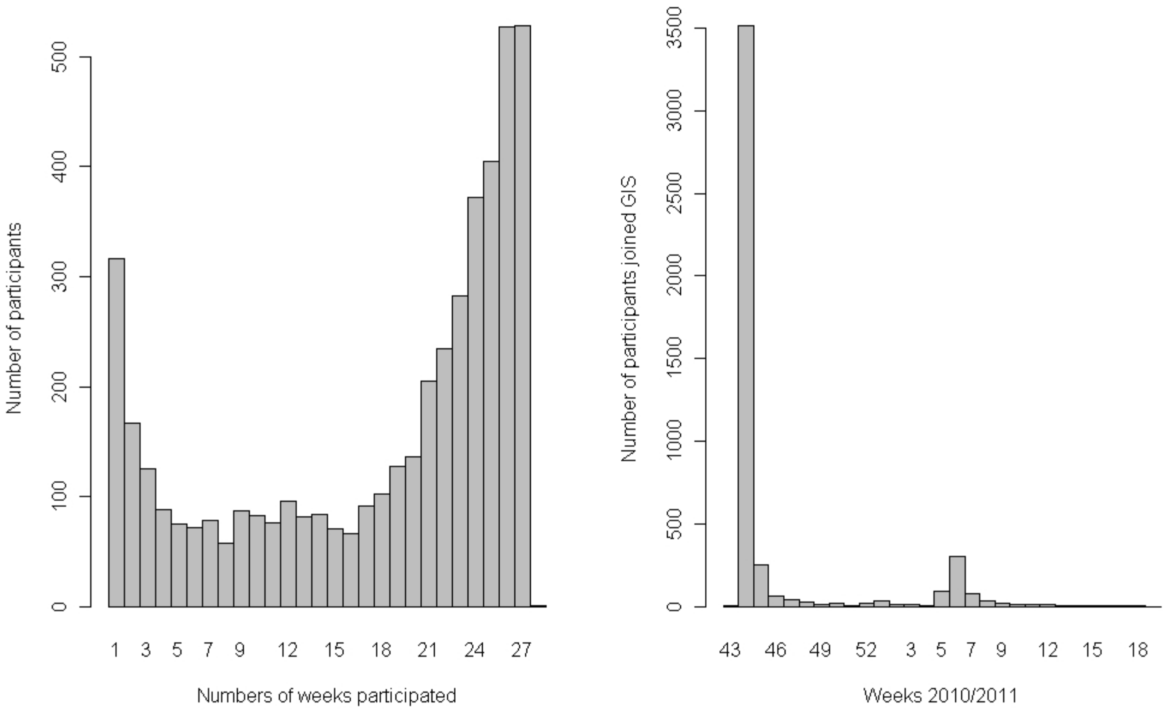

Figure 3.1: *Total number of participants by the number of weeks they participated (left). The week at which participants filled in their first symptoms questionnaire (right).*

3.2 Results

3.2.1 Explorative Analysis

In total 4634 individuals participated in the GIS in 2010/2011. From figure 3.1 (left) it can be observed that the number of weeks participants filled in the symptom questionnaires is bimodal. 316 (6.82%) volunteers only participated once. The percentage of individuals that participated at least three weeks is 89.58% (4151 participants) and 58.05% participated at least 20 weeks. Most of the participants (81.18%) filled in their first symptom questionnaire already in week 44 or 45 in 2010 (figure 3.1 right). In week 6 in 2011 another increase is present.

Table 3.3 presents some characteristics of the GIS. On average 2934 symptom questionnaires were returned every week. Volunteers participated on average for 18.36 weeks. The influenza season of 2010/2011 was one with a prevalence of only 4.08% according to the GIS. Only 4 participants experienced ILI multiple times, which is defined as a recurrent episode of ILI when the participant had reported at least one week without ILI between both episodes. Of the ILI patients, 63.25% visited a general practioner for their symtoms. A larger percentage, namely 78.04%, stayed home.

During each week of the influenza season in 2010/2011 an almost equal amount of symptom questionnaires were filled in (figure 3.2 left). In the first week only 4 question-

Table 3.3: *Characteristics of the GIS in 2010/2011.*

Characteristic	
Number of participants	4634
% of participants with ILI	4.08%
Number of symptom questionnaires	85092
Mean Number of symptom questionnaires per participant	18.36
Mean number of symptom questionnaires per week	2934.21
% of ILI patients staying home	78.04%
% participants with multiple ILI	0.09%
% of ILI patients visiting a GP	63.25%

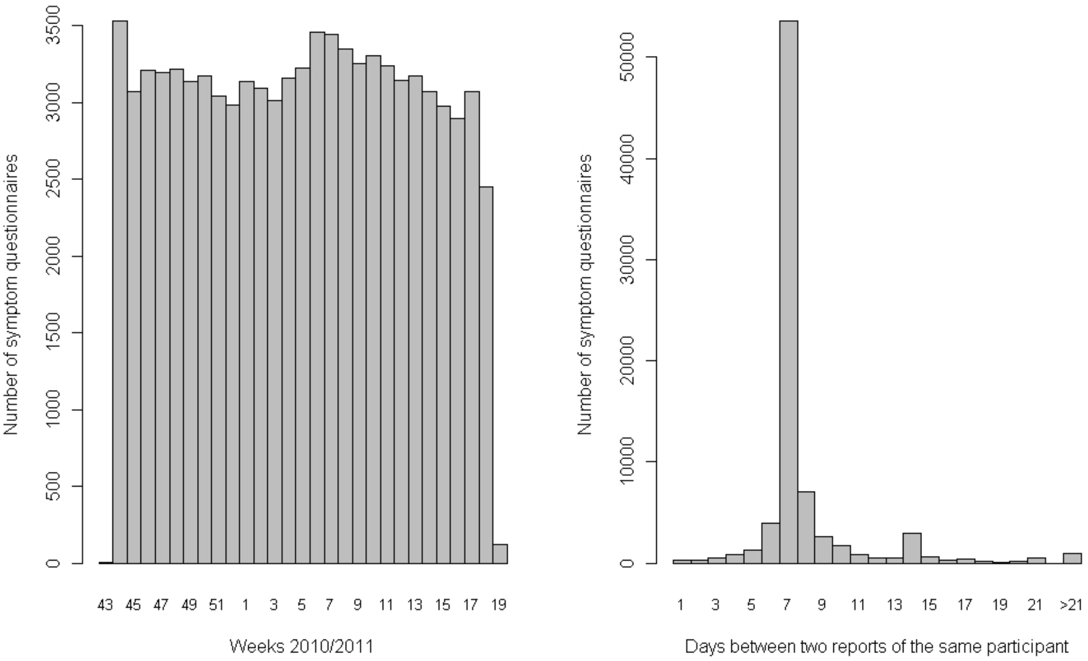

Figure 3.2: *Total number of symptoms questionnaires returned each week (left). The number of days between two consecutive symptoms questionnaires of the same participant (right).*

naires were returned, this week was the beginning of the season. Thereafter, around 3000 symptoms questionnaires were at disposal every week. In the last two weeks, the number of reports diminished again strongly. Most participants filled in the symptom questionnaire every seven days (figure 3.2 right). The number of days between two consecutive reports of the same participant was rarely larger than 14 days. One can conclude that participants filled in the symptoms questionnaires at a regular basis.

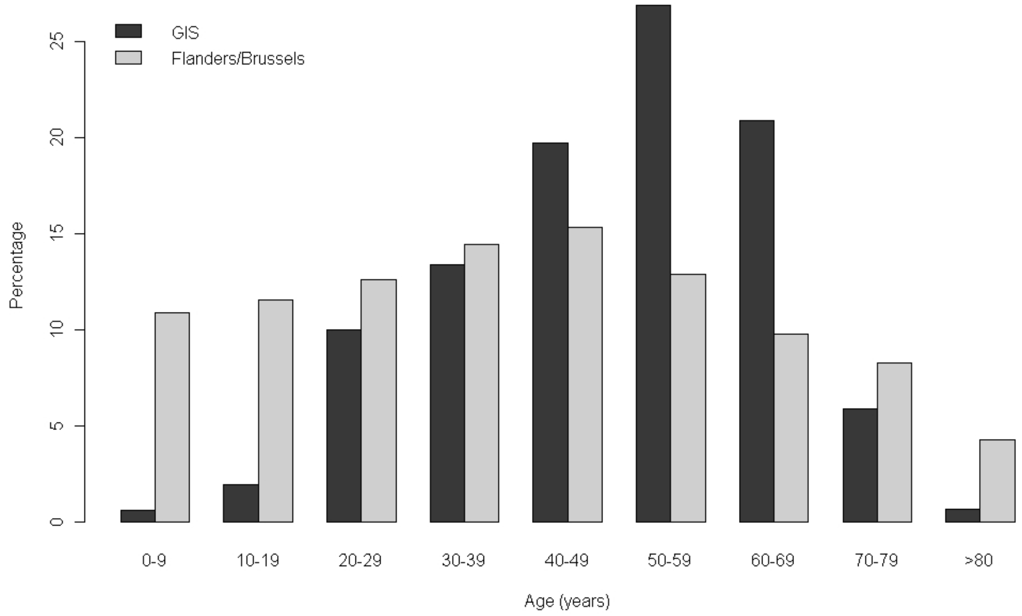

Figure 3.3: *Age distribution of the GIS population and the Flanders/Brussels population in age groups of 10 years.*

3.2.2 Demographics of the GIS Population

The age distributions of the GIS population and the Flanders/Brussels population are presented in figure 3.3. As expected, the age groups 0-9 and ≥80 years are seriously underrepresented in the GIS population. This can be explained by the limited internet usage of these two groups. For the 10-19 years old there is also a large underrepresentation. The age groups 20-29 and 70-79 are moderately underrepresented. A large overrepresentation is observed for the age groups between 40-69 years. Especially the males are underrepresented for the age groups 20-29 and 30-39 years (figure B1 in appendix B). Whereas, males are largely overrepresented for the 60-69 years old. It is interesting to note that in the GIS population females are more represented than males until 49 years, whereafter they are less represented.

The results of the demographics of the GIS and the Flanders/Brussels population are summarized in table 3.4. There was an overrepresentation of the males in the GIS population, because the male-female ratio in the Flanders/Brussels population is 49/51%. Good similarities are seen between the prevalences of asthma and diabetes for both population. For asthma their is a difference of 2.27% in favor of the GIS and of 1.19% for diabetes. For the total population the vaccination rate for influenza in the GIS population is higher as for the Belgium population. For the 65+ population it is also larger. Also for other agegroups the vaccination rates in the GIS population are higher as those in the general population (table B1 in appendix B).

The province of Antwerp is overrepresented in the GIS population, whereas the Brussels plus the Vlaams-Brabant region is underrepresented (Table 3.5). The latter is

Table 3.4: *Demographic results for the GIS and the Flanders/Brussels (F/B) population.*

	GIS	*F/B*
% Males	55.31%	49.13%
% Females	44.69%	50.87%
Prevalence asthma	5.48%	3.21%[a]
Prevalence diabetes	4.53%	3.34%[a]
Vaccination rate total population	44.07%	29.14%[a]
Vaccination rate ≥65 years	74.18%	64.53%[b]

[a] A weighted prevalence from the Flanders and Brussels prevalence.
[b] In the Belgium population.

Table 3.5: *Comparison of the spatial distribution according to provinces in Flanders/Brussels of the GIS population and the Flanders/Brussels (F/B) population.*

Province	*GIS*	*F/B*
Antwerpen	31.10%	23.79%
Brussels + Vlaams-Brabant	20.86%	29.00%
Limburg	11.83%	11.49%
Oost-Vlaanderen	20.80%	19.58%
West-Vlaanderen	15.41%	16.15%

mostly due to the serious underrepresentation in Brussels. The other provinces are almost similarly distributed in the GIS as in the total Flanders/Brussels population.

3.2.3 ILI Trends in the GIS and Comparison with the Belgium EISN Data

The incidence curves per 100,000 persons are presented in figure 3.4. Definition $D3$ is used as denominator. The incidence trend for the EISN network is shown, together with the incidence curves for the complete and restricted dataset, with and without weighting the sample. It is interesting to note that the curves follow the same trend throughout the ILI season. In the first period, only a small number of cases occurs, followed by a sharp increase to reach its peak after about one month and then goes down again to a small number of cases. At the beginning and the end of the ILI season the GIS curves estimate a higher incidence as the EISN network. It seems that the complete dataset is more capable of following the rising and declining of the EISN incidence curve than the restricted dataset. The weighted datasets overestimate the EISN ILI incidence curve, whereas the unweighted datasets underestimate it. The trends estimated by the restricted and complete unweighted datasets show almost no difference, except that during the peak of ILI incidence the incidence is somewhat higher for the complete dataset. This can

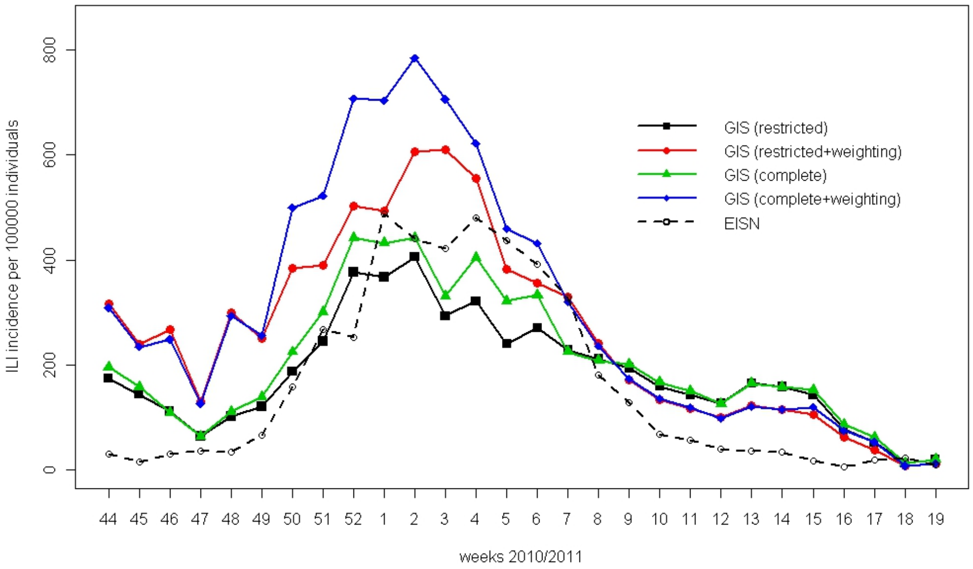

Figure 3.4: *Incidence rates of ILI per 100,000 individuals for different datasets. The EISN incidence curve is plotted and four different GIS incidence curves. All curves of the GIS are plotted with a 3-week moving average.*

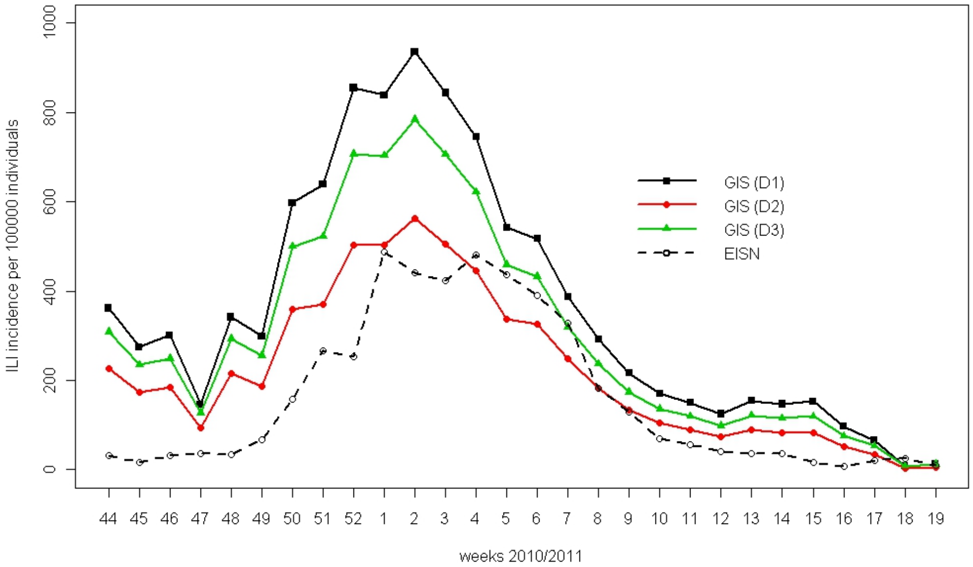

Figure 3.5: *Incidence rates of ILI per 100,000 individuals for different denominators. The complete weighted dataset is used.*

be explained by the fact that the age distribution in the restricted dataset is somewhat dissimilar as in the complete dataset (Figure B2 in appendix B). The ages that are more

Table 3.6: *Pearson correlation and cross correlation coefficients between different ILI incidence curves of the GIS and the EISN system. Coefficients are provided for different datasets and different denominators.*

	Pearson corr.(95% CI)	*Cross correlation*		
		lag=-1	*lag=-2*	*lag=+1*
Different datasets (denominator=D3)				
Restricted	0.744 (0.513 - 0.874)	0.722	0.616	0.645
Restricted+weighting	0.663 (0.385 - 0.830)	0.720	0.682	0.595
Complete	0.779 (0.573 - 0.893)	0.771	0.677	0.681
Complete+weighting	0.774 (0.563 - 0.890)	0.800	0.827	0.659
Different denominator (weighted complete dataset)				
D1	0.766 (0.550 - 0.886)	0.796	0.822	0.654
D2	0.779 (0.572 - 0.892)	0.797	0.812	0.671

represented in the complete dataset are mostly at a higher risk to obtain ILI (see chapter 4), which yields the higher incidence when estimated using the complete dataset. This gap between the restricted and complete dataset is even larger when using the weighted datasets. The same reason as above can be used to explain this fact and by weighting the sample this difference is enlarged, because especially those age groups at a higher risk get higher weights.

Table 3.6 presents the Pearson and cross correlation coefficients for these incidence curves. The Pearson correlation coefficients are higher for the complete datasets than for the restricted datasets. Not much difference is seen between the Pearson correlation coefficients for the unweighted complete dataset, 0.779 (95% CI: 0.573 - 0.893), and the weighted complete dataset, 0.774 (95% CI: 0.563 - 0.890). This indicates the GIS network matches the trend of the EISN network reasonably close. The highest cross correlation is observed for the completed weighted dataset with a lag time of -2 weeks, with a value of 0.827, indicating that the EISN system has to shift two weeks back in order to match the GIS system. From this, it can be concluded that estimating the ILI trend with the weighted complete dataset follows the rising and declining of the EISN trend the best.

In figure 3.5 the impact on the ILI incidence curves, due the different definitions of the denominator, is presented for the weighted complete dataset. As would be expected, using all participants (denominator $D2$) results in a lower ILI incidence than the other two. Nevertheless, this denominator definition seems to fit the EISN data the best in terms of the height of the incidence curve. Both denominator definitions $D1$ and $D3$ yield a higher incidence than is estimated by the EISN network. Using the active participants as denominator produces estimates in between the other two definitions. In terms of rising and declining of the curves not much differences are observed between the different denominators. From table 3.6 minor differences are observed for the Pearson and cross correlation coefficients for the different denominator definitions. The highest cross correlation is again observed at a lag time of -2 weeks for denominator definition three, although the cross correlation for denominator definition one is almost similar. For this reason, using denominator definition $D1$ or $D3$ is ougth to yield the best results in this case.

3.3 Discussion

The Great Influenza Survey had around 4600 participants in Flanders/Brussels during the 2010/2011 influenza season. Most volunteers were regular participants, with over 89% participating at least three times, this is in accordance to other influenza seasons (van Noort *et al.*, 2007). This percentage of participation is a serious improvement as compared to the first monitoring season of the GIS in 2003/2004, when only 53% responded at least three times (Marquet *et al.*, 2006).

The participants in Flanders/Brussels were not equally distributed with regard to age and gender. Especially the younger and elderly are underrepresented, whereas participants between 40 and 69 years are overrepresented. Age distributions were also different between males and females. Similar age and gender distributions were also found in the Netherlands (Marquet *et al.*, 2006 and Friesema *et al.*, 2009) and in the UK (Tilston *et al.*, 2010). A disproportionate distribution of this kind could cause problems for ILI, because some age groups are more susceptible for ILI. By weighting the GIS population it was made more similar to the general population. The region of Brussels is underrepresented in the GIS. This could be explained by the fact that many inhabitants of Brussels are French speaking, so are not attracted by the GIS. The vaccination rates are higher in the GIS population than in the general population. This can be partly explained by the fact that Brussels is underrepresented and the vaccination rates in this region is lower than in Flanders (Gisle *et al.*, 2010). Another reason is that the younger are seriously underrepresented in the GIS and vaccination rates in this group is small. A higher vaccination rate could also be an indicator of a relative high sense of health and more healthy behaviour in the GIS population.

There was a good association between the trends of ILI monitored by the GIS and the traditional surveillance network of the EISN. This is in accordance to what is found in the Netherlands (Friesema *et al.*, 2009) and in the UK (Tilston *et al.*, 2010). The major difference between the GIS and EISN network are observed before the start and at the end of the ILI incidence peak, with higher incidence levels for the GIS. A shortcoming of this analysis is that the incidence rates of the GIS are based only on the Flanders and Brussels region, whereas the EISN network is for whole Belgium. It was found that using all participants in estimating the ILI incidence rates did not yield biased results when compared with the analysis using the restricted dataset. This could be explained by the fact that the influenza season of 2010/2011 is already the eight GIS season and thus is already a well established surveillance network with mostly serious participants. This was certainly not the case for the first GIS season (Marquet *et al.*, 2006). In terms of the definitions of the denominators, not much differences are observed between the trends of ILI incidence, but only in the height of the rate. This is in contrast with what has been observed in the UK (Tilston *et al.*, 2010). This can be explained by the fact that participants in this study participated on a very regular basis and therefore not many differences are observed in terms of the denominators over time.

There was evidence that the incidence of ILI by the GIS is monitored $1-2$ weeks ahead of the EISN network. This means that the GIS incidence of week$_n$ associates well with EISN incidence of week$_{n+1}$ and week$_{n+2}$. An explanation for this could be that most people will not go to the GP on the first day that they experience symptoms, while in the GIS this first day is reported. However, this delay does not imply that the GIS is able to detect ILI trends earlier than the EISN system. The speed of detecting trends depends on when participants report their symptoms to the GIS and the rate and time they visit their GP (van Noort *et al.*, 2011).

Several extensions to the analysis performed in this chapter could be made. Another approach to calculate the nominator in the ILI incidence rate is to calculate a daily instead of weekly incidence (van Noort *et al.*, 2007). However, this method shows many similarities with the nominator that was used, because the daily incidences are transformed to weekly incidences for each day. The nominator for the daily incidence is determined by the number of participants with an onset of ILI on a given day. The nominator for the weekly incidence for each day is next determined by the total number of participants with an onset of ILI in the previous seven days. Post-stratification weighting of the dataset was only considered for the age and gender distribution. Weighting for other factors, such as asthma, diabetes and province is also possible. The problem of post-stratification weighting of many factors is that the strata rapidly have too few observations and too detailed information is necessary, which is often not available. The first problem was encountered in this study when the spatial distribution, in terms of provinces, was also considered in the weighting procedure. A possible method to take into account the spatial distribution in the post-stratification weighting is by using raking methods, also know as iterative proportion weighting or rim weighting (Kalton and Flores-Cervantes, 2003), but is out of the scope of this thesis. This method only takes the marginal distributions into account for constructing the weights.

The analysis in this chapter showed that the GIS in Flanders and Brussels recruits a high number of regular participants. The demographics of the GIS population and the general population were dissimilar, certainly for the age and gender distributions. The ILI trends monitored by the GIS paralleled the trend obtained by the traditional EISN surveillance network well. This association was best when correcting for the disproportionate age and gender distribution by weighting and using all participants. Care needs to be taken with respect to the obtained results and conclusion, because they are based on information on only one influenza season. To validate the obtained results, it is interesting to also investigate the other influenza seasons in Belgium. A major advantage of the GIS system as compared to the EISN network is the fact that individual data and characteristics are available, whereas EISN is based on aggregated data. This fact is used in the next chapter, where the ILI trends for different subgroups of the GIS population is investigated.

Chapter 4

Semiparametric Models to Analyse the Great Influenza Survey (GIS)

In this chapter the GIS is analysed using a model based approach. Due to the findings in the previous chapter the complete dataset is used. The response variable of interest is having influenza-like illness or not. Based on semiparametric logistic regression models, the response variable is modelled over time, where penalized splines are used to capture the trend over time. This modelling approach easily lends itself to investigate and infer differences in time trends of ILI between several subgroups. Firstly, semiparametric (logistic) regression models based on truncated power basis splines are explained. The connection of semiparametric models with mixed models is discussed next. O'Sullivan splines are introduced to overcome some difficulties encountered with truncated power basis splines. Finally, the discussed methodology is used to analyse the GIS.

4.1 Introduction and Overview

Parametric regression methods are well documented in literature. Parametric regression models all assume a linear (or some parametric) form for the covariate effects. This assumption is too restrictive for many practical applications. This led to the development of nonparametric and semiparametric regression methods, within which the linear or parametric form of (some of) the covariates are replaced by a flexible function. The methods can be broadly classified into kernel methods and spline smoothing. In what follows I will focus on spline methods.

Spline functions are piecewise polynomials, with the polynomial pieces joining at the knots and fulfilling continuity conditions for the spline itself and some of its derivatives (Costa, 2008). The knots cover and are chosen within the domain of the covariate X. A spline function $f(x)$ with knots at k_i, i=1,...,K, can be defined as

$$f(x) = \sum_{i=1}^{L} \theta_i B_i(x),$$

where θ_1,...,θ_L represent smoothing parameters and $B_1(x)$,...,$B_L(x)$ are the spline basis functions, where some basis functions are constructed using the K knots k_1,...,k_K. There are several types of splines and these can be roughly classified as smoothing splines, regression splines and penalized splines (P-splines).

Smoothing splines have a knot at each unique observation or design point of the variable X. The most commonly used smoothing spline is the natural cubic smoothing spline. The natural cubic spline arises as the solution of the penalized residual sum of squares criterion (Hastie *et al.*, 2009)

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int [f^{''}(x)]^2 dx,$$

where y_i, i=1,...,n, are the responses and $\lambda \geq 0$ is a smoothing penalty. When λ=0 the natural cubic spline estimator interpolates the data, and it yields a linear fit if $\lambda \rightarrow 0$. Because for smoothing splines the number of smoothing parameters to be estimated is as large as the number of unique observations, they are computationally intensive.

Regression splines, on the contrary, only use a small, but carefully chosen number of knots. Hence, their positions on the domain play a crucial rule. In regions where the function to be estimated has greater flexibility more knots need to be placed. Friedman (1991) proposed a data-driven method for knot selection and placement, namely Multivariate Adaptive Regression Splines (MARS).

Penalized splines can be seen as a hybrid of smoothing and regression splines and were proposed by Eilers and Marx (1996). The idea is the same as in smoothing splines, namely penalizing the spline to prevent overfitting. However, the number of knots used for penalized splines is typically far less than the number of unique observations or design points, but much larger than the number of knots used in regression splines. Hence, they are computationally more efficient than smoothing splines. Another advantage is that they are less sensitive to the placements of the knots as compared to regression splines. Eilers and Marx (1996) proposed the use of difference penalties for controlling the smoothness of the spline using a B-spline basis. The focus in this thesis is on penalized splines, as described in the book of Ruppert *et al.* (2003). In the next section penalized splines using the truncated power basis are formally introduced for a simple case, later the scope is enlarged to encompass more complicated models.

4.2 Semiparametric Models Using Penalized Splines

Suppose the simple case, where data (x_i, y_i), with x_i univariate and y_i continuous, is available. Consider the model

$$y_i = f(x_i) + \epsilon_i, \tag{4.1}$$

where f is a smooth function giving $E(y_i|x_i)$ and ϵ_i, i=1, ..., n, are independent mean zero errors with constant covariance σ^2_ϵ. For inference it is mostly assumed that $\epsilon_i \sim \mathcal{N}(0,\sigma^2_\epsilon)$. For the spline f, consider the truncated power basis of degree p with K knots k_1,...,k_K

$$1, x, ..., x^p, (x - k_1)^p_+, ..., (x - k_K)^p_+,$$

where $(x)_+$ is a truncated function equal to x if x is positive and equal to 0 otherwise. Figure 4.1 present some of these basis function of degree 1 and 2. Using the truncated

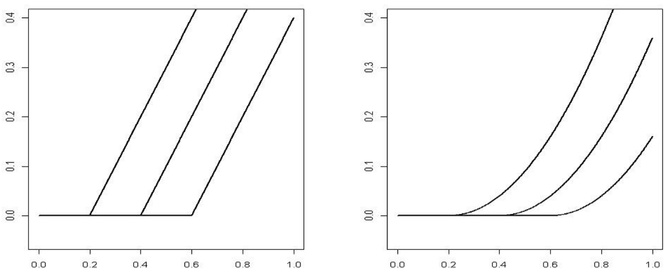

Figure 4.1: *Truncated power basis functions of degree 1 (left) and 2 (right), with knots at 0.2, 0.4 and 0.6.*

power basis of degree p, $f(x)$ in model (4.1) becomes

$$f(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{m=1}^{K} b_m (x - k_m)_+^p. \tag{4.2}$$

The smoothness of the spline is controlled by penalizing the squares of the smoothness parameters b_m, m=1,...,K. Model (4.1) is then fit by using the penalized least squares criterion

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda^{2p} \sum_{m=1}^{K} b_m^2, \tag{4.3}$$

where $\lambda \geq 0$ is a smoothing parameter.

Using the following notations

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^p & (x_1 - k_1)_+^p & \cdots & (x_1 - k_K)_+^p \\ \vdots & & & & & & \vdots \\ 1 & x_n & \cdots & x_n^p & (x_n - k_1)_+^p & \cdots & (x_n - k_K)_+^p \end{pmatrix}$$

and

$$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p, b_1, \ldots, b_K)^T,$$

it is easy to see that (4.3) becomes

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda^{2p} \boldsymbol{\beta}^T \mathbf{D} \boldsymbol{\beta}, \tag{4.4}$$

where $\mathbf{D}$=diag($\mathbf{0}_{p+1}$,$\mathbf{1}_K$). The penalization term in (4.4) could be generalized to $\alpha\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}$ for some positive semidefinite matrix $\mathbf{D}$ and scalar $\alpha \geq 0$.

For the selection of the smoothing parameter λ cross-validation, generalized cross-validation, Mallow's C_p criterion and many others could be used (see Ruppert *et al.*, 2003). The maximum likelihood approach for the selection of the smoothing parameter is considered here, which is based on the mixed model representation explained in the next section. For the selection of the locations of the knots the recommendations of Ruppert (2002) are followed. The smoothing is done with K equally spaced knots, selected as quantiles of the covariate. For the selection of K it is better to have too many knots than too few knots. Ruppert (2002) and Ngo and Wand (2004) provide recommendations for choosing the number K.

4.3 Semiparametric Models in the Linear Mixed Model Framework

Brumback *et al.* (1999) noticed the close link between penalized splines and the optimal predictor in a linear mixed model. This link is extremely useful, there semiparametric regression analysis can be fit using widely available mixed model software. Fitting penalized splines by a linear mixed model has some other appealing advantages, such as the automatic determination of the smoothing parameter, a unified framework for inference and the ease of extending the model (Maringwa *et al.*, 2008b).

The general linear mixed model can be represented as (Verbeke and Molenberghs, 2000)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{4.5}$$

where

$$\mathrm{E}\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathrm{Cov}\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}.$$

The vector $\mathbf{y}$ is the response vector, the matrix $\mathbf{X}$ contains the values of the covariates, $\boldsymbol{\beta}$ contains the fixed effects parameter, $\mathbf{u}$ is the random effects vector, $\mathbf{Z}$ is the design matrix for the random effects and $\boldsymbol{\epsilon}$ is an error vector. Here, it is assumed that $\mathbf{G}=\sigma_u^2\mathbf{I}$ and $\mathbf{R}=\sigma_\epsilon^2\mathbf{I}$. Making the distributional assumptions that

$$\mathbf{y}|\mathbf{u} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}) \text{ and } \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}),$$

and maximizing the likelihood of the ($\mathbf{y}$,$\mathbf{u}$) over the unknows $\boldsymbol{\beta}$ and $\mathbf{u}$, leads to the minimization criterion (Ruppert *et al.*, 2003)

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}^T\mathbf{G}^{-1}\mathbf{u}. \tag{4.6}$$

From (4.6) the Best Linear Unbiased Prediction (BLUP) $(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}})$ for $(\boldsymbol{\beta}, \mathbf{u})$ can be obtained. Estimates $\hat{\sigma_u}^2$ and $\hat{\sigma_\epsilon}^2$ of the covariance parameters σ_u^2 and σ_ϵ^2 are obtained via maximum likelihood (ML) or restricted maximum likelihood (REML). Using $\hat{\sigma_u}^2$ and $\hat{\sigma_\epsilon}^2$, Estimated BLUPs (EBLUP) of $(\boldsymbol{\beta}, \mathbf{u})$ can be formed and are denoted as $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}})$. For more information see Verbeke and Molenberghs (2000).

The connection between penalized splines and the linear mixed model is now shown. Consider again the spline from (4.2)

$$f(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{m=1}^{K} u_m (x - k_m)_+^p.$$

Let

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{and} \quad \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_K \end{pmatrix},$$

be the coefficients of the polynomial function and truncated functions. Construct the matrices

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & & & \vdots \\ 1 & x_n & \cdots & x_n^p \end{pmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} (x_1 - k_1)_+^p & \cdots & (x_1 - k_K)_+^p \\ \vdots & & \vdots \\ (x_n - k_1)_+^p & \cdots & (x_n - k_K)_+^p \end{pmatrix}$$

The penalized spline criterion (4.4), when divided by σ_ϵ^2, can then be written as

$$\frac{1}{\sigma_\epsilon^2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\right\|^2+\frac{\lambda^{2p}}{\sigma_\epsilon^2}\mathbf{u}^T\mathbf{u}.$$

Notice that this expression equals the criterion (4.6) from the linear mixed model, when treating $\mathbf{u}$ as a set of random effects with $\text{Cov}(\mathbf{u})=\sigma_u^2\mathbf{I}$ and when $\lambda^{2p}=\frac{\sigma_\epsilon^2}{\sigma_u^2}$. This connection enables one to fit penalized splines in the mixed model framework. One just needs to construct the matrices $\mathbf{X}$ and $\mathbf{Z}$, and let $u_m \overset{ind}{\sim} \mathcal{N}(0,\sigma_u^2)$, m=1,...K. The smoothing parameter λ is automatically estimated by ML or REML. Linear mixed model software is widely available, in particular one can use the function `lme()` in R and the MIXED procedure in SAS to fit LMMs.

Until now only one covariate X was considered. Extensions to include more covariates are straightforward due to the mixed model representation. A good overview of these models and their implementation in mixed model software can be found in Ngo and Wand (2004). Additive models can be used for regression problems where some covariates enter the model linearly and others nonparametricaly. Additive mixed models can be used to model longitudinal data, where subject specific random effects are included, to account for the clustered nature of observations. Applications of these type of models can for example be found in Maringwa *et al.* (2008b) and Maringwa *et al.* (2008c).

4.4 Semiparametric Models in the Generalized Linear Mixed Model Framework

Notice that until now the response has been considered to be continuous. For non-Gaussian data equivalent nonparametric and semiparametric approaches exist, together with a connection to the generalized mixed model framework. Consider a response y from the 1-parameter exponential family of distributions, it has a density of the form

$$f(y;\eta)=\exp\left(\frac{y\eta-b(\eta)}{\phi}+c(y,\phi)\right),$$

for some functions $b(\eta)$ and $c(y,\phi)$, where ϕ is a dispersion parameter. The parameter η is called the natural parameter. It can be shown that $\mu\equiv\text{E}(y)=b^{'}(\eta)$ and $\text{var}(y)=\phi b^{''}(\eta)=\phi v(\mu)$, with $v(.)$ a specified variance function (McCullagh and Nelder, 1989). In Generalized Linear Models (GLM), proposed by Nelder and Wedderburn (1972), it is assumed that $g(\mu_i)=\eta_i$ and g is an appropriately chosen link function. It is assumed that the natural parameter, η_i, depends on a vector of covariates $\mathbf{X}_i$, namely by the linear predictor $\eta_i=\mathbf{X}_i\boldsymbol{\beta}$.

For analysis of clustered data, it is useful to incorporate random effects into the GLM. The resultant model is known as a Generalized Linear Mixed Model (GLMM). The set up of a GLMM is of the same form as in GLMs, where additionally a random effects vector $\mathbf{u}$ and the random effects design matrix $\mathbf{Z}$ are incorporated. In particular, it is assumed that $g(E[y_i|\mathbf{u}_i])=\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{u}_i$. The resulting density is of the form

$$f(\mathbf{y};\boldsymbol{\beta}|\mathbf{u})=\exp\left(\frac{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})-\mathbf{1}^Tb(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})}{\phi}+\mathbf{1}^Tc(\mathbf{y},\phi)\right). \tag{4.7}$$

Mostly, it is assumed that $\mathbf{u}{\sim}\mathcal{N}(0,\mathbf{G}_\theta)$, where $\boldsymbol{\theta}$ are variance parameters. For the conditional variance of this model, one can write var($\mathbf{y}|\mathbf{u}$)=$\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$. The matrix $\mathbf{A}$ is a diagonal matrix and contains the variance functions $v(.)$ of the model. The matrix $\mathbf{R}$ models the so called R-side covariance structures. It is used to model overdispersion or to specify a heterogeneous variance model. When only overdispersion is modelled, it is assumed that $\mathbf{R}{=}\phi\mathbf{I}$. For more information on GLMMs see McColluch and Searle (2001).

In a similar fashion as in the Gaussian case, where penalized splines can be represented by a linear mixed model, responses from the 1-parameter exponential family of distribution can be represented in the GLMM framework (Ruppert *et al.*, 2003). Consider for example, non-Gaussian data $\mathbf{y}{=}(y_1,...,y_n)$, which is modelled on two covariates S and T. Using the truncated power basis of degree p to represent the spline, the design matrices

$$\mathbf{X} = \begin{pmatrix} 1 & s_1 & \cdots & s_1^p & t_1 & \cdots & t_1^p \\ \vdots & & & & & & \vdots \\ 1 & s_n & \cdots & s_n^p & t_n & \cdots & t_n^p \end{pmatrix}$$

and

$$\mathbf{Z} = \begin{pmatrix} (s_1-k_1^s)_+^p & \cdots & (s_1-k_K^s)_+^p & (t_1-k_1^t)_+^p & \cdots & (t_1-k_K^t)_+^p \\ \vdots & & & & & \vdots \\ (s_n-k_1^s)_+^p & \cdots & (s_n-k_K^s)_+^p & (t_1-k_1^t)_+^p & \cdots & (t_n-k_K^t)_+^p \end{pmatrix}$$

are obtained. Construct the fixed and random effects vectors as

$$\boldsymbol{\beta} = (\beta_0,\beta_1^s,...,\beta_p^s,\beta_1^t,...,\beta_p^t)^T \qquad \mathbf{u} = (u_1^s,...,u_K^s,u_1^t,...,u_K^t)^T.$$

Assuming that

$$\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_s^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_t^2\mathbf{I} \end{bmatrix}\right),$$

a generalized linear mixed model is obtained as in (4.7). These models are easily extended to include covariates where some covariates enter the model linearly and others nonparametricaly. Also the incorporation of subject specific random effects, to model clustering of observations, is straightforward. An application of a semiparametric model with Poisson counts in the GLMM framework can be found in Maringwa (2008a). GLMMs can be fit using the GLIMMIX procedure in SAS.

Estimation of model parameters ($\boldsymbol{\beta}$,$\boldsymbol{\theta}$) in a GLMM typically involves maximum likelihood. Maximization of the likelihood is often hindered by the presence of a high-dimensional integral, which is often intractable for direct calculation. A first method, that is considered here, to overcome this problem involves pseudo-likelihood (PL) estimation techniques (Wolfinger and O'Connell, 1993). In this method the model is approximated by pseudo-data based on Taylor series expansions. A second method, that is considered, is to approximate the integral by Laplace approximation (Raudenbush *et al.*, 2000). More details are provided in appendix D. For more information on both methods, see, for example, Molenberghs and Verbeke (2005). The advantage of integral approximation methods is that likelihood ratio tests among nested models can be performed and likelihood-based fit statistics can be computed. An advantage of pseudo-likelihood techniques is that R-side covariance structures can be incorporated. The disadvantages of this approach is that a true objective function is absent. For this reason pseudo-likelihood values should not be compared across different statistical models, even if the models are nested. Because the methods are based on approximations, potentially biased estimates can be obtained. For more information on advantages, disadvantages and notes on bias of these methods, see SAS Insitute Inc. (2008).

4.5 O'Sullivan Penalized Splines

Until now, only the truncated power basis function for penalized splines was considered. While this basis is conceptually simple, it is often scorned because of their numerical instability (Hastie *et al.*, 2009) and they do not have attractive features that other spline basis functions exhibit. A spline basis that enjoys numerical stability and other attractive features, is the O'Sullivan penalized spline basis (O'Sullivan, 1986). In contrast to the truncated power basis functions, the O'Sullivan penalized spline basis functions are bounded, which gives them more stable numerical properties. O'Sullivan splines are a direct generalization of smoothing splines in the sense that the latter arise when the maximal number of spline basis functions is included (Wand and Ormerod, 2008). O'Sullivan penalized splines also have a close connection with the P-splines of Eilers and Marx (1996), although they perform better in terms of boundary conditions than P-splines. The latter have a tendency to deviate from the natural behaviour of smoothing splines, whereas O'Sullivan splines do not (Wand and Ormerod, 2008).

The cubic O'Sullivan penalized splines are used here, as described in Wand and Ormerod (2008). Consider again the simplest nonparametric model given in (4.1). Suppose that an estimate of f is required over $[a, b]$, an interval containing the x_i's. For a number K, define the knot sequence, $k_1, ..., k_{K+8}$, such that

$$a = k_1 = k_2 = k_3 = k_4 < k_5 < ... < k_{K+4} < k_{K+5} = k_{K+6} = k_{K+7} = k_{K+8} = b.$$

Let B_1,...,B_{K+4} be the cubic B-spline basis functions defined by these knots. A cubic B-spline basis function $B_{i,m}(x)$ is defined recursively in terms of divided differences as followed (Hastie *et al.*, 2009):

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } k_i \leq x < k_{i+1} \\ 0 & \text{otherwise} \end{cases},$$

for i=1,...,$K+7$. Then $B_{i,m}(x)$ is defined as

$$B_{i,m}(x) = \frac{x - k_i}{k_{i+m-1} - k_i} B_{i,m-1}(x) + \frac{k_{i+m} - x}{k_{i+m} - k_{i+1}} B_{i+1,m-1}(x),$$

for i=1,...,$K+8-m$. A cubic B-spline is of order m=4. Set up an n x $(K+4)$ design matrix $\mathbf{B}$ with $(i$,$k)$th entry B_{ik}=$B_k(x_i)$. Define the $(K+4)$ x $(K+4)$ penalty matrix $\boldsymbol{\Omega}$ with $(k$,$k')$th entry

$$\boldsymbol{\Omega}_{kk'} = \int_a^b B_k^{''}(x) B_{k'}^{''}(x) dx.$$

Then an estimate of $f(x)$, given a smoothing parameter λ>0, is given by

$$f_\lambda(x) = \mathbf{B}_x \hat{\boldsymbol{\nu}}, \text{ where } \hat{\boldsymbol{\nu}} = (\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{B}^T\mathbf{y}.$$

where $\mathbf{B}_x$=$[B_1(x), ..., B_{K+4}(x)]$. The computation of $\boldsymbol{\Omega}$ can be performed by matrix calculations, for more information see (Wand and Ormerod, 2008).

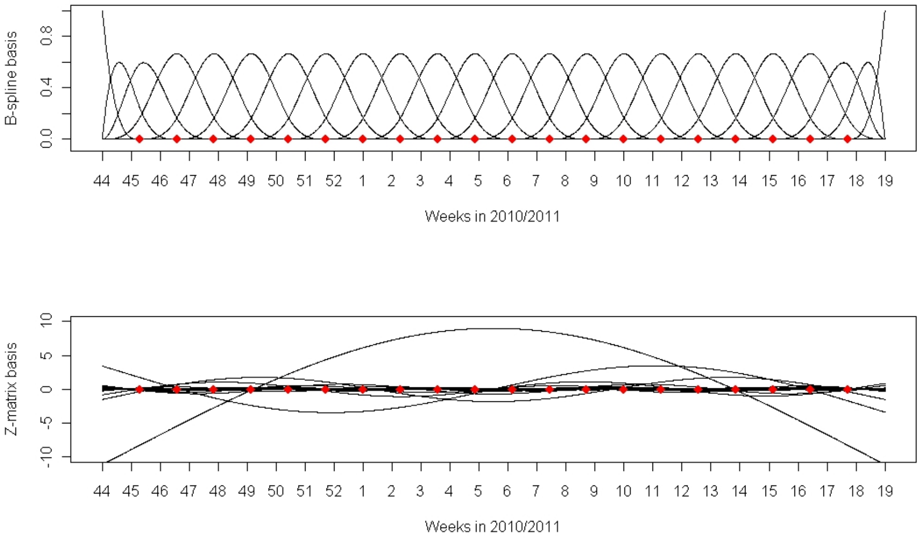

Figure 4.2: *Comparison of the B-spline basis, corresponding to* $\mathbf{B}$ *and corresponding to* $\mathbf{Z}$*, for the GIS data.*

The O'Sullivan penalized splines can also be presented in a mixed model formulation. For this purpose a linear transformation matrix $\mathbf{L}$ needs to be constructed, such that

$$\mathbf{L}^T\boldsymbol{\Omega}\mathbf{L} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

The method for obtaining $\mathbf{L}$ is spectral decomposition. The spectral decomposition of $\boldsymbol{\Omega}$ is of the form $\boldsymbol{\Omega}$=$\mathbf{U}$diag($\mathbf{d}$)$\mathbf{U}^T$, where $\mathbf{U}^T\mathbf{U}$=$\mathbf{I}$ and $\mathbf{d}$ is a vector with exactly two zero entries and $K+2$ positive entries, denote these $K+2$ positive entries as $\mathbf{d}_Z$. Let $\mathbf{U}_Z$ be the submatrix of $\mathbf{U}$ with columns corresponding to the positive entries of $\mathbf{d}$ and let $\mathbf{U}_X$ be the submatrix corresponding to the other two columns. Then the linear transformation matrix $\mathbf{L}$ is

$$\mathbf{L} = \left[\mathbf{U}_X|\mathbf{U}_Z\mathrm{diag}(\mathbf{d}_Z^{-1/2})\right].$$

The fixed and random effects design matrices are then

$$\mathbf{X} = \mathbf{B}\mathbf{U}_X \quad \text{and} \quad \mathbf{Z} = \mathbf{B}\mathbf{U}_Z\mathrm{diag}(\mathbf{d}_Z^{-1/2}).$$

It can be shown that $\mathbf{B}\mathbf{U}_X$ is the basis for the space of straight lines, so often $\mathbf{X}$=$[1\ x_i]_{1\leq i\leq n}$ is used instead, without affecting the fit. For more details see Wand and Ormerod (2008).

Figure 4.2 presents the B-spline basis, corresponding to $\mathbf{B}$ and corresponding to $\mathbf{Z}$ for the GIS data. Twenty equally spaced knots are placed, selected as quantiles of the time variable. It can be seen that there is damping of the $\mathbf{Z}$ matrix basis functions with increasing oscillation.

4.6 Model Selection, Inference and Confidence Intervals

Model selection is based on the popular Akaike's Information Criterion (Akaike, 1974). The smaller the AIC value, the better the model. Because the AIC is based on the likelihood value, Laplace approximation is used, because it yields likelihood values. The idea behind the AIC is to penalize the loglikelihood with the number of parameters, namely AIC=-2LL+2p with p the number of parameters in the model (fixed effects and variance components). The usage of the AIC in this form may not be appropriate in case of semiparametric models (Maringwa *et al.*, 2008c) and instead an adjusted AIC, abbreviated as AIC$_{adj}$, should be used. The penalty term of the AIC$_{adj}$ takes the effective number of parameters into account, which generally is higher than p because the smoothing is accounted for. Let $\mathbf{C}$=[$\mathbf{X}$ $\mathbf{Z}$] be the design matrix with the fixed effects and random effects corresponding to the penalized spline, then the effective number of parameters for the generalized model is (Ruppert *et al.*, 2003)

$$E_p = \text{trace}\left\{(\mathbf{C}^T\mathbf{W}\mathbf{C} + \frac{1}{2}\mathbf{\Lambda})^{-1}\mathbf{C}^T\mathbf{W}\mathbf{C}\right\},$$

where $\mathbf{W}$=var($\mathbf{y}|\mathbf{X},\mathbf{Z},\mathbf{u}$) and $\mathbf{\Lambda}$=diag$(0, 0, \frac{1}{\sigma_u^2}, ..., \frac{1}{\sigma_u^2})$ with the number of non-zero entries equal to the number of columns in $\mathbf{Z}$. The adjusted AIC is then given by AIC$_{adj}$= -2LL+2E_p.

Next to model selection, formal tests are required. Likelihood ratio test can be used to compare a model with the corresponding null model. For the use of likelihood ratio tests, again Laplace approximation is required. For tests involving fixed effects only, the appropriate chi-square distribution is one with degrees of freedom equal to the number of parameters in the null model. Testing for zero variance components is a non trivial situation, because one is on the boundary of the parameter space. Hence conventional chi-squared null distributions do not apply. In the case of testing the variance parameter controlling for the amount of smoothing in semiparametric regression, Crainiceanu *et al.* (2003) showed that the asymptotic theory of Stram and Lee (1994) for boundary problems does not apply. The appropriate finite sample and asymptotic distributions were derived by Crainiceanu and Ruppert (2004). For a more detailed exposition on this matter, see Crainiceanu *et al.* (2005).

Construction of bias-adjusted simultaneous confidence bands require the use of the following variance-covariance matrix (Ruppert *et al.*, 2003; SAS Insitute Inc. (2008))

$$\mathbf{V} = \text{Cov}\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{b}} - \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{S}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{S}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{S}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{S}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-}.$$

The matrix $\mathbf{S}$ is the conditional variance of the pseudo-data generated during the fitting with pseudo-likelihood. Evaluation of bias-adjusted confidence bands require the following standard deviation around a fitted value $\hat{f}(x)$

$$\widehat{st.dev}\left\{\hat{f}(x) - f(x)\right\} = \sqrt{\mathbf{C}_x\mathbf{V}\mathbf{C}_x^T}, \tag{4.8}$$

where $\mathbf{C}_x$=[$\mathbf{X}_x$ $\mathbf{Z}_x$] is the vector of fixed and random effects evaluated at x. For simultaneous confidence bands over a grid of M x-values $(x_1, ..., x_M)$, consider

$$\mathbf{f_x} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_M) \end{bmatrix},$$

and notice that it can be assumed that

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{b}} - \mathbf{b} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}). \tag{4.9}$$

Simultaneous confidence bands for $\mathbf{f_x}$ can then be obtained as

$$\left[\hat{f}(x_i) \pm h_{(1-\alpha)} \widehat{st.dev} \left\{ \hat{f}(x_i) - f(x_i) \right\}\right]_{1 \leq i \leq M}, \tag{4.10}$$

where the standard deviation is calculated in the same way as in (4.8) and $h_{(1-\alpha)}$ is the $1 - \alpha$ quantile of (Ruppert *et al.*, 2003)

$$\max_{1 \leq i \leq M} \left| \frac{\left(\mathbf{C_x} \begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{b}} - \mathbf{b} \end{bmatrix} \right)_{x_i}}{\widehat{st.dev} \left\{ \hat{f}(x_i) - f(x_i) \right\}} \right|. \tag{4.11}$$

To obtain $h_{(1-\alpha)}$ one simulates from (4.9) and computes (4.11) for N times. The value with rank $(1 - \alpha)N$ becomes $h_{(1-\alpha)}$. I shall consider N=10000. The construction of simultaneous confidence bands in the generalized case can be performed on the scale of the linear predictor and then transformed to the original scale of the response.

4.7 Analysis of the GIS

4.7.1 Sample

In chapter 3 it was shown that use of the the complete dataset, without exclusion of participants, yielded accurate results. For this reason this dataset is also used here. Although, some additional knowledge is used to complete the dataset in case of missing observations. Firstly, it is assumed that having had ILI, provides immunity for the rest of the ILI season. Consequently one can assume that if a participant reported having ILI that in the rest of the weeks he/she must have reported not having ILI, due to immunity. This is equivalent to assuming a SIR model. Using this assumption one can add information for each week for those participants that experienced ILI. Secondly, the fact that the symptoms questionnaires yields information from the 14 days before the date of the symptoms questionnaire is used. Suppose, for example, that for a participant the symptoms questionnaire of week i is available, then this can be used to yield information for week $i - 1$, if it would be missing. In this manner the number of symptoms questionnaires is increased from 85092 to 95180. Similar as in chapter 3 an analysis is done without and with post-stratification weighting of the sample. The weights are used to make the age and gender distributions of the GIS population similar to the age and gender distribution of the Flanders/Brussels population. Remember that the post-stratification weight for each participant is assigned according to the formula:

$$w_i = \frac{p_i^{F/B}}{p_i^{GIS}}$$

where, w_i is the weight of participant i, $p_i^{F/B}$ is the proportion of the Flanders/Brussels population in the same age and gender category as participant i and p_i^{GIS} is the proportion of the GIS population in the same age and gender category as participant i.

4.7.2 Models for the Trend

Using semiparametric models the overall trend of ILI in the season 2010/2011 is investigated, as well as differences in trends for different subgroups. In this manner risk factors for ILI can be identified. The interest in this thesis is to investigate differences in trends for vaccination status, gender, having asthma and/or diabetes, having one or more allergies (see table A1), living with at least one child, smoking and age. The time variable in the models, namely the week of the symptoms questionnaire, is modelled by using a penalized spline based on the truncated power basis of degree one and the 0'Sullivan basis. Firstly, each variable is investigated univariately. All variables , except for age, are dichotomous and five different models will be considered for these variables. These models, with hypothetical examples, are displayed in figure 4.3. In model 1 (panel A) the two groups of a variable have the same trend. This is equivalent to obtaining an overall trend of ILI for the whole GIS population. In model 2 (panel B) the trends for the groups differ only by a constant. In model 3 (panel C) the trends for both groups differ in their linear component but the smooth part of the trend is identical. In panel D the two groups show differences both in the linear and the smooth component of the fit. Two cases are considered for panel D, namely the smooth components of the two groups have the same level of smoothing (model 4) and a different level of smoothing (model 5). E.g., the linear component of model 5 for gender has the following form

$$\eta_{ij} = \beta_0 + \beta_1 G_i + \beta_2 W_{ij} + \beta_3 G_i W_{ij} + \sum_{k=1}^{K} u_k^M B_k(W_{ij}) + \sum_{k=1}^{K} u_k^F B_k(W_{ij}), \tag{4.12}$$

where G_i is the gender of participants i, W_{ij} is the time variable in week j, u_k^M and u_k^F, for k=1, ..., K , are the smooth effects for males and females having normal distributions with different variances and $B_k(W_{ij})$, for k=1, ..., K, represents the smooth basis function for the time variable. In model 4 the assumption is that u_k^M and u_k^F have a normal distribution with similar variance. Model 3 further assumes that u_k^M=u_k^F for each k. Model 2 is a further simplification of model 3 with β_3=0. In model 1, β_1 is also equal to zero.

For the influence of age on the ILI trend three univariate models are considered. In model 1 the age effect is neglected, so again an overall trend is obtained. In model 2, the age variable enters the model in a linear way, whereas in model 3 a penalized spline is used to model the effect of age. The O'Sullivan basis is used for this. The linear component of model 3 for age is

$$\eta_{ij} = \beta_0 + \beta_1 A_i + \beta_2 W_{ij} + \sum_{k=1}^{K} v_k B_k(A_i) + \sum_{k=1}^{K} u_k B_k(W_{ij}), \tag{4.13}$$

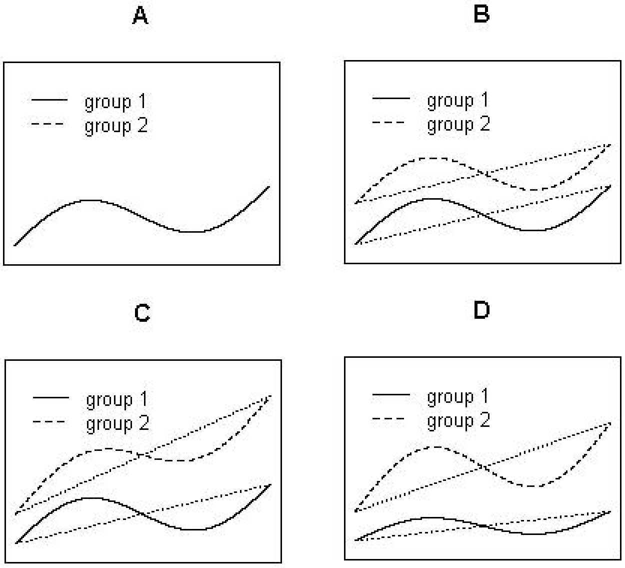

Figure 4.3: *Hypothetical examples of the semiparametric models. The models illustrate how the group specific curves could possible differ.*

where A_i is the age of participant i and v_k, for for k=1, ..., K, are the smoothing effects corresponding to the basis functions $B_k(A_i)$ for the age variable. For model 2, it is assumed that the variance corresponding to the v_k is zero. In model 1, β_1 is also assumed to be zero. An interaction between time and age was investigated using a tensor product basis (see Ruppert *et al.*, 2003), however, this model was computationally not feasible.

Besides the univariate analysis, a multiple semiparametric regression model is constructed. To this purpose the effects are entered into the model based on their corresponding best model in the univariate analysis. It is interesting to investigate whether the effect of several variables is dependent on the vaccination status, for this purpose an interaction effect of that variable with the vaccination status is considered.

In all models the logit link is used as link function. For reasons of obtaining enough flexibility, smoothing for the penalized splines, both for the time as age effect, is done with 20 equally spaced knots, selected as quantiles of the time and age variable. The models are fit using the GLIMMIX procedure in SAS. To perform likelihood ratio tests and to obtain the likelihood based information criteria AIC and AIC_{adj}, the Laplace estimation technique for GLMMs is used. To construct simultaneous confidence bands pseudo-likelihood estimation is used.

4.7.3 Results

The overall incidence trend of ILI (model 1) is shown in figure 4.4, together with the 95% simultaneous confidence bands, for the truncated power basis of degree one and the O'Sullivan basis. It can be seen that the trend using the O'Sullivan basis is more smooth than the truncated power basis, especially when considering the confidence bands. The confidence bands from the O'Sullivan basis are very smooth, whereas the bands of the truncated power basis show bumps. There were problems with convergence using

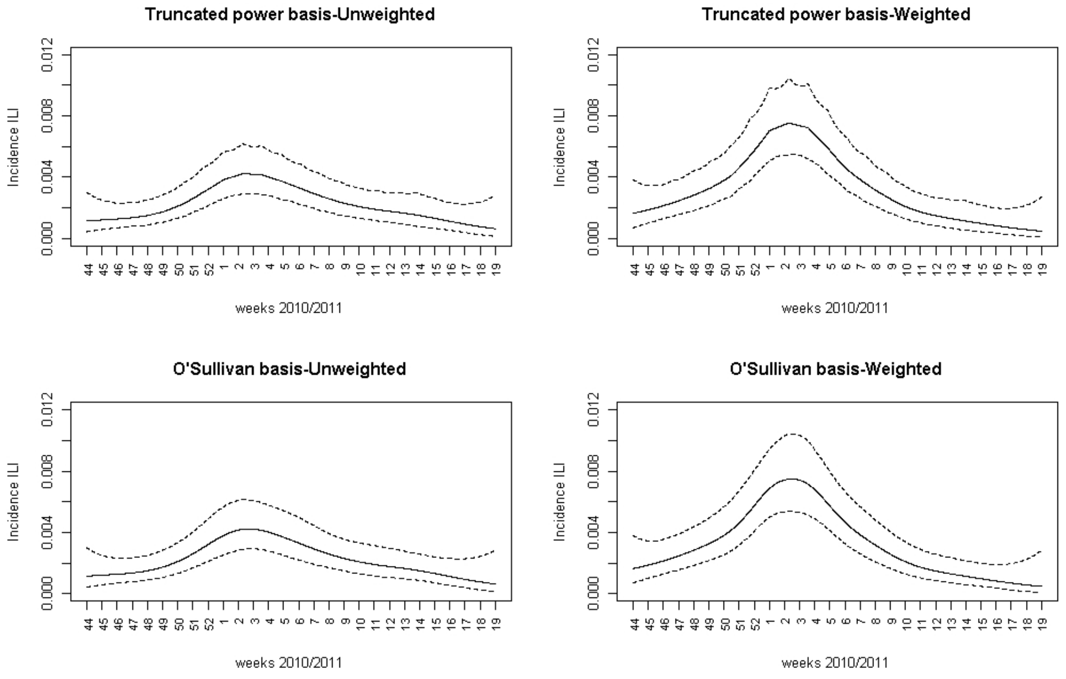

Figure 4.4: *Estimated overal trend of ILI, together with 95% simultaneous confidence bands (dashed line) using the truncated power basis of degree one and the O'Sullivan basis, both for the unweighted and weighted dataset.*

the truncated power basis and some differences in estimates were obtained between the Laplace and pseudo-likelihood estimation. Using the O'Sullivan basis these problems were not encountered. For this reason, the results that follow are based on the analysis using the O'Sullivan basis. The overal trend using the weighted dataset is again higher, for similar reasons as described in chapter 3.

The results of the univariate analysis for the unweighted dataset are presented in table 4.1 and for the weighted dataset in table 4.2. Firstly, the results of the unweighted dataset are discussed. In table 4.1, it can be observed that a separate smoothing effect (model 4 and model 5) for the different subpopulations under consideration is not necessary. The difference between the AIC and AIC_adj can for example be observed for the gender variable. Using the marginal AIC it seems that using model 4 yields a better fit than model 1, however, when correcting for the effective degrees of freedom, this is not case anymore. It can be observed that for each dichotomous variable, except for having asthma and/or diabetes, model 2 yields the best fit in terms of AIC_{adj}. Note that, in this case, the same models are selected when based on the AIC. These models are selected as the best fitting models and are compared with model 1 using a likelihood ratio test. In this manner it can be tested if their is a significant difference in trends between the subpopulations. All the likelihood ratio test statistics follow asymptotically a chi-square distribution with one degree of freedom. On the 5% significance level the effects of gender (p=0.0029), allergy (p=0.0004), living with at least one child (p=0.0374) and smoking (p=0.0302) are all significant. Vaccination status (p=0.0581) has a small marginal effect. An increased risk is observed for females with an odds ratio of 1.50 [1.15 - 1.97], not having allergies

Table 4.1: *Results of the univariate analysis for the GIS using the unweighted dataset. Minus twice the log likelihood values, the marginal AIC, the adjusted AIC (AIC_{adj}) and the effective number of parameters E_p are presented for the models.*

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*
Vaccination Status					
$-2LL$	2997.61	2994.02	2994.00	2996.97	2995.64
AIC	3003.61	3002.02	3004.00	3006.97	3007.64
AIC_{adj}	3010.41	$\boxed{3008.45}$	3010.43	3017.72	3017.41
E_p	6.40	7.22	8.22	10.38	10.88
Gender					
$-2LL$	2997.61	2988.75	2987.59	2991.33	2990.79
AIC	3003.61	2996.75	2997.59	3001.33	3002.79
AIC_{adj}	3010.41	$\boxed{3003.18}$	3004.04	3011.39	3010.62
E_p	6.40	7.22	8.23	10.03	9.92
Asthma or/and Diabetes					
$-2LL$	2997.61	2996.68	2996.09	2999.85	2999.65
AIC	3003.61	3004.68	3006.09	3009.85	3011.65
AIC_{adj}	$\boxed{3010.41}$	3011.10	3012.52	3019.23	3018.92
E_p	6.40	7.21	8.21	9.69	9.63
Allergies					
$-2LL$	2997.61	2984.90	2984.36	2987.82	2987.10
AIC	3003.61	2992.90	2994.36	2997.82	2999.10
AIC_{adj}	3010.41	$\boxed{2999.32}$	3000.79	3008.50	3007.37
E_p	6.40	7.21	8.21	10.34	10.13
At Least One Child at Home					
$-2LL$	2997.61	2993.28	2992.69	2997.11	2997.07
AIC	3003.61	3001.28	3002.69	3007.11	3009.07
AIC_{adj}	3010.41	$\boxed{3007.71}$	3009.13	3017.70	3018.04
E_p	6.40	7.21	8.22	10.29	10.49
Smoking					
$-2LL$	2997.61	2992.91	2991.06	2994.78	2994.67
AIC	3003.61	3000.91	3001.06	3004.78	3006.67
AIC_{adj}	3010.41	$\boxed{3007.34}$	3007.50	3015.12	3015.02
E_p	6.40	7.21	8.22	10.17	10.17
Age					
$-2LL$	2997.61	2942.89	2942.57		
AIC	3003.61	2950.89	2952.57		
AIC_{adj}	3010.41	$\boxed{2957.54}$	2961.39		
E_p	6.40	7.32	9.41		

Table 4.2: *Results of the univariate analysis for the GIS using the weighted dataset. Minus twice the log likelihood values, the marginal AIC, the adjusted AIC (AIC_{adj}) and the effective number of parameters E_p are presented for the models.*

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*
			Vaccination Status		
$-2LL$	3827.37	3811.88	3811.13	3805.88	3802.77
AIC	3833.37	3819.88	3821.13	3815.88	3814.77
AIC_{adj}	3841.18	$\boxed{3827.54}$	3828.79	3829.39	3832.85
E_p	6.90	7.83	8.83	11.76	15.04
			Gender		
$-2LL$	3827.37	3827.35	3823.15	3827.10	3823.74
AIC	3833.37	3835.35	3833.15	3837.10	3835.74
AIC_{adj}	3841.18	3842.99	$\boxed{3840.84}$	3854.82	3865.79
E_p	6.90	7.82	8.85	13.86	21.03
			Asthma or/and Diabetes		
$-2LL$	3827.37	3825.87	3820.60	3824.91	3824.90
AIC	3833.37	3833.87	3830.60	3834.91	3836.90
AIC_{adj}	3841.18	3841.52	$\boxed{3838.28}$	3845.83	3845.81
E_p	6.90	7.83	8.84	10.46	10.45
			Allergies		
$-2LL$	3827.37	3815.92	3815.91	3818.48	3815.57
AIC	3833.37	3823.92	3825.91	3828.48	3827.57
AIC_{adj}	3841.18	$\boxed{3831.54}$	3833.53	3851.32	3847.83
E_p	6.90	7.82	8.81	16.42	16.13
			At Least One Child at Home		
$-2LL$	3827.37	3778.54	3778.13	3768.67	3760.31
AIC	3833.37	3786.54	3788.13	3778.67	3772.31
AIC_{adj}	3841.18	$\boxed{3794.19}$	3795.79	3814.60	3798.90
E_p	6.90	7.82	8.83	22.96	19.30
			Smoking		
$-2LL$	3827.37	3824.60	3824.59	3827.34	3827.33
AIC	3833.37	3832.60	3834.59	3837.34	3839.33
AIC_{adj}	3841.18	$\boxed{3840.26}$	3842.25	3851.51	3851.69
E_p	6.90	7.83	8.83	12.09	12.18
			Age		
$-2LL$	3827.37	3687.41	3643.82		
AIC	3833.37	3695.41	3653.82		
AIC_{adj}	3841.18	3703.14	$\boxed{3682.75}$		
E_p	6.90	7.86	19.46		

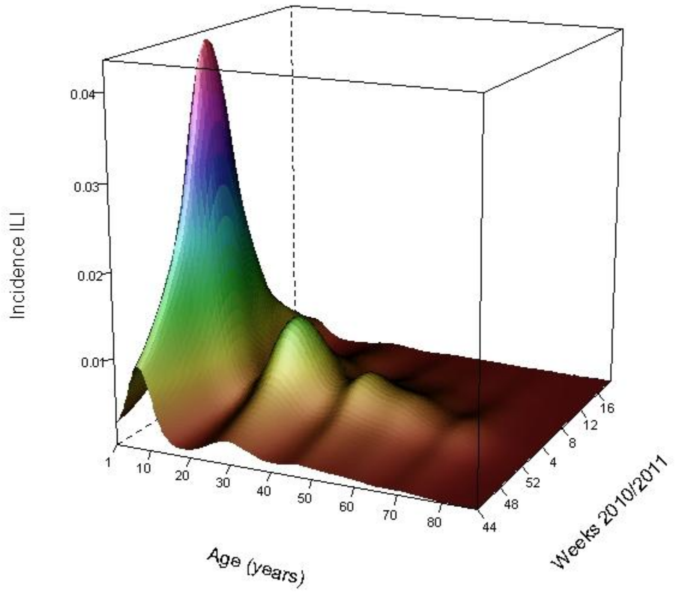

Figure 4.5: *Three-dimensional representation of the influence of age on the incidence of ILI for the influenza season 2010/2011.*

reduces the risk of ILI with an odds ratio of 0.58 [0.44 - 0.78] and smoking increases the risk with an odds ratio of 1.48 [1.05 - 2.07]. Living with at least one child increases the odds by 1.37 [1.02 - 1.83] and being vaccinated reduces the risk by an odds ratio of 0.75 [0.55 - 1.01]. From table 4.1 it can be observed that, by comparing model 2 with model 3 for age, no spline for the age variable is necessary. Therefore, the effect of age is linear (p <0.0001). When age increases with one year the odds decrease by 0.97 [0.96 - 0.98].

The univariate analysis for the weighted dataset is summarized in table 4.2. There is clearly a difference between the model selection as compared to the unweighted analysis. The difference between AIC and AIC_{adj} can be clearly illustrated for vaccination status. Using the marginal AIC model 5 would be depicted, using AIC_{adj} model 2 is shown as the best model under consideration. Also for allergies, living with at least one child and smoking, model 2 is chosen as being best based on AIC_{adj}. For gender and having asthma and/or diabetes model 3 is chosen. Again it can be formally tested whether there is a difference between the subgroups by comparing the best chosen model with model 1. The effects of vaccination status and living with at least one child become very significant in this case (both have p <0.0001). The effect of gender (p=0.1212) and smoking (p=0.0960) become unsignificant on the 5% significance level. The effects of allergies (p=0.0007) and having asthma and/or diabetes (p=0.0339) stay significant. The odds on ILI decrease when vaccinated for influenza by 0.56 [0.41 - 0.76]. Living with at least one child increase the odds by 2.28 [1.81 - 2.87]. Having allergies, has increased effect by an odds ratio of 1.54 [1.21 - 1.97]. Because the effect of asthma and/or diabetes differs over time, no single odds ratio can be provided. From table 4.2 it is clear that a spline for age is required. The effect of age on the incidence of ILI is represented in figure 4.5. The risk is highest for young children, aged 3-10 years. A second peak is observed around the age of 25 to 35 years. This can be explained by the fact that individuals in this age group are often parents of young children and due to contacts with their children have a higher risk for ILI.

Table 4.3: *Result of the multiple semiparametric logistic regression model for the analysis of the GIS. The odds ratio (95% confidence intervals) are given for the covariates in the model.*

	Unweighted	*Weighted*
Having any allergies	1.49 [1.12 - 1.99]	1.40 [1.09 - 1.80]
Smoking	1.60 [1.14 - 2.25]	1.96 [1.39 - 2.78]
Age (increase of 1 year)	0.97 [0.96 - 0.98]	
$-2LL$	2929.36	3624.43
AIC	2941.36	3638.43
AIC_{adj}	2947.57	3666.68
E_p	9.10	21.13

A least small peak is observed around 50 to 70 years, which could possible be explained by the contacts between young children and grandparents.

A multiple semiparametric logistic regression model was also constructed. Both for the unweighted and weighted analysis interactions between risk factors and vaccination status were insignificant on the five percent level. For this reason risk factors entered the model based on the best fitting model in the univariate case. Using likelihood ratio tests model building was performed. The final model only included allergy status, smoking status and age, as these were found to be significant in the model building steps. In the unweighted case, age is incorporated linearly and in the weighted case using a smoothing spline. The results of the analysis can be found in table 4.3. Having any allergies and smoking increases the risk on ILI incidence. The form of the spline for age in this multivariate weighted analysis in this case is almost identical as presented in figure 4.5.

4.7.4 Model Extensions

A first model extension that is considered, is the use of a R-side covariance effect. In this manner, it is possible to model an association between observations from one week and those of another week. A first-order autoregressive, AR(1), structure is considered, which recognizes that observations which are more proximate, in terms of time, are more correlated than observations that are more distant. The final multiple semiparametric logistic regression model using the weighted dataset, discussed in the previous section, is fitted using this AR(1) covariance structure. It was found that the association in this model was only minor, with a correlation equal to 0.06558. Also, in terms of the estimated trend as well as the estimated effect of risk factors, no quantitatively difference is observed.

A second model extension is considered by the inclusion of province specific random intercepts in the final multiple semiparametric logistic regression model using the weighted dataset. This gives the possibility to investigate in which province ILI incidence was higher/lower in the season 2010/2011, however no formal test is considered here. The region of Brussels and Vlaams-Brabant are merged together, because the Brussels region has too few observations to obtain a reliable estimate. Figure 4.6 shows the estimated trend for each province. It can be seen that four provinces are clustered together and

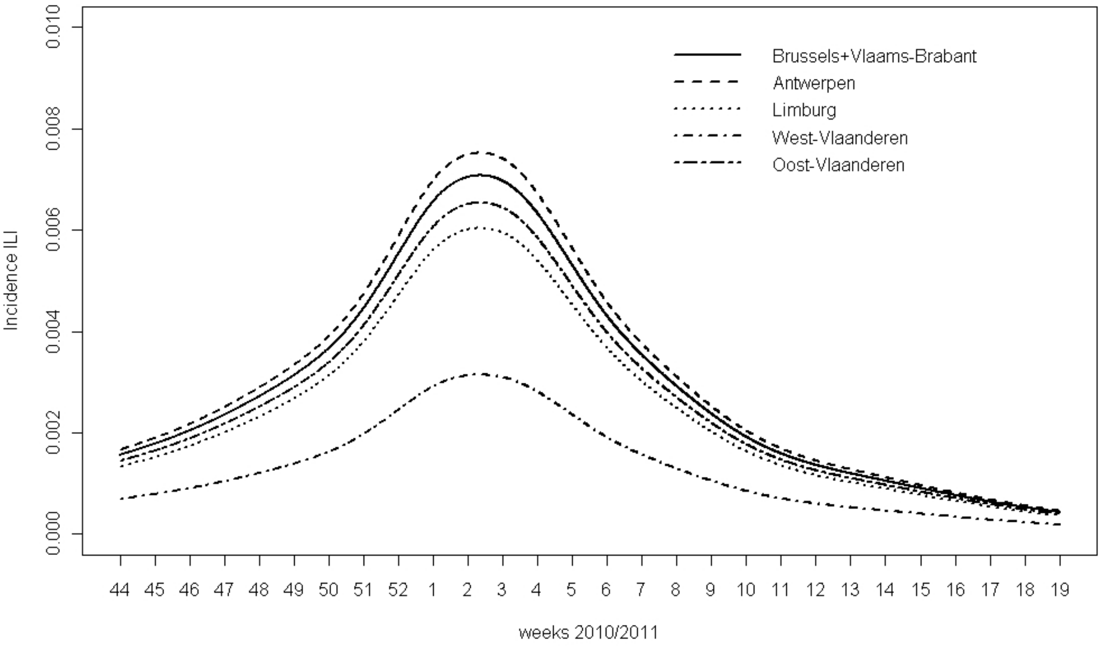

Figure 4.6: *Estimated trend for each province, by the incorporation of province specific random intercepts. The other effects are held fixed at no smoking, no allergies and age equal to 35 years.*

the province of West-Vlaanderen has a lower incidence trend of ILI. The incidence is highest for the provinces of Antwerpen and Vlaams-Brabant(+ Brussels) and lowest for the provinces of Limburg and West-Vlaanderen. A possible explanation for this could be that the population density per square kilometer is highest for the first two provinces and lowest for the latter two. In this model each province had the same trend over time and difference was only due to a random intercept. A further extension is to consider a model in which the trend for each province is modelled separately, however this model would be very intensive in terms of computation.

4.8 Discussion

This chapter dealt with the estimation of influenza-like-illness trends using generalized semiparametric logistic regression models. Using penalized splines the trend was modelled to allow for a flexible shape. The connection with the generalized linear mixed model framework allows for convenient fitting of these models using widely available commercial software. The considered models can equally well be fitted within the Bayesian framework (see Ruppert *et al.*, 2003) and consequently, WinBUGS, a freely available software package can be used. Using the O'Sullivan spline basis, numerically more stable results were obtained. This is due to the fact that this set of basis functions is bounded, whereas this is not the case for the truncated power basis. The fit and simultaneous confidence bands are less smooth using the truncated power basis of degree one than the O'Sullivan basis. The problem with smoothness can be overcome by using the truncated power basis of degree two or higher, although these bases suffer from great numerical instability. Analysis was based with and without post-stratification weighting the sample based on the age and gender distribution.

Estimating the overall trend of ILI yielded comparable results as obtained in chapter 3, both for the weighted and unweighted analysis, in the sense that that the trend followed the same pattern as the EISN data. The Pearson correlation between the overall trend based on the weighted analysis and the EISN data is 0.937 (95% CI: 0.867 - 0.971). The peak of ILI incidence was in the first few weeks of 2011, with incidences of around 400 individuals per 100000 individuals in the unweighted case and around 750 individuals per 100000 individuals in the weigthed case. In comparison with the analysis of chapter 3, the trend is much smoother in this case. From figures 3.4 to 3.5 it was clear that the trend based on the GIS, showed an elevation of ILI in the beginning of the season. Using penalized splines this elevation has disappeared, which is comparable to the EISN data. The dissappearance of this elevation can be explained by the fact that the same smoothing is used throughout the time frame. If seperate smoothing is needed dependent on the location in the time frame, local penalty smoothers can be used (Ruppert *et al.*, 2003). The analysis using the weighted dataset yielded an higher overall trend for ILI. This is explained by the fact that children and young adults are underrepresented and ILI incidence is higher in these subgroups.

The univariate analysis revealed different risk factors for ILI incidence based on the unweighted and weighted analysis. For the weighted analysis the following risk factors were found to be significant on the five percent level: vaccination status for influenza, having asthma and/or diabetes, having at least one allergy (hay fever, dust mite allergy and allergy for pets), living with at least one child and age. Smoking was significant on the ten percent level. Similar results were found in the unweighted case, however gender was significant there and having asthma and/or diabetes not. These results are consistent with existing ILI literature, see for example van Noort *et al.* (2011) for a discussion. The multiple semiparametric regression model yielded the significance of only the risk factors: allergies, smoking and age. Where having any allergies (hay fever, dust mite allergy and allergy for pets) and smoking increase the risk on ILI incidence. In the unweighted analysis ILI risk decreases linearly with age. In the weighted case the influence of age is modelled through a penalized spline and the following interesting result was found: The first and highest peak for ILI incidence was observed for children aged 4 to 10 years. A second peak was observed for adults aged from 25 to 35 years. This could be explained by the fact that individuals in this latter age group often have young children and due to their extensive contacts with young children, acquire more risk on ILI. A third, but smallest, peak was between the ages of 50 to 70 years, which could be explained by the contacts of grandparents with their grandchildren. This explanation is based on the fact that in contact surveys, often similar patterns are observed (see for example Hens *et al.*, 2009). It may seem contra intuitive that elders do not have a higher risk of obtaining ILI, however this thinking may stem from the result that elderly are at a greater risk for complications, not for infection (Monto, 2004). van Noort *et al.* (2011) discuss that this result can also be explained by the possibility that for elderly, participating in the GIS, the risk of ILI is not independently associated with internet use, resulting in strong sampling bias.

These results are only based on the GIS data from the season 2010/2011, to confirm the obtained results it would be very interesting to investigate the previous influenza seasons. By taking up subject specific random effects into the model, the fact that not every participant is equally susceptible for ILI can be incorporated. However, this approach was computationally not feasible in our case. An interaction between the age and time variable was computationally not feasible based on the product tensor basis. Another basis which is often used to model interactions, namely the radial smoothing basis (Ruppert *et al.*, 2003), could be used instead. The results presented in chapter 4.7 were based on the use of the logit link function. As a sensitivity analysis the complementary log-log link was also used. The obtained results with this link function did not differ quantitatively.

In the analysis model 4 and model 5 were never chosen as being optimal. This seems reasonable when investigating only one influenza season. However, model 4 and model 5 should certainly not be depreciated, as they could be very useful to compare ILI trends over several years or between different countries for example. The presented methodology is certainly not limited to be used for ILI trends, but it could be applied for many diseases for which an estimate of the trend is needed.

Chapter 5

Conclusion and Further Research

5.1 Conclusion

This thesis handled the analysis of the Great Influenza Survey (GIS) of the influenza season 2010/2011 in Belgium, restricting to the Flanders and Brussels region. This kind of analysis was not yet performed in detail for Belgium. The GIS attracts many volunteers that participate more than three times (89 %). The GIS population is not similarly distributed with respect to age and gender as the Flanders/Brussels region. Using several assumptions ILI trends based on the GIS were compared with Belgium Sentinel data, coming from the EISN surveillance system. Pearson and cross correlation coefficients were used to quantify the similarity between the GIS and the EISN data. Strong associations were observed between the two trends. There is evidence that the incidence of ILI in the GIS is monitored 1-2 weeks ahead of the EISN system. After correcting for the demographic bias, coming from the age and gender distribution of the GIS population, by appropriately post-stratification weighting of the data, such that the age and gender distribution is similar to the age and gender distribution in the Flanders/Brussels region, even better associations were observed. During the influenza season 2010/2011, the ILI incidence was highest in the first few weeks of 2011 with estimated incidences of around 400 (unweighted analysis) and around 800 (weighted analysis) individuals per 100,000 indivduals.

Semiparametric logistic models were used to estimate the overall trend of ILI incidence, as well as for several subgroups of the population. The time variable was modelled using penalized splines in order to flexibly capture the trend over time. It was shown how the semiparametric models cast into the mixed model framework, which facilitates fitting and inference using the standard methodology of mixed models. It was observed that estimation using O'Sullivan penalized splines was computationally more stable than using the truncated power basis of degree one. The estimated trend using the latter basis was less smooth, especially in terms of simultaneous confidence bands. The overall trend based on the semiparametric models showed great similarities with that of the EISN system. Based on univariate analysis the following factors were observed to increase the risk on ILI incidence: not being vaccinated for seasonal influenza, having asthma and/or diabetes, having one or more allergies (hay fever, dust mite allergy and allergy for pets), living with at least one child and smoking. The difference in trends for these factors was only by an additive effect in the linear component. This implicates that the different ILI trends only differ in their relative heights, but not in their time points of rising and declining. This seems reasonable when considering only one influenza season. The influence of age was modelled through a penalized spline. It was found that young children (4-10 years) have the highest risk to obtain ILI. A second peak of ILI incidence risk was found for adults aged 25-35 years and a last, but smaller, peak for elderly aged 50-70 years. This result could possibly be explained by the contacts between parents and grandparents with their children and grandchildren respectively. Based on a multiple semiparametric model, only the risk factors allergies, smoking and age were identified.

The Great Influenza Survey offers a good surveillance system in Belgium that yields comparable results for the incidence of ILI as compared to the trend coming from the Belgium Sentinel practice. The advantage of having individual data in the GIS can be exploited to estimate ILI trends for different subgroups. The semiparametric regression approach yields a well established framework for this, where inference and fitting can be done using widely availabe commercial software.

5.2 Further Research

It would be interesting to compare the GIS trend with data of Google Flu Trends, see for example van Noort *et al.* (2011). Using semiparametric models it can then be formally tested if there is a difference in trends between these two surveillance systems. As was discussed in chapter 3 and 4, analyzing more influenza seasons would be necessary to confirm or reject the obtained results. If indeed more years are under investigation, a multivariate analysis of those participants participating every year could be performed. In this manner it would be possible to investigate whether some individuals are more inclined to have recurrent episodes of ILI throughout the different seasons.

In this thesis only the temporal aspect of ILI trends was analysed. It would be interesting to perform a spatio-temporal analysis for ILI. In this manner it could be investigated whether ILI trends occur randomly in the area under investigation or whether their is a spatial trend. For example, a spatio-temporal analysis of influenza and norovirus using telehealth data of the United Kingdom is already performed by Cooper *et al.* (2008). When analyzing the spatio-temporal distributions of ILI trends, it is of interest to known whether preferential sampling occurs. Preferential sampling arises when the process that determines the data locations and the process being modelled are stochastically dependent (Diggle *et al.*, 2010). For ILI, it is not unlikely that more GIS participants fill in the symptoms questionnaire in a region where ILI incidence is higher. If preferential sampling is present, this should be taken up in the analysis. This is a topic which I will investigate in the next few years.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

Brumback, B., Ruppert, D. and Wand, M.P. (1999). Comment on "Variable selection and function estimtion in additive nonparametric regression using data-based prior" by Shively, Kohn and Wood. *Journal of the American Statistical Association*, **94**, 794-797.

Cooper, D.L., Smith, G.E., Regan, M., Large, S. and Groenewegen, P.P. (2008). Tracking the spatial diffusion of influenza and norovirus using telehealth data: A spatiotemporal analysis of syndromic data. *BMC Medicine*, **6(16)**.

Costa M.J., (2008). Penalized spline models and applications. *University of Warwick, Ph.D. Thesis.*

Crainiceanu, C.M., Ruppert, D. and Vogelsang, T.J. (2003). Some properties of likelihood ratio tests in linear mixed models. Avaialble at www.orie.cornell.edu/∼davidr/papers.

Crainiceanu, C.M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B*, **66**, 165-185.

Crainiceanu, C.M., Ruppert, D., Claeskens, G. and Wand, M.P. (2005). Exact likelihood ratio tests for penalized splines. *Biometrika*, **92**, 91-103.

Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction.* Oxford: Clarendon.

Diggle, P.J., Menezes, R. and Su, T.-L. (2010). Geostatistical inference under preferential sampling. *Applied Statistics*, **59(2)**, 191-232.

ECDC, European Centre for Disease Prevention and Control (2009). *Annual meeting of the European Influenza Surveillance Network.* Meeting Report: Stockholm.

ECDC, European Centre for Disease Prevention and Control (2010). *Main surveillance developments in week 40/2010 (04 10 October 2011).* Surveillance report, weekly influenza surveillance overview: Stockholm.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89-121.

Flahault, A. (2006). Global monitoring of influenza: potential contribution of national networks from a French perspective. *Expert Review of Anti-Infective Therapy*, **4(3)**, 387-393.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1-67.

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 populations: experience of five influenza seasons in the Netherlands. *Vaccine*, **27**, 6353-6357.

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

Gisle, L., Hesse, E., Drieskens, S., Demarest, S., Van der Heyden, J. and Tafforeau, J. (2010). *Gezondheidsenquête België, 2008. Rapport II Leefstijl en Preventie.* Brussel: Wetenschappelijk Instituut Volksgezondheid.

Gosling, S.D., Vazire, S., Srivastava, S. and John, O.P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, **59**, 93-104.

Harder, K.M., Andersen, P.H., Baehr, I., Nielsen, L.P., Ethelberg, S., Glismann, S. and Molbak, K. (2011). Electronic real-time surveillance for influenza-like illness: experience from the 2009 influenza A(H1N1) pandemic in Denmark. *Euro Surveillance*, **16(3)**, pii=19767.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning.* New York, NY: Springer.

Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Damme, P. and Beutels, P. (2009). Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC Infectious Diseases*, **9(5)**.

Jansen, A.G.S.C., Sanders, E.A.M., Hoes, A.W., Van Loon, A.M. and Hak, E. (2007). Influenza- and respiratory syncycial virus-associated mortality and hospitalisations. *European Respiratory Journal*, **30(6)**, 1158-1166.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, **19(2)**, 81-97.

Maringwa, J.T. (2008a). Flexible modelling techniques and use of historical controls in animal studies. *Hasselt University, Ph.D. Thesis.*

Maringwa, J.T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Van Ammel, K., Teisman, A. and Bijnens, L. (2008b). Analysis of cross-over designs with serial correlation within periods using semi-parametric mixed models. *Statistics in Medicine*, **27(28)**, 6009-6033.

Maringwa, J.T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Van Ammel, K., Teisman, A. and Bijnens, L. (2008c). Application of semi-parametric mixed models and simultaneous confidence bounds in a cardiovascular safety experiment with longitudinal data. *Journal of Biopharmaceutical Statistics*, **18(6)**, 1043-1062.

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**.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* New York: Wiley.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models.* New York: Wiley.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Monto, A.S. (2004). Occurrence of respiratory virus: time, place and person. *The Pediatric Infectious Disease Journal*, **23**, 58-64.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of* *the Royal Statistical Society, Series A*, **135**, 370-384.

Ngo, L. and Wand, M.P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, **9**, 1-56.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, **1**, 505-527.

Raudenbush, S.W., Yang, M.-L. and Yosef, M. (2000). Maximum likelihood estimation for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, **9**, 141-157.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735-757.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric regression.* Cambridge: Cambridge University Press.

SAS Institute Inc. (2008). *The GLIMMIX Procedure.* Cary, NC: SAS Institute Inc.

Simonsen, L., Clarke, M.J., Williamson, G.D., Stroup, D.F., Arden, N.H. and Schonberger, L.B. (1997). The impact of influenza epidemics on mortality: introducing a severity index. *American Journal of Public Health*, *87(12)*, 1944-1950.

Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171-1177.

Tilston, N.L., Eames, K.T.D., Paolotti, D., Ealden, T. and Edmunds, J.W (2010). Internet-based surveillence of influenza-like-illness in the UK during the 2009 H1N1 influenza pandemic. *Public Health*, **10**.

Truyers, C., Lesaffre, E., Bartolomeeusen, S., Aertgeerts, B., Snacken, R., Brochier, B., Yane, F. and Buntinx, F. (2010). Computerized general practice based networks yield comparable performance with sentinel data in monitoring epidemiological time-course of influenza-like illness and acuter respiratory illness. *BMC Family Practice*, **11(24)**.

Van der Heyden, J., Gisle, L., Demarest, S., Drieskens, S., Hesse, E. and Tafforeau, J. (2010). *Gezondheidsenquête België, 2008. Rapport I - Gezondheidstoestand.* Brussel: Wetenschappelijk Instituut Volksgezondheid.

van Noort, S.P., Muehlen, M., Rebelo de Andrade, H., Koppeschaar, C.E., Lima Lourenco, J.M. and Gomes, M.G.M. (2007). Gripenet: An internet-based system to monitor influenza-like illness uniformly across Europe. *Euro Surveillance*, **12(7)**, pii=722.

van Noort, S.P., Codeco, C.T., Koppeschaar, C.E., van Ranst, M., Faustino, V. and Gomes, M.G.M. (2011). The Influenzanet self-reporting system warrants consistency monitoring across countries and seasons (Peer review). *Emerging Infectious Diseases*, unpublished.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Wand, M.P. and Ormerod, J.T. (2008). On semiparametric regression with O'Sullivan penalized splines. *The Australian and New Zealand Journal of Statistics*, **50(2)**, 179-198.

Wolfinger, R.D. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **4**, 233-243.

Zeger, S.L., Irizarry, R. and Peng, R.D. (2006). On time series analysis of public and biomedical data. *Annual Reviews of Public Health*, **27**, 57-79.

Appendix

A Questionnaires of the GIS

Table A1: *The intake questionnaire. Each participant is asked to complete this questionnaire upon registration for each influenza season. Where appropriate, participants can select multiple answers.*

Question	*Answer*
Postal code	
Birth data	
Sex	1.Male, 2.Female
Daily occupation	1.School, 2.Work, 3.Home, 4.Reformed, 5.Other
Daily means of transport	1.Bicycle, 2.Motorcycle, 3.Car, 4.On foot, 5.Public transport
How many colds per year?	1.Less than two, 2. Between two and five, 3. More than five
Did you receive a flu vaccine for the current season?	1.Yes, 2.No
Reason for vaccination	1. GP recomendation, 2.To protect me, 3.To protect other, 4. Part of a risk group, 5. Company vaccination
Reason for no vaccination	1.GP recomendation, 2.No protection, 3.I will get the flu if I take it, 4. Side effects, 5. Will get later 6. Not part of risk group
Chronic diseases	1.Asthma, 2.Diabetes, 3.Heart disease, 4.Kidney disorder, 5.Auto-immune
Allergies	1.Hay fever, 2.Dust mite allergy, 3.Allergy for pets
Smoking habits	1.Daily, 2.Sometimes, 3.Never
Fruit and vegetable intake	1.Regularly, 2.Rarely, 3.Hardly ever
Vitamin supplement intake	1.Regularly, 2.Rarely, 3.Hardly ever
Do you follow a diet?	1.Vegatarian, 2.Vegenistic, 3.Low calorie, 4.Other
Hours of sport per week	1.Less than one, 2.Between one and four, 3. More than four
Household characterization	1.Alone, 2.Only adults, 3.With children
Where do the children spend most of their day?	1.Home, 2.Nursery, 3.School
Pets at home?	1.Cats, 2.Dogs, 3.Birds, 4.Other

Table A2: *The symptoms questionnaire. Each participant is reminded weekly to complete this questionnaire. Where appropriate, participants can select multiple answers.*

Question	*Answer*
Did you experience any of the following symptoms since yourlast vist?	1.No symptoms, 2.Cough, 3.Running nose, 4.Headache, 5.Sore throat, 6.Chest pain, 7.Muscle pain, 8.Diarrhea, 9.Abdominal pain, 10.Cold shivers, 11.Nausea, 12.Irritated eyes, 13.Vomitting
When did the symptoms start?	
Body temperature	Smaller than 37°C/ Between 37°C-40°C in steps of 0.5°C
When did the fever start?[a]	
Did the fever start abruptly?[a]	1.Yes, 2.No, 3.Don't know
Did you see a GP?[b]	1.Yes, 2.No
Did you have to alter your daily routine?[b]	1.Yes, I stayed home, 2.Yes, but I went to work/school, 3. No, 4. Stayed at home but worked at home
If you stayed at home, how many days?[b]	

[a] Only asked when body temperature is higher than 38°C.

[b] Only asked when symptoms or fever is present.

B Demographics of the GIS Population

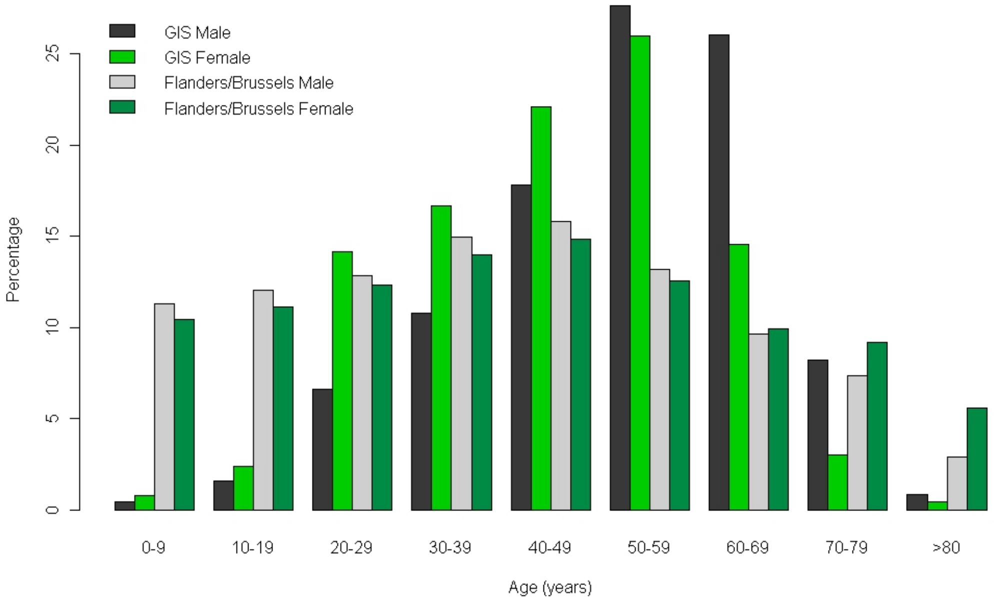

Figure B1: *Age distribution of the male and female GIS population and the Flanders/Brussels population in age groups of 10 years.*

Table B1: *Vaccination rates for influenza for certain age groups of the GIS and Belgium population.*

Age group	*GIS*	*Belgium*
0-18 years	2.88%	1.24%[a]
19-49 years	31.21%	11.40%
50-64 years	49.46%	27.90%
65-84 years	74.06%	63.50%
$\geq$ 85 years	85.71%	74.60%

[a] By weighting the agegroups 0-23 months, 2-5 years, 6-12 years and 13-18 years.

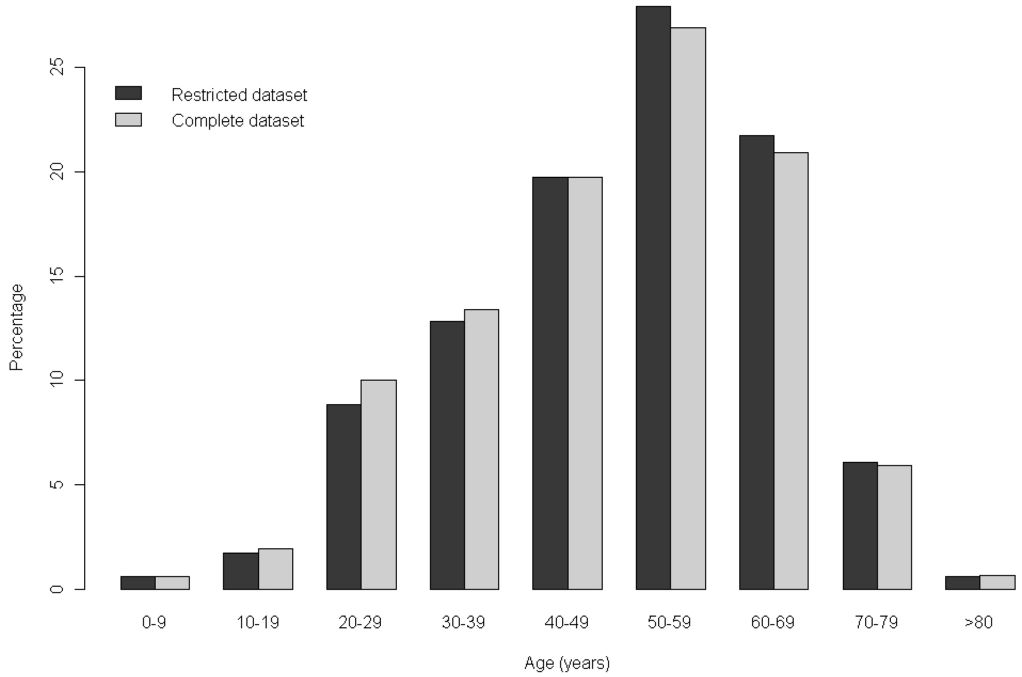

Figure B2: *Age distribution, in age groups of 10 years, of the GIS participants in the restricted dataset and in the complete dataset.*

C Age and Gender Distribution Throughout the Influenza Season 2010/2011

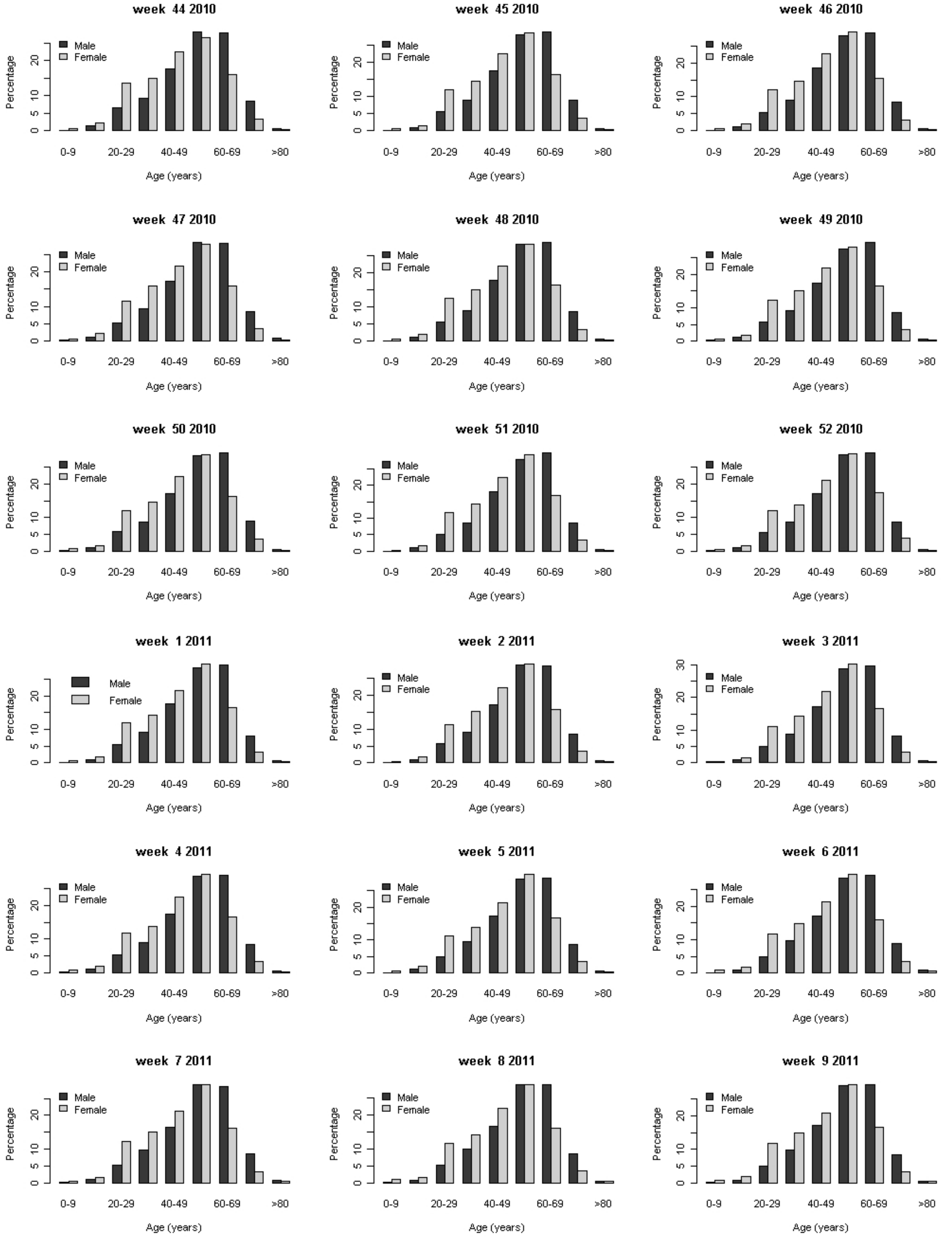

Figure C1: *Age and gender distribution of the GIS population in age groups of 10 years for weeks 44/2010 to week 9/2011.*

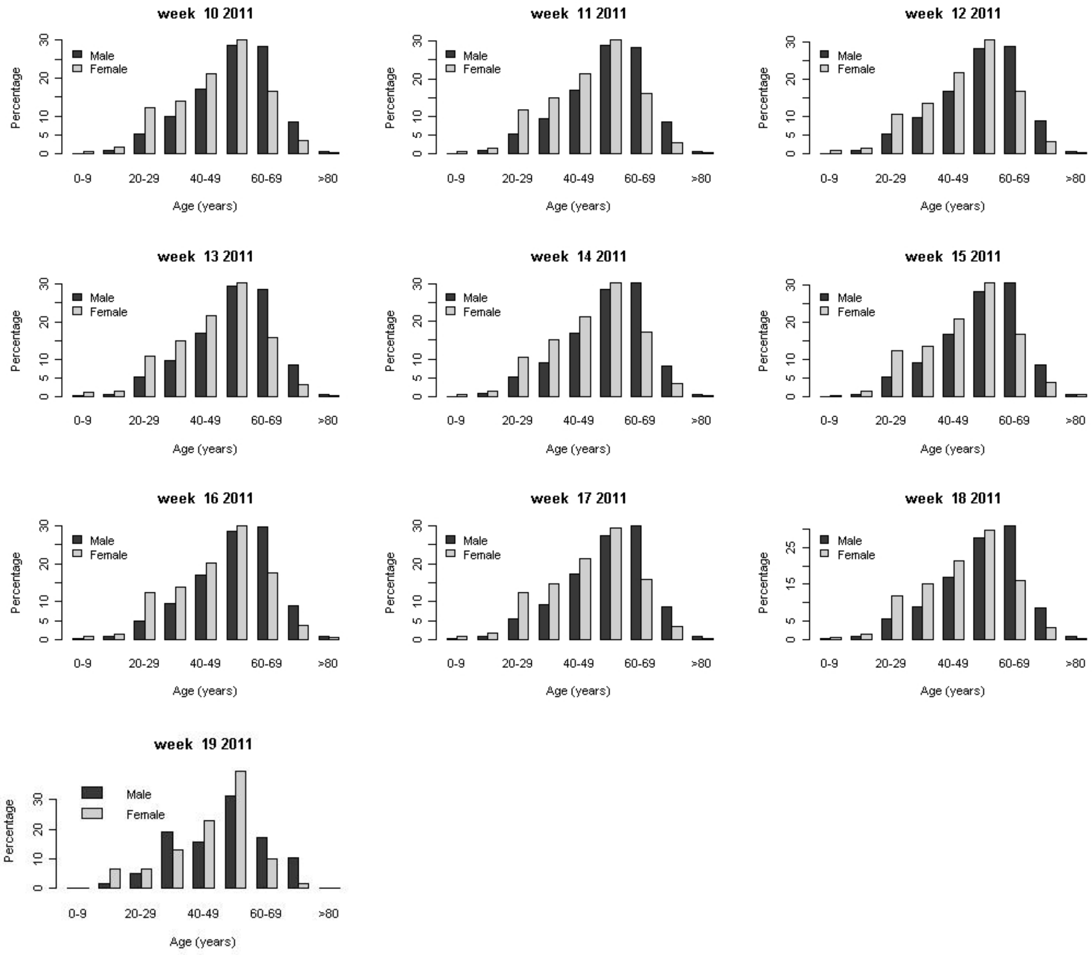

Figure C2: *Age and gender distribution of the GIS population in age groups of 10 years for weeks 10/2011 to week 19/2011.*

D Estimation in Generalized Linear Mixed Models

Pseudo-Likelihood Estimation Based on Linearization

In this section the pseudo-likelihood estimation technique, as implemented in the GLIMMIX procedure in SAS, is explained. Recall from section 4.4 that

$$\boldsymbol{\mu} = E[\mathbf{y}|\mathbf{u}] = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) = g^{-1}(\boldsymbol{\eta}),$$

where $\mathbf{u} \sim \mathcal{N}(0, \mathbf{G})$ and var$[\mathbf{y}|\mathbf{u}]$=$\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$. Assume that the ϕ=1. A first order Taylor series expansion of $\boldsymbol{\mu}$ around $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ yields

$$\boldsymbol{\mu} = E[\mathbf{y}|\mathbf{u}] \approx g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}) + \hat{\boldsymbol{\Delta}}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\Delta}}\mathbf{Z}(\mathbf{u} - \hat{\mathbf{u}}),$$

where

$$\hat{\boldsymbol{\Delta}} = \left(\frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right)_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}$$

is a diagonal matrix of derivatives of the conditional mean evaluated at the expansion locus $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$. Rearranging the terms yields

$$E[\mathbf{y}|\mathbf{u}] - g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}) \approx \hat{\boldsymbol{\Delta}}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\Delta}}\mathbf{Z}(\mathbf{u} - \hat{\mathbf{u}}),$$

from which it follows that

$$\hat{\boldsymbol{\Delta}}^{-1}\left\{E[\mathbf{y}|\mathbf{u}] - g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}})\right\} + \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}} \approx \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}. \tag{D1}$$

Define

$$\mathbf{P} \equiv \hat{\boldsymbol{\Delta}}^{-1}\left\{\mathbf{y} - g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}})\right\} + \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}},$$

then the expected value of $\mathbf{P}$, conditional on $\mathbf{u}$, is equal to the left hand side of (D1). The conditional variance of $\mathbf{P}$ is given by

$$\text{var}\left[\mathbf{P}|\mathbf{u}\right] = \hat{\boldsymbol{\Delta}}^{-1}\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}\hat{\boldsymbol{\Delta}}^{-1} \tag{D2}$$

From this, one can consider the linear mixed model

$$\mathbf{P} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{D3}$$

for which $\mathbf{P}$ is the pseudo response and var$[\boldsymbol{\epsilon}]$=var$[\mathbf{P}|\mathbf{u}]$ defined in (D2). The marginal variance in model (D3) is given by

$$\text{var}\left[\mathbf{P}; \boldsymbol{\theta}\right] = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \hat{\boldsymbol{\Delta}}^{-1}\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}\hat{\boldsymbol{\Delta}}^{-1},$$

where $\boldsymbol{\theta}$ is a vector containing all unknown variance parameters in $\mathbf{G}$ and $\mathbf{R}$. Assuming that $\boldsymbol{\epsilon}$ has a normal distribution, model (D3) can be fit by using standard techniques for linear mixed models. First, the variance parameters $\boldsymbol{\theta}$ are estimated by using profiled maximum likelihood or restricted maximum likelihood. When estimates $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ are available, the fixed effects parameters $\boldsymbol{\beta}$ and random effects $\mathbf{u}$ are estimated. With these statistics, the pseudo-responses and linearized model (D3) are recomputed and fitted again. This process continues until the relative change between parameter estimates at two successive iterations is sufficiently small. When $\phi \neq 1$ this estimation procedure needs to be adjusted slightly. For more information see SAS Insitute Inc. (2008).

Maximim Likelihood Estimation Based on Laplace Approximation

Let $\mathbf{x}$ denote a q-dimensional vector, N the number of data points and $h(\mathbf{x})$ a scalar function of $\mathbf{x}$. The multivariate Laplace approximation of the integral

$$\int_{\mathbb{R}^q} \exp\{-Nh(\mathbf{x})\}$$

is given by (see, for example, Molenberghs and Verbeke, 2005):

$$\exp\{-Nh(\hat{\mathbf{x}})\}\,(2\pi)^{q/2}|\Sigma|^{1/2}N^{-q/2}. \tag{D4}$$

In expression (D4) $\hat{\mathbf{x}}$ satisfies the condition $\frac{\partial h(\mathbf{x})}{\partial \mathbf{x}}|_{\hat{\mathbf{x}}}{=}0$ and $\Sigma{=}(\frac{\partial^2 h(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T}|_{\hat{\mathbf{x}}})^{-1}$ is the inverse of the Hessian of $h(.)$ evaluated at $\hat{\mathbf{x}}$.

Consider longitudinal data with m independent subjects and $\mathbf{y}_i{=}(y_{i1}, ..., y_{in_i})^T$ denotes the response vector for subject i, i=1,...,m. Assuming conditional independence leads to

$$p(\mathbf{y}_i|\mathbf{u}_i) = \prod_{j=1}^{n_i} p(y_{ij}|\mathbf{u}_i),$$

so that the marginal distribution of the data is expressed by

$$\begin{aligned} p(\mathbf{y}) &= \prod_{i=1}^{m} p(\mathbf{y}_i) \\ &= \prod_{i=1}^{m} \int p(\mathbf{y}_i|\mathbf{u}_i)p(\mathbf{u}_i)d\mathbf{u}_i \\ &= \prod_{i=1}^{m} \int \exp\{n_i f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{u}_i)\}\, d\mathbf{u}_i, \end{aligned} \tag{D5}$$

where

$$\begin{aligned} n_i f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{u}_i) &= \log\{p(\mathbf{y}_i|\mathbf{u}_i)p(\mathbf{u}_i)\} \\ &= \sum_{j=1}^{n_i} \log p(y_{ij}|\mathbf{u}_i) + n_i \log p(\mathbf{u}_i). \end{aligned}$$

Using (D4), the Laplace approximation to the i'th individuals marginal probability density function is

$$\begin{aligned} p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\theta}) &= \int \exp\{n_i f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{u}_i)\}\, d\mathbf{u}_i \\ &= \left(\frac{2\pi}{n_i}\right)^{q/2} |-f^{''}(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \hat{\mathbf{u}}_i)|^{-1/2} \exp\{n_i f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \hat{\mathbf{u}}_i)\} \\ &= \frac{(2\pi)^{q/2}}{|-n_i f''(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \hat{\mathbf{u}}_i)|^{1/2}} \exp\{n_i f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \hat{\mathbf{u}}_i)\}\,, \end{aligned} \tag{D6}$$

where $\hat{\mathbf{u}}_i$ satisfies the condition

$$\left(\frac{\partial f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{u}_i)}{\partial \mathbf{u}_i}\right)|_{\hat{\mathbf{u}}_i} = 0. \tag{D7}$$

Consequently, combining (D5) and (D6),the Laplace approximation for the marginal log-likelihood is

$$\log\left\{L(\boldsymbol{\beta}, \boldsymbol{\theta}; \hat{\mathbf{u}}, \mathbf{y})\right\} = \sum_{i=1}^{m} \left\{ n_i f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \hat{\mathbf{u}}_i) + \frac{q}{2}\log(2\pi) - \frac{1}{2}\log(|-n_i f''(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}; \hat{\mathbf{u}}_i)|) \right\}.$$

The objective function for optimization consequently is $\log\left\{L(\boldsymbol{\beta}, \boldsymbol{\theta}; \hat{\mathbf{u}}, \mathbf{y})\right\}$. In the GLIMMIX procedure in SAS first a suboptimization problem is performed to determine for given values of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ the random-effects solution vector $\hat{\mathbf{u}}_i$, in particular the solution satisfying expression (D7). For more information see SAS Insitute Inc. (2008).

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Estimating trends of influenza-like illness based on an observational study

Richting: **master of Statistics-Epidemiology & Public Health Methodology**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt
behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -,
vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten
verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Vandendijck, Yannick

Datum: **12/09/2011**