Stockholms universitet

Regression analysis of mortality with respect to seasonal influenza in Sweden 1993-2010

Achilleas Tsoumanis

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2010:6
Matematisk statistik
December 2010

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2010:6**
http://www.math.su.se

Regression Analysis of Mortality with Respect to Seasonal Influenza in Sweden 1993-2010

Achilleas Tsoumanis*

December 2010

Abstract

Influenza is widely considered as a cause of substantial morbidity and mortality nearly every year. Apart from influenza, several studies account also Respiratory Syncytial Virus (RSV) and Norovirus (NoV) as responsible for the amount of excess deaths and hospitalizations every year. Poisson regression models were constructed to predict the excess mortality, caused by these three infections and to quantify the burden of each infection to excess mortality. The data are weekly number of reported deaths and laboratory confirmed cases of the viruses in Sweden for the period 1993-2010. Generalized linear models and generalized additive models were used, with number of deaths as response variable and reported cases, along with week and season number as explanatory variables to capture the seasonal variability of mortality. Baseline mortality was proposed, by setting the infections effects to zero and excess mortality was calculated. The amount of excess mortality varies according to the different approaches for each infection. All three viruses contributed to excess mortality. Week was a good predictor to capture the seasonal variation of the data and GLM provided more accurate predictions than GAM. In summary, every year in Sweden there are approximately 1400 excess deaths attributed to influenza, 200 attributed to RSV and 300 attributed to NoV on average. These numbers change slightly if they refer only to elderly people.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:achillestsoumanis@hotmail.com . Supervisor: Mikael Andersson.

Acknowledgments

I would like to thank those that encouraged me and supported me during my work for this Master thesis. I would like to sincerely thank my supervisor Mikael Andersson for his encouragement, guidance and patience throughout the writing of this thesis. I would also, like to thank Rolf Sundberg for his valuable advice and my colleagues, Ilias Galanis and Fatemeh Zamanzad, for their feedback in the final stages of the thesis and their general support.

List of Tables

List of Figures

Contents

1 Introduction

Influenza is a virus that is widely considered as a cause of substantial morbidity[1] and mortality[2], nearly every year [1-7]. Especially, elderly people and people with certain medical conditions are at increased risk of developing serious complications from the influenza virus [1-4, 6, 8-13]. During a regular influenza season, about 90% of deaths occur in people older than 65 years of age [14]. It is difficult, though, to estimate the influenza-associated health-care burden accurately because relatively few hospitalizations or deaths are specifically coded as influenza related [15]. A virological diagnosis of influenza is often not sought and even when it is, influenza viruses may no longer be detectable after secondary bacterial infection has supervened [8]. Thus, the morbidity and mortality caused by influenza is often attributed to secondary bacterial infection and the primary viral illness goes unrecognized. Nonetheless, winter time influenza epidemics is proved to be associated with increased hospitalizations and mortality for many diagnoses, including congestive heart failure, chronic obstructive pulmonary disease, pneumonia, and bacterial superinfections [2-4, 9, 11, 15]. Despite a lack of coded cases, a rise in mortality is observed within the influenza season, particularly in seasons with high influenza incidence [9].

An indirect approach involving statistical modeling has long been used to estimate the influenza-associated burden [1, 2]. The concept of excess mortality was established as early as in the 1850's. Excess mortality during an influenza season is calculated as the difference between the number of deaths observed and the expected baseline mortality[2]. By the concept baseline mortality we conceive the hypothetical number of deaths occurred in the absence of circulating diseases. Statistical models have been used to predict seasonally adjusted baseline trends in mortality [1]. From the model introduced by the classical paper of Serfling (1963) [1], all deaths in excess of a seasonal baseline based on years of low influenza activity are attributable to influenza. Serfling's approach has been further developed by Simonsen et al (1997) [2].

An influenza severity index has also been suggested to adjust for the varied pathogenicity of different types and subtypes of influenza [2]. The primary index for assessing the severity of influenza epidemics has long been based on national levels of pneumonia- and influenza-related deaths. However, pneumonia and influenza excess mortality estimates account for only a subset of influenza-associated deaths and are not a good measure of the total burden of influenza on mortality. Another measure, the excess in mortality due to any cause of death (all-cause excess mortality), potentially captures all influenza-related deaths, but these seasonal estimates may not be as accurate as pneumonia and influenza excess mortality estimates [2].

The contribution of influenza to seasonal excess deaths remains a subject of controversy among researchers, with some claiming that there is an underestimation

[1] *Morbidity = incidence of ill health*

[2] *Mortality = incidence of death in a population*

of the excess mortality caused by influenza and others arguing that the number of excess deaths attributed to influenza is overestimated [5]. The reason behind these arguments is that apart from influenza, other viral epidemics occur in the winter period, such as Respiratory Syncytial Virus (RSV) and Noroviruses (NoV). RSV is a respiratory virus that infects the lungs and breathing passages and Noroviruses are a group of viruses that cause gastroenteritis in people [16]. Respiratory Syncytial Virus epidemics often overlap with influenza epidemics, and are recognized as a cause of excess winter morbidity and mortality particularly in young children [17, 18] and more recently in older adults[3, 13, 15, 19, 20]. Like influenza, RSV infections can precipitate both cardiac and pulmonary complications and are rarely diagnosed in adults [3]. In part, this happens because available rapid antigen-detection tests are insensitive in adults and few tests for RSV are requested for this age group by medical practitioners [3]. It is likely that some deaths previously attributed to influenza are actually associated with RSV infection [3, 19, 20]. Noroviruses are also suggested to contribute to winter excess morbidity and mortality [13]. If not included in the analysis, these simultaneous events may give falsely high numbers for influenza related mortality. Also, excessively cold periods and other season related factors contributing to the excess winter mortality may confuse the picture[5, 13].

In the present study, mortality data from Sweden are analyzed with respect to influenza, RSV and Norovirus. The data contain information about mortality from all causes and influenza, RSV and Norovirus incidence, which are reported to the Swedish Institute for Infectious Disease Control (SMI) [21] weekly, for the whole population and for persons aged 65 and older, separately. Different regression models are employed to explore excess mortality attributable to influenza. The approaches include generalized linear models (GLM) and generalized additive models (GAM) [22, 23].

First, an introduction about the nature of influenza, RSV and NoV is given. The concepts of baseline and excess mortality are introduced. The objectives of the study are presented in Chapter 2. The structure of the data used is specified in Chapter 3. Information about the variables, data transformation and the special features of the dataset are also given here. In Chapter 4, the theoretical background of the analysis is described. The main analysis and its results are presented in Chapter 5. There, the different models fitted are presented and their properties are compared. This chapter also contains details about characteristics of each model and its ability to describe the relationship between mortality and the viruses. The conclusions along with the discussion of the thesis are given in Chapter 6. Finally, additional results, figures and tables, along with some theoretical topics are presented in the Appendix.

2 Key Questions

1. What is the relation between influenza and excess mortality in Sweden?
2. How many deaths are attributable to influenza each year?
3. Which are the factors that affect mortality in Sweden?
4. What level of seasonality is efficient in explaining this relation, months or weeks?
5. Are there differences between Influenza A and B?
6. What happens in the elderly people?

3 Data

The dataset analyzed consists of data about mortality and virus incidence. The mortality data come from Sweden Statistics [24]. All deaths are reported to Sweden statistics, and the Swedish Institute for Infectious Disease Control (SMI) [21] gets an update of the number every second week, including age and sex of the diseased. The completeness of the reporting of deaths to Sweden statistics increases with time, and the reporting is relatively complete after one month. Since the considered data period ended in week 20 (end of May) 2010, we regard data to be complete. Data regarding influenza, Respiratory Syncytial Virus (RSV) and Norovirus are reported weekly by all laboratories performing diagnostic tests for infections in Sweden to SMI. One thing to be noted is that reporting is voluntary, which means that there exist an unknown number of cases that go unnoticed. However, the reporting system is rather consistent, so the measurements of disease activity in the population can be considered reliable. Reports contain also the age and sex of the patients. Influenza A and B are separated by the diagnosing laboratories. A subset of the specimens is sent to SMI for influenza A sub-typing and genetic characterization. Weekly data for aggregated cases of influenza are available since 1993 and data for RSV, NoV and influenza A and B separated are available since 2003.

Our dataset consisted initially of 869 weeks and 13 variables:
Year (Week): It defines the year and the week of the report that was sent to SMI. It starts from 1993(40) until 2010(20). From this column, we created two different columns to represent the year and the week of the report (see variables Year and Week). Except the information derived from this variable, Year(Week) was not used in the rest of the analysis.
Week: It defines the number of the week. It goes from 1 until 52, or 53 in some years. It is used to describe the seasonal variation of influenza and the other viruses. It was treated as a categorical variable with values 1 to 52/53.
Number of deaths: The number of deaths from all causes in Sweden.

Number of deaths (65+): The number of deaths from all causes in Sweden only for those of age 65 and older.
Number of influenza cases: The number of influenza cases reported to SMI.
Number of influenza cases (65+): The number of influenza cases in the 65+ age group.
Number of RSV cases: The number of Respiratory Syncytial Virus cases reported to SMI.
Number of RSV cases (65+): The number of Respiratory Syncytial Virus cases in the 65+ age group.
Number of Norovirus cases: The number of Norovirus cases reported to SMI.
Number of Norovirus cases (65+): The number of Norovirus cases in the 65+ age group.
Influenza A: The number of influenza A cases.
Influenza B: The number of influenza B cases.

Data on RSV cases for the whole population are available since 2001(43), while data for Norovirus, Influenza A and B, as well as the 65+ age group are available only since 2003(43). This makes it a little difficult to see and work with our dataset as a whole.

In addition to the original variables stated above, two more variables were created to help us understand better the seasonal variation of influenza: **Season** and **Month**. The variable Season is an indicator variable for each year, that helps us capture the long term trends in mortality rates and population growth. Influenza usually starts in the late autumn and is at its peak during the beginning of the calendar year. In order to capture the complete influenza season in one year, we considered the “influenza-year” to start in week 27 (start of July) and to finish in week 26 the following year. Season was treated as a categorical variable with values 1 to 17, with 1 corresponding to the 1993-1994 influenza-year and 17 to 2009-2010. Corresponding to Season, the “influenza-year” starts in July and finishes in June. Another approach, suggested in order to explain the seasonal variation of influenza, is Month. By using ISO 8601 standard [25] and the concept of fiscal year we were able to transform the week numbers into months. Month was treated as a categorical variable as well.

We also consider the concept of “influenza season”, which is the time period where the majority of influenza cases are reported. According to general practice, we use the surveillance period from week 40 until week 20 (beginning of October until mid-May). Almost all cases of influenza in our dataset are reported during this period (99.1%), while the percentages of the other illnesses were similarly large (99.2% for RSV and 93.8 for Norovirus). In the original dataset, there were a lot of empty cells in all variables that measured virus cases. That was because outside the surveillance period, no reporting was made for these viruses to SMI and thus no data were available. We assumed that outside the influenza season, there is no incidence of these viruses, and even if there are some cases they can be considered as isolated cases. For this reason, the value zero was given to all empty cells outside influenza season for all viruses.

From the database, we had to remove the 2004(52) line. That was the week of the 2004 Indian Ocean tsunami with a very high death toll for Sweden, a fact that had no connection with the subject of this study. The extremely high mortality of that week would affect the results and it was considered wiser to remove the particular week from the dataset. Also, there was one case of missing value (in week 53 1998) and so the whole week was excluded from the analysis.

The statistical analysis was conducted by using R language (R version 2.11.1, 2010-05-31 [26]) and a first crude exploratory analysis was conducted by the help of Microsoft Excel 2007.

4 Methods

The main purpose of the thesis is to examine the relationship between mortality and the counts of the different viruses and to find a model that explains adequately this relationship. Two different approaches were suggested in order to fit the data into regression models, generalized linear models (GLM) and generalized additive models (GAM) [22, 23]. In this chapter, the theoretical background and the methods behind the analysis of the data will be presented. The theory was retrieved from [22], [27], [28] and [29].

4.1 Generalized Linear Models

Definition of Generalized Linear Models

Generalized linear models are a generalization of linear models that allow the response distribution to be other than normal and the relationship between the mean value and the linear predictor to be other than linear. They are of the structure

$$g(E(Y_i)) = \mu = \mathbf{X}_i^T\{\beta\},$$

where g is a smooth monotonic, differentiable link function, $\mathbf{Y}$ a set of independent random variables from the exponential family, the transposed vector ${\mathbf{X}_i}^T$ is the ith row of a design matrix $\mathbf{X}$, and $\boldsymbol{\beta}$ is a vector of unknown parameters.

Estimation and inference with GLMs is based on the maximum likelihood estimation theory, although the maximization of the likelihood requires an iterative least squares approach. For more technical details on these topics see section A.5 of Appendix.

Exponential family of distributions

The response variable in a GLM can follow any distribution from the exponential family. The exponential family of distributions includes many distributions, such as the Poisson, the Binomial, the Gamma and the Normal distribution. If $\mathbf{Y}$ is a random variable whose probability function depends on a parameter vector $\boldsymbol{\vartheta}$, then the distribution of $\mathbf{Y}$ belongs to the exponential family if it can be written in the form

$$\mathrm{f}(\mathrm{y};\theta)=\alpha(\theta)\mathrm{h}\,(\mathrm{y})\exp\{\theta^{\mathrm{T}}\mathrm{t}\,(\mathrm{y})\},$$

where θ is the canonical parameter vector $\theta=(\theta_1,...,\theta_k)$ and t(y) is the sufficient statistic $t(y)=(t_1(y),...,t_k(y))$ and α and h are known functions. Instead of function $\alpha(\theta)$, the quantity 1/ C(θ) is usually used as a normalizing constant. The model above can be re-written in the form

$$\mathrm{f}(\mathrm{y_i};\theta_\mathrm{i},\varphi)=\exp\left\{\frac{\theta_\mathrm{i}t_i(\mathrm{y_i})-\log(\mathrm{C}(\theta_\mathrm{i}))}{\varphi}\right\}\mathrm{h}(y_\mathrm{i};\varphi),$$

if we want to incorporate the dispersion parameter, φ. Note that if $\varphi=1$ then the two equations are equivalent.

Poisson regression

In this study, the variable of interest is the number of deaths occurring in Sweden every week or month. We can assume that the number of deaths each week or month occur independently of each other. Since the data represent independent counts, Poisson regression is the most appropriate approach to model the data. Poisson regression models are generalized linear models with the logarithm, as the canonical link function, and response variable assumed to follow the Poisson distribution. The assumptions made in Poisson regression include:

- The changes in the rate from combined effects of different exposures are multiplicative
- At each level of the covariates the number of cases has variance equal to the mean.
- Observations are independent

The formula for the Poisson distribution is

$$P(Y_i|x_i)=\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

where, Y_i is the random variable representing the number of occurrences, λ_i is the parameter that represents the expected value of the count i, where y_i represents the observed number of cases. The effect of explanatory variables on the response variable Y is modeled through the parameter λ. Since the logarithmic function is the natural link function for the Poisson distribution, a log linear model is employed,

$$\log \mu = \log \lambda_i = x_i^T \beta, \tag{1}$$

where $\mathbf{x}_i^T \beta$ is the usual linear combination of predictors for case *i.* The expected number of events is given by$E[yi|xi] = \lambda_i = e^{x_i^T \beta}$.
The parameter vector β of the model is estimated by using Maximum Likelihood method (see Appendix, section A.4). Because of the structure of Equation (1), parameter estimates are often interpreted as odd ratios in the exponential scale e$^{\beta}$. The major assumption of the Poisson model is that $E[Y] = \mu = \lambda = Var[Y]$.

Overdispersion

As stated above, the Poisson distribution assumes that the variance is equal to the mean. In real situations though, this can fail when there is positive dependence between the observations, or incomplete information about all relevant covariates. Thus, it can often be observed that the variance is larger than the mean. In these cases, the data are said to be *overdispersed.* Parameter estimates in Poisson regression models on overdispersed data have standard errors and p-values that are too small.

One way to check if overdispersion is present is to divide the deviance statistic by its degrees of freedom. If there is no overdispersion, the ratio will be close to 1. Values greater than 1 indicate overdispersion, that is, the true variance is bigger than the mean. The Pearson's Chi-Square has also been suggested to capture the excess variability by some statisticians [30].

Evidence of overdispersion indicates inadequate fit of the Poisson model. Quasi-Poisson and Negative binomial regression are typically used when there are signs of overdispersion in Poisson regression. Negative binomial regression uses a different probability model which allows for more variability in the data.

Quasi-Poisson and Negative Binomial regression

Two alternative approaches when there are signs of overdispersion are quasi-Poisson and negative binomial regression models. Both these models can be framed as generalized linear models. Quasi-Poisson and Negative Binomial models allow for greater variance in the data than in the Poisson model. Both have an extra parameter that accounts for dispersion.

Let Y be a random variable that is assumed to follow the quasi-Poisson model, $Y_i \sim Poi(\mu_i, \theta)$. Then, $E[Y] = \mu$ and $Var[Y] = V_{Poi}(\mu) = \theta\mu$, where E(Y) is the expected value of Y, Var(Y) is the variance of Y and $\theta > 1$, is called the overdispersion parameter.

The quasi model formulation has the advantage of leaving parameters in a natural, interpretable state and allows standard model diagnostics without a loss of efficient fitting algorithms. In the quasi-Poisson models, p-values for parameter estimates are based on t statistics instead of z statistics. On the other hand, AIC cannot be computed because the likelihood is not defined, and also the residual deviance is the same for the Poisson and quasi-Poisson models, so there is not much information to allow for a comparison between the Poisson and quasi-Poisson models.

We will denote a random variable Y having a negative binomial distribution as $Y \sim NB(\mu_{\mathrm{i}}, \theta)$. The Negative binomial model can be derived by letting the mean of the Poisson distribution to vary according to a random parameter γ that follows the Gamma distribution.

$$\Upsilon_i|\gamma_i \sim \text{Poisson}(\gamma_i\mu_i),$$

$$\gamma_i \sim \frac{1}{\theta}Gamma(\theta)$$

The marginal distribution of Y_i is then the negative binomial with mean $E[Y] = \mu$ and variance $Var[Y] = V_{NB}(\mu) = \mu + \theta\mu^2$, where $\mu > 0$ and $\theta > 0$.
Here, the overdispersion is the multiplicative factor $1 + \theta\mu$, which depends on μ (in contrast to the quasi-Poisson).

The mean, for both Quasi-Poisson and Negative binomial models, is a single parameter that can vary as a function of covariates. For quasi-Poisson regression, we assume $Y_i \sim Poi(\mu_i, \theta)$ and for NB we assume $Y_i \sim NB(\mu_i, \theta)$, where we let the mean μ_i for the i^{th} observation vary as a function of the covariates for that observation in both models. Because the mean $\mu_i > 0$, it is natural to model μ_i as $g(\mu) = \mathbf{X}\beta$, which is the standard form of generalized linear models, with log as the link function.

But, how much are the models affected after being fitted by these two methods? This is a natural question. While they often give similar results, there can be striking differences in estimating the effects of covariates. The variance of a quasi-Poisson model is a linear function of the mean while the variance of a negative binomial model is a quadratic function of the mean. These variance relationships affect the weights in the iteratively weighted least-squares algorithm of fitting models to data, since these weights are inversely proportional to the variance. Because the variance is a function of the mean, large and small counts get weighted differently in quasi-Poisson and negative binomial regression. Thus, negative binomial and quasi-Poisson will weight observations differently.

4.2 Generalized Additive Models (GAM)

Definition of GAMs

A generalized additive model is a generalized linear model with a nonlinear predictor in the form of a sum of smooth functions of covariates. In general the model has a structure

$$g(\mu_i) = X_i\theta + \sum_{j=1}^{k} f_j(x_{ji})$$

where $\mu_i = E(Y_i)$ and the response variable, Y_i follows some exponential family distribution. The smooth monotonic function g is called the 'link function', X_i is a row of the model matrix for any strictly parametric components and θ is the corresponding parameter vector. Last, f_j is a parametric smooth function of the covariates x_k. In other words, instead of a single coefficient for each variable (additive term) in the model, in generalized additive models an unspecified function is estimated for each predictor. In this paper, thin plate regression splines will be used to estimate the smooth functions in the models. To fit a GAM, the penalized iterative re-weighted least squares (P-IRLS) method will be used. P-IRLS method is further explained in section A.6 in the Appendix.

Thin plate splines

Smoothing splines provide an excellent means for estimation and inference with models like

$$y_i = f(x_i) + \varepsilon_i$$

or

$$y_i = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + ... + \varepsilon_i,$$

where y is the response variable, x_1, x_2, x_3, ... are the covariates, f' s are smooth functions and ε is a random variable, independent for each different i.

First, let us consider the first model, because it is simpler. The model can be estimated by finding the function from an appropriate reproducing kernel in Hilbert space which minimizes

$$||y - f||^2 + \lambda \int f''(x)^2 dx, \tag{2}$$

where $\mathbf{y}$ is a vector of $y_i's$, f is the corresponding vector of $f(x_i)$-values and $||\centerdot||$ is the Euclidean norm. λ is called smoothing parameter, which must be chosen appropriately in order to achieve the right balance between maximizing the model's goodness of fit as measured by the first term and the model's 'wiggliness' as measured by the second. The result of this minimization turns out to be finite dimensional and is a cubic spline, which is a special case of a thin plate spline.

In general the smoothing functions are obtained as the solution of the generalization of expression (2) to problems in which f is a function of any finite number

$d \geq 1$ of covariates and the order m of differentiation in the wiggliness penalty can be any integer satisfying $2m > d$, where d is the number of covariates. A further straightforward generalization of Equation (2) is the replacement of the least squares term in the objective with a negative log-likelihood based on an exponential family.

There are two disadvantages, though, in the use of thin plate spline smoothers and their widespread adoption in practical statistical work. The first is computational. To fit a thin plate spline to n data points requires the estimation of n parameters and an additional smoothing parameter. This has often a large computational cost. Except in the case d = 1 this involves $\mathcal{O}(n^3)$ operations, which is frequently prohibitive. The second obstacle is the fact that their use requires a change in modeling methodology relative to conventional linear or generalized linear modeling: the flexibility of a fitted model must be selected by adjusting the smoothing parameter λ, rather than by adding or dropping model terms.

One way to reduce the computational cost is to employ regression splines. The basis implied by solving the spline smoothing problem for a small representative data set is found and this small basis is used to construct a model for the full data set of interest. The model is typically fitted as a linear or generalized linear model without imposing a wiggliness penalty. The covariate points that are used to obtain the reduced basis are known as the 'knots' of the regression spline. The number of knots controls the flexibility of the model, but unfortunately their location also tends to have a marked effect on the fitted model. Theoretically, conventional hypothesis-testing-based model selection can be used to determine the appropriate flexibility for regression spline models, but in practice there are difficulties. If the knots of order k and order *k-1* regression spline models for a data set are arranged to ensure the best performance of both models, then the two models will not generally be nested. Alternatively, if knots are not moved, but some knots are simply dropped during model selection, then nesting is maintained, but very uneven knot spacing can result and this has undesirable approximation theoretic consequences. Another more subtle problem with the latter strategy is 'knot confounding'. Finally, when d > 1, even deciding where to place knots so that they appear evenly spread through the covariates can become problematic. Some of the problems with knot placement can be partially alleviated by abandoning pure regression splines in favor of penalized regression splines. But in this case model flexibility is again controlled by a smoothing parameter λ, rather than the basis dimension, so that some conventional (generalized) linear modeling methods are once again inapplicable.

5 Results

In this chapter the results of the statistical analysis performed will be presented. First, an exploratory analysis was performed in order to better understand the nature of the data. Different kinds of statistics were calculated for each variable and an effort was made to reveal the relations among the variables. In the second part, different models were employed to fit the relationship between mortality and the other variables. Generalized linear models and generalized additive models were used in this part. Last, baseline mortality was calculated according to the different models and the amount of excess mortality was specified. It should be noted that all hypothesis testing was performed at the $\alpha=$ 0.05 significance level.

5.1 Exploratory Analysis

As a first step of understanding our data better, we calculate various descriptive statistics, such as the mean value, the median, the variance, minimum and maximum values, as well as, the skewness and the kurtosis of each variable. All these statistics can be found in Table 1. Also, histograms of each variable (Figures 20, 21 in A.4 section in Appendix) were created in order to help us better understand Table 1 visually.

Table 1: **Descriptive Statistics**

Variable	**Mean**	**Median**	**Variance**	**Min**	**Max**	**Skewness**	**Kurtosis**
Deaths	1773	1734	25489.8	1495	2634	1.59	4.17
Influenza	22.02	0	2467.12	0	355	3.24	11.44
RSV	38.97	12	3411.62	0	277	1.98	3.47
Norovirus	86.45	34	11909	0	505	1.67	2.13
Influenza A	22.51	2	2734.32	0	285	3.18	9.8
Influenza B	5.18	0	155.89	0	69	3.37	11.31
Deaths65	1526.48	1492	22989	1270	2363	1.59	4.15
Influenza65	11.07	0	582.94	0	140	2.75	7.47
RSV65	0.94	0	4	0	13	2.82	8.21
Norovirus65	63.64	20	7471.99	0	397	1.7	2.19

Deaths have a mean value of 1773, while all the viruses have much smaller mean values, which vary from 0.94 to 86.45. A first interesting element that can be spotted from the above table is the differences of the mean values regarding the age groups (whole population and elderly people). For example, the mean values of deaths do not differ much in the two categories, which means that the majority of deaths occur in the 65+ age group. In fact, from our data it turns out that around 86% of all deaths come from people of age 65 and older. The corresponding percentages are 40% for influenza, 2% for RSV and 74% for Norovirus. This 2% percentage explains very well the mean value difference for RSV (38.9 in the whole population and 0.94 in the elderly people). It seems that RSV does not affect the elderly age group much. RSV typically affects infants and young children mostly, and they in return infect the elderly people (e.g. their grandparents) upon contact.

In general, we can see that all the variables have high variances. This may imply overdispersion in our data. Also, except deaths, all other variables have a big count of zeros, especially influenza that has more than half (51%) of its observations with the value zero. The rest of the variables have smaller proportions of their observations at value zero (22% Influenza65, 27% RSV65, 22% Influenza B). By checking the histograms we can verify the above statement. All variables except deaths show large frequencies at value zero.

By looking at the histograms, we can visually observe the skewness of each variable as given in Table 1. By skewness, we define the asymmetry of a distribution. Positive values of skewness show that the distribution's right tail is longer than the left one and that more values lie to the left of the mean, including the median. Values of skewness near zero indicate that the values are relatively evenly distributed on both sides of the mean. Influenza (both for the total counts, as well as for the two subtypes, A and B) has the largest skewness (11.44, 9.8 and 11.31 respectively) among all variables. This corresponds to the fact that these variables have a big mass of observations on value zero, as stated above. The same stands for RSV and Norovirus (both for the whole population and the above 65 group), but to a smaller extent.

The *excess kurtoses of all variables are* positive, which suggests that we have *leptokurtic* distributions. Basically, this means that the distributions of all variables have a "sharp" peak and "heavy" tails, something that can be seen from the histograms. Influenza and RSV65+ show the largest kurtosis. The large values of kurtosis confirm our hypothesis of presence of overdispersion in our data.

In order to examine the relationship among all variables, we calculate Spearman's correlation coefficient for all pairs of variables. As can be easily seen both from Table 2 and the scatter-plot (Figure 22 in A.4 in Appendix) all variables are highly correlated with each other. The values of the correlation coefficient vary around 0.6-0.7 and reveal a strong correlation. To visualize the results of Table 2 we create the scatter-plots for all the possible pairs of variables. From the scatter-plots, a linear relationship is suggested.

Table 2: **Spearman Correlation Coefficient**

	Influenza	**RVS**	**Norovirus**	**Influenza A**	**Influenza B**
Deaths	0.689 *	0.716 *	0.679*	0.642 *	0.503 *
Influenza		0.656 *	0.657*	0.943 *	0.703 *
RVS			0.701 *	0.529 *	0.669 *
Norovirus				0.545 *	0.612 *
Influenza A					0.517 *

5.2 Modeling

The main objective of the present paper is to examine the relationship between mortality and the counts of the different viruses and to find a model that explains this relationship adequately. Two different approaches were used in order to fit the data into regression models, generalized linear models (*glm* function in R) and generalized additive models (*gam* function from *mgcv* package in R).
The response variable in all models is the number of weekly reported deaths. Categorical variables season and either week or month were used as explanatory variables to describe the seasonal variation in mortality. The number of reported influenza, RSV and NoV cases were the other explanatory variables used in the models. All interaction terms among reported cases and seasonal effects were also considered as possible explanatory variables and were included in the fitting process.

Because of the structure of our data, many different models were created to explain the relationship between mortality, and viruses and seasonal parameters. The models were created by the *stepwise* method. Variables are added one by one to the model, and the χ^2 statistic for a variable to be added must be significant according to a certain level (here, $\alpha = 0.05$). After a variable is added, the *stepwise* method looks at all the variables already included in the model and deletes any variable that does not produce a significant χ^2 statistic. After this check and the necessary deletions are accomplished, another variable is attempted to be added to the model. The stepwise process ends when none of the variables outside the model has a significant χ^2 statistic and every variable in the model is significant, or when the variable to be added to the model is the one just deleted from it. The results of the modeling procedure are given in this chapter.

Data from the period 2003 until 2010 were analyzed first, because in these seasons there was systematic reporting of all viruses. The data consists of

342 weeks and can be treated as complete. Either week or month is used as explanatory variable to capture the seasonal effect of mortality. Models using Poisson regression and Generalized Additive Poisson regression were employed with either week or month as a seasonal effect.

By using the stepwise method as described above, and week as seasonal effect, the GLM and GAM models produced respectively are

Model 1:

$$log(\mu_{ij}) = \beta_0 + \beta_j^w + \beta_i^s + \beta_i^I x_{ij} + \beta_i^R y_{ij} + \beta_i^N z_{ij}$$

Model 2:

$$log(\mu_{ij}) = \beta_0 + f(\beta_j^w) + \beta_i^s + \beta_i^I x_{ij} + \beta_i^R y_{ij} + \beta_i^N z_{ij}$$

where β_0 is the intercept, β_j^w is a vector with the 53 levels of factor *week*, β_i^s is the vector with the 7 levels of factor *season* and x_{ij}, y_{ij}, z_{ij} are the observations of influenza, RSV and NoV respectively in each week j and season i. Note that the model accounts for the differences in each season. The parameter vectors of these interaction terms are denoted by β_i^I, β_i^R, β_i^N and f in Model 2 is the smoothing function for *week*.
The corresponding models with month as the seasonal effect are

Model 3:

$$log(\mu_{ij}) = \beta_0 + \beta_k^m + \beta_i^s + \beta_i^I x_{ij} + \beta_i^R y_{ij} + \beta_i^N z_{ij}$$

Model 4:

$$log(\mu_{ij}) = \beta_0 + f(\beta_k^m) + \beta_i^s + \beta_i^I x_{ij} + \beta_i^R y_{ij} + \beta_i^N z_{ij}$$

with β_k^m the monthly seasonal effect.

The validity of the models was examined visually through fitted values, residuals and residual deviance. The residuals should be evenly distributed above and below zero when plotted against fitted values if the hypotheses of residual independence and constant variance are to be held (upper left and lower left plots in Figure 23). Moreover the standardized residuals should be normally distributed and unrelated to all explanatory variables. From the Figures 23-27 it is shown that models with week have better fit than models with month. GLMs have better fit than GAMs. Some statistics of the models are presented in Table 3. By looking at the Table, we can see that the models that involve *week* as the seasonal effect are better than the models with *month.* The Analysis of deviance and Analysis of Variance tables of Model 1 and 2 respectively are given in the Appendix (section A.4, Tables 7, 8).

Table 3: **Model Statistics (a)**

Model	**AIC**	**Deviance**	**Degrees of freedom**	**Deviance explained**
Model 1 (GLM -week)	3668.6	326.42	262	91.3%
Model 2 (GAM-week)	3691.6	435.94	305.3	88.3%
Model 3 (GLM-month)	3852.4	592.24	303	84.1%
Model 4 (GAM-month)	3848.6	593.76	305.6	84.1%

The plots of the fitted values along with the observed deaths are used to understand the predictive power of each model. In all models the observed deaths are depicted as dots and the predicted mortality as a line. As can be seen from Figures 1-4, GLMs with weekly seasonal effect describe the mortality better than any of the other models. All models seem to be unable to predict some peaks in the mortality however, which is usually the matter of interest in most cases. Models with month seem appear less capable of predicting mortality than models with week. GAMs offer a more smooth way to predict mortality, without the wiggliness of GLMs. If they are used with week, they have better prediction abilities than both GLMs and GAMs with month, but worse than GLMs with week.

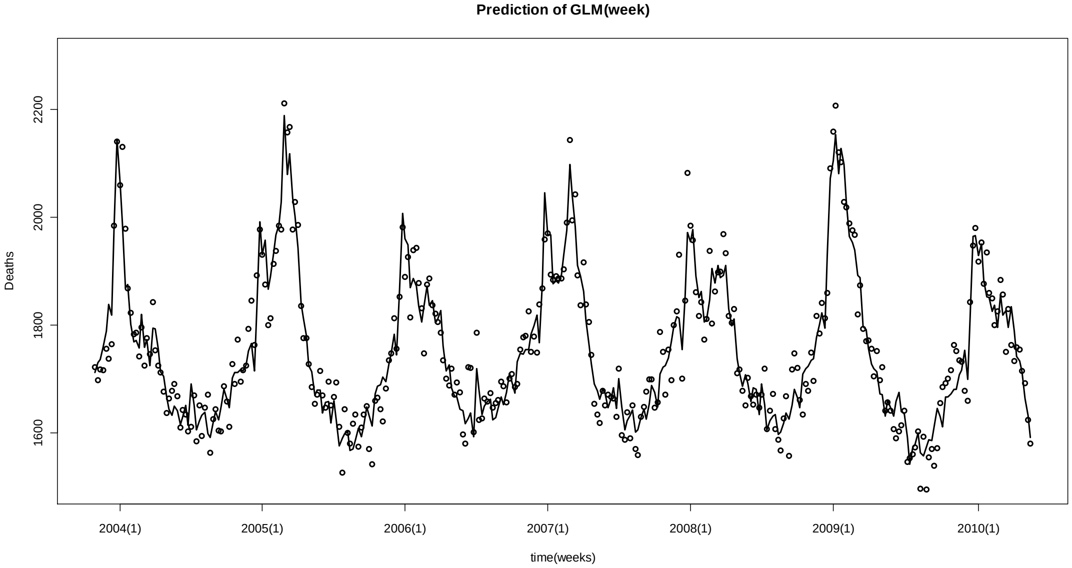

Figure 1: Predicted mortality according to Model 1 (GLM with week)

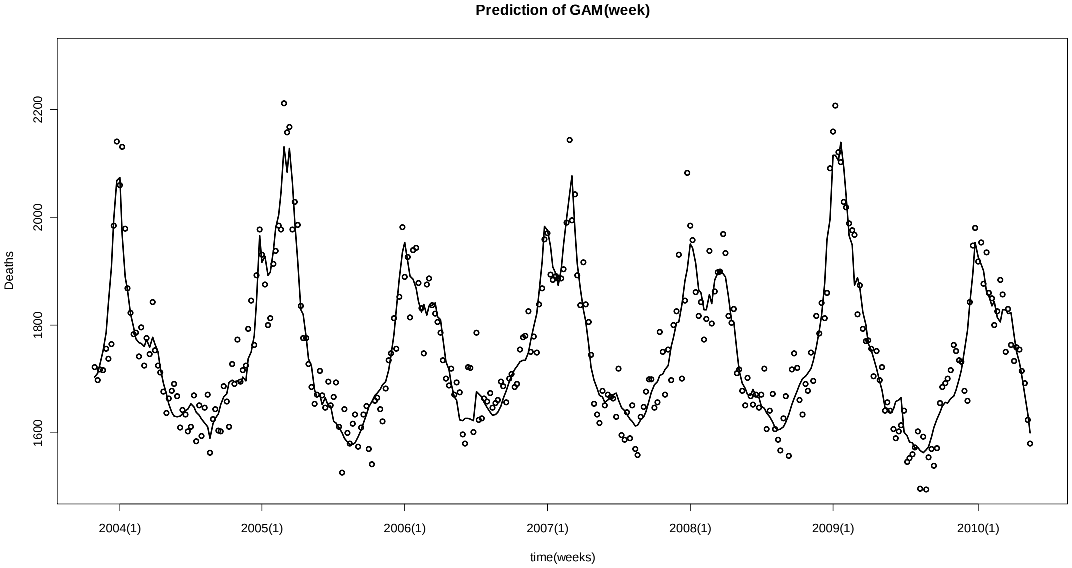

Figure 2: Predicted mortality according to Model 2 (GAM with week)

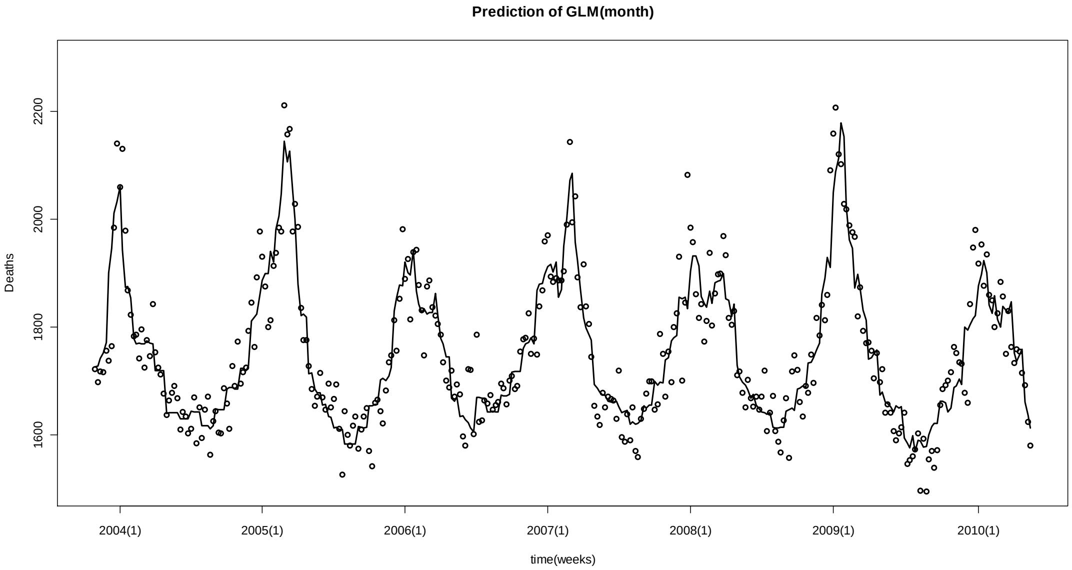

Figure 3: Predicted mortality according to Model 3 (GLM with month)

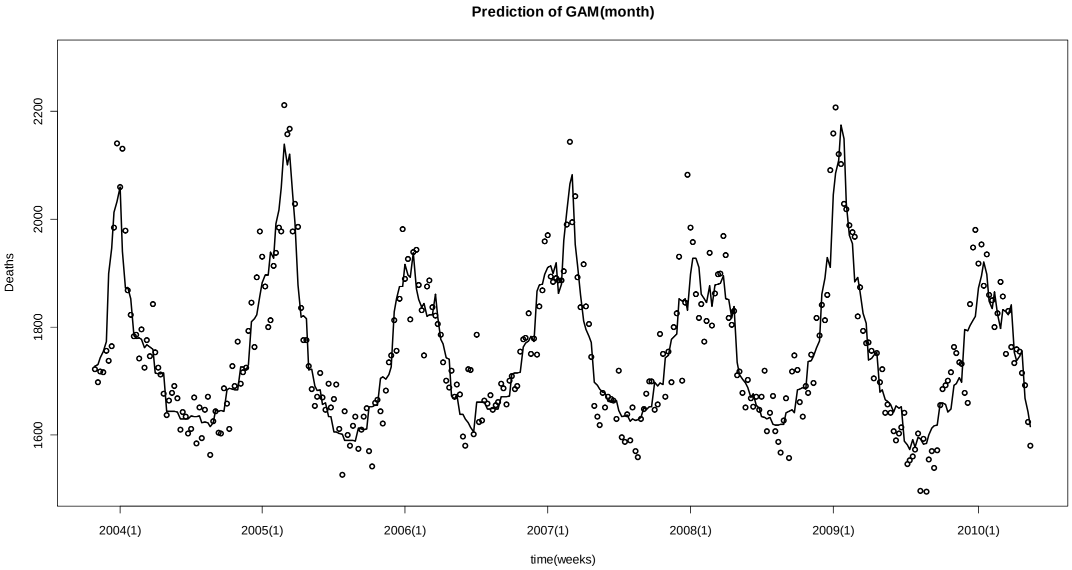

Figure 4: Predicted mortality according to Model 4 (GAM with month)

If the objective is to reduce the large number of parameters in the model ($341 - 262 = 79$ for week) by using month (with 38 parameters), then the alternative of generalized additive models should be considered, as their results are better than the models with month, but in addition, they use fewer parameters (36 parameters). Moreover, the GAMs with week reduce the parameters of week from 53 to 8.9, while the GAMs with month reduce the parameters of month from 12 to 8.6, which cannot be considered as a great improvement.

The two different influenza strains (influenza A and influenza B) are used in a more detailed approach in comparison to the total influenza reports. Another 4 models were produced, with very similar results as before, as one can see in Table 4.

Table 4: **Model Statistics (b)**

Model	AIC	Deviance	Degrees of freedom	Deviance explained
Model 5 (GLM-week)	3664.2	321.94	262	91.4%
Model 6 (GAM -week)	3686.24	430.51	305.2	88.5%
Model 7 (GLM -month)	3842.6	568.4	296	84.8%
Model 8 (GAM -month)	3838.76	570.29	299	84.7%

The predictive power is again given in Figures 5-8 and one can easily notice that once more the models with month do not capture the mortality as well as the GLMs with week. In addition, the prediction of models with month is worse than the prediction of GAMs with week, so the latter is to be preferred instead, as a more flexible alternative model.

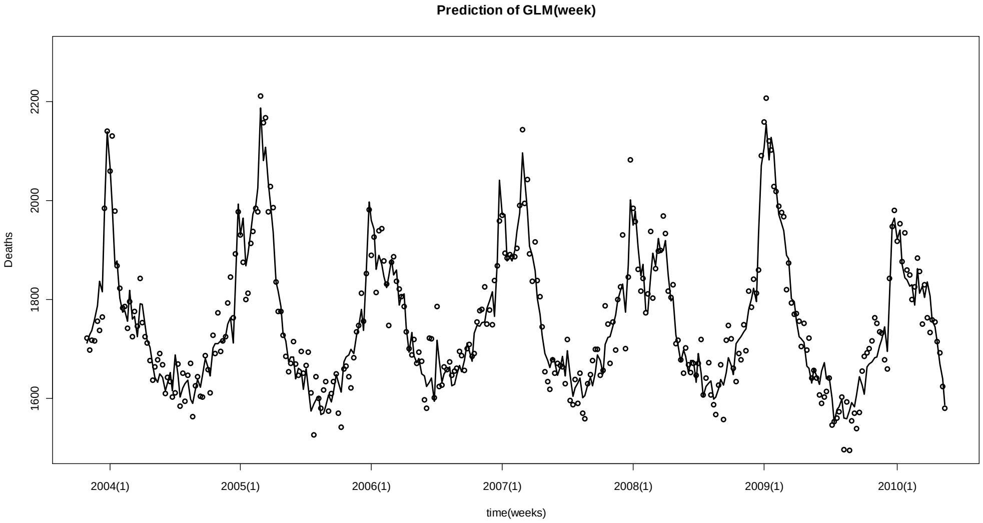

Figure 5: Predicted mortality according to Model 5 (GLM with week)

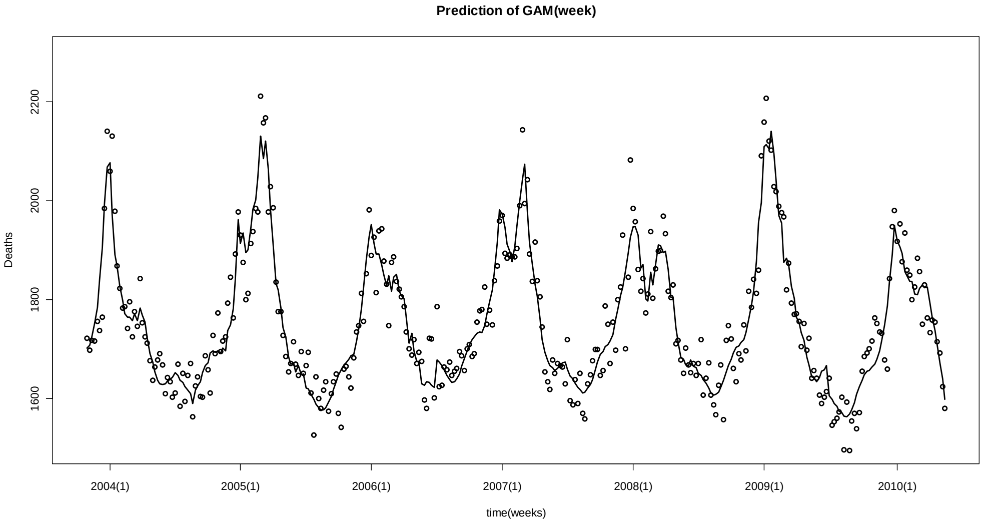

Figure 6: Predicted mortality according to Model 6 (GAM with week)

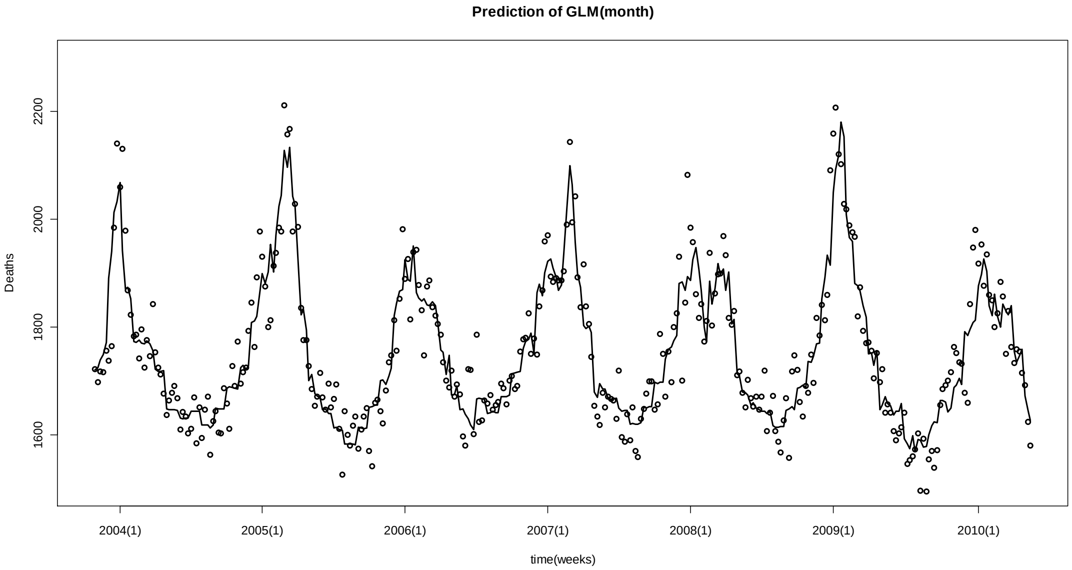

Figure 7: Predicted mortality according to Model 7 (GLM with month)

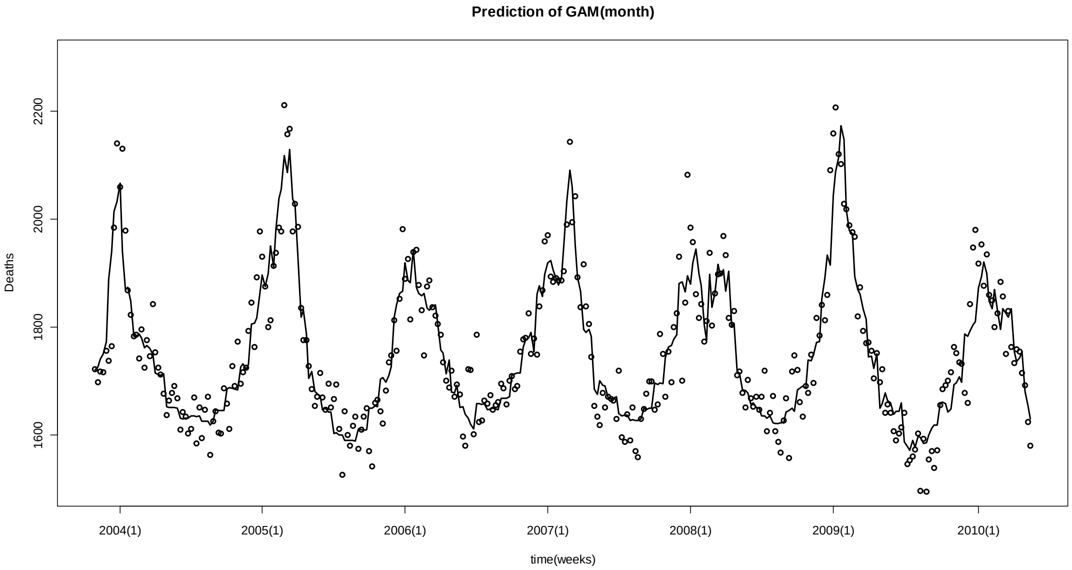

Figure 8: Predicted mortality according to Model 8 (GAM with month)

All models were examined for overdispersion. One way to spot overdispersion is to compare the residual deviance with the degrees of freedom. A ratio larger than 1 indicates overdispersed data, or a model that cannot explain the data well enough. In the GLMs with week, the ratio is around 1.25, which indicates that these models explain sufficiently well the variation in the data. In the other cases, quasi-Poisson and Negative Binomial regression models were fitted, in order to capture this excess variation. The results of the two approaches were very similar and the approach of quasi-Poisson was preferred because of convenience. The overdispersion parameter for the GAM was 1.43 and for models with month 1.95 for both models. The results were very similar for the case of models with separate influenza strains. The existence of overdispersion implies that some important covariate was left out of the model. Indeed, some covariates (e.g. interaction of week and influenza) decreased the value of the overdispersion parameter and improved the fit. The problem with these models is the difficulty in their interpretation. For, this reason, we will stay with our current models.

In many studies it has been proven that influenza is a major cause of excess mortality every winter. Although, there are studies that claim that the effect of influenza in mortality has been overestimated and that other causes of mortality should be considered. A regression model was produced that includes only influenza reports along with seasonal effects in a try to see if RSV and NoV should be considered important factors in mortality. Data from the whole dataset were used in this model. The model constructed was

<u>Model 9</u>:

$$log(\mu_{ij}) = \beta_0 + \beta_j^w + \beta_i^s + \beta_i^I x_{ij}$$

This simpler model does not fit the data well enough as implied by Figure 9. Especially, the scale-location plot implies that the data are overdispersed. A quasi-Poisson regression model is tried instead, but the fit of the data remains unsatisfactory.

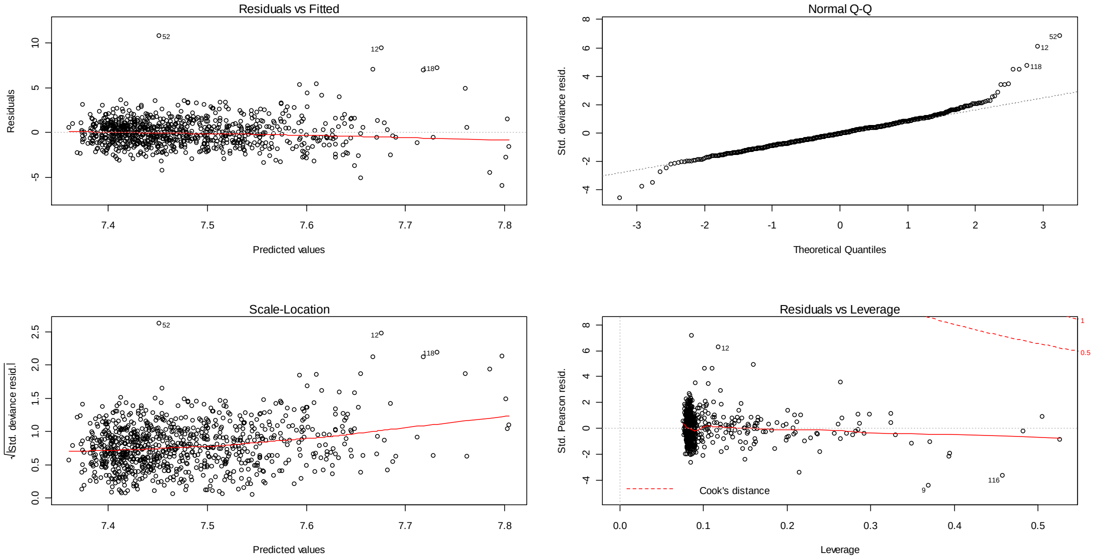

Figure 9: Goodness of Fit for Model only with influenza

As seen from Figure 10 there is still some variation to be explained.

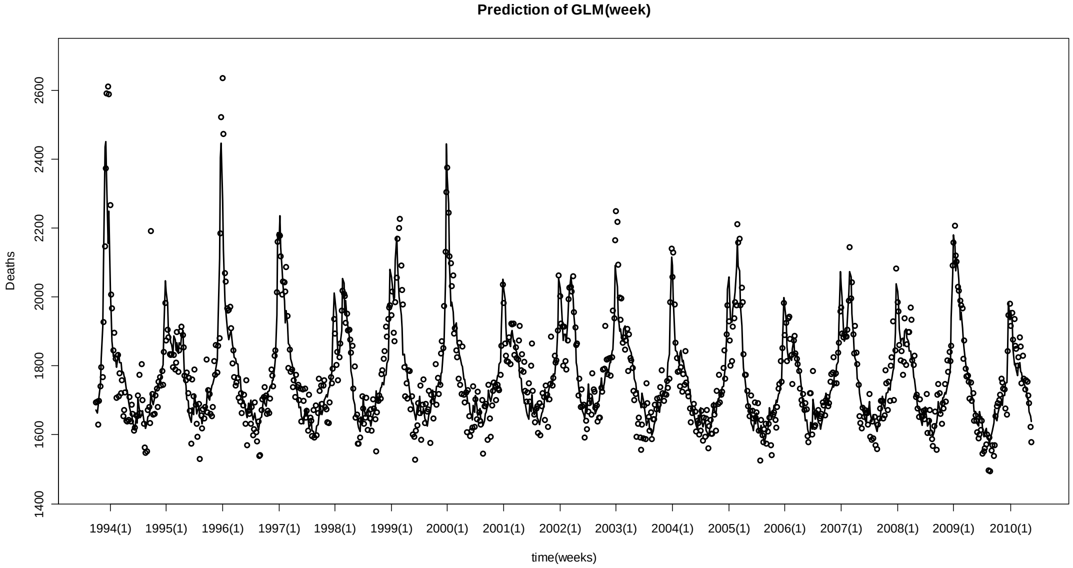

Figure 10: Predicted mortality according to Model 9 (only influenza)

The deviance explained by the simpler model is 82.3%, which is less than for the first models. The overdispersion parameter for this model is 2.7, which is another clue that influenza alone is unable to explain mortality sufficiently. In addition, the effect of influenza on mortality is overestimated, with the magnitude of coefficients in the model being 10 times higher for all seasons than the previous models. Thus, models that do not include diseases other than influenza, cannot predict mortality, sufficiently enough and the effect of influenza is overestimated. Thus, RSV and NoV have to be included in our models.

The original models were fitted also to data that distinguish cases that refer to elderly people. The results of the models for people of age 65 and above were compared to the results of the original models, to check the effect of the viruses in a more vulnerable age group. Again, GLMs with week had better predictive power and better fit than GAMs or GLMs with month. GAMs had again similar results with models with month and will be preferred to models with month because they use less parameters. The fit of GLM with week and its predictive power against the corresponding GAM are presented in Figures (24, 25 (section A.4 in Appendix) and 11 and 12 respectively.

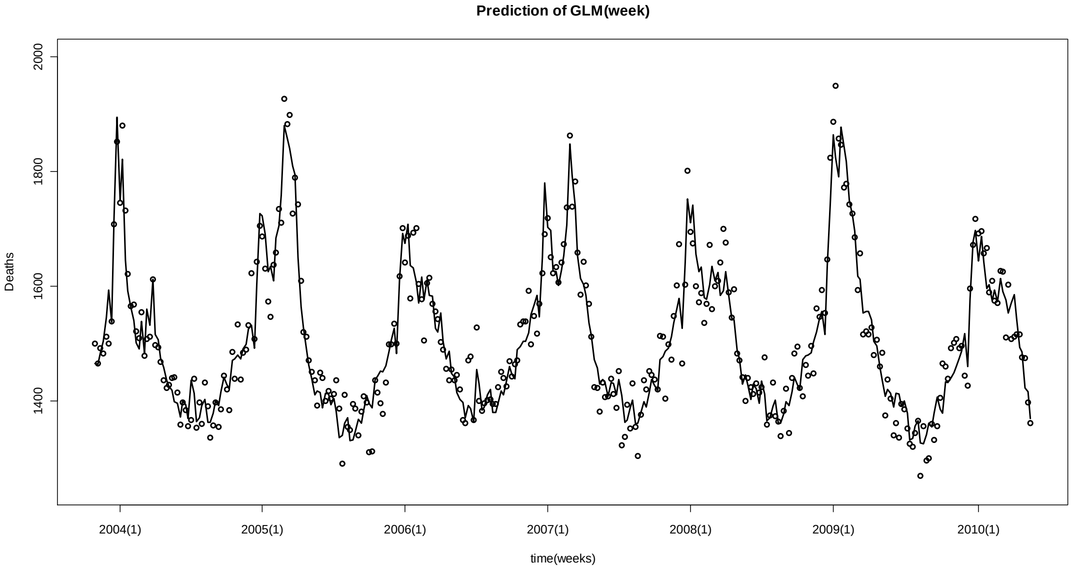

Figure 11: Predicted mortality for individuals of 65+ years with GLM (week)

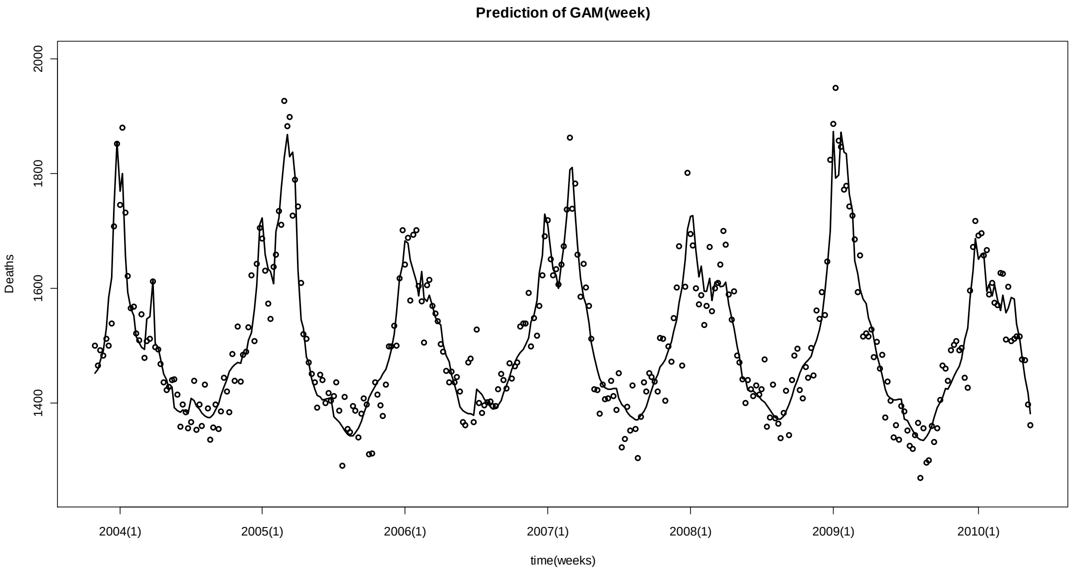

Figure 12: Predicted mortality for individuals 65+ years with GAM (week)

As a first conclusion, Model 1 can be considered as the most appropriate model among all the other models fitted if we compare the figures and the statistics. Model 5 gives slightly better results and can be equivalently used, especially when the effect of each influenza strain is of interest. The corresponding GAMs (models 2 and 6) do not seem to predict mortality as efficiently as the previous models, but can be used when more flexible and smooth models are needed. The disadvantage of models 1 and 5 is the wiggliness in their predictions, while models 2 and 6 do not capture the peaks of mortality very accurately.

It has been shown that peak mortality lags roughly 1-2 weeks behind the peak of influenza activity[5]. This is likely the result of several different time delays from different influenza mortality pathways. For this reason, alternative model forms were tested for possible delays. No obvious lags were detected between mortality and viral reporting. Models, in which deaths lagged laboratory-confirmed cases by approximately 1 week, were found to be a slightly better than the original models when month was used as a seasonal variable, but decreased the fit when week was used. Thus, the original models were considered as the most appropriate ones. All other alternatives decreased the model fit, or produced equivalent models.

5.3 Estimating Excess Mortality

The effect of all three infections: Influenza, RSV and Norovirus changes dramatically every season. For this reason, we employed a model that accounts for changes of the three diseases over the seasons. The problem with that approach is that some coefficients were negative, which means that the presence of a particular disease in that particular season had a protective effect on the population. The above statement is obviously wrong, because the presence of infections in the population adds an extra burden in mortality. In order to avoid unrealistic results, we set the negative coefficients to zero, claiming that we had no effect from the particular disease on that season. The problem with the models described above is that they may predict quite well the seasonal variation of influenza but they are too complex to interpret easily. A simpler model can be used instead in order to overcome this problem. If we neglect the fact that the intensity of each disease changes every season we could have a simple model:

$$log(\mu_{ij}) = \beta_0 + \beta_j^w + \beta_i^s + \beta^I x_{ij} + \beta^R y_{ij} + \beta^N z_{ij}$$

This simple model explains 88% of the deviance. This is somewhat lower than our original model but not that different. The coefficients of RSV and NoV were found to be insignificant, which means that these two viruses have very small or no effect in mortality. Also, the overdispersion parameter was 1.6, which was higher than before. This means that some important covariate has been left out of the model. An intermediate approach would be to account for seasonal changes only for influenza, because influenza is the main cause of excess mortality. An intermediate model would be

$$log(\mu_{ij}) = \beta_0 + \beta_j^w + \beta_i^s + \beta_i^I x_{ij} + \beta^R y_{ij} + \beta^N z_{ij}$$

This model explains 89% of the variation which is a slight improvement. Again, the coefficients of RSV and NoV were not significant, which means that these diseases change seasonally. An encouraging fact was that all coefficients of influenza were positive, except for 2009-2010 season, but that was a problematic season from the beginning. The overdispersion parameter was 1.46 which was again higher than the original model.

The average of the coefficients of influenza in Model 1 is $20.11/10^5$ of population. This means that almost 20 deaths are related to each reported case of influenza per 100000 of population. The rates of the GAM and the models with *month* (Models 2-4) are similar to each other, but seem to overestimate the effect of influenza in mortality ($35/10^5$). The model that accounts the different influenza strains give an average coefficient of $19.4 / 10^5$for Influenza A, which means that influenza A is responsible for the majority of deaths due to influenza. If only influenza reports are used to predict mortality, then the effect of influenza is overestimated, with a rate of $86.2/10^5$. Also, in the simple and intermediate model introduced late in the Results chapter, the effect of influenza is higher than the original model ($46/10^5$ and $25/10^5$ respectively).

For persons over 65 years the coefficient of influenza is $93/10^5$ for the GLM with week and $99/10^5$ for the GAM with week. For GLM with month the effect of influenza is $107/10^5$, which is again much higher than the rest of the models. As expected, influenza has a bigger impact in older people. In the simple model the effect of influenza is again higher, $124/10^5$ and for the intermediate model, the corresponding effect is $94/10^5$.

The concept of baseline mortality was discussed in the introduction of this paper. It is the expected mortality if influenza or the other viruses were not circulating. Baseline mortality is calculated by setting the coefficients of all terms that include viruses in the model to zero. In the graph below, one can observe the baseline mortality as calculated by the original model introduced. Baseline mortality appears as a periodic function that captures the low levels of mortality, but not its peaks. The number of deaths that are not explained by baseline mortality is called excess mortality and these deaths are attributed to influenza and the other viruses every year.

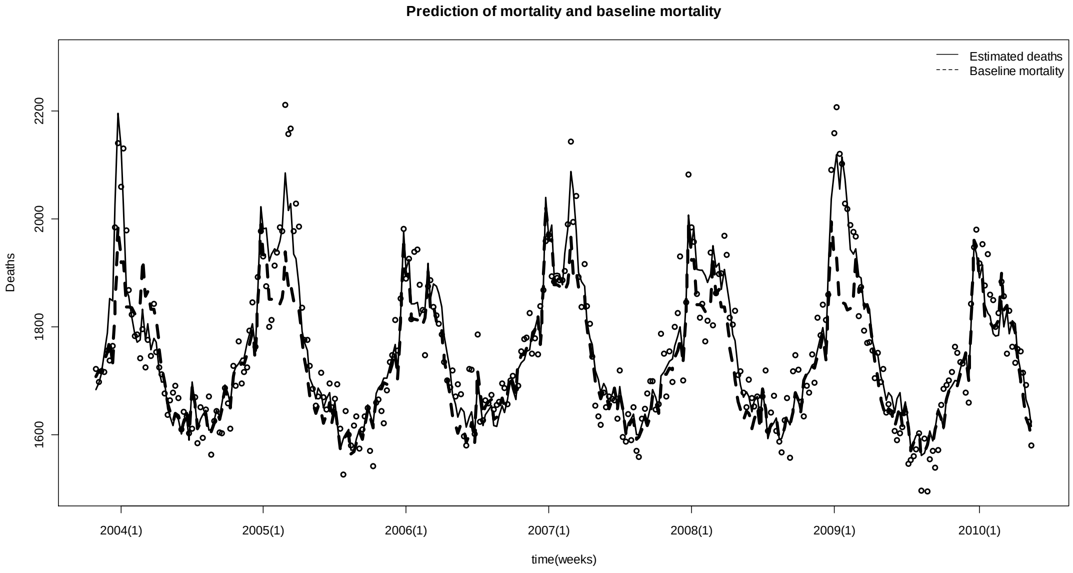

Figure 13: Predicted and baseline mortality

In an effort to quantify the burden of each of the diseases the aggregated number of excess deaths for each infection is calculated. Except the original model, the simple and the intermediate approach are also used. Furthermore, the corresponding GAMs are employed as well. The coefficients that correspond to all other viruses except the virus of interest are set to zero. The predicted values of these modified models give the number of deaths by a specific cause. If, then, baseline mortality is subtracted from this vector, the number of excess deaths by a specific cause is obtained.In the Tables 5 and 6 one can see the estimated excess deaths attributed to each virus. In average we have approximately 850 excess deaths every year according to the original model and 1050 according to its corresponding GAM attributed to influenza. Around 300 are attributed to RSV every year and around 1500 to NoV. These results seem unrealistic and so, the more simple approaches should be considered instead. With the simpler approaches around 1400 excess deaths are attributed to influenza every season, 200 to RSV and 300 to NoV.

Table 5: **Excess mortality attributed to each infection (GLM)**

	Complex Model			**Intermediate Model**			**Simple Model**		
Season	Influenza	RSV	NoV	Influenza	RSV	NoV	Influenza	RSV	NoV
2004	865	0	769	1020	167	37	1415	34	45
2005	1591	1441	0	2662	89	138	1775	18	170
2006	0	561	1297	485	284	105	823	58	130
2007	941	93	0	1307	105	330	1225	21	406
2008	0	0	4017	559	298	269	1111	60	331
2009	1687	0	2741	2252	199	477	1890	40	586
2010	0	0	1938	0	295	347	225	60	427
Average	847	299	1537	1381	205	243	1373	42	299

Table 6: **Excess mortality attributed to each infection (GAM)**

	Complex Model			**Intermediate Model**			**Simple Model**		
Season	Influenza	RSV	NoV	Influenza	RSV	NoV	Influenza	RSV	NoV
2004	810	0	0	1028	163	31	1437	38	39
2005	2317	1565	0	2682	87	116	1803	20	148
2006	0	369	968	501	277	88	836	64	113
2007	1383	332	0	1354	102	277	1244	24	353
2008	0	0	3085	572	291	225	1128	67	288
2009	1806	0	1729	2243	194	399	1920	45	509
2010	0	0	1434	0	288	291	229	67	371
Average	1053	324	1031	1397	200	204	1395	46	260

One interesting thing to be noted first is that in none of the seasons we have excess mortality because of all three causes together. This may be a sign of overfitting. The model cannot account for all three causes simultaneously in each season and that's why it has negative coefficients for some viruses in some seasons. As seen from the Tables above, the number of excess deaths change according to each approach, but not dramatically. The only case that we see a big difference is in the estimations of RSV for the simple model. The estimated number of deaths is around 40, which is considerably lower than for the other models. This can be explained if we consider the fact that the coefficients for RSV were not significant and very close to zero. That means that the model attributed only a small part of excess mortality to RSV. On the other hand, the original model gives some very large numbers for RSV and especially NoV. Because of the negative coefficients in some seasons, the model overestimates the excess deaths for the other causes and shows extremely large numbers of excess deaths in certain seasons and cannot be trusted blindly. Excess mortality for each cause according to complex and simple model is shown in Figures 14 and 15.

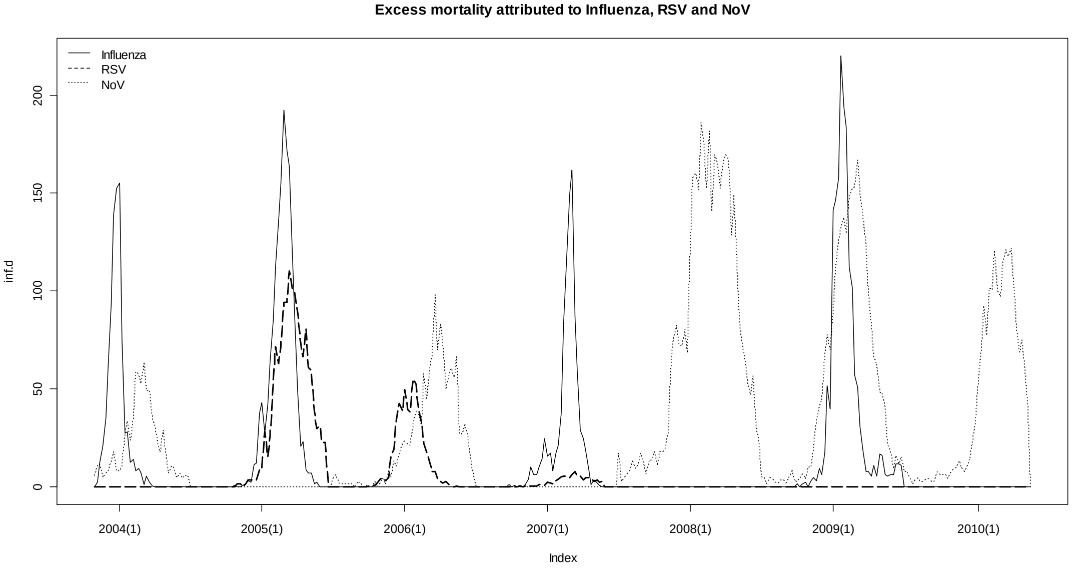

Figure 14: Excess mortality attributed to Influenza, RSV and NoV

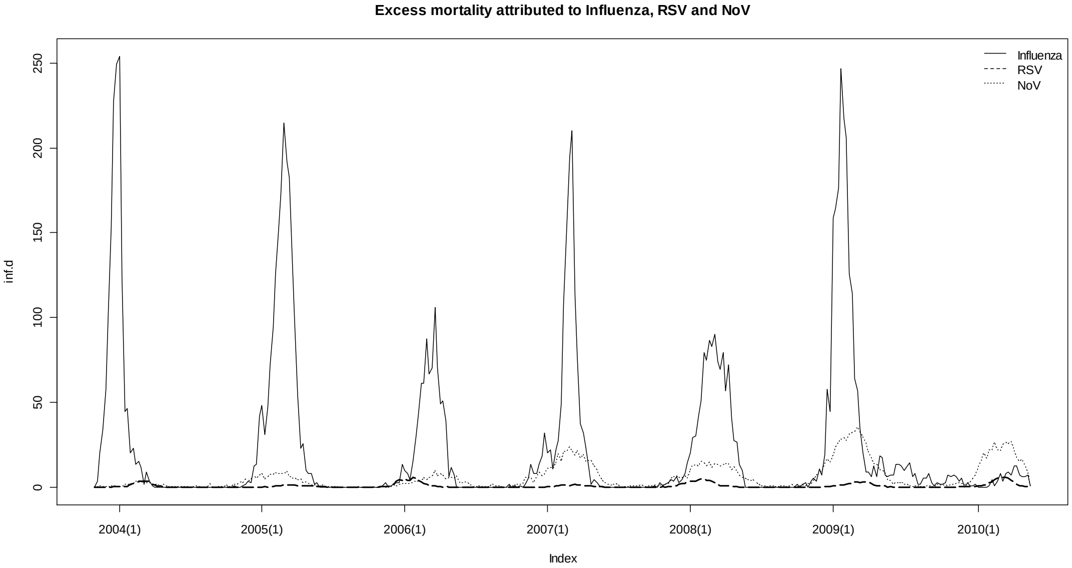

Figure 15: Excess mortality attributed to Influenza, RSV and NoV (Simple model)

The same procedure is followed for the persons in the 65+ age group. For this group, the results of all three approaches are very similar. In average, there are 1100 deaths attributed to influenza, 200 deaths attributed to RSV and 400 deaths attributed to NoV. The excess mortality from each cause can be also depicted in Figure 16 and 17.

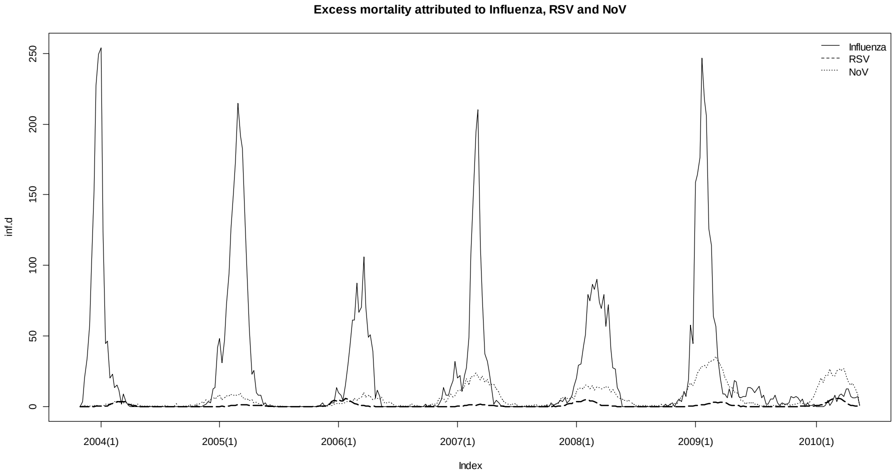

Figure 16: Excess mortality attributed to Influenza, RSV and NoV for 65+

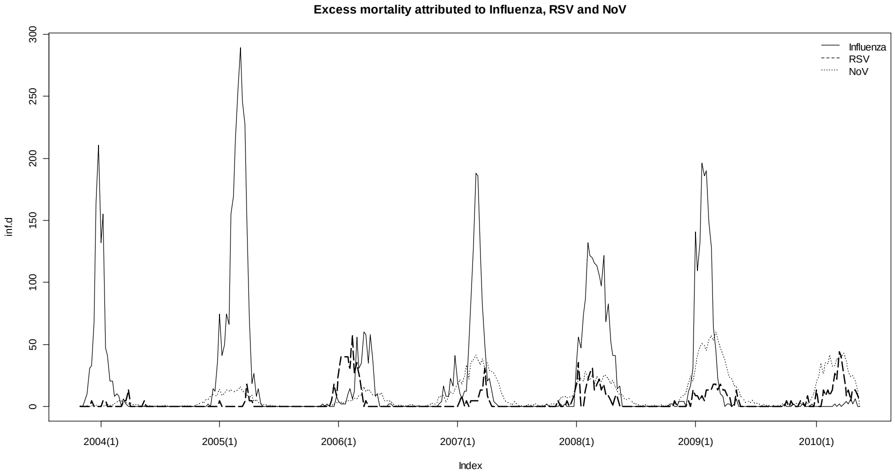

Figure 17: Excess mortality attributed to Influenza, RSV and NoV for 65+ (Simple model)

6 Conclusions/Discussion

Various approaches for predicting mortality in Sweden have been developed in this paper. Generalized linear models and generalized additive models were the two methods used to explain the counts of deaths. Models with influenza, RSV and NoV reported cases, as explanatory variables, were constructed, accounting *week* as the seasonal effect and factor *season* to capture long term trends and population growth. The relationship between mortality and the variables seems linear for all cases. The slope for all graphs in the scatter-plot is positive, which means that all infections contribute to the increase of mortality. The slope of the graphs varies for the different variables, larger for influenza and influenza A and almost zero for influenza B.

A generalized linear model that accounted the differences of the three infections each season was considered as the most appropriate to predict the mortality in Sweden. Although this model provided better fit to the data than the corresponding generalized additive model, GAM could still be employed as a rather smoother approach to explain mortality, since it uses much fewer parameters than the GLM (36 parameters instead of 79). Models with *month* instead of *week* were used in order to decrease the number of parameters, but both their fit and predictive power were insufficient. Models that accounted for the different influenza strains (influenza A and B) and week as seasonal effect were fitted, to get a more detailed view of the influenza effect on mortality and they provided slightly better results than the original model. In many seasons, we observe two peaks in the same season. Generalized linear models with week, that accounted either the total cases of influenza or the different strains, capture these two peaks quite accurately, with the latter models to be slightly better. Models without RSV and NoV cannot predict mortality satisfactory enough, and also overestimate the burden in mortality caused by influenza.

The problem with our model is that although it predicts well the mortality, it is not easy to interpret. There is probably an overfitting problem in our model, because some coefficients of the diseases in some seasons are negative and thus, in these seasons, it is implied that the presence of the viruses have a protective effect. It is not possible with the current model to predict excess mortality from all causes for all seasons. Simpler models can provide a solution to this problem, but their predictive power is not as good as the original model's and also their estimates are not as accurate because of the presence of overdispersion. An intermediate model that accounted for seasonal changes only for influenza was fitted, but again the fitting of the data was not satisfactory. The problem of finding a suitable model is yet to be solved.

The average of the coefficients of influenza in Model 1 is $20.11/10^5$ of population. This means that almost 20 deaths are related to each reported case of influenza per 100000 of population. This is in a reasonable agreement with estimates found in other studies. Thompson et al. [3] found the same rate ($20/10^5$) while Schanzer et al [6] found a lower rate ($13/10^5$).The rates of the other models were reasonably higher. Models that did not account for seasonal changes of some or all of the diseases, seem to overestimate the effect of influenza. For persons over 65 years the coefficient of influenza is $93/10^5$ for the GLM with week. As expected, influenza has a bigger impact in older people. Our findings are close to the findings of other studies. Andersson et al [13] found a rate of $90/10^5$, Thompson et al [3] a rate of $133/10^5$ and Schanzer et al [6] a rate of $108/10^5$ for the 65+ age group.

The baseline mortality calculated, represents the theoretical mortality without the circulation of the three viruses in the population. Baseline mortality captures the low points of mortality quite well, with the exception of the beginning of last season (2010). At that point, we observe an unexpected low number of deaths. This is probably because the year before that (2009) was a year with high influenza activity which resulted in high mortality and eliminated many individuals, especially from the high risk groups. So, the following year, the population had less of these individuals, which means that there were more 'healthy' people in the population, and thus fewer deaths. This phenomenon is called the "harvesting effect". In average we have approximately 1400 excess deaths every year attributed to influenza, 200 attributed to RSV and 300 attributed to NoV. For the 65+ age group, we have approximately 1100 deaths attributed to influenza, 200 deaths attributed to RSV and 400 deaths attributed to NoV. From the figures 14-17 one can recognize the seasons when there was a severe epidemic of a certain virus. For example, 2006 season was a severe influenza season, just like 2009. Also, 2008 and 2009 was a severe season for NoV.

As a general comment we could say that Norovirus can be considered as a substantial cause of excess mortality. So far, not many studies investigated the effect of Norovirus in excess mortality. From our findings, it is suggested that NoV is a more serious cause of mortality than RSV. Also, NoV is a serious threat for the persons above 65 years of age, along with influenza.

From our study, it turns out that influenza is the main cause of excess mortality both in the whole population and in the 65+ age group. Along with influenza, RSV and NoV also contribute to a substantial number of excess deaths every year. Influenza A is the main cause of mortality if the two different influenza strains are considered. Influenza B does not seem to have a significant effect on mortality in Sweden.

The seasonal variation is sometimes modeled by harmonic functions, like sine or cosine, which reduces the number of parameters in the model, but imposes strong restrictions on the form of the seasonal component. Week proved an efficient way to capture seasonality, with the cost of 52 extra parameters. A smooth function of week reduces the number of parameters used, but at the same time decreases the predictive ability of the model.

The residual winter mortality that remains to be explained could be attributed to various reasons, such as extreme cold, lack of access to doctors and/or antibiotics during the holidays and other respiratory tract viral infections that might circulate in the same period.

In the data set there seems to be no outliers. This is verified from the residual plots in the models. In each model, different observations appear to have large residuals, so the existence of possible outliers can be ruled out.

One of the restrictions of the study is that we only have seven seasons of data. It would be interesting to see, how the models behave if more seasons were added, especially, seasons with more varied patterns of outbreaks of infections. Moreover, in the present paper, no simulation techniques were performed. Possibly a simulation study with the bootstrap method or permutation tests would strengthen the validity of our results. Also, the models might work better if other kinds of data were available. Since there are big geographical differences throughout Sweden, mainly in the temperature, perhaps different data for each region would provide more clear results. Another aspect of the subject of the thesis that has not been mentioned at all is vaccination and vaccine matching. Both these factors are very important and should not be forgotten in a future expansion of the thesis, especially for the elderly people. Data containing how many people were vaccinated and/or whether the vaccine was a good match every season, might be very informative.

7 Appendix

A.1. Graphical presentation of the data

The relation of mortality and influenza is depicted quite clearly in Figure 18. One thing that is important to be mentioned is that across all seasons, the peaks in mortality coincide with the peaks in influenza reported cases.

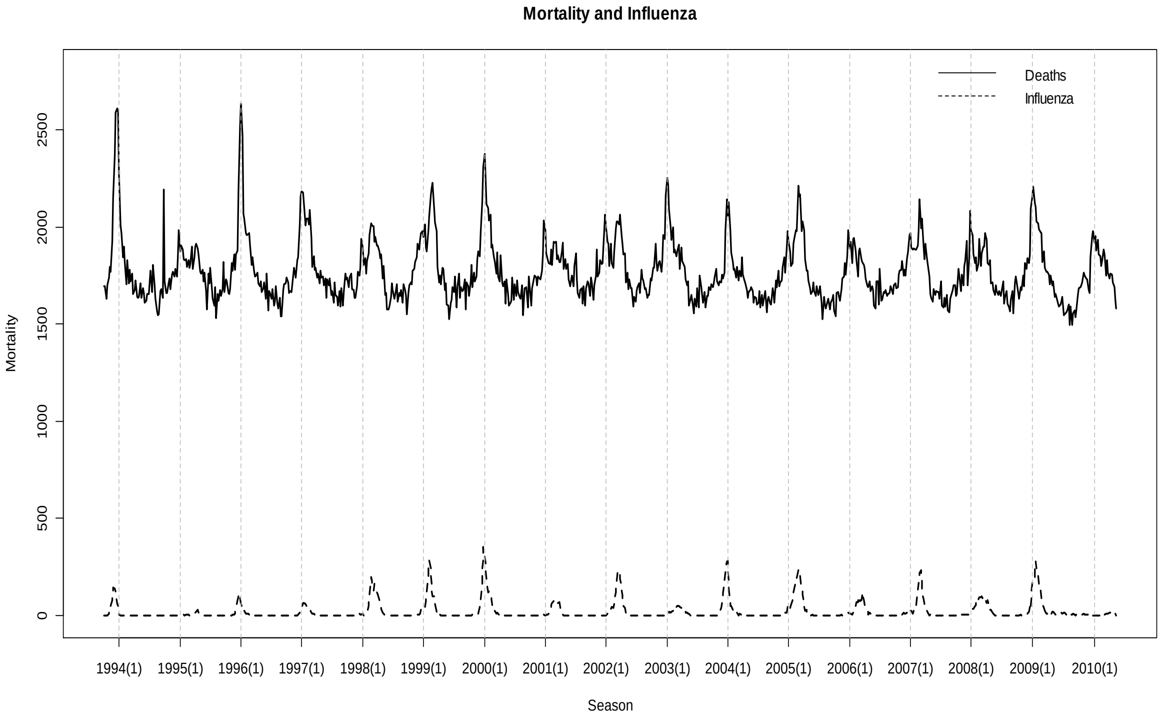

Figure 18: Observed Deaths and Influenza reported cases

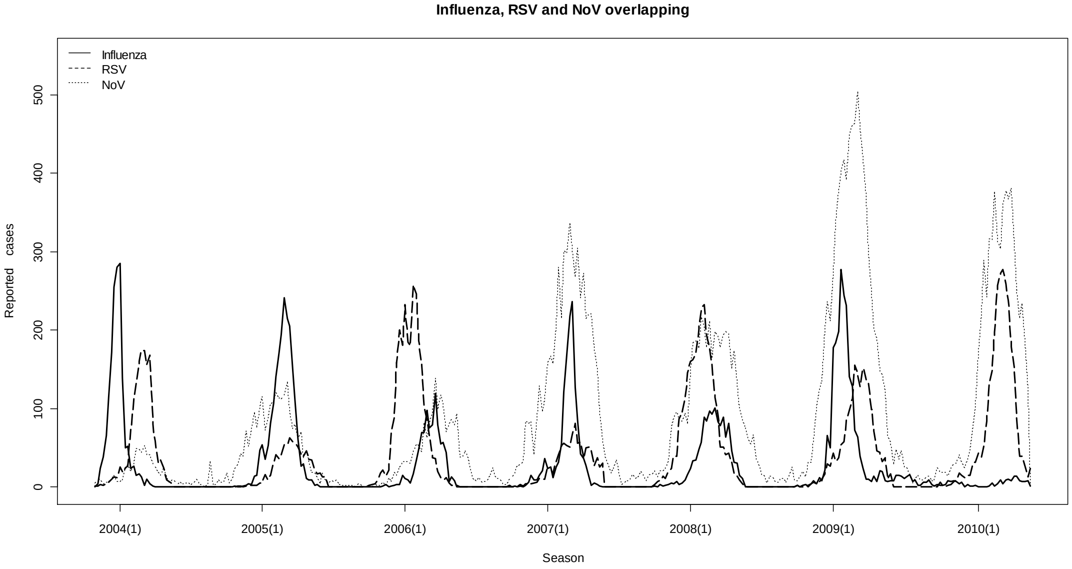

Figure 19: Influenza, RSV and NoV reported cases

In Figure 19 one can observe the overlapping of the three viruses of the study. Notice that although the intensity of the different viruses changes across the seasons, they all peak around the same time for the given period. Data for all three viruses are only available for the last seven seasons, because there was no systematic reporting for RSV and Norovirus in the past. Another thing that is interesting in this graph is that in the last season (2010) there were too few reported cases of influenza, because of the H1N1 flu pandemic.

A.2. Histograms of all variables

From the histograms we can see the distribution of each variable. The large value of skewness is quite clear for all observations, but also expected for this kind of data. The mass of observations on value zero can be easily spotted in the histograms for the diseases.

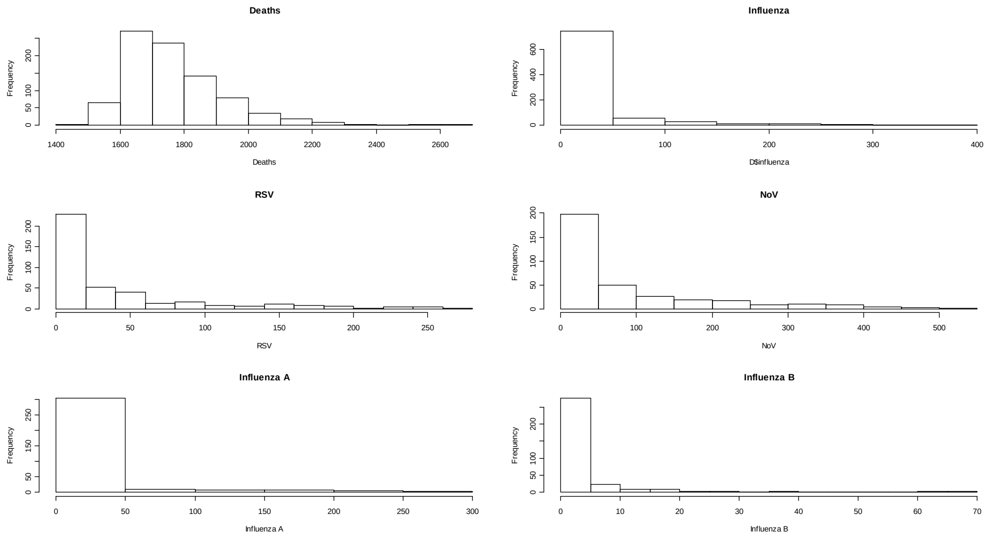

Figure 20: Histograms of the variables (a)

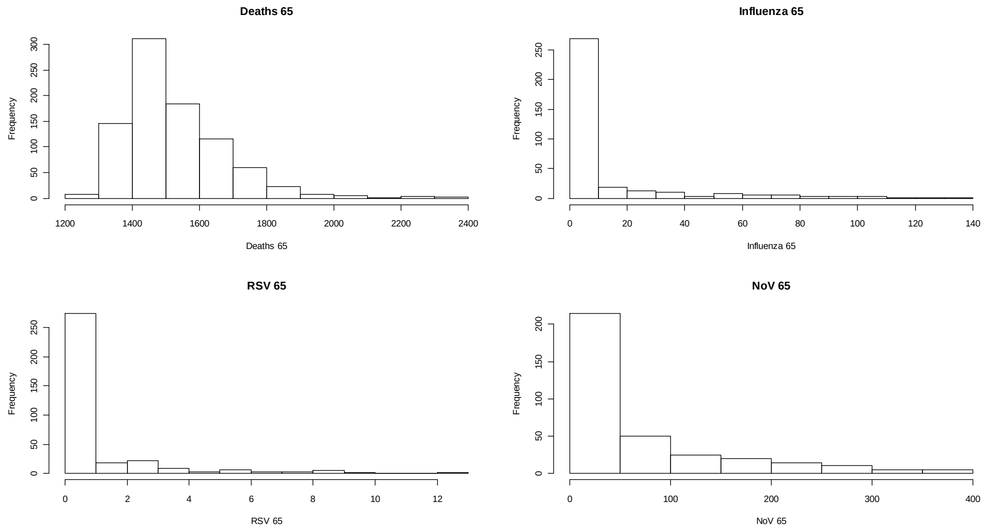

Figure 21: Histogram of the variables (b)

A.3. Scatter-plot

From the scatter-plot we can see that deaths are correlated with all the diseases. The relationship seems linear for all cases. The fact that a large proportion of the counts of the diseases equals zero, makes this conclusion less clear. The slope for all graphs is positive. The slope of the graphs varies for the different variables, larger for influenza and influenza A and closer to zero than to one for the other viruses. This means that all diseases contribute, even in a smaller or larger extend to mortality.

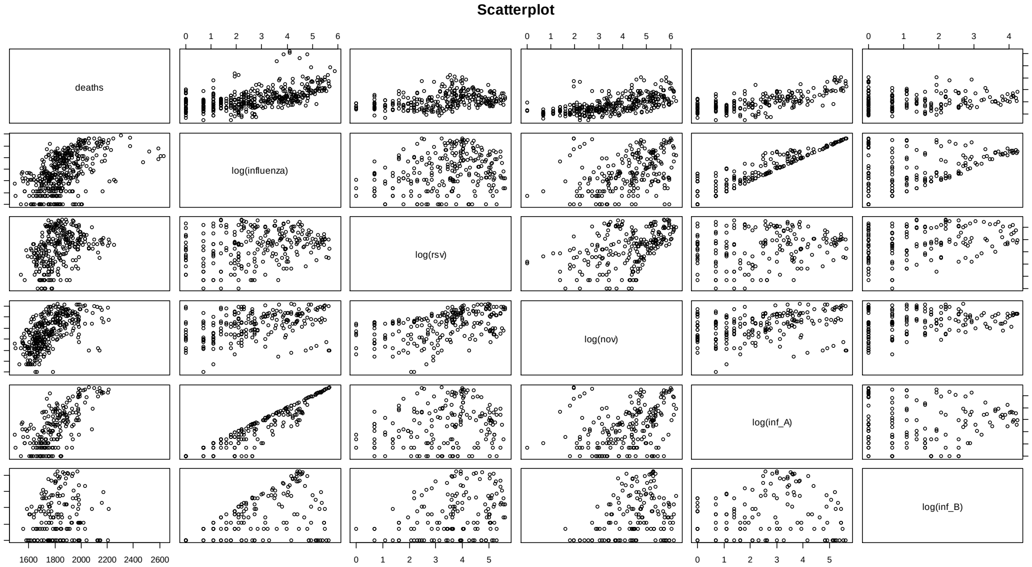

Figure 22: Scatter-plot of all variables

A.4. Goodness of fit diagnostics

Model 1

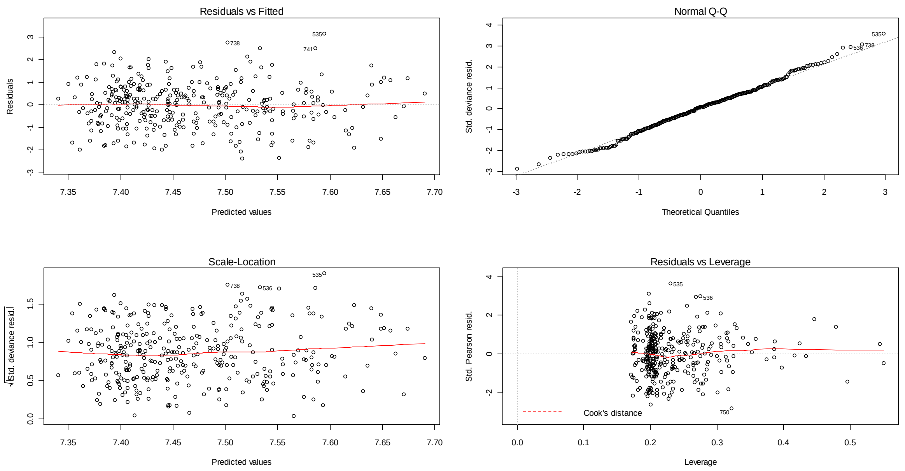

Figure 23: Model fit for Model 1 (a)

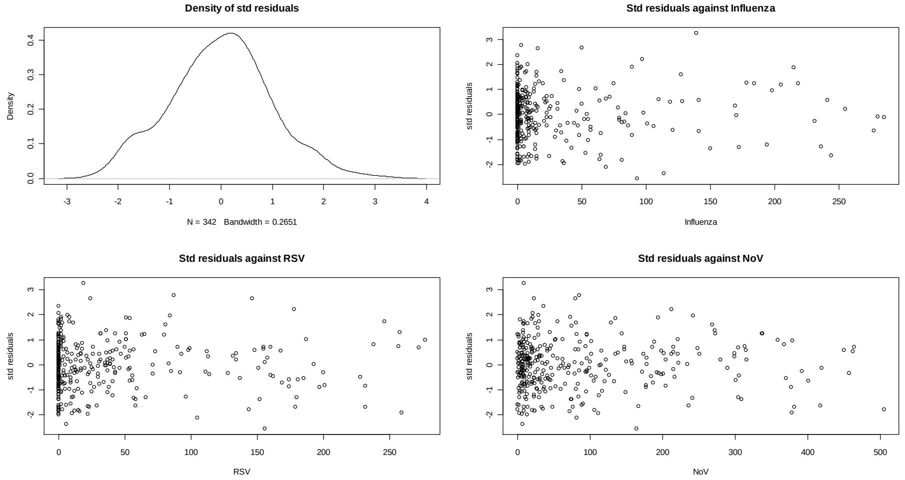

Figure 24: Model fit for Model 1(b)

Table 7: **Analysis of Deviance Table**

Response: deaths

Terms added sequentially (first to last)

	df	Deviance	Resid. Df	Resid.Dev	P(>|Chi|)
NULL			341	3736.4	
week	52	2890.49	289	845.9	< 2.2e-16 ***
season	6	160.74	283	685.2	< 2.2e-16 ***
influenza	1	240.21	282	444.9	< 2.2e-16 ***
season*influenza	6	43.47	276	401.5	9.401e-08 ***
season*rsv	7	42.08	269	359.4	5.022e-07***
season*nov	7	32.97	262	326.4	2.687e-05 ***
—					

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Model 2

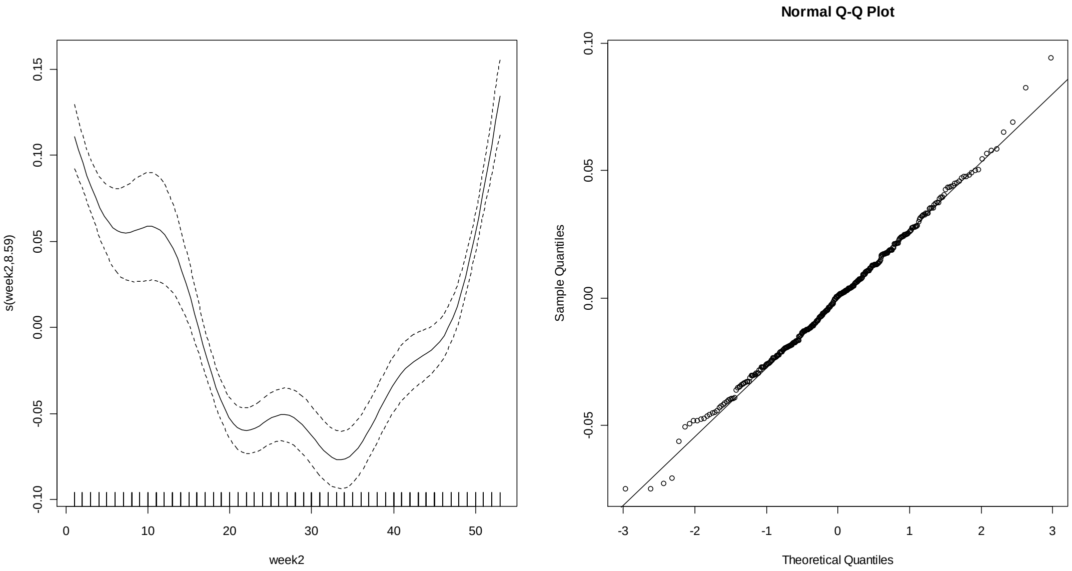

Figure 25: Smooth function for week and QQ plot for Model 2

Table 8: **Analysis of Variance Table**

Family: quasipoisson			
Link function: log			
deaths ~ f(week2)+ season + season:influenza + season:rsv + season:nov			
Parametric Terms:			
	df	F	p-value
season	6	3.999	0.000718***
season:influenza	7	9.445	1.30e-10***
season:rsv	7	3.164	0.003033**
season:nov	7	3.195	0.002796**
Approximate significance of smooth terms:			
	Edf	F	p-value
f(week2)	8.589	40.61	<2e-16***
—			
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Model 3

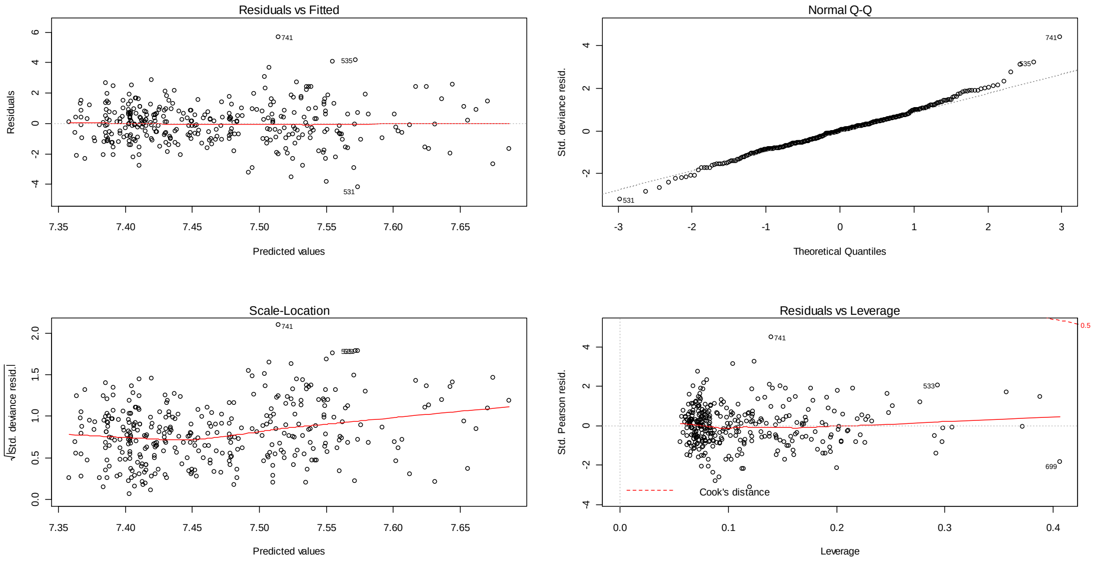

Figure 26: Model fit for Model 3 (a)

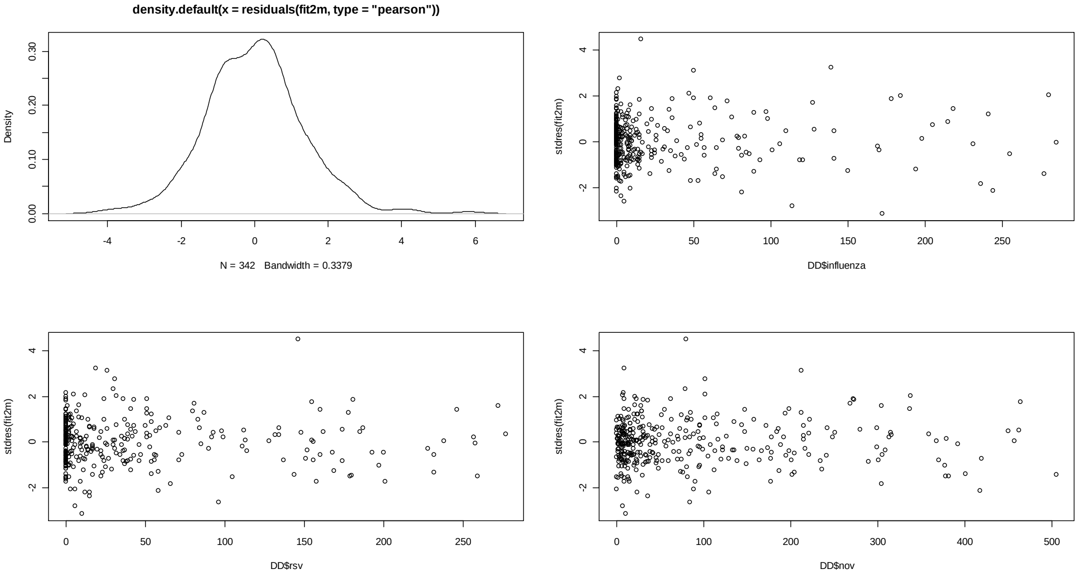

Figure 27: Model fit for Model 3 (b)

GLM with week for the 65+ age group

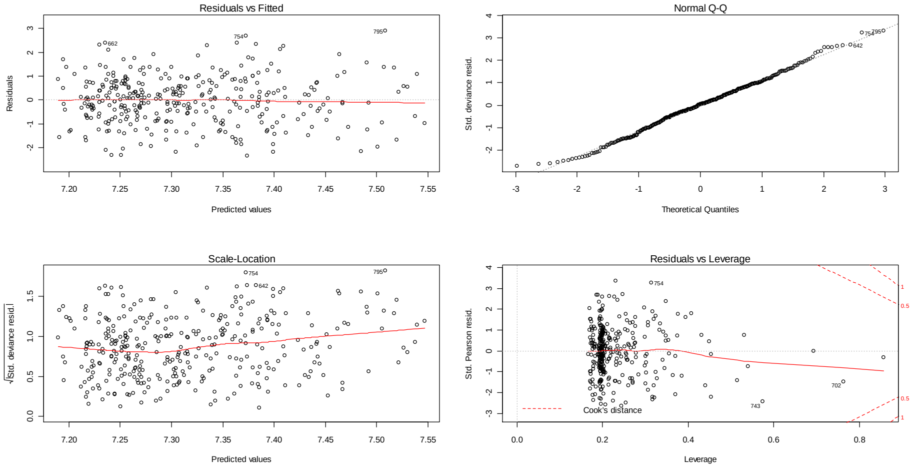

Figure 28: Model fit for 65+ age-group

GAM with week for the 65+ age group

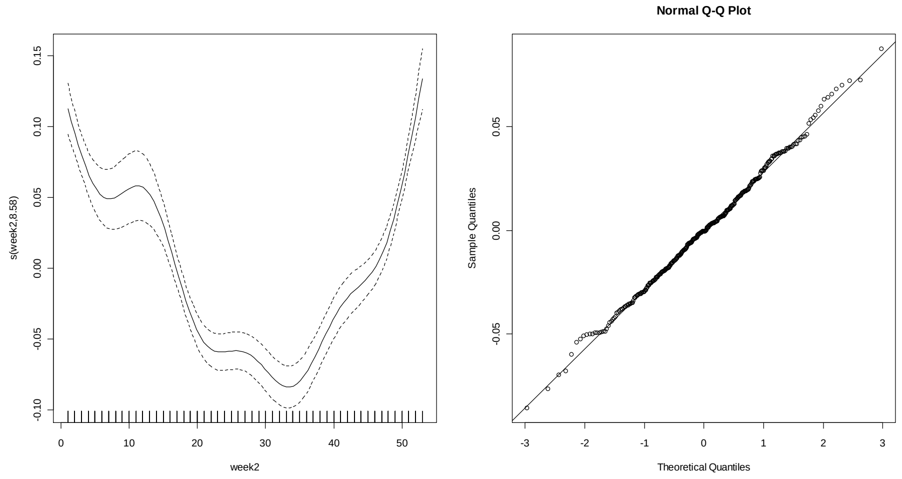

Figure 29: Smooth function for week and QQ plot

A.5. Fitting Generalized Linear Models

The fitting of GLM is achieved by using a method called *Iteratively Re-weighted Least Squares*. The main hypothesis for generalized linear models is that the response variable follows an exponential family distribution and the elements Y_i of the vector $\mathbf{Y}$ are mutually independent. The likelihood of the parameter vector β is

$$L(\beta) = \prod_{i=1}^{n} f_{\theta_i}(y_i),$$

and hence, the log-likelihood of β is

$$\log L(\beta) = \sum_{i=1}^{n} \log[f_{\theta_i}(y_i)] = \sum_{i=1}^{n} \{y_i\theta_i - b_i(\theta_i)\}/a_i(\varphi) + c_i(\varphi, y_i),$$

where the equation depends on β is through the dependence of the $\theta_i \text{on} \beta$.

The estimation of β requires the maximization of the log-likelihood, which is achieved by partially differentiating logL with respect to each element of β, setting the resulting expressions to zero and solving for β. After some calculations the equation to solve for β is

$$\frac{\partial \log L}{\partial \beta_j} = \frac{1}{\varphi} \sum_{i=1}^{n} \frac{[y_i - b_i^{'}(\theta_\iota)]}{b_i^{''}(\theta_\iota)\alpha(\varphi)/\varphi} \frac{\partial \mu_\iota}{\partial \beta_j} = 0$$

which can be re-written as

$$\sum_{i=1}^{n} \frac{[y_i - \mu_\iota]}{V(\mu_\iota)} \frac{\partial \mu_\iota}{\partial \beta_j} = 0, \forall j$$

However, these equations are exactly the same as if non-linear weighted least squares method was used, if the weights $V(\mu_i)$ were known in advance and were independent of β. In this case, the least squares objective would be

$$S = \sum_{i=1}^{n} \frac{(y_i - \mu_\iota)^2}{V(\mu_\iota)}$$

where μ_i depends non-linearly on β, but the weights $V(\mu_i)$ are treated as fixed. To find the least squares estimates involves solving$\partial S/\partial \beta_j = 0, \forall j$. This requires the use of an iterative method until convergence.

The above equation can be written in matrix form

$$S = \left\| \sqrt{V_{[k]}^{-1}} [y - \mu(\beta)] \right\|^2,$$

where $\mathbf{V}_{[k]}$ is the diagonal matrix such that $V_{[k]ii} = V(\mu_i^{[k]})$. If μ is replaced by its first order Taylor expansion around $\hat{\beta}^{[k]}$

$$S \approx \left\| \sqrt{V_{[k]}^{-1}} [y - \mu^{[k]} - J(\beta - \hat{\beta}^{[k]})] \right\|^2$$

where $\mathbf{J}$ is the 'Jabocian' matrix, with$J_{ij} = \partial \mu_i / \partial \beta_j |_{\hat{\beta}^{[k]}}$. Now,

$$g(\mu_i) = X_i \beta \Rightarrow g'(\mu_i) \frac{\partial \mu_i}{\partial \beta_j} = X_{ij}$$

and hence

$$J_{ij} = \left. \frac{\partial \mu_i}{\partial \beta_j} \right|_{\hat{\beta}^{[k]}} = X_{ij} / g'(\mu_i^{[k]}).$$

So, by defining $\mathbf{G}$ as the diagonal matrix with elements$G_{ii} = g'(\mu_i^{[k]})$, $\mathbf{J} = \mathbf{G}^{-1}\mathbf{X}.$

Hence,

$$S \approx \left\| \sqrt{V_{[k]}^{-1}} G^{-1} [G(y - \mu^{[k]}) + \eta^{[k]} - X\beta] \right\|^2 = \left\| \sqrt{W^{[k]}} (z^{[k]} - X\beta) \right\|^2$$

where $z^{[k]}$ are called 'pseudo-data' with definition $z^{[k]} = g'(\mu^{[k]})(y_i - \mu_i^{[k]}) + \eta_i^{[k]}$ and $\mathbf{W}^{[k]}$ is the diagonal weight matrix with elements

$$W_{ii}^{[k]} = \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2}.$$

The following step are iterated to convergence.

Using the current $\mu^{[k]}$ and $\eta^{[k]}$ calculate pseudo-data $z^{[k]}$ and iterative weights $\mathbf{W}^{[k]}$.

Minimize the sum of squares $\left\|\sqrt{W^{[k]}}(z^{[k]} - X\beta)\right\|^2$ with respect to β, in order to obtain $\hat{\beta}^{[k+1]}$, and hence $\eta_i^{[k+1]} = X\hat{\beta}^{[k+1]}$and $\mu^{[k+1]}$. Increase k by one.

The converged $\hat{\beta}$ is the maximum likelihood estimator of β.

Notice that to start the iteration we only need $\mu^{[0]}$ and $\eta^{[0]}$ values, but not $\hat{\beta}^{[0]}$. Hence, the iteration is usually started by setting $\mu_i^{[0]} = y_i$ and $\eta_i^{[0]} = g(\mu_i^{[0]})$.

A.6. Penalized Iterative Re-weighted Least squares (P-IRLS)

In order to fit a generalized additive model, penalized iterative re-weighted least squares method is iterated to convergence. The steps of the method are the following:

1. Given the current linear predictor estimate $\eta^{[k]}$, and corresponding estimated mean response vector, $\mu^{[k]}$, calculate:

   $$\omega_i \propto \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2}$$

   and

   $$z_i = g'(\mu_i^{[k]})(y_i - \mu_i^{[k]}) + X_i\beta^{[k]}$$

   where Var(Y$_i$) = φ V($\mu^{[k]}$) and $\mathbf{X}_i$ is the ith row of $\mathbf{X}$.

2. Minimize the quantity

   $$||\sqrt{W}(z - X\beta)||^2 + \lambda_1\beta^{\mathrm{T}}S_1\beta + \lambda_2\beta^{\mathrm{T}}S_2\beta$$

   with respect to β to obtain $\beta^{[k+1]}$, and hence $\eta^{[k+1]} = \mathbf{X}\beta^{[\mathbf{k+1}]}$. $\mathbf{W}$ is a diagonal matrix such that $W_{ii} = w_i$.

Step 2 can be replaced by the equivalent:

Minimize

$$\left\| \begin{bmatrix} \sqrt{W} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} z \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ B \end{bmatrix} \beta \right\|^2$$

with respect to β to obtain $\beta^{[k+1]}$, and hence $\eta^{[k+1]} = \mathbf{X}\beta^{[\mathbf{k+1}]}$. $\mathbf{B}$ is a matrix square root such that $\mathbf{B}^T\mathbf{B} =\lambda_1\mathbf{S}_1 + \lambda_2\mathbf{S}_2$.

8 References

[1] Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports* 1963;78:494-506.

[2] Simonsen L et al. The impact of influenza epidemics on mortality: introducing a severity index. *American Journal of Public Health* 1997;87:1944-1950

[3] Thompson WW et al. Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA* 2003;289:179-186.

[4] Glezen WP, Payne AA, Snyder DN, et al. Mortality and influenza. *J Infect Dis* 1983;146:313-321

[5] Donaldson GC, Keatinge WR. Excess winter mortality: Influenza or cold stress? *BMJ* 2002;324:89-90

[6] Schanzer DL et al. Influenza – attributable deaths, Canada 1990-1999. *Epidemiology and Infection* 2007;135:1109-1116.

[7] Dushoff J et al. Mortality due to Influenza in the United States-An annualized regression approach using multiple-cause mortality data. *American Journal of Epidemiology* 2006;163:181-187.

[8] Wong CM et al. Influenza-associated mortality in Hong Kong. *Clinical Infectious Diseases* 2004;39:1611-1617.

[9] Newall AT et al. Influenza-related hospitalization and death in Australians aged 50 years and older. *Vaccine* 2008;26:2135-2141.

[10] Nichol KL, Goodman M. The health and economic benefits of influenza vaccination for healthy and at-risk persons aged 65 to 74 years. *Pharmacoeconomics* 1999;16:63-71.

[11] Nichol KL, Nordin J, Mullooly J, et al. Influenza vaccination and reduction in hospitalizations for cardiac disease and stroke among the elderly. *N Eng J Med* 2003;348:1322-1332.

[12] Glezen WP. Serious Morbidity and Mortality Associated with Influenza Epidemics. *Epidemiologic Reviews* 1982;4:25-44.

[13] Andersson M et al. Excess Mortality Related to Outbreaks of Influenza, Respiratory Syncytial Virus and Norovirus in Sweden 2004-2007. *Submitted to Epidemiology and Infection.*

[14] Centers for disease control and prevention (www.cdc.gov)

[15] Jansen AG et al. Influenza- and respiratory syncytial virus-associated mortality and hospitalizations. *The European Respiratory Journal* 2007;30:1158-1166.

[16] World Health Organization (www.who.org)

[17] Fleming DM, Pannell RS, Cross KW. Mortality in children from influenza and respiratory syncytial virus. *J Epidemiol Community Health* 2005;59:586-590.

[18] Izurieta HS, Thompson WW, Kramarz P, et al. Influenza and the rates of hospitalization for respiratory disease among infants and young children. *N Engl J Med* 2000;342:232-239

[19] Fleming DM, Cross KW. Respiratory syncytial virus or influenza? *Lancet* 1993;342:1507-1510.

[20] Gay NJ et al. Estimating deaths due to influenza and respiratory syncytial virus. *JAMA* 2003;289:2499.

[21] Swedish Institute for Infectious Disease Control (http://www.smittskyddsinstitutet.se/)

[22] Wood SN. Generalized additive models: An introduction with R. CRC-Press,2006.

[23] Zeileis et al. Regression Models for Count Data in R. *Journal of Statistical Software* 2008;27(8):1-25.

[24] Statistics Sweden (www.scb.se)

[25] International Organization for Standardization (http://www.iso.org/iso/date_and_time_format)

[26] Everitt BS, Hothorn T. A handbook of statistical analyses using R. CRC-Press,2009.

[27] Sundberg R. Lecture Notes on Statistical Modeling by Exponential Families. Stockholm University, March 2010.

[28] Dobson A.J. An introduction to Generalized Linear Models. Second Edition. Chapman & Hall/CRC Press, 2001

[29] Hastie T, Tibshirani R. Generalized Additive Models. *Statistical Science* 1986; 1.3:297-318.

[30] Hilbe J, Negative Binomial Regression. Cambridge University Press, 2007, p 73.