University of Strathclyde
Department of Mathematics and Statistics

Extending Scottish Exception Reporting Systems Spatially and Temporally

by

Adam Wagner

A thesis presented in fulfilment of the
requirements for the degree of
Doctor of Philosophy

2010

For Dad: I'm sorry you can't be here to see this, but I hope you'd be proud. Thank you for your all love and guidance – Ads

> *'We are now cruising at a level of two to the power of twenty-five thousand to one against and falling, and we will be restoring normality as soon as we are sure what is normal anyway, thank you.'*
> Douglas Adams' 'Hitchhikers' Guide to the Galaxy', Radio episode II

> *'It is known that there is an infinite number of worlds, but that not every one is inhabited. Therefore there must be a finite number of inhabited worlds. Any finite number divided by infinity is as near to nothing as makes no odds, so if every planet in the universe has a population of zero then the entire population of the universe must also be zero, and any people you may actually meet from time to time are merely the products of a deranged imagination.'*
> Douglas Adams' 'Hitchhikers' Guide to the Galaxy', Radio episode V

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: Date: March 30, 2010

Acknowledgements

As I sit here finishing off my thesis the night before I submit, it is finally time to thank the many people that have helped get me here. It humbles me that so many people have given of themselves, and of their time, to help me academically, personally and in so many other ways. For any one I neglect to mention, forgive me; I'm blessed in knowing so many good people that it is hard to recall everyone.

To Profs. McKenzie and Robertson – Eddie and Chris – my thanks for all your time and effort given over to helping me... And of course, for your patience!

To my family: Mum, your patience molded me to get me here; Uncle Mike, your strength, support and intellect reassure me that there is always someone looking out for me; Aunt Karen, your compassion, empathy and level headedness help me know there is always someone that 'gets me', even in my most neurotic moments; Gran, Granpa, Nan, and Grandad, your shared experiences help broaden my horizons and keep me balanced. Thank you all for your love.

To David Ault – Dave – the best friend: there's so much to thank you for – the encouragement, support and many long phone calls – but most of all, thank you for treating me like a brother! Long may it continue.

My friends: the Edinburgh IRC gang – Hugh Griffiths, Kate Ho and Will Foote – you kept me sane during the long hours of coding – thanks for the chat; friends from Woodlands Church – Robert MacBean, Natalie Astridge (and of course Amelie!), Jacelyn Ong, Jamie Catlow, Anna Fisk, Jenny Hunter – thanks for the many fine Sunday lunches and support; Emma McDonald – thanks for all your support and letting me rant; Liyang Hu, thanks for all the holiday breaks and random chat. Thank you all for your friendship – it means a lot to me.

Friends and fellow partners-in-crime from the department – thanks go to: my long suffering office mates, Anny (Traiani) Stari, Lukas Spruch, Graham Smith (forever the 'Wee-one'), Matina (Stamatiki) Rassias (fondly, the 'Smiley-one') and Karen Lamb; Ian Thurlbeck, for all the computer help and chat when work drives me spare; Lynne Westwood for all the 'free food'; and all the other members of the department that make it a fun and enjoyable place to work.

Contents

List of Tables

List of Figures

Abstract

Title: 'Extending Scottish Exception Reporting Systems Spatially and Temporally'

Abstract: Exception reporting systems allow medical conditions and micro-organisms to be automatically monitored for unusually high levels. Typically, statistical models are used to predict expected levels. Where observed levels exceed the predicted ones by some pre-determined amount, an 'exception' is reported to give warning. We focus on developing suitable models for use in the predictive component of such systems.

Two particular systems are extended spatially to monitor counts at the regional health board level. The first of these uses call data from the 24-hour medical helpline NHS24, to monitor particular medical syndromes. The second uses counts of positive lab identifications of micro-organisms collected by Health Protection Scotland (HPS). Regional incidences tend to have very small counts, and for these, we use negative binomial Generalized Linear Models (GLMs). However, GLMs assume that observations are independent, which is rarely, if ever, the case in the systems we consider. Two approaches are investigated for dealing with serial correlation and capturing local trend, both of which improve the models. We also investigate links between the health boards and investigate if these links can be used to further improve the models.

A new system is produced for monitoring daily all-cause mortality in Scotland, using data collated by the General Register Office. Fitting models to this data is challenging because of the sharp peaks present in the annual seasonality; to address this, we use Generalized Additive Modelling. There is also a marked delay in the reporting of deaths, which must be dealt with if the system is to detect unusually high levels of mortality in a timely fashion. We present a straight forward 'correction' to do this. Combining these elements, a mortality surveillance system is produced, which has been used by HPS to monitor mortality during the swine flu pandemic (2009).

Chapter 1

Introduction

McCabe (2004) notes at the start of his thesis that: 'In recent years there has been an increasing expectation that institutional bodies should be adequately prepared for, and respond rapidly to, events that impinge on public life' (p.1). This continues to be increasingly true and has been borne out recently in the actions of many governments to support their banking systems in response to the economic downturn. Within the health area, bodies responsible for monitoring public health face greater demands from governments, who in turn are influenced by ever greater public expectations. Reasons for the increase in expectations are myriad. Advances in information technology contribute to increasing the expectation of what it is possible to do (Lombardo 2007). Threats to public health, such as bio-terrorism and epidemics, also lead to governments putting pressure on public health bodies to provide them with information quickly and efficiently, so that the public can be kept informed in a timely fashion. Systems used for the surveillance of public health must continue to adapt, evolve and extend themselves, if they are to remain useful and fit for purpose. In this context, our research looks at extending surveillance systems used in Scotland.

1.1 Exception Reporting Systems

Among the tools used in the surveillance of disease and medical conditions are exception reporting systems. These systems allow streams of data to be automatically monitored for unusually high levels of occurrences. So, for example, if

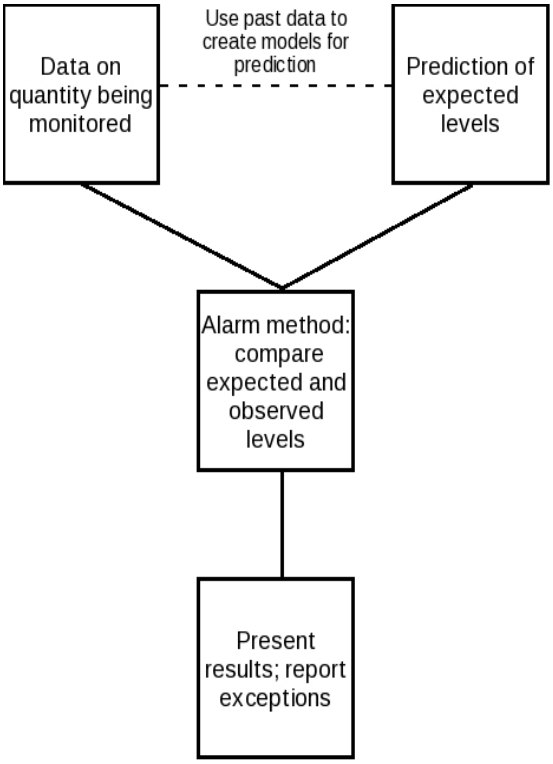


Figure 1.1: The separate elements of an exception reporting system.

monitoring the number of cases of *Salmonella*, such a system can help indicate when more than the expected number of cases have been reported – an 'exception' – ideally, taking into account such things as seasonal variation. Sometimes, an 'exception' can be indicative of things like an outbreak of a disease, while at other times, an exception can just be a high value caused by statistical variation. This distinction between the different sources of exceptions is the reason that exceptions are not simply labelled as 'outbreaks', or, other suitable equivalent (McCabe 2004): exceptions must be investigated to distinguish between those corresponding to genuine biological events or changes, and those caused by statistical variation. However, the strength of these systems lies in their ability to help focus the attention of a user on unusually high values, particularly crucial when monitoring many data-streams. Depending on the implementation, these systems can also give an indication of how large the exception is compared to usual values.

Typically, an exception reporting system can be sub-divided into several dis-

tinct elements. One such division is shown in Figure 1.1. Typically, data about the quantity being monitored will be passed to the exception surveillance system at regular intervals (usually, daily or weekly). A prediction is then made for the quantity. The predicted level is then compared to the observed one using an alarm method to determine if an exception should be reported. A report is then made to the users of the system, detailing the results of the alarm method.

1.1.1 Data Sources

Many data sources can be monitored by these systems. The choice of data obviously determines the nature of what can be reported. Some data sources are collected solely for the purpose of surveillance, while in others it can be a secondary consideration. The collection of general practitioner consultation rates (for particular medical conditions) gives an example of the former. In contrast, the primary purpose of the NHS24 telephone line (introduced in Chapter 2) is to provide members of the public with triage level medical advice; however, its log of telephone calls can be used to monitor levels of illness in the population. Often, data from those systems where surveillance is a tertiary function can give early warning of exceptions and outbreaks, but with less specificity about the underlying medical cause. For example, the number of calls relating to fever in NHS24 data can be monitored; an exception would potentially indicate an outbreak of some infectious disease but would not tell us *which* disease was the cause. Thus, it is useful to be monitoring a range of data sources, from those that can give non-specific early warning, to others that can give specific, but more delayed, information.

For our research we were provided with particular data-sets and did not have to make decisions on how data should be collected. For a longer discussion on data sources, see Babin et al. (2007); principles of how data should be stored are discussed in Wojcik et al. (2007). For examples of different data sources, see: Magruder (2003), sales from pharmacies; Effler, Ching-Lee, Bogard, Ieong, Nekomoto, and Jernigan (1999), laboratory test data; Cooper, Smith, Baker, Chinemana, Verlander, Gerard, Hollyoak, and Griffiths (2004), tele-health call logs.

1.1.2 Prediction Methods

There are many ways of forming predictions for the number of cases expected. Our research focuses on fitting suitable regression models to past data. These models are then used to predict expected future levels of the quantity being monitored. Regression models allow for the inclusion of elements that can address such things as seasonality and linear trend, which help improve predictions. We generally use Generalised Linear Models (GLMs) as they allow us to fit count distributions (for example, Poisson and negative binomial distributions) to data. For a discussion on the use of regression models for prediction, see Burkom (2007).

Auto-regressive modelling is another approach that can be used for creating predictions. Developed in Box and Jenkins (1976), it is well suited to modelling data with changing trends. However, this modelling is generally considered to require high levels of expertise to use and so is rarely used in this area (Burkom (2007); Chatfield (1978)). The modelling also assumes that the modelled data is normally distributed; when the counts are very small, this is not a reasonable approximation – we consider this point further in Chapter 2.

A simple, but robust, approach is given by exponential smoothing. One extended form is know as Holt-Winters smoothing, which incorporates elements to model seasonality and local trend (Holt 2004; Winters 1960). Burkom (2007) commends this relatively straight forward method for the quality of its predictions, which are aided by it being able to adapt to changes in data. We consider ways of combining Holt-Winters predictions with GLMs to improve models (see Chapters 3, 4, 5). A review of exponential smoothing is given in Appendix B.

1.1.3 Alarm Methods

Once we have an observed and expected level, it is necessary to compare the two values to determine if an exception should be reported. In our research, we have used a version of exceedance alarms as developed in McCabe (2004) and Farrington, Andrews, Beale, and Catchpole (1996). Using GLMs for prediction means that we actually predict a distribution of future values. By comparing the actual count observed that day with this distribution, we can compute an associated 'p-value', i.e. the probability of obtaining a value as large as that observed under the predicted distribution. If this p-value is larger than, say, 5%,

we say an exceedance has occurred and report an exception. Exceedance alarm methods are best suited to detecting 'sharp' or 'spiky' changes in data (Burkom 2007). They also have the advantage that they can be presented easily to a non-statistical audience. For a longer consideration of exceedance alarm methods, see Burkom (2007), McCabe (2004).

Other alarm methods include control charts. These were initially used by industry to monitor manufacturing. Typically, they would monitor the mean of some manufacturing process and check for violations of certain rules (Ryan 1989); should the rules be violated, the process would be considered out of control and marked for investigation. Usually, the means are displayed on a chart and the rules are shown graphically to allow for easy interpretation. Control charts can also be applied to analyse the difference between observed and predicted expected values (forecast residuals: $observed - expected$) (Mandel 1969). For example, Page (1954) gives a CUSUM (cumulative sum) control chart applied to forecast residuals: an increasing cumulative sum would suggest that the underlying quantity being monitored is increasing. Control charts are better suited to detecting small and distributed changes (Brillman et al. (2005); McCabe (2004)). There are also a range of other control charts; for details of these and a longer consideration of control charts see McCabe (2004), Ryan (1989).

1.1.4 Reporting

Having compared the prediction and the observed value, we then need to present the user of the exception reporting system with these results. There are many ways in which this can be done, including tables, charts and maps. It is important that the chosen method is accessible to the user and is reasonably intuitive. Tables of p-values may be reasonably clear to statistically skilled users, but they will not be clear to the general user. Charts showing the expected value, the observed value and exception limits will be more widely accessible. It is important that the presentation method is chosen wisely, as otherwise it will not matter how good a system is, since users of the system will simply disregard it because of impenetrable output. For a longer consideration of these issues, see Hauenstein et al. (2007), McCabe (2004).

1.2 Thesis Outline

In this thesis, we deal with three exception reporting systems for use in Scotland. In Chapter 2, we introduce NHS24, a twenty-four hour medical advice telephone line. We aim to develop regional exception reporting systems for monitoring levels of particular medical syndromes using the call log data. We begin by exploring the call log and fitting ARIMA models to the daily national totals of calls relating to vomiting. From these, we suggest a potential system for monitoring the national levels of calls relating to vomiting. However, at the regional level, the counts of calls are not large enough to approximate as being normally distributed, violating an underlying assumption of ARIMA modelling. Thus, we switch to using Generalised Linear Models (GLMs) in Chapter 3. We begin by fitting a GLM to the national counts and then fit GLMs to the regional health board counts. We propose a way of dealing with serial correlation in the counts, through the use of a variable based on exponential smoothing. We then investigate the use of this variable as a way of linking the separate health board models together, in the hope that this will improve the models. Since we only have two years of data from the start of NHS24, it is not possible to easily test the developed systems.

Next, we consider developing a regional exception reporting system for monitoring lab confirmed samples of micro-organisms, which are collated by Health Protection Scotland (HPS). HPS suggested six organisms that we could investigate doing this with; we chose *Cryptosporidium*, a leading cause of water borne infection. As with the NHS24 data, we begin by fitting a model at the national level (Chapter 4) and then at the regional level (Chapter 5). We again find there to be serial correlation in the counts and so try different approaches for dealing with this. Once we have the regional models, we again look for links between the health boards, but in a more careful fashion than was done with the NHS24 data. We find that the relationships between neighbouring boards are generally complex and cannot be modelled easily using separate models for each health board. In Chapter 6, we develop uniform regional models and combine these with an alarm method to produce regional exception reporting systems for monitoring *Cryptosporidium*. We then compare the results of using these systems with the current national system applied to the regional counts of *Cryptosporidium* cases

from 2006. We find that our systems do as well as, or better than, the national system applied to the regional counts.

The last exception reporting system we develop is a national all-cause mortality surveillance system, developed to monitor the mortality of swine flu. In Chapter 7, we fit models to the daily counts of deaths occurring in Scotland. The seasonality present in the counts of deaths cannot be captured easily using trigonometric harmonics, so we use a Generalized Additive Model (GAM) to more effectively model the seasonality. There is a significant delay associated with the reporting of deaths, so an approach for dealing with this is presented in Chapter 8. In the same Chapter, we draw all the necessary elements together for a complete mortality exception reporting system.

Finally, we summarise the work covered in this thesis in Chapter 9 and suggest directions for future work.

Chapter 2

NHS24 Preliminary Modelling

2.1 Introduction

NHS24 is a confidential telephone service that provides 24-hour free health advice every day of the year to the people of Scotland (NHS24 2008). Anyone can call up with a medical complaint seeking advice. The urgency of the ailment is assessed, and then the call is triaged appropriately, from the caller being directed to self care to the emergency services being contacted directly. Calls that are of a non-critical nature are mostly referred to another health service such as a general practitioner. Often, the staff at NHS24 will give details of these referral services, sometimes calling ahead to make appointments for the patient.

Aside from dealing with particular ailments, NHS24 also serves as a source of information for health related issues (Wilson, Smith, Meyer, Robertson, Baxter, Cooper, and McMenamin 2007). Health information advisers at NHS24 can provide information about local services and provide information for various health queries. In some situations the adviser will even research information for the caller. This is particularly useful for those that cannot get to a general practitioner's surgery within normal office hours.

NHS24 is the Scottish equivalent of the longer running NHS Direct service, which serves the same role for England and Wales. NHS Direct began taking calls in 1998, while NHS 24 began receiving calls on January 1st 2004 (NHS Direct 2009). NHS24 is more heavily used than NHS Direct (43 compared to 21 per 100,000 population for 2004 (Health Protection Agency: Primary Care

Surveillance Team 2005)) and tends to receive more calls from the elderly and triage fewer calls to more serious out-comes (for example, calling the emergency services) (Health Protection Agency: Primary Care Surveillance Team 2005)).

Each call received by NHS24 is logged, with important details about the call being recorded. This log of calls forms a very large data-set that may allow the development of a regional exception reporting system to detect unusually high reporting levels of particular syndromes. We discuss the choice of monitoring syndromes in Section 2.2, but note the substantial development of a syndrome surveillance system for England and Wales by NHS Direct and the Health Protection Agency (Cooper, Smith, Baker, Chinemana, Verlander, Gerard, Hollyoak, and Griffiths 2004; Doroshenko, Cooper, Smith, Gerard, Chinemana, Verlander, and Nicoll 2005; Cooper 2007). The somewhat generic nature of the categories in a syndromic surveillance system mean that they are better suited to capturing widely dispersed 'health events', rather than localised outbreaks of disease (Henning 2004; Doroshenko, Cooper, Smith, Gerard, Chinemana, Verlander, and Nicoll 2005; Cooper 2007). However, the users of the NHS Direct syndromic surveillance system have suggested that their system might be able to detect local outbreaks if they had a higher rate of callers and some sort of spatial relationships were included in their models (Smith, Cooper, Loveridge, Chinemana, Gerard, and Verlander 2006; Cooper, Verlander, Smith, Charlett, Gerard, Willocks, and O'Brien 2006). With the NHS24 data, we have twice the utilisation of the phone-line by the population, so the first of these issues may be addressed. To consider the spatial relationships we can split up the NHS24 call totals between regions (health boards) in Scotland and see if this added information can improve surveillance models. To begin with, we consider some models that could be used for prediction at the national level in this chapter and then proceed to developing a potential system at the regional level.

At the time of this analysis only two years of data were available from when NHS24 went 'online'. Thus, the ability to test models and the strength of the models developed here will be affected by this. However, some interesting directions were still found.

2.2 Syndromic Surveillance

The threat of terrorist attack became a greater reality to most people after the attacks in America during 2001 (The Washington Post 2006). However, even before this time, the Centers for Disease Control and Prevention (CDC) in America had set about plans to develop systems for early detection of biological or chemical terrorist attacks (Henning 2004). While there is no precise agreement on a definition of syndromic surveillance, most sources cite detecting biological attacks as a motivating (if not the main) reason for developing syndromic surveillance (Robertson 2006; Henning 2004).

In syndromic surveillance 'non-specific symptoms or "health events"' are typically monitored to detect unusual patterns that might be indicative of a terrorist attack or other threat that users wish to be warned of (Robertson, Kavanagh, McKenzie, and Wagner 2008). The data monitored can be very variable, ranging from over-the-counter prescription sales to school absenteeism (Henning 2004). Typically, the data used for surveillance is not collected with that purpose in mind (i.e. school absenteeism is monitored to indicate problems with schools or domestic situations, rather than as a metric for health surveillance) (Babin, Magruder, Hakre, Coberly, and Lombardo 2007). This is also true of NHS24: its primary goal is to help address the medical needs of individuals calling it, rather than to collect data for surveillance purposes. While such data comes with the disadvantage of being non-specific, it has the significant advantage of being very timely; lag times between data-recording and results of analysis are increasingly being measured in terms of hours (Robertson, Kavanagh, McKenzie, and Wagner 2008). This short lag time is a crucial quality in any system whose purpose is minimise the damage done by a terrorist attack, but is of course an important quality in any surveillance system.

NHS24 and NHS Direct are among the only tele-health services in the world that are national (Cooper 2007). This allows them to carry out much broader syndromic surveillance than most other systems. This allows them to effectively go beyond merely monitoring for terrorist attacks (Cooper 2007). Their surveillance data and methods have also been shown to be useful for reassuring the public that some perceived threat is in fact safe (Smith, Cooper, Loveridge, Chinemana, Gerard, and Verlander 2006). Further, the daily data will become an increasingly

rich resource as time goes on, allowing for all manner of different health planning (Smith, Cooper, Loveridge, Chinemana, Gerard, and Verlander 2006). Syndromic surveillance began in earnest in Scotland in 2005, as a way of monitoring for terrorist attacks on the G8 summit (Robertson, Kavanagh, McKenzie, and Wagner 2008; Robertson 2006). Since then, syndromic surveillance of NHS24 has become routine at HPS.

2.3 Operation and the Produced Data-set

Calls to NHS24 are dealt with by nurses. Assuming the reason for a call is an ailment, as opposed to a request for information, a nurse answering a call can choose to use the NHS Clinical Assessment System (CAS) to triage the call. This system contains over two hundred clinical 'algorithms that form tree-like structures of questions relating to the symptoms of the person about whom the call is made' (Cooper, Smith, Baker, Chinemana, Verlander, Gerard, Hollyoak, and Griffiths 2004). The algorithm will specify the advice that a nurse should give to a patient, which can be briefly summarised into one of the following call outcomes: advice for self care; routine doctor referral; an urgent doctor referral; emergency department referral; or a paramedic dispatch (the nurse will request an ambulance to attend the caller). However, it is important to note that the nurses 'triage rather than diagnose illness in callers' (Cooper, Smith, Baker, Chinemana, Verlander, Gerard, Hollyoak, and Griffiths 2004): calls are "classified ('triaged') on the basis of described symptoms to determine priority of need and appropriate place of treatment" (Cooper and Chinemana 2004). Diagnosis of specific illness would require the use of doctors and possibly lab confirmation. The possibility of obtaining biological samples from callers has been explored (Cooper, Smith, Chinemana, Joseph, Loveridge, Sebastionpillai, Gerard, and Zambon 2008). Calls that are diagnosed using the CAS system will be referred to as 'algorithmic calls'. The nurse can also choose to dispense with the use of an algorithm and deal with the call from their own personal knowledge and experience. Some research has been done into how different experiences of nurses can affect the types of advice they give and whether they use the CAS system (O'Cathain, Nicholl, Simpson, Walters, McDonnell, and Munro 2004). In either treatment process, a range of information about each call is collected, including the fields detailed in Figure 2.1.

- Date and time of call
- Name of caller
- Age of caller
- Postcode location of caller
- The diagnosis of any clinical algorithm used
- Call reason - free text field detailing the presenting complaint

Figure 2.1: A selection of data collected by call-handlers from callers to NHS24.

- Cold & flu
- Coughs
- Diarrhoea
- Difficulty breathing
- Double vision
- Fever
- Eye problems
- Lumps
- Rash
- Vomiting

Figure 2.2: The syndromes that the clinical algorithm can diagnose calls into, which we will be considering.

All the call details taken together form a call-log that is monitored by HPS.

From the call-log, those calls diagnosed by a clinical algorithm were extracted to form a data-set to be analysed. For reasons of data protection, that is, to preserve the anonymity of callers, the data-set analysed did not contain the name of the caller or the complete postcode. The clinical algorithms are grouped into ten general syndromes given in Figure 2.2. These particular syndromes were chosen to be monitored as they could be indicative of a rise in common infections or diseases, or provide early indicators of a biological or chemical terrorist attack (Cooper 2007; Baker et al. 2003). These syndromes represent the prodromal stages of disease, typically before the disease can be tested for; potentially, this gives authorities earlier warning of health events than they would gain from other forms of surveillance (Robertson, Kavanagh, McKenzie, and Wagner 2008; Cooper, Verlander, Smith, Charlett, Gerard, Willocks, and O'Brien 2006). During 2005 the algorithms were changed, which may affect the data linked with each syndrome (Advances in Disease Surveillance 2007).

Another way to categorise the call-log into various syndromes would have been

to group calls using the free-text 'call reason' field (see Figure 2.1). However, there are no checks on data entered into this field and so the quality and range of data contained therein is very variable. For instance, 'diarrhoea' can easily be mis-spelt. In such a case, if we were interested in the counts of calls pertaining to this syndrome, some might be easily missed. This would subsequently affect any models we might form to model the counts of calls relating to diarrhoea. Further, if a call reason mentions both fever and vomiting, it is not clear which syndrome group the call should be counted in. Thus, the algorithmic calls were preferred, as they are likely to be more consistent. Later work has looked at using the 'call reason' field (Williamson 2006) and a syndromic exceedance system is now in place at HPS using the call reason field (Kavanagh, Robertson, and McMenamim 2007).

When the data were analysed, there were two years of data, starting from 2004 when NHS24 went 'live'. Using data from the start of the system contributes to a number of problems in analysing the data. Certainly for 2004 the system had not reached equilibrium; indeed, there is an increasing trend in the number of calls taken that year, as shown in Figure 2.3. Another problem is caused by the different health board regions not going completely 'live' from the start of 2004. For some areas, such as the Borders health board, a number of months passed before a sizable number of calls were received from this region (see Figure 2.4). Scotland is presently divided up into fourteen such regions: Highland and Argyll; Grampian; Tayside; Fife; Lothian; Borders; Forth Valley; Greater Glasgow and Clyde; Lanarkshire; Ayrshire and Arran; Dumfries and Galloway; Orkney; Shetland; and the Western Isles. A further complication is given by the proportion of calls diagnosed by algorithm generally decreasing from the inception of NHS24, as observed in Wilson (2006). Figure 2.5 shows a clear downward trend for the proportion of calls diagnosed algorithmically for 2005. This downward trend seems to have continued, motivating Kavanagh et al. (2007) to use the call reason field. These factors present significant difficulties when trying to fit models to the data.

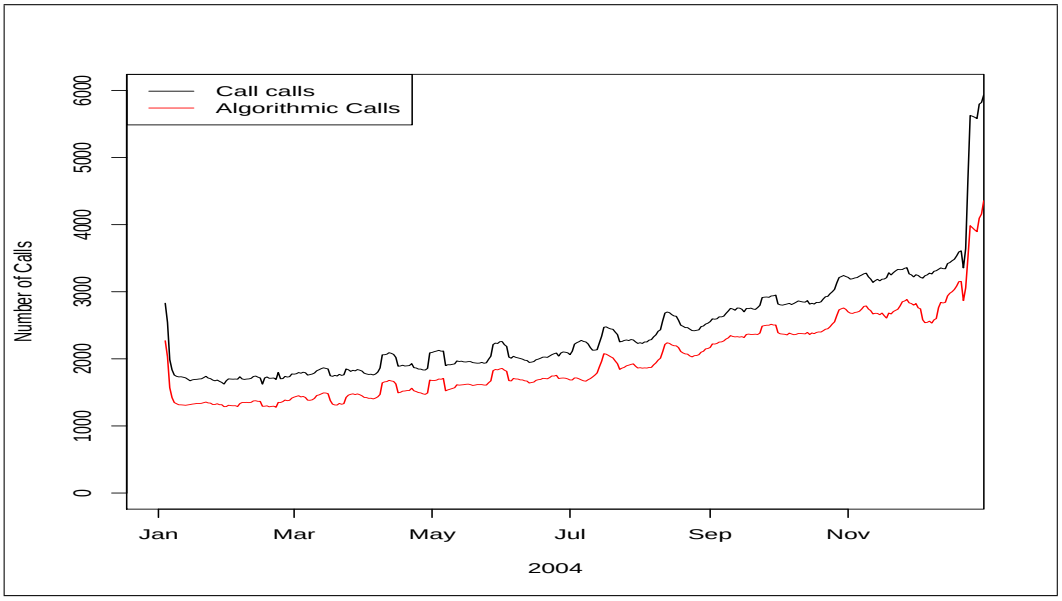


Figure 2.3: The counts of calls to NHS24 for 2004, after a moving average of order seven has been used on the data to remove weekly (day-of-the-week effects) seasonality. There appears to be quite a strong increasing trend.

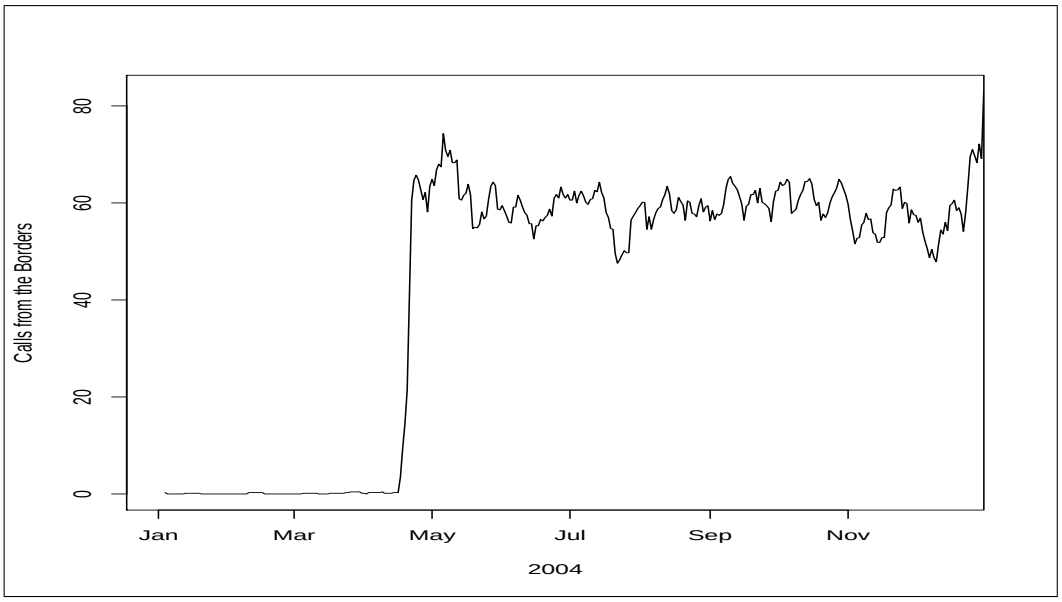


Figure 2.4: The counts of calls from the Borders health board district for 2004. A moving average of order seven has been run through the data to remove weekly seasonality. Barely any calls were received from the Borders health board until halfway through April.

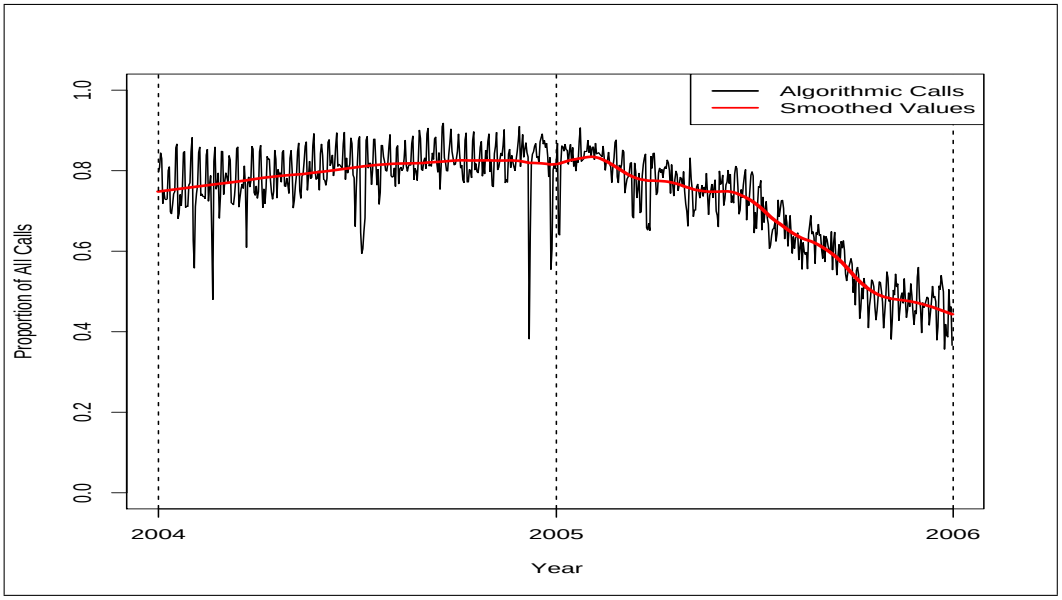


Figure 2.5: The changes in proportions of all recorded calls that are diagnosed by clinical algorithms over 2004 and 2005. The red line is given by Friedman's Supersmoother being applied to the proportions. We find that the proportion of total calls diagnosed by clinical algorithm decreases over time.

2.4 Data Preparation and Database-usage

Even if we restrict our attention solely to the algorithmic calls, there are still a very large number of records – over one and a half million calls for 2004 and 2005. Given the amount of data to be considered, a database presents the only feasible option for managing the data. Further, by using a database, future data can be added easily and calls can be categorised into different diseases or geographical areas with ease. Wojcik, Hauenstein, Sniegoski, and Holtry (2007) give a brief introduction to databases and extol their virtues for such data management.

However, before the data – see Figure 2.6 for a sample of it – could be incorporated into a database, it was necessary to 'clean' the data of certain errors and strange characters that were present in the original text file format. The biggest inconsistencies in the data occurred in the postcode field. A number of postcodes had the form 'XX0X' which needed to be corrected to 'XXX', or had one too many digits, corrected by the last digit being removed. These corrections reduced the number of calls with inaccurate postcodes from 15% to around 1%.

```
2004-03-21,08:02:01,AB41, 12,Attend PCEC within 12 Hrs,Female,
          Cough Child (Age 5-16 years), 2
2004-03-21,08:14:45,G81, 0,Attend PCEC within 1 Hr,Female,Cough
          Infant (Age 0-1 year), 2
2004-03-21,08:12:15,KY6, 0,Attend PCEC within 4 Hrs,Male,Unwell
          or Crying Infant (Age 0-1 year),11
2004-03-21,08:13:39,G31, 7,Home Visit within 2 Hrs,Female,
          Breathing Difficulty Child (Age 5-16 years), 4
```

Figure 2.6: A sample of the text-file containing the algorithmic calls to NHS24 during 2004 and 2005. This is after it has been cleaned and categorised into the syndromes we are interested in (Figure 2.2). The fields are in the following order: Date, Time, Post code, Age, Outcome, Sex, Protocol, Syndrome.

Such changes were carried out over the original text files, using `Perl` scripts, as shown in Appendix A.

The simple database we use consists of three tables, as shown in Figure 2.7. The `calls` table is the primary one, containing the data from NHS24, with the additional fields of 'Syndrome' and 'Valid_PC'. The former field is an integer value that corresponds to the syndromes given in Figure 2.2; this correspondence is encoded in the `syndromes` table and is primarily for convenience. The 'Valid_PC' field is a boolean value indicating if the post code contained for any given row in the `calls` table is valid. The `hb_pc` table is a look-up table, allowing one to take each valid post code in `calls`, and find which health board the call originated from. This feature will be essential when we look at modelling counts of the syndromes at the regional level of health boards. However, since we only have the post code district due to data protection (see Section 2.3), mapping from the post code to health board region is only approximate.

The 14 health board divisions of Scotland can be seen in Figure 2.8. Initially at the start of 2004, the beginning of the data-set, there were 15 health boards in Scotland (left in Figure 2.8). However, during the period under study, the Argyll and Clyde health board became bankrupt (BBC News 2005). The region that it previously covered was divided up between the Highlands and Greater Glasgow health boards, creating the map in the right of Figure 2.8 (Scottish Parliament 2006). This change was easy to effect in the database, as only those rows in `hb_pc` that corresponded to Argyll and Clyde needed to be changed. The ease of using

`calls` table:

Date	Time	PC	Age	Outcome	Sex	Protocol	Synd.	Valid_PC

`hb_pc` table:

Health_board	Post Code

`syndromes` table:

Id	Description

Figure 2.7: The three tables, with their various fields, that constitute the database containing the records of algorithmic calls to NHS24 during 2004 and 2005.

a database was also evident when new data was received from NHS24; new data was easily added to the database and queries simply re-executed to get updated counts of the syndromes.

2.5 National Counts of Syndromes

The daily counts of NHS24 data have not received much analysis. Thus, before considering a regional exception reporting system, we begin by looking at modelling the national counts and consider options for a national exception system.

2.5.1 Data Used

Within the data-set, not all of the calls had a valid postcode for each caller's location. The status of the postcode is recorded in the log file when certain of the Perl scripts are executed (see Section 2.4). Thus there are a number of calls that cannot be attached to any particular geographical area. Obviously, this does not present a problem when looking at the calls at the national level, as we do not need to know where a call originates from. However, this presents a concern when we start considering the calls at a regional level. Fortunately, only $16,652$ of the $1,576,282$ calls, just over 1%, either had no postcode or an invalid entry. Therefore, when we look at the regional counts, those calls with no valid postcode are simply ignored. Generally, this approach is suitable, but might have ramifications for those syndromes with very low counts (such as double vision), or areas with very low call rates (for instance the Scottish islands). For consistency,

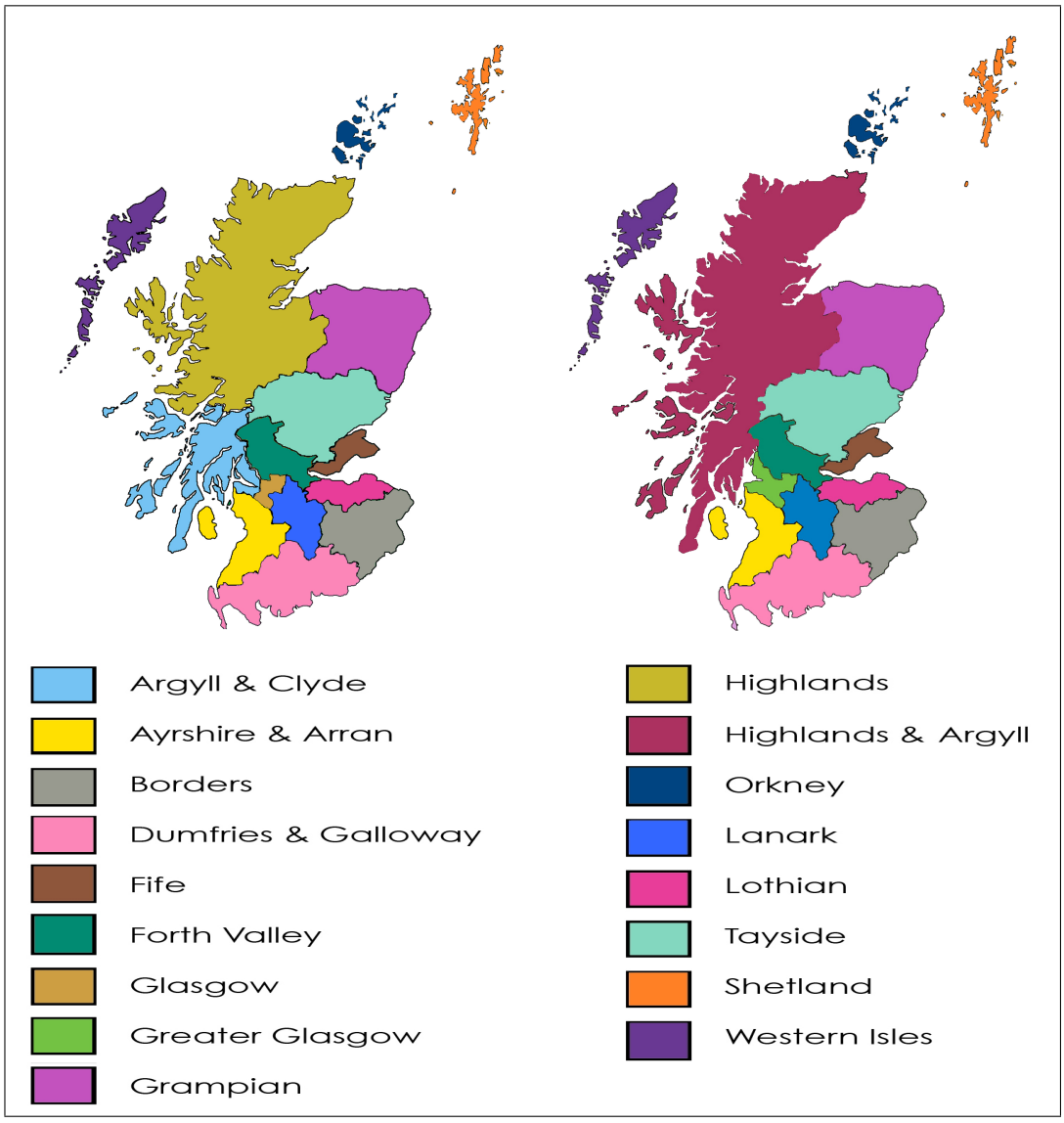


Figure 2.8: The health boards in Scotland. *Left*: The 15 health boards, before Argyll and Clyde closed. *Right*: The 14 remaining health boards after this closure.

in the national analysis of counts here, we restrict ourselves to those calls with a valid postcode.

2.5.2 Choice of Syndrome

As there are ten syndromes, it was felt best to focus on one of them in this exploratory work. We choose between them by looking at the counts of calls and the proportion of total calls relating to each syndrome over the two years for which we have data. As will be considered in greater detail in Section 2.6, working with proportions has a number of advantages.

Double vision is eliminated as a candidate since there are very few calls received that relate to it. In fact, for 633 days of the 731 under consideration, no calls were algorithmically diagnosed as relating to double vision. On the other days only one or two calls were received. This would mean that any model we fit would have to deal with large numbers of zeros, and is likely not to be suitable for application to other syndromes.

Other syndromes such as lumps, eye problems, diarrhoea and cold & flu, are also discounted because of their size. Their counts are plotted in Figure 2.9. A moving average of order seven has been applied to the counts, since there is a marked difference in the numbers of calls received at the weekend, and during the week, as demonstrated in Figure 2.10. The difference is primarily caused by doctors' surgeries being closed at the weekend, leading to more people calling NHS24 for advice at these times. Of these smaller counts, Diarrhoea has some of the largest ones and yet still only averages around sixty calls each day. If we proceeded to consider this at a regional level, each health board would get only four calls per day, as a rough estimate. This would present further difficulties, with an already problematic data-set. Thus, we choose from those syndromes with larger counts. The remaining syndromes are plotted in Figures 2.11 and 2.12.

Another criterion to help choose between the syndromes is consistency. Since we only have the first two years of data, it is sensible to pick the syndrome which looks like it has the most consistent annual seasonal pattern. When the remaining counts are considered, little annual pattern can be found within them. This leads us to consider each syndrome's proportion of total algorithmic calls, for each day

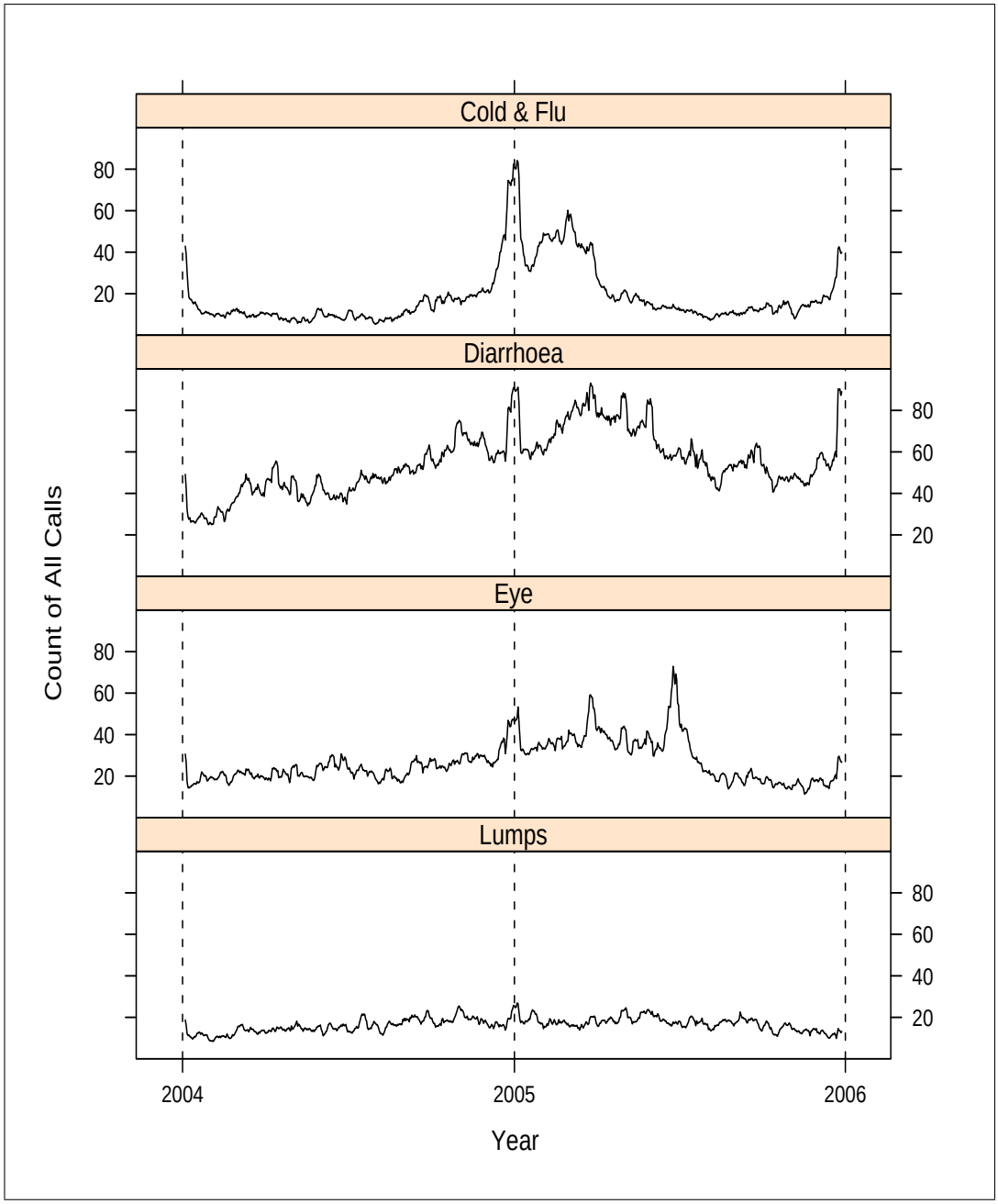


Figure 2.9: The counts of calls relating to lumps, eye problems, diarrhoea and cold & flu, for each day of 2004 and 2005. A moving average of order seven has been applied to the data to remove weekly seasonality.

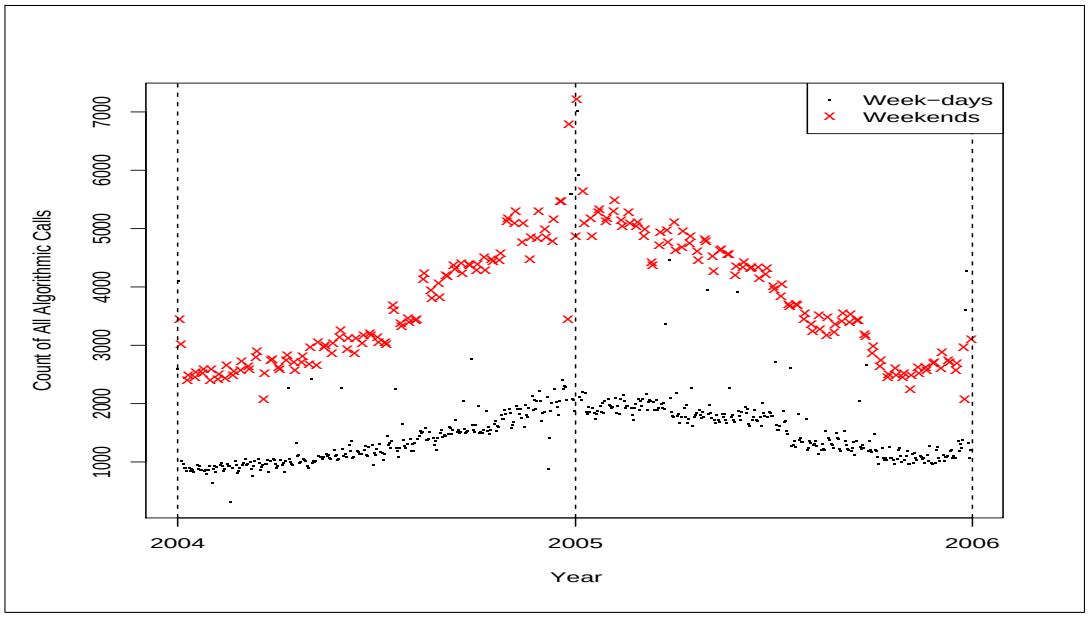


Figure 2.10: The total number of all calls received by NHS24 that are diagnosed by clinical algorithms.

over the two years. We hope this will remove any weekly seasonal patterns or trending effects present in the data, allowing any annual pattern to be seen more clearly. We return to dealing with proportions in Section 2.6, but briefly consider them further here to justify our choice in focusing on those calls that relate to vomiting.

A plot of each of the remaining syndrome's daily proportion of all algorithmic calls is shown in Figure 2.13. There appears to be some trend remaining in rash and difficulty breathing. In coughs, there is local peak around March in 2005 which does not occur in 2004. These factors would present difficulty when trying to find an annual pattern, particularly as we only have two years of data. We eliminate these options and focus on the remaining vomit and fever syndromes.

Finally, the choice was between fever and vomiting, as shown in Figure 2.14. Here, a moving average of order seven is applied to data, since it appears in Figure 2.13 that there is still weekly seasonality present in the proportions. It was hoped that by working with the proportions that the weekly seasonality would be removed – see Section 2.6. It seems that the vomit syndrome gives the most consistent annual pattern and so presents the best modelling opportunity. Thus,

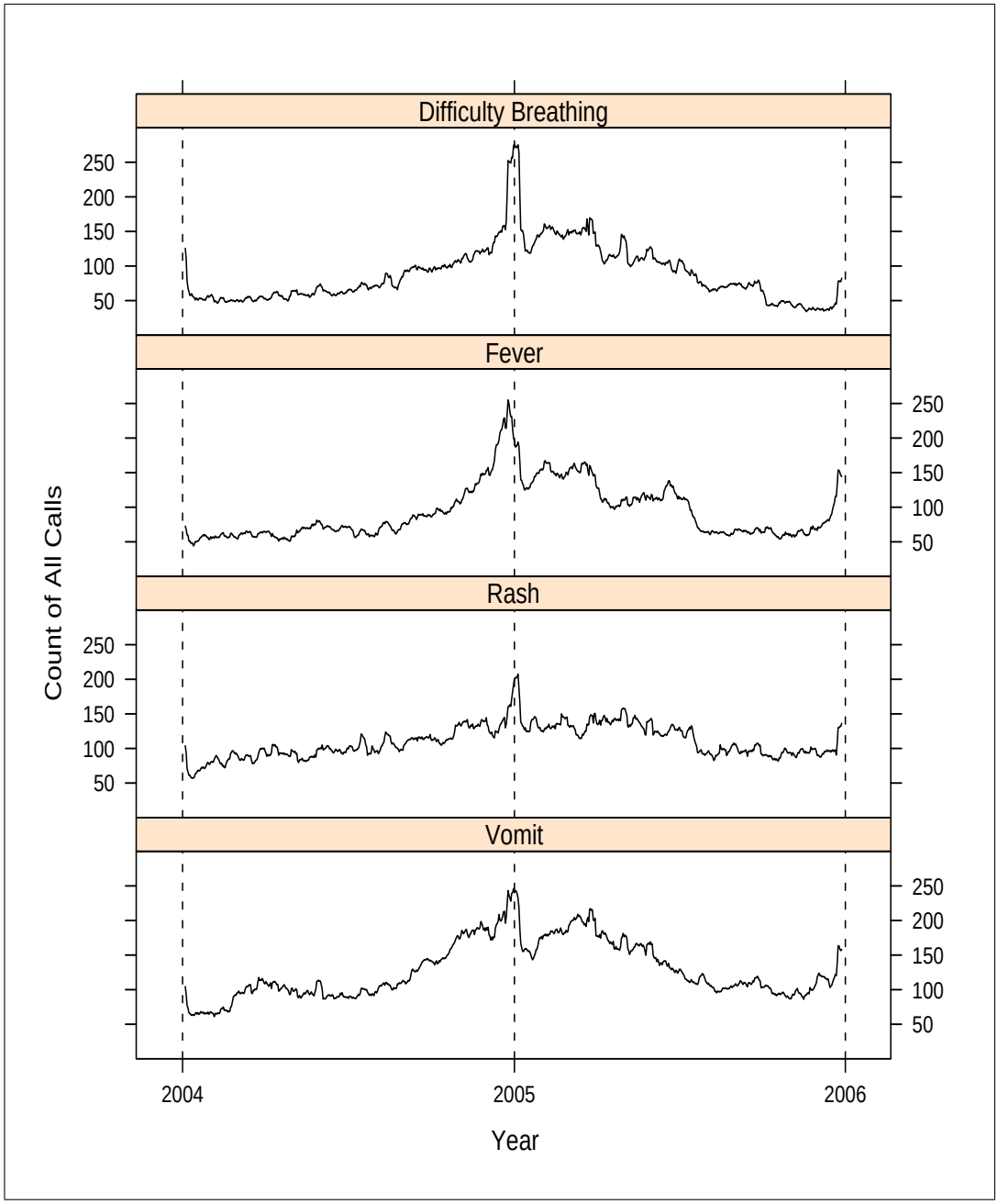


Figure 2.11: The counts of calls relating to rash, vomit, fever and difficulty breathing, for each day of 2004 and 2005. A moving average of order seven has been applied to the data to remove weekly seasonality.

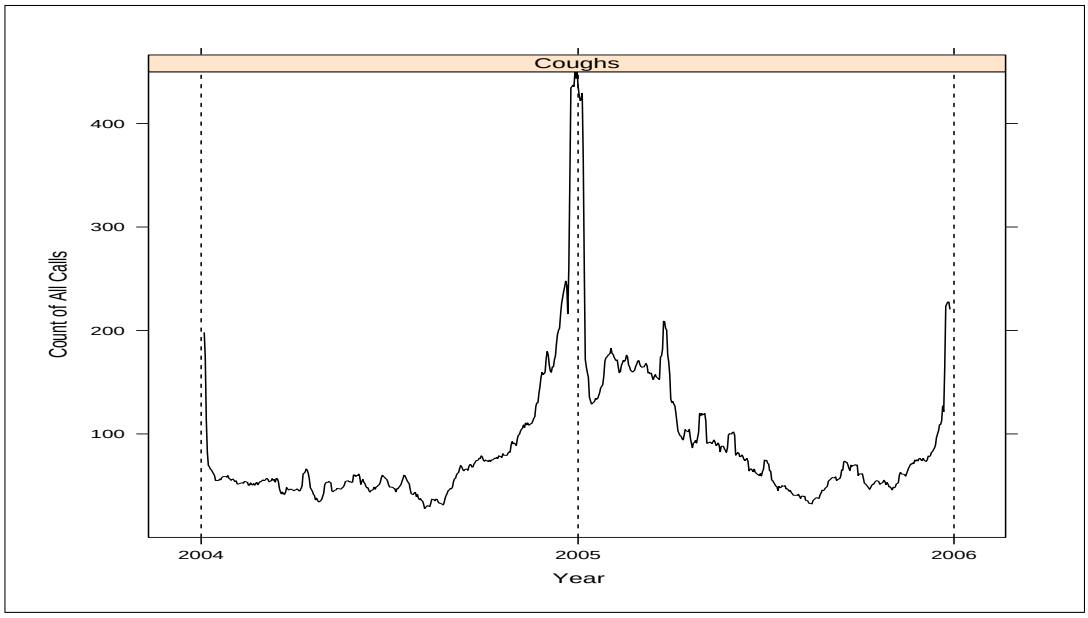


Figure 2.12: The counts of calls relating to coughs, for each day of 2004 and 2005. A moving average of order seven has been applied to the data to remove weekly seasonality.

we proceed by fitting some exploratory models to the counts of calls relating to people complaining of vomiting, before turning to the more consistent proportions associated with this syndrome.

2.5.3 Exploratory Modelling of Counts Relating to Vomiting

A plot of the counts of calls pertaining to 'vomit' over the two years can be found in figure 2.15. The most striking feature in the counts is the difference in level between week-days and weekends; the effect of this is most clearly seen on a log-scale, as in shown figure 2.16. It can be seen from this graph that the log-level at the weekend is nearly a constant distance above that of the log-level during the week. To find this constant difference, we fit the following linear model:

$$\ln(vc) = \beta_0 + daytype + \varepsilon,$$

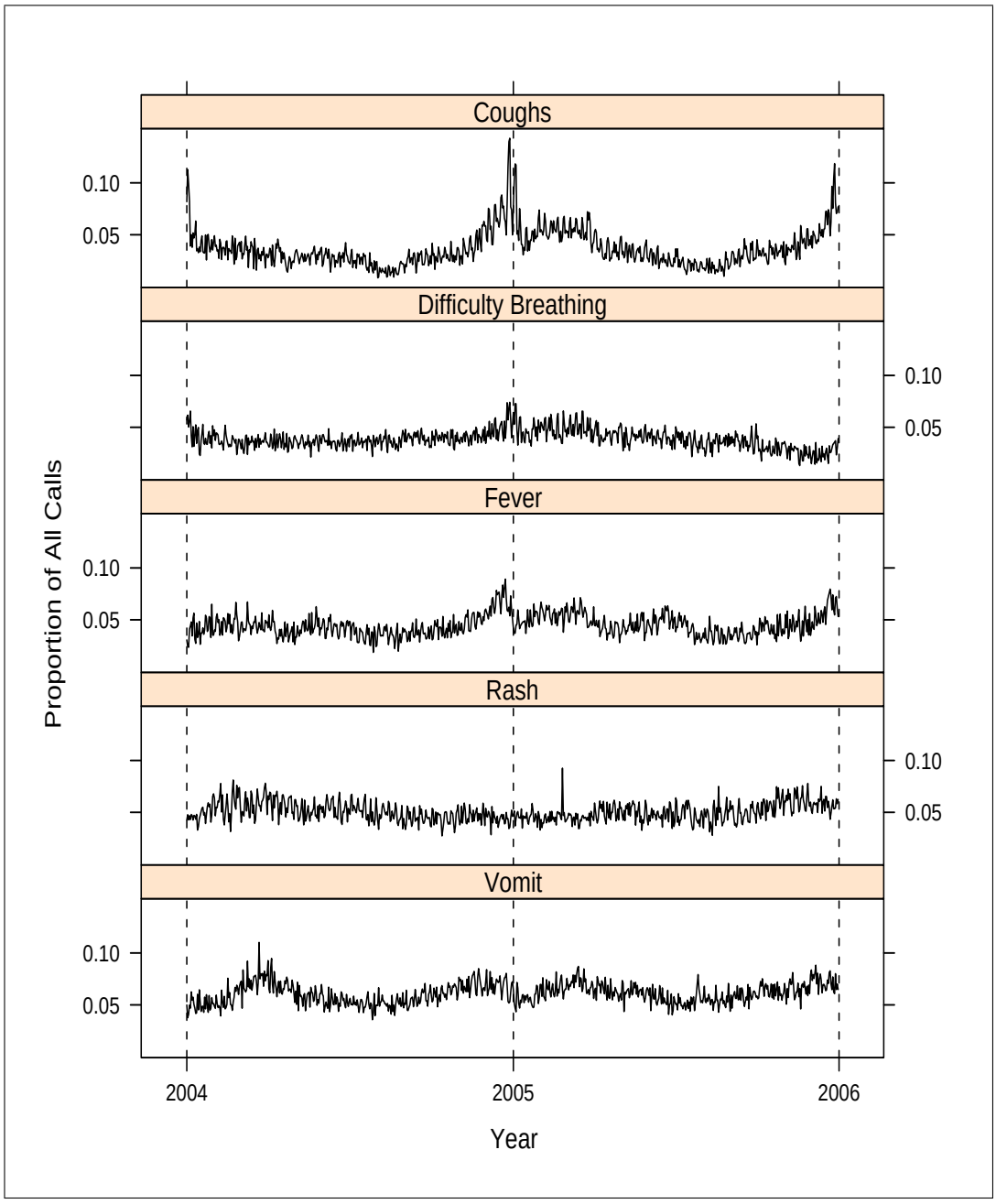


Figure 2.13: The proportions of algorithmic calls relating to vomit, rash, fever, difficulty breathing and coughs, for each day of 2004 and 2005.

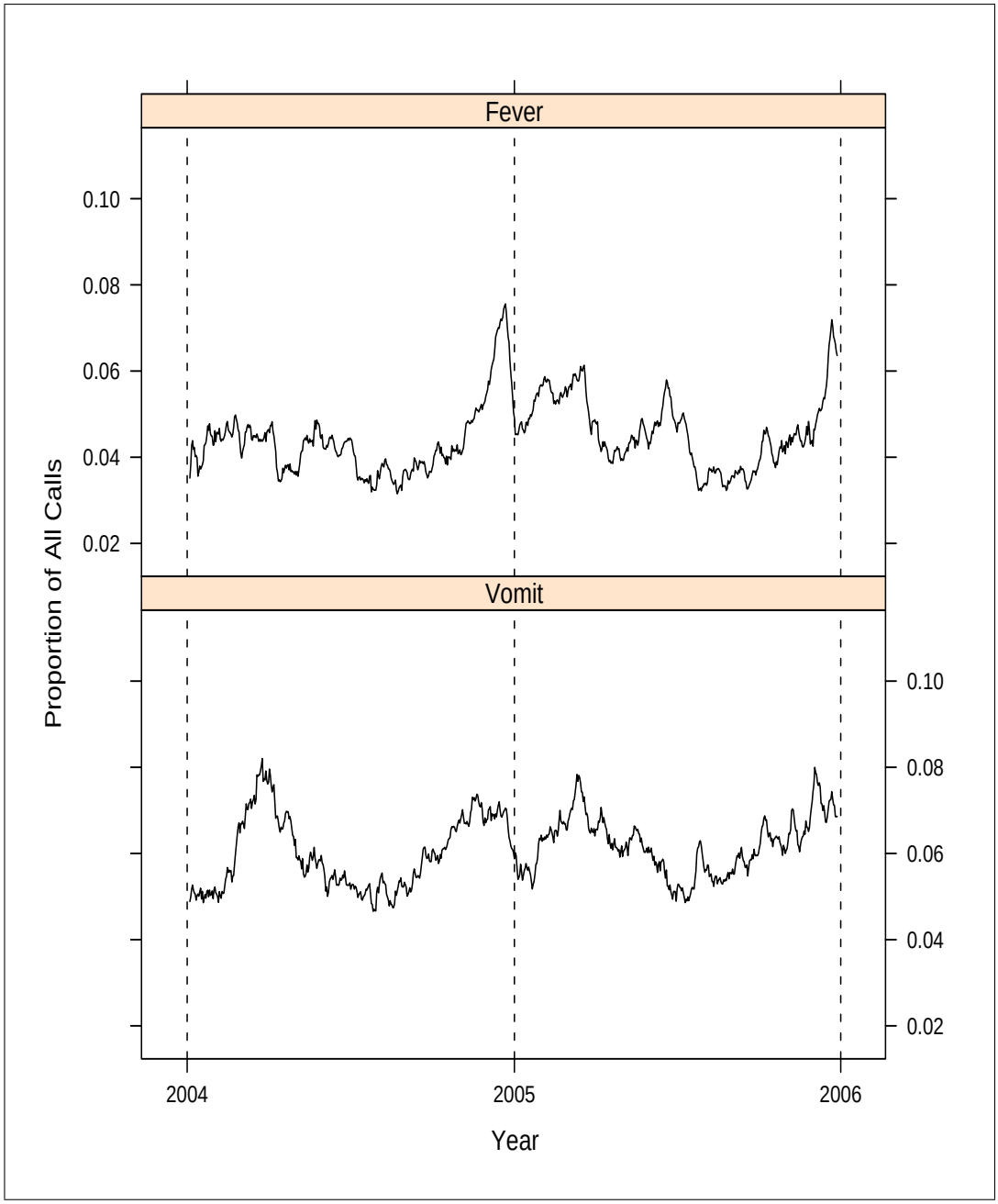


Figure 2.14: The proportion of algorithmic calls relating to vomit and fever, for each day of 2004 and 2005. A moving average of order seven has been applied to the proportions.

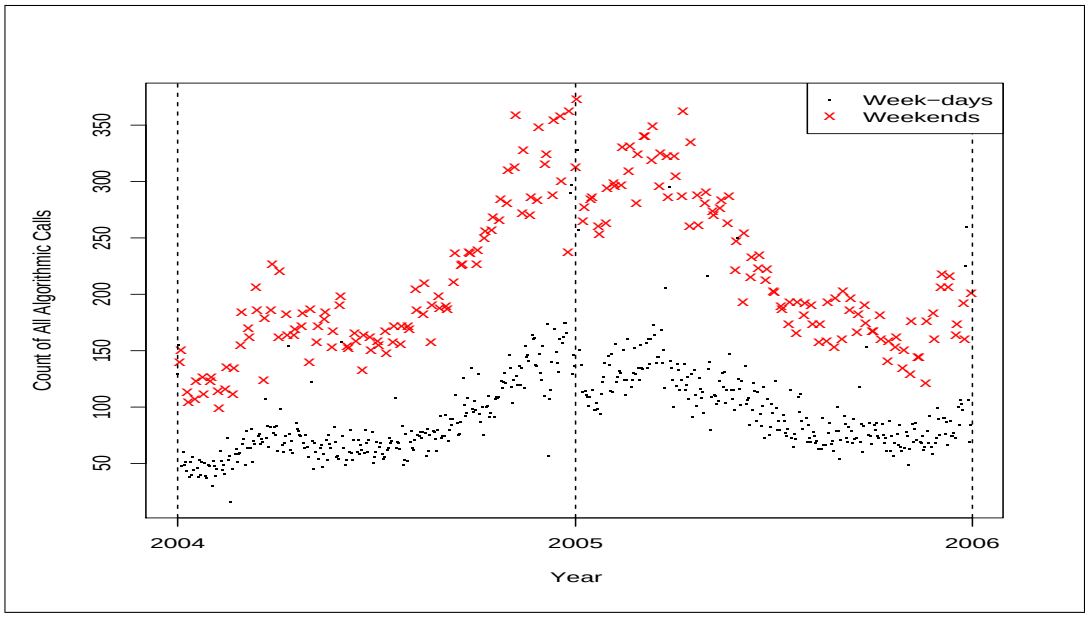


Figure 2.15: The counts of those calls that mention vomiting for 2004 and 2005.

where vc is the daily number of calls from people suffering with vomiting, $daytype$ is a factor with two levels, indicating a week-day or weekend, and ε are the residuals. Fitting this model, we find the following relation:

$$\ln(vc) = 4.468 + daytype + \varepsilon,$$

with

$$daytype = \begin{cases} 0 & \text{Week} - \text{days}, \\ 0.884 & \text{Weekends}. \end{cases}$$

It is known that at peak times of demands, such as at weekends when general practitioners' surgeries are closed, more staff are on duty at NHS24 allowing more calls to be taken. This observation about staffing levels becomes apparent on considering the other syndromes in a similar fashion. When this is done, a comparable difference between week-day and week-end levels is found. Further, the factor 2.3 (i.e. $\exp(0.884)$), appears constant over 2004-2005, when we might expect the number of calls to rise. This may indicate capacity rather than demand, if, say, there are about 2.3 times as many staff on duty during the weekend.

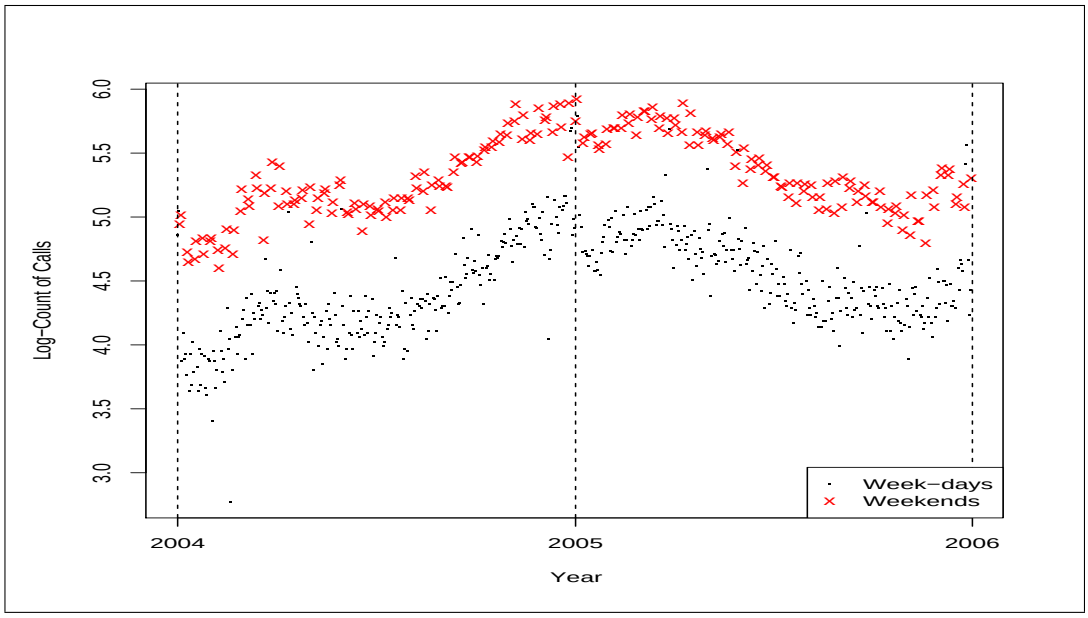


Figure 2.16: The log-counts of those calls that mentioning vomiting for 2004 and 2005.

For both years, there seems to be a local peak around April, perhaps corresponding to an annual outbreak of vomiting. There are also local peaks at the beginning and end of the two years. One of the problems presented by this dataset is that while there are a large number of observations, 731 in total, (one for each day of the two years), they are from a period of only two years. Thus, any conclusions about annual seasonal effects cannot be made with great confidence. For instance, we cannot know for sure if the local peaks during April each year is just an artifact of these two years, or if they are part of an annual seasonal pattern.

We find that the counts generally increase for 2004, and mostly decrease for 2005. Stabilisation and trend may also be factors for this data. It is not obvious how long it will take the system, after the start in January 2004, to reach an equilibrium state. Further, if NHS24 increases in popularity with the public, more people will call causing an increase in the long term trend. Also, as noted in section 2.3, the proportion of calls being algorithmically diagnosed is decreasing. For these reasons, we consider the proportion of algorithmic calls related to vomiting instead of the counts. Thus, the effects of a possible long

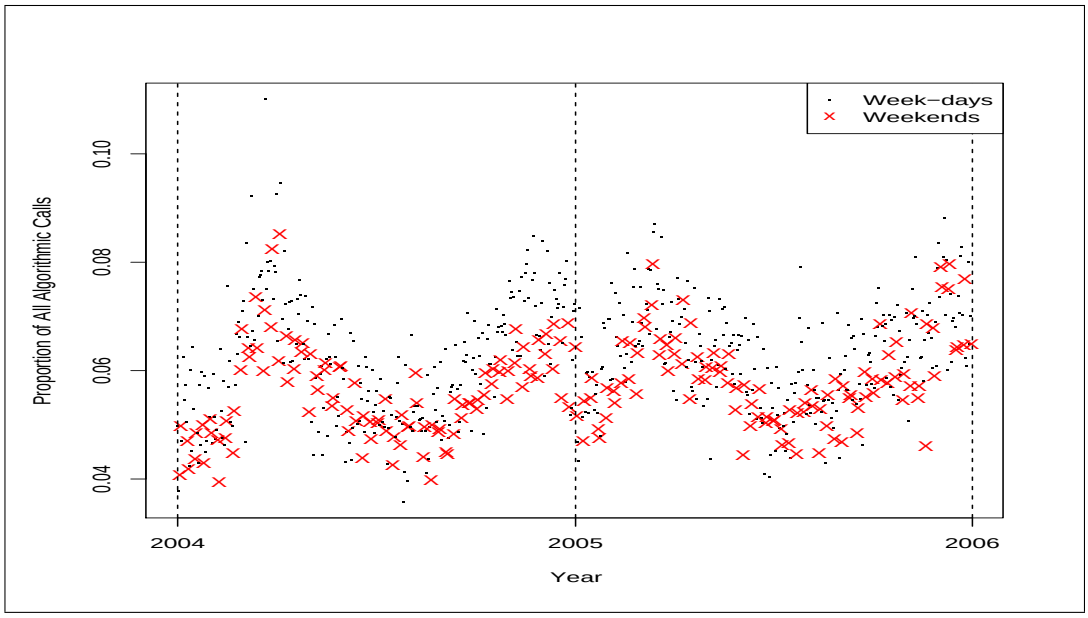


Figure 2.17: The proportion of algorithmic calls diagnosed each day as pertaining to people vomiting.

term (rising) trend as the public become familiar with NHS24 might be removed or greatly reduced. This may also be true of other effects related to modelling the counts directly.

2.6 Proportions

A plot of the proportion of algorithmic calls related to vomiting, hereafter referred to as VP, can be found in Figure 2.17. Surprisingly, perhaps, the day-of-the-week effect still seems to be present in the data, although the difference appears less distinct. However, the day-of-the-week effect is reversed: in general, VP is lower at the weekends than during the week. This reversal is caused by some of the other syndromes being proportionately more prevalent at the weekend, as shown in Figure 2.18 and Figure 2.19. Calls at the weekend will generally cover a wider array of conditions than during the week, since anyone that might want a GP has little choice but to call NHS24.

Working with proportions complicates modelling as the syndromes are now *coupled*. During an outbreak, the proportions will be negatively correlated: as

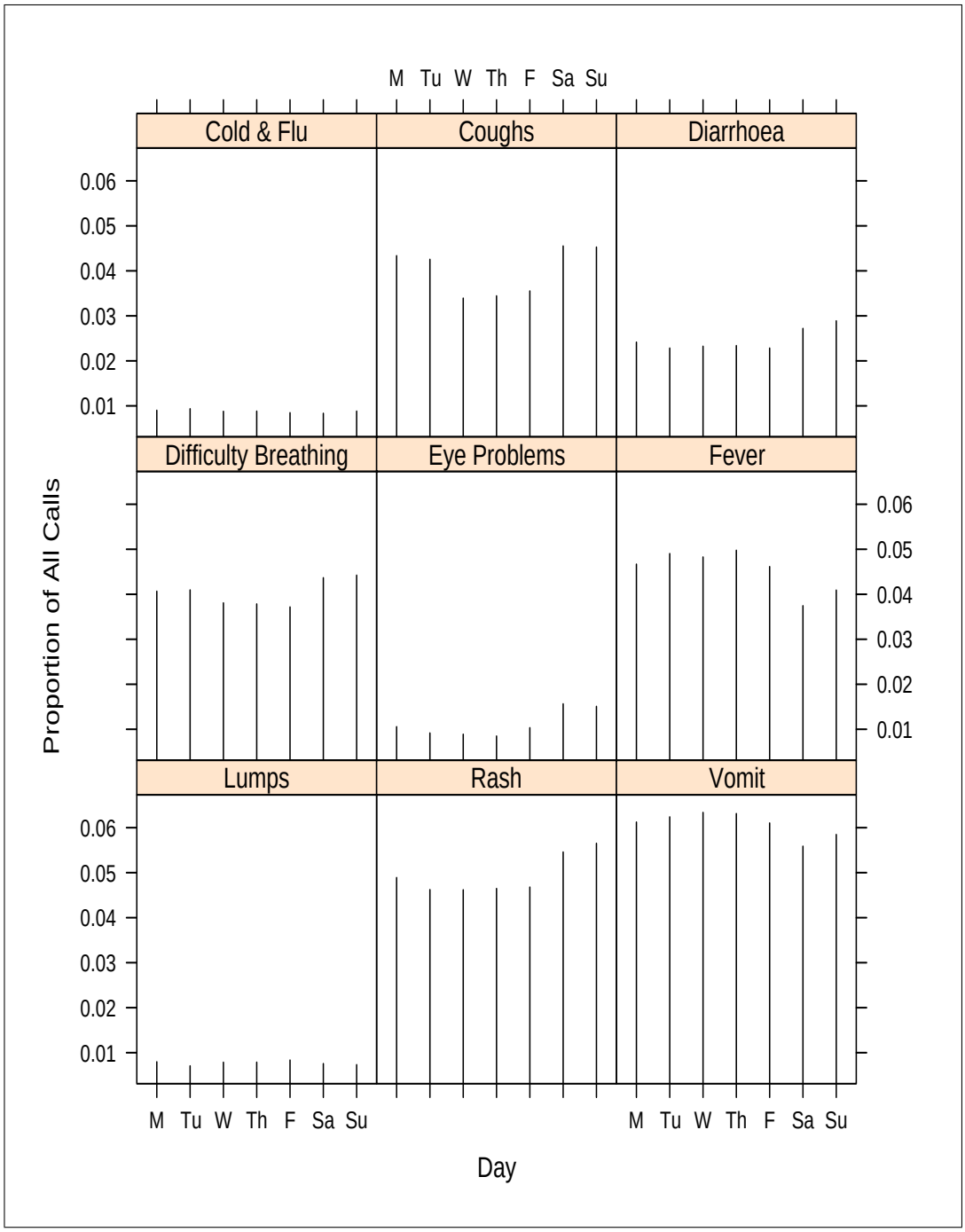


Figure 2.18: The graphs show each syndrome's proportions of all algorithm calls for 2004/2005, for each day of the day of the the week. For example, just under one percent of all calls received on Mondays during 2004/2005 were related to vomiting. Double vision has been omitted as the number of calls relating to this syndrome are very small; see Section 2.5.2. The weekly pattern for the 'other' syndrome is shown in Figure 2.19.

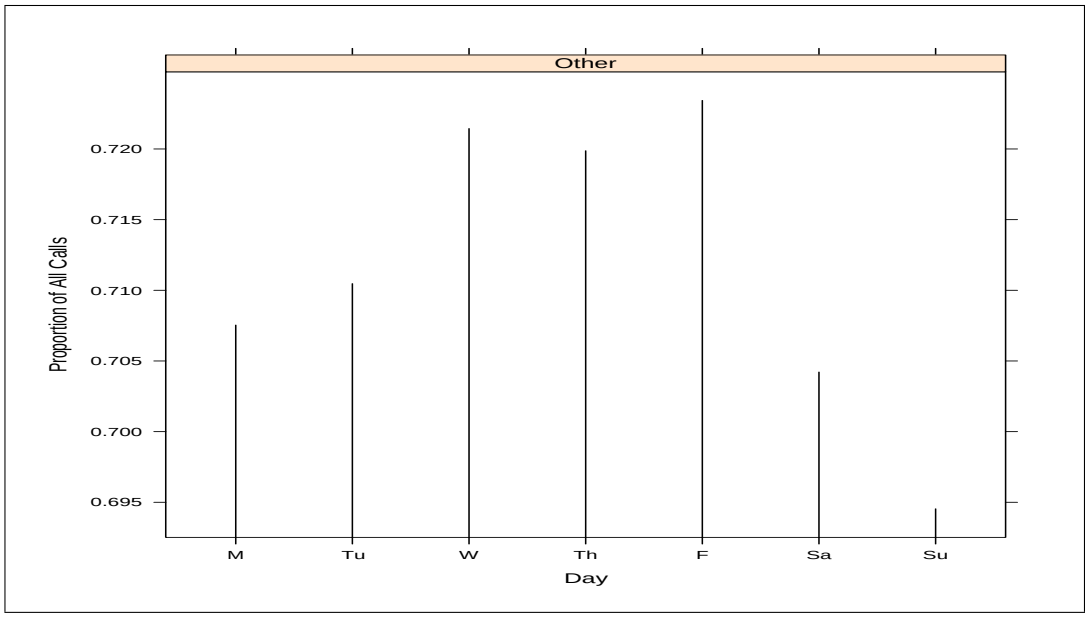


Figure 2.19: The proportion of algorithm calls diagnosed as 'other', of all algorithm calls for 2004/2005, for each day of the week.

one syndrome increases others will decrease. So, for instance, a fall in VP may indicate that another syndrome is occurring more often or maybe in outbreak. However, it was hoped that the stability gained by working with the proportions would make this added difficulty worthwhile. A correlation matrix between the different proportions of each syndrome can be found in Figure 2.20. It is worth noting that fever has the closest positive correlation to vomit. This can be seen clearly in Figure 2.18, where these two syndromes have a similar weekly pattern, if at different levels.

2.6.1 Annual Pattern

As with the counts, there appear to be two annual peaks within VP. The first is a local peak around March-April each year and the second occurs near the end of the year. By looking at the proportions, it appears that these are part of an annual cycle. Thus, we try to capture this annual pattern or *profile*. However, at the time of analysis there existed only two years of data, meaning that it is quite plausible that any pattern found within in the data was local.

From Figure 2.17, we can see that at the beginning of the year, VP starts

Cold & Flu	1										
Cough	0.70	1									
Diarrhoea	0.03	0.15	1								
Diff. Breathing	0.31	0.41	-0.13	1							
Double Vision	-0.06	-0.02	0.01	-0.04	1						
Fever	0.47	0.37	-0.15	0.08	-0.06	1					
Eye Problems	0.00	0.15	0.24	0.17	0.13	-0.16	1				
Lumps	-0.26	-0.31	0.10	-0.32	0.01	-0.27	-0.04	1			
Rash	-0.09	0.01	0.37	-0.23	0.02	-0.21	0.30	0.13	1		
Vomit	0.23	0.14	0.20	-0.12	-0.07	0.42	-0.23	-0.06	-0.02	1	
Other	-0.70	-0.84	-0.35	-0.40	0.04	-0.54	-0.25	0.24	-0.23	-0.46	1

Figure 2.20: The correlation matrix for the different syndrome proportions.

quite low and rises to a local peak during March-April. The level of VP then drops more slowly to a low during August and rises from there to another high towards the end of the year. The low during August is attributable to the 'other' syndrome, being proportionately higher during this period. If a call cannot be diagnosed into one of the syndromes of Figure 2.2, it is then labelled as falling into the catch-all category of 'other'. Around 65-75% of algorithmic calls each day fall into this category. From Figure 2.21, it is easy to see that the highs during August for 'other' occur approximately around the same time as VP is low during. The 'other' category may be particularly high during this period as a number of the other syndromes, (again, see Figure 2.2), are typically associated with winter, and so are low at this time. Further, a number of other conditions (i.e. allergies, sun burn) not captured by the syndromes under study are likely to be prevalent during this time too. Thus, when people call up with such conditions during the Summer, the number of calls falling into the 'other' category will increase. This demonstrates some of the complexity mentioned previously in working with the proportion of algorithmic calls corresponding to a particular syndrome instead of the counts of each syndrome.

The usual methods of finding an annual seasonal pattern are differencing or seasonal decomposition. Differencing is not practical here, since a year of data would have been lost, which as there are only two years of data is not tenable. Seasonal decomposition would only really work with more years of data. With the data-set considered here, seasonal decomposition would come down to averaging two numbers for each day, which is unlikely produce an accurate result. Both approaches also suffer from a calendar problem: while August 12^{th} of one year

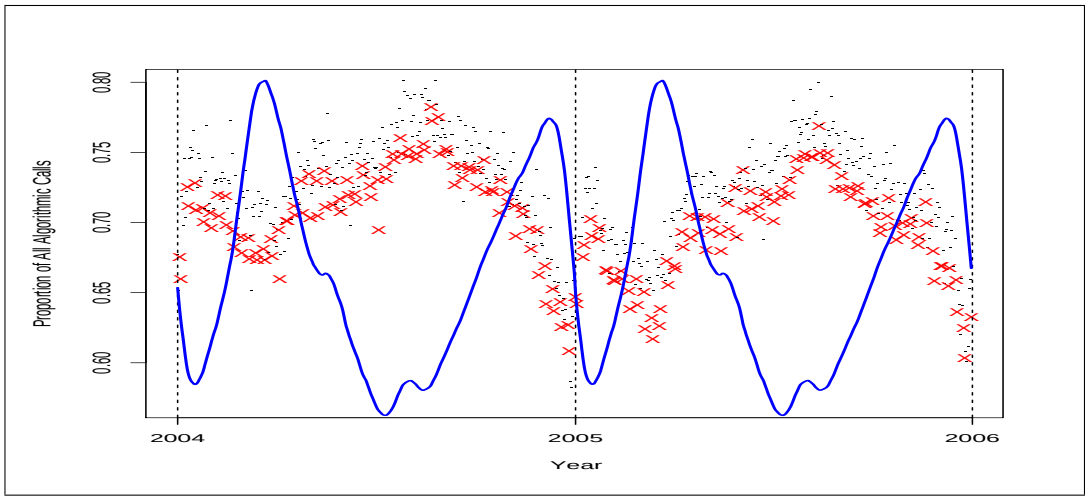


Figure 2.21: The points represent the proportion of all calls that fall into the catch-all category of 'other'. The blue line gives the annual seasonal pattern as we went on to find by running a smoother through VP, scaled to allow comparison with 'other'. The scaling was found by finding the ratio between the maximum proportion value for 'other' and the maximum value of the annual seasonal pattern. This was then used as a scale factor to multiply the annual pattern of VP by, to get it on the same scale as 'other'.

will have a strong link with August 12$^{\text{th}}$ next year, the level will be affected by the day of the week (for instance, the difference in being a week-day or weekend), and any public holidays nearby. Instead of linking only days 365 days apart, we want to link larger periods together and find the more general level for a number of days. This can be done using a periodic smoother: any two days 365 days apart will be linked, but they will also be augmented by days around them. Doing so returns a more general level. Thus, we proceed by using a smoother to find the annual profile.

To calculate a profile, a smoother is applied to the data. We use `R`'s implementation of Friedman's 'super-smoother', `supsmu`, as it is robust and allows the data to be treated as periodic (Friedman 1984). Treating the data as periodic ensures that the beginning and end of the data match in level, as would be expected in an annual profile. The resulting profile plotted against the original data is shown in Figure 2.22.

The annual profile found seems to follow the pattern in the data for the two years reasonably well. However, the smoother does not get far into the local

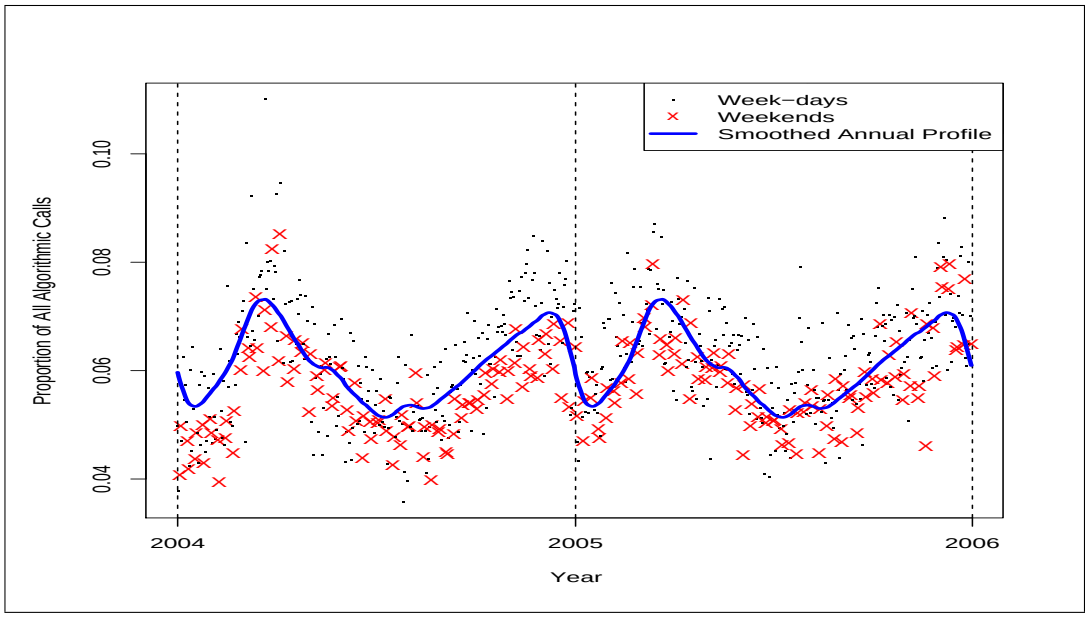


Figure 2.22: The proportion of national calls diagnosed algorithmically each day that mention 'vomit', with the smoother capturing the annual profile in blue.

peaks in March-April, nor at the end of the year. Never the less, we will use this as the basic annual profile and attempt to model deviations from it. In practice, of course, this profile would be re-estimated each year.

We define some terms:

$$X_t = VP \text{ on day } t, \tag{2.1}$$

$$A_t = \text{Annual profile for day } t, \tag{2.2}$$

$$Y_t = X_t - A_t. \tag{2.3}$$

Thus Y_t are our residuals, plotted in Figure 2.23, along with their estimated auto-correlation function, *acf*. It is clear that Y_t is not white noise. There is an evident pattern repeated every seven lags, strongly suggesting weekly seasonality. The earlier observation that weekends tend to be lower than week-days agrees with this. We now consider some time series models that could be fit to Y_t and thus allow predictions to be made, a necessary stage in a exceedance reporting system.

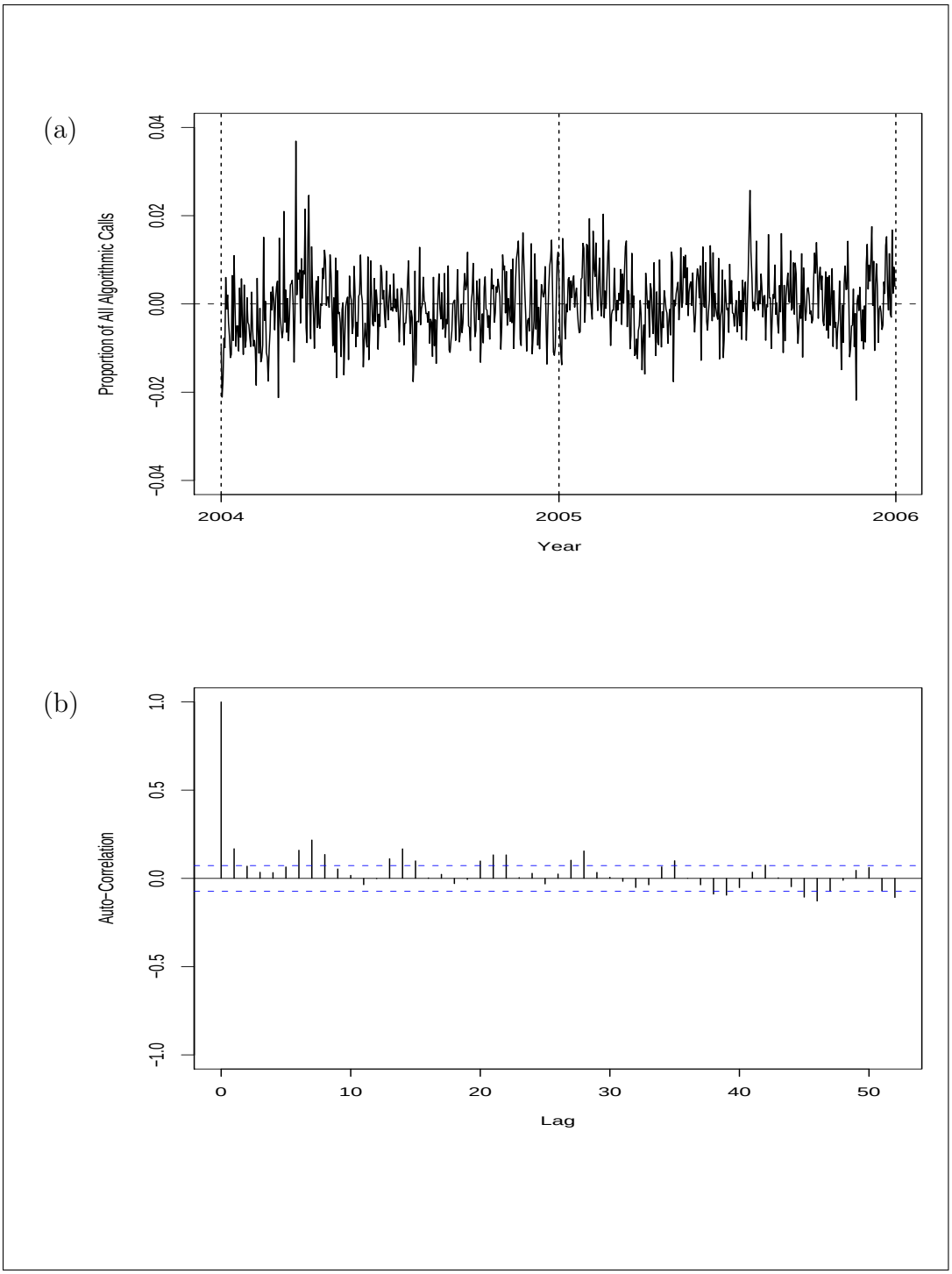


Figure 2.23: Diagnostic plots of the residuals Y_t – that is the difference between the annual seasonal profile (A_t) found by Friedman's Supersmoother and the proportion of algorithmic calls related to vomiting that day: (a) the residuals, (b) acf of the residuals. The blue lines on the acf correspond to the 95% confidence intervals for white noise.

2.6.2 Modelling VP With Annual Seasonality Removed

To gain an insight into trend and seasonality contained in Y_t, we first carry out a seasonal decomposition on Y_t in Section 2.6.3. We then try fitting two types of ARIMA models: a multiplicative seasonal ARIMA model in Section 2.6.4 and a non-seasonal ARIMA model with seasonal factors in Section 2.6.5. ARIMA models are less popular tools for modelling time series in the surveillance context (Burkom 2007). Part of this reticence may be linked with the perceived need for substantial amounts of data and more advanced statistical expertise (Chatfield 1978). However, Reis and Mandl (2003) give an example of syndromic surveillance that utilises ARIMA models. Such models have the advantage of dealing well with short term trend (Burkom 2007); since it would not be surprising if there were serial correlation between successive daily calling proportions, ARIMA modelling seems a reasonable approach to explore. Even if the ARIMA models we develop here are not entirely satisfactory, they are likely to provide insights into the data that will be useful when we try to develop a regional syndromic surveillance system in Chapter 3. Finally, we consider fitting a simple linear regression model with seasonal factors to Y_t in Section 2.6.6 to contrast with the ARIMA models.

2.6.3 Seasonal Decomposition

In seasonal decomposition, a time series, such as Y_t, is broken up into a number of simpler parts. Thus,

$$Y_t = S_t + T_t + R_t,$$

where

$$\begin{aligned} S &= \text{Weekly (day-of-the-week) seasonal pattern,} \\ T &= \text{Trend,} \\ R &= \text{Residuals.} \end{aligned}$$

We calculate an initial estimate of the trend $\widehat{T}$ by running a moving average of order seven through the data. By subtracting $\widehat{T}$ from Y_t, a de-trended series results. The seasonal factors, (that is a level for each day of the week), are found by calculating a simple mean for each day in turn i.e. all Mondays, then all Tuesdays and so on. Finally, the seasonal factors are subtracted from the original

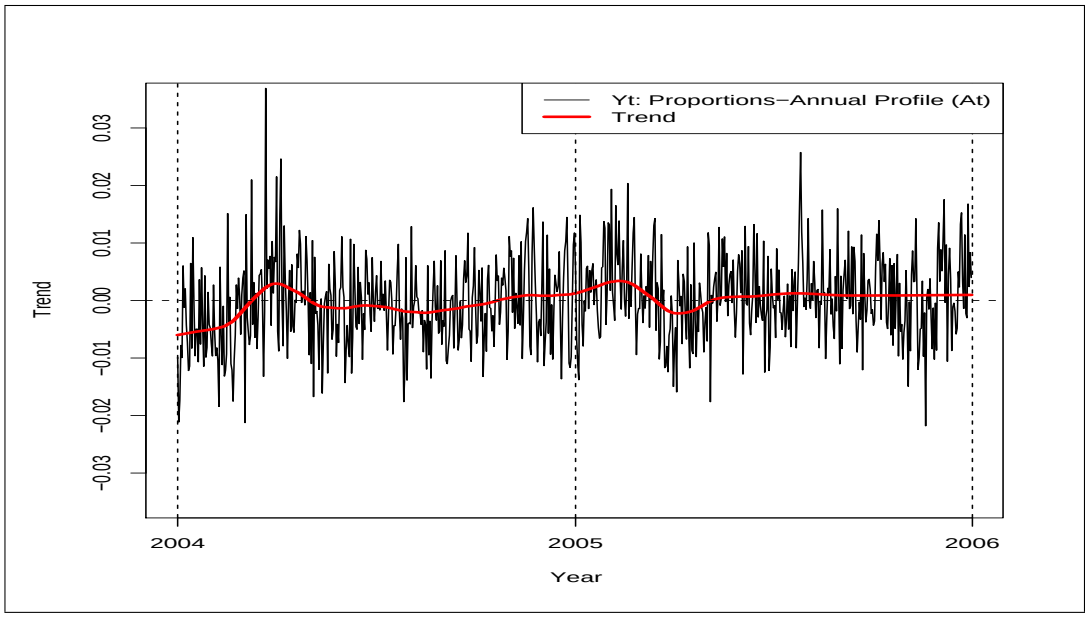


Figure 2.24: Y_t (the residuals after the annual profile has been removed VP) in black. The trend, as calculated by seasonal decomposition in Section 2.6.3, is shown in red.

data, Y_t, yielding a de-seasonalised version of the data, from which a smooth trend was obtained by using Friedman's super-smoother (Friedman 1984).

A number of results become apparent when we decompose Y_t. Firstly, there does appear to be a trend still remaining within the data – see Figure 2.24. It was hoped that by dealing with proportions of calls, instead of with the direct counts of calls, that most trend effects would be removed. The trend is generally increasing slowly for the two years, with a reasonably small magnitude. The trend starts at its lowest value at the beginning of 2004, rising relatively quickly for the next four months, possibly corresponding to the calling levels approaching equilibrium. A local peak occurs in April 2004 and March 2005, with the interesting artifact that in 2005 there follows a local minimum in April. The absolute magnitudes of these turning points are not particularly large, and correspond to the local peaks in X_t (VP before the annual profile has been removed) where the smoother used to produce the annual profile does not follow these peaks particularly closely. It is reasonable to assume from Figure 2.24 that the trend is stochastic: thus, this model could not be used for prediction.

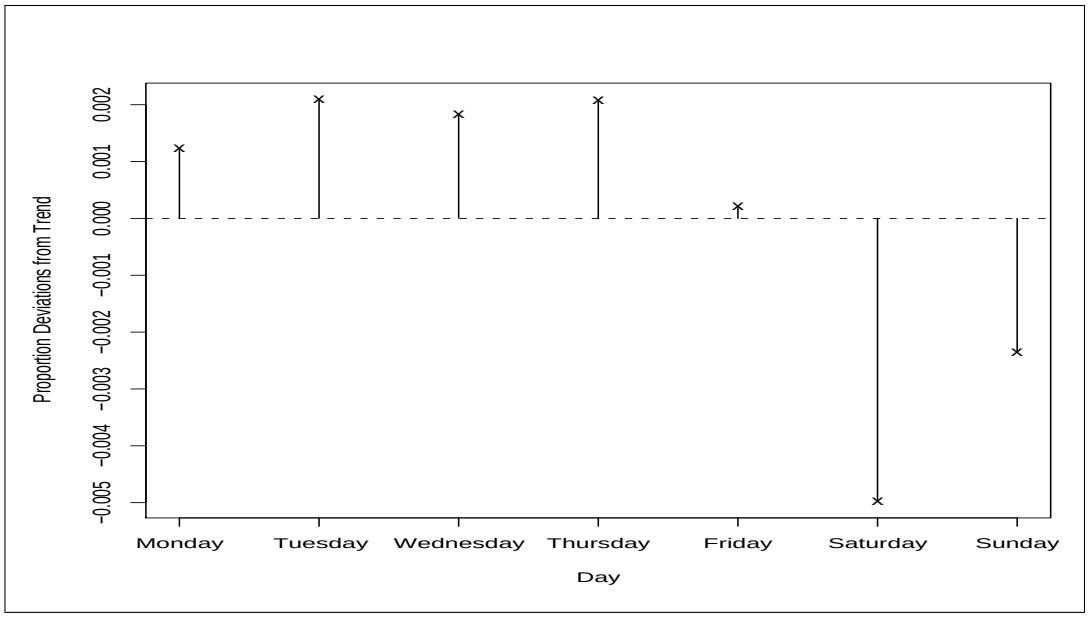


Figure 2.25: The weekly seasonal pattern in Y_t, calculated by using seasonal decomposition in Section 2.6.3. Shown here are the deviations from the trend T as shown in Figure 2.24.

Secondly, this decomposition gives an insight into the seasonal pattern for the days of the week – from Monday to Thursday, 0.001 to 0.002 above trend, Friday about on trend, Saturday 0.005 and Sunday 0.002 below trend (see Figure 2.25). In later models, this suggested that a three level (week-day, Saturday, Sunday) factor might be used to capture weekly seasonality.

The residuals appear well behaved and the qq-plot suggest that they are very close to being normally distributed. A quantile-quantile-plot, qq-plot allows the residuals to be checked for normality. The quantiles in the residuals are plotted against the theoretical quantiles of the normal distribution; if the resulting points approximately follow a 45 degree straight line, there is good evidence for considering the residuals to be normally distributed.

2.6.4 Multiplicative Seasonal ARIMA Model

The general form of the seasonal ARIMA model, as described in Box and Jenkins (1976), is given by:

$$\phi_p(B)\,\Phi_P\left(B^S\right)\nabla^d\nabla_S^D Y_t = \theta_q(B)\,\Theta_Q\left(B^S\right)a_t, \tag{2.4}$$

where a_t is zero-mean white noise, B is the back-shift operator (i.e. $B^k Y_t = Y_{t-k}$), S the period of the seasonality (so for a weekly seasonality, $S = 7$), ϕ and θ are polynomials in B of order p and q respectively (satisfying stationarity and invertability conditions), similarly Φ and Θ are polynomials in B^S of order P and Q (once again, satisfying stationarity and invertability conditions). The difference operator ∇ is defined as $1-B$, so here $\nabla^d = (1-B)^d$, and $\nabla_S^D = (1-B^S)^D$, where d and D the orders of differencing and seasonal differencing respectively. This describes a 'multiplicative seasonal' model, denoted ARIMA $(p,d,q)\times(P,D,Q)_S$. Those parts of Equation (2.4) that involve B^S are referred to as seasonal parts, since B^S links those observations that are separated by the length of the seasonal cycle. We proceed to fit a model of this form to the proportion of algorithmic calls pertaining to vomiting, once the annual profile (A_t) has been removed.

From the auto-correlation function (acf) of Y_t shown in Figure 2.23, we observe a repeating cycle every seven lags corresponding to weekly seasonality. To remove this seasonality, we seasonally difference Y_t (so $\nabla_7 Y_t$). When we consider the acf of $\nabla_7 Y_t$, we find there is a significant correlation at lag seven. The corresponding partial auto-correlation function (pacf) has significant decreasing correlations at intervals of seven, reflecting the significant correlation on the acf. These suggests a multiplicative seasonal ARIMA $(0,0,0)\times(0,1,1)_7$ model.

The acf and pacf of the residuals that result from fitting the ARIMA $(0,0,0)\times(0,1,1)_7$ model are given in Figure 2.26. There appear to be no patterns in correlations seven lags apart, suggesting that the seasonality has been removed. However, all the correlations up to lag 28 on the acf have a positive correlation, while nearly all the lags after 28 have a negative correlation. This suggests that the $(0,0,0)\times(0,1,1)_7$ model should be modified to include a non-seasonal differencing term of order one to address the trend. The pacf also suggests that the non-seasonal part of the model should be adapted to include auto-regressive terms. We try each of these approaches and then choose between them.

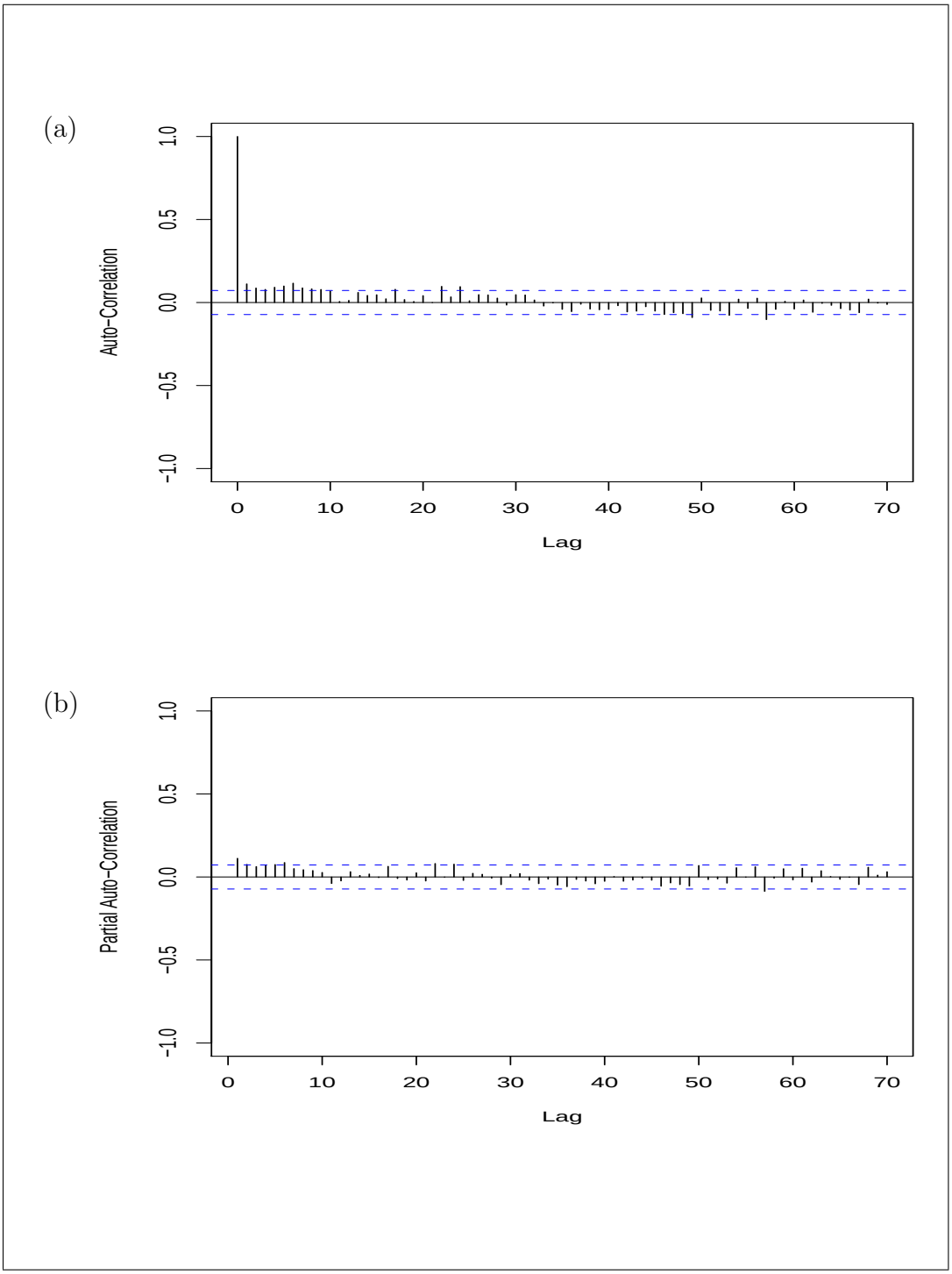


Figure 2.26: The acf (a) and pacf (b) of the residuals of the ARIMA $(0,0,0) \times (0,1,1)_7$ model fitted to Y_t.

Parameter	Value	Standard Error
θ_1	-0.9163	0.0175
Θ_1	-0.9795	0.0161
a_t standard deviation	0.0070	

Table 2.1: A multiplicative ARIMA $(0,1,1) \times (0,1,1)_7$ model is fit to Y_t in Section 2.6.4. It has the following structure: $\nabla\nabla_7 Y_t = (1+\theta_1 B)(1+\Theta_1 B^7)a_t$.

Adapting ARIMA $(0,0,0) \times (0,1,1)_7$ with non-seasonal differencing

Adapting the $(0,0,0)\times(0,1,1)_7$ model to include non-seasonal differencing changes the model to an ARIMA $(0,1,0)\times(0,1,1)_7$. The acf of the residuals for this model have a large negative correlation at lag one of around -0.5. This suggests further adapting the model to include a non-seasonal moving average component of order one: an ARIMA $(0,1,1) \times (0,1,1)_7$. The only significant correlations in the residuals of the ARIMA $(0,1,1) \times (0,1,1)_7$ are at lag 57 on both the acf and pacf. This lag is isolated and bears no obvious relationship with the data source, so we disregard it and believe the residuals free of significant serial correlation. The qq-plot of the residuals also suggest the model is a reasonable fit. Thus, we have a reasonable model in this direction, with its parameters given in Table 2.1.

Adapting ARIMA $(0,0,0) \times (0,1,1)_7$ with auto-regressive terms

Instead of adapting the ARIMA $(0,0,0) \times (0,1,1)_7$ model to use differencing, we can also adapt it to use auto-regressive terms in the non-seasonal component. The acf in Figure 2.26 suggests that an $AR(2)$ or $AR(6)$ may fit the data well. The $AR(2)$ is a reasonable option as the correlation at lag 3 is barely significant, and the $AR(6)$ is explored, as the magnitude of correlations after lag 6 drop almost to in-significance. To decide between the high or low order model, a likelihood ratio test, LRT, can be carried out between the $AR(2)$ and the $AR(6)$. If there is no significant improvement in using the bigger model, (the null hypothesis), the LRT statistic has a χ^2 distribution with k degrees of freedom. In this case, $k = 4$, as there is a difference of 4 between the orders of the two models. When calculated, the LRT has a high significance giving a p-value of 0.001 and so we accept the alternative hypothesis that the $AR(6)$ is more suitable.

Parameter	Value	Standard Error
ϕ_1	0.0928	0.0371
ϕ_2	0.0680	0.0374
ϕ_3	0.0577	0.0373
ϕ_4	0.0666	0.0374
ϕ_5	0.0731	0.0375
ϕ_6	0.0898	0.0374
Θ_1	−0.9632	0.0149
a_t standard deviation	0.0070	

Table 2.2: A multiplicative ARIMA $(6,0,0) \times (0,1,1)_7$ model is fit to Y_t in Section 2.6.4. It has the following structure: $(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_4 B^4 - \phi_5 B^5 - \phi_6 B^6)\nabla_7 Y_t = (1 + \Theta_1 B^7)a_t$.

The only significant correlations in the residuals of the ARIMA $(6,0,0) \times (0,1,1)$ model occur at lag 57 and as above we disregard this correlation since it bears no obvious relationship to the data-source. Thus, these residuals appear to be free of serial correlation. A plot of the residuals suggests they are well behaved. The qq-plot shows some curvature at either 'end' of the data, but this is minor, so the residuals seem reasonably normally distributed. Thus, this seems to be another viable model. Thus, we have arrived at the ARIMA $(6,0,0) \times (0,1,1)_7$ model as the best multiplicative seasonal ARIMA model using auto-regressive terms in its non-seasonal component, with parameters given in Table 2.2.

Final model selection for the multiplicative seasonal ARIMA model

We now have two models to choose between: an ARIMA $(0,1,1) \times (0,1,1)_7$ or an ARIMA $(6,0,0) \times (0,1,1)_7$. The decision depends on whether the residuals of the ARIMA $(0,0,0) \times (0,1,1)_7$ model suggest that the non-seasonal component should contain a differencing term. It is possible to test for a series having a unit root, and so allowing us to test if the series should be differenced. We use the Phillips-Perron unit root test, which has the null hypothesis that a series has a unit root against a stationary alternative – see Perron (1988) for details. When this test is carried out on the residuals of the ARIMA $(0,0,0) \times (0,1,1)_7$, a p-value of 0.01 is found, leading us to reject the null hypothesis that the series has a unit root. This leads us to reject the differenced model, and so we choose the

$AR(6)$ as the most suitable multiplicative seasonal ARIMA.

The acf and pacf of the residuals have already been considered, but we show them in Figure 2.27 for completeness and to demonstrate how these functions behave when they are showing white noise. A plot of the residuals can be found in Figure 2.28. Finally, a qq-plot of the residuals is shown in Figure 2.29.

In Table 2.2, we see that the seasonal parts of the equation are very close to redundancy since $\Theta_1 = -0.963 \approx -1$. This suggests that the seasonal part of the model is unstable and might better be estimated by a deterministic component. This leads to the next model, where seasonality is dealt with by using regressors, instead of seasonal differencing.

2.6.5 Non-Seasonal ARIMA Model with Seasonal Factors

We now deal with the seasonality deterministically. Thus, the proportions of algorithmic calls relating to people complaining of vomit after the annual profile (A_t) has been removed (Y_t) is modelled by:

$$Y_t = \beta_0 + \beta_1 \texttt{Saturday}_t + \beta_2 \texttt{Sunday}_t + x_t, \tag{2.5}$$

where: the β_0 is a constant; β_1 and β_2 are coefficients; $\texttt{Saturday}_t$ and $\texttt{Sunday}_t$ take the value 1 if day t is a Saturday or Sunday respectively and 0 otherwise; x_t is an ARIMA (p,d,q) series:

$$\phi_p(B)\,\nabla x_t = \theta_q(B)\,a_t, \tag{2.6}$$

where the terms are defined as in Equation (2.4).

When an ARIMA model with regressors is fitted in `R`, a linear regression is fitted with an ARIMA model for the error term. Thus, we begin by fitting a linear model and then considering its residuals to determine the ARIMA model that should be fitted to the residuals. The following simple linear model is fit:

$$Y_t = \beta_0 + \beta_1 \texttt{Saturday}_t + \beta_2 \texttt{Sunday}_t + \epsilon_t, \tag{2.7}$$

where the terms are as defined above and ϵ_t are the residuals. This model means we are treating the seasonality as having three levels: weekdays, Saturdays and

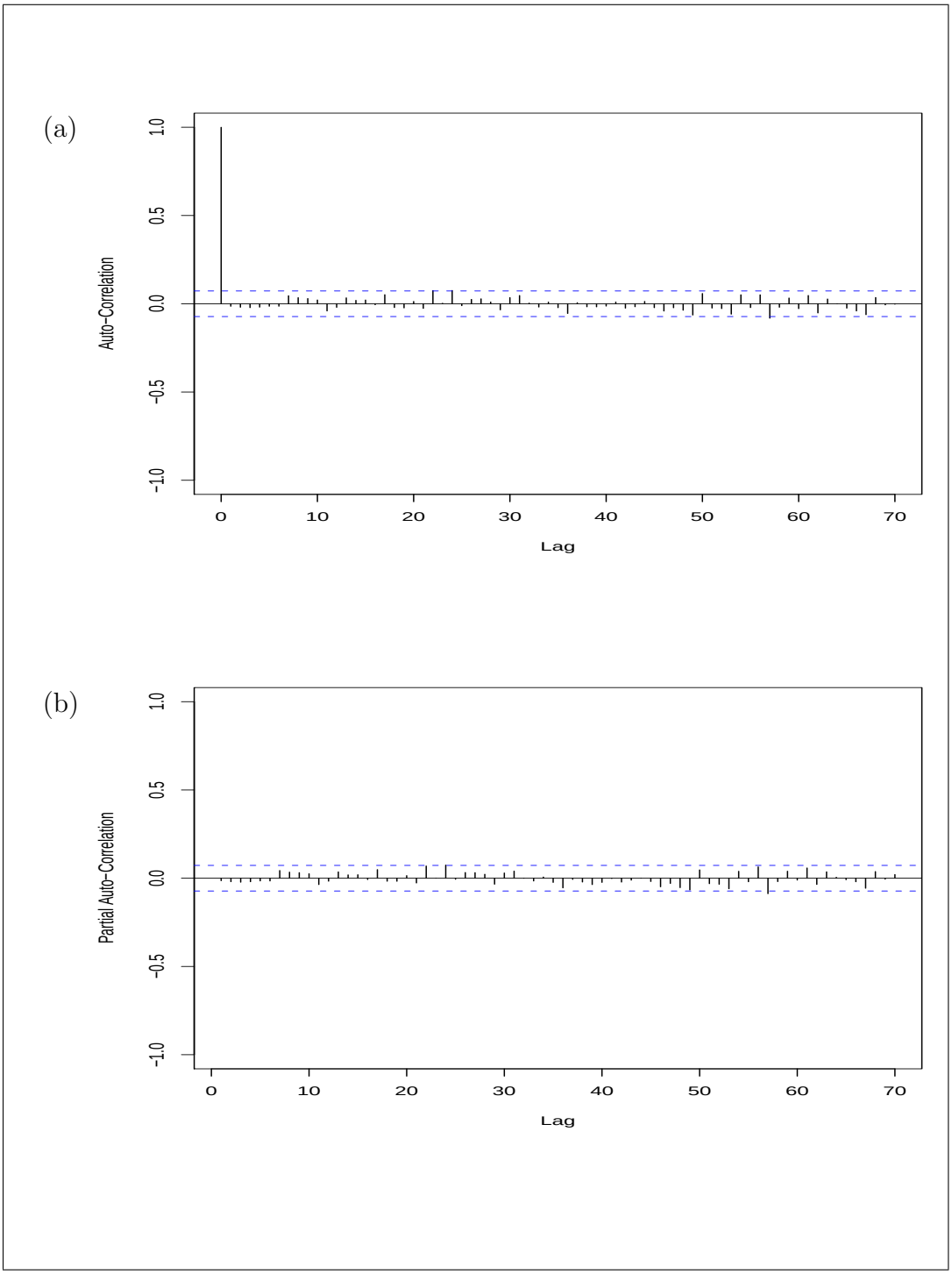


Figure 2.27: The acf (a) and pacf (b) of the residuals of the ARIMA $(6,0,0) \times (0,1,1)_7$ model fitted to Y_t.

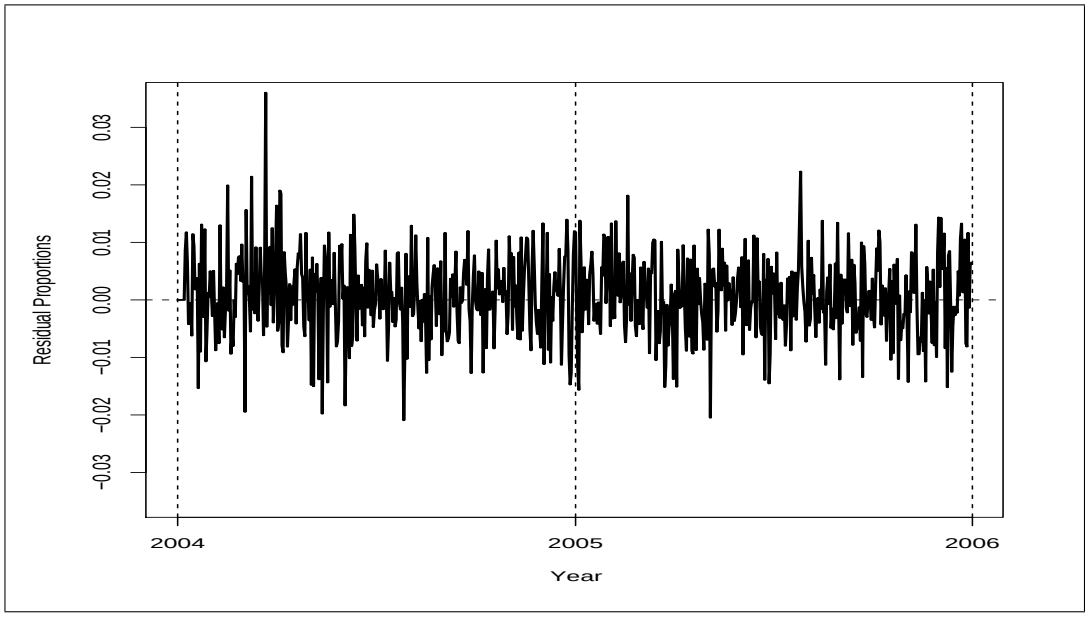


Figure 2.28: The residuals of the ARIMA $(6,0,0) \times (0,1,1)$ model fitted to Y_t.

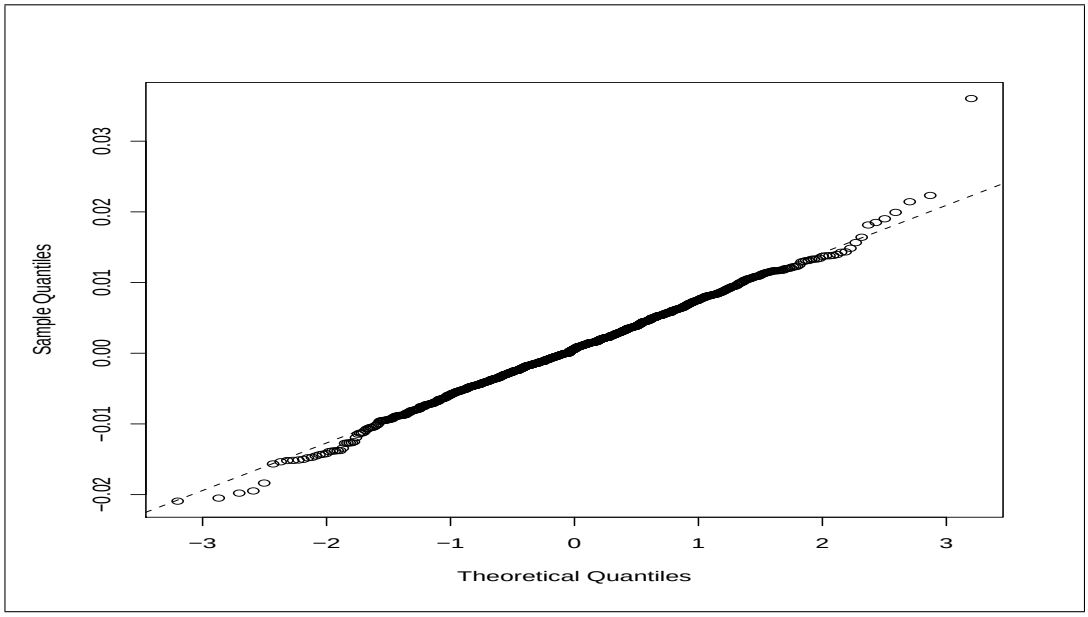


Figure 2.29: A normal quantile-quantile plot, qq-plot, of the residuals of the ARIMA $(6,0,0) \times (0,1,1)_7$ model for Y_t. The dashed line gives the ideal theoretical fit.

Sundays. This is a reasonable initial number of levels given the evidence of Figure 2.25, found when we worked with the seasonal decomposition of Y_t in Section 2.6.3. The pacf and acf of the residuals that result from fitting this model have no patterns in the correlations seven lags apart; this suggests that seasonality has been modelled sufficiently well. As with the model fitting in Section 2.6.4, the residuals of Y_t, once its seasonality has been modelled, suggest two directions of modelling. Firstly, the acf seems to exhibit linear decay, and so is suggestive of the residuals having trend. This is consistent with Y_t having a noticeable trend component, as was also found in the time series decomposition carried out in Section 2.6.3. Thus, since the model specified by Equation 2.7 only addresses seasonality, it is not surprising to find trend in its residuals. Thus, in the first modelling direction, we difference x_t before fitting further ARIMA models. However, the pacf of the residuals ϵ_t also suggest that an auto-regressive model could be fit to x_t, so we take that as our second modelling direction.

Differencing the linear model residuals x_t

In this first modelling direction, we fit the seasonal factors and then an ARIMA $(0, 1, 0)$ to the resulting residuals x_t. The acf of the residuals for this model have a significant correlation at lag one, suggesting that the series should have a moving average of order one fitted ($MA(1)$). Thus, we apply an ARIMA $(0, 0, 1)$ to x_t

On fitting an ARIMA $(0, 1, 1)$ to x_t, we find there is little evidence of serial correlation remaining in the residuals. The only significant correlation on both the acf and pacf occurs at lag 57. It is only just significant. As before, since this lag is unimportant in terms of this system, we disregard it. A plot of the residuals and their qq-plot suggest they behave reasonably well as white noise. Now that we have a suitable model for x_t, we have a suitable model for Y_t, which we detail in Table 2.3.

Linear regression with AR errors

The residuals of the model we initially considered (defined by Equation 2.7) have significant correlations between lags one and seven inclusive on the pacf. Now, the correlation at lag 3 is only just significant, suggesting an auto-regressive model of order 2, $AR(2)$, might be sufficient. However, the significance of the correlations

Parameter	Value	Standard Error
β_0 $(\times 10^{-5})$	1.4637	4.2633
β_1	−0.0065	0.0007
β_2	−0.0038	0.0007
θ_1	−0.9203	0.0171
a_t standard deviation	0.0070	

Table 2.3: A non-seasonal ARIMA model with seasonal factors is fit to Y_t in Section 2.6.5. In this model, a linear regression is used to deal with seasonality: $Y_t = \beta_0 + \beta_1 \texttt{Saturday}_t + \beta_2 \texttt{Sunday}_t + x_t$, and an ARIMA $(0, 1, 1)$ model is fit to x_t: $\nabla x_t = (1 + \theta_1 B) a_t$.

at lags at 4, 5 and 6 grow consecutively larger, so an $AR(6)$ might be another reasonable model. Finally, we might consider an $AR(7)$, since this is the last low order correlation that is significant.

When the $AR(2)$ model is fit, a number of the low order correlations are significant in both the acf and pacf of the residuals, suggesting that a higher order AR series might be better. The acf and pacf of the residuals of the $AR(6)$ and $AR(7)$ both suggest that the residuals are white noise, with little difference between them. To choose between the low or high order models, a LRT is carried out between the $AR(2)$ and the $AR(7)$, in the same fashion as in Section 2.6.4. When calculated, the LRT has a high significance, giving a p-value of 0.0001, indicating the higher order models are preferable.

It now remains to decide between an $AR(6)$ and an $AR(7)$. To decide, we consider the coefficient of the seventh auto-regressive term. The coefficient is 0.0802 and has a standard error of 0.0374. Since the coefficient is more than twice the size of the standard error we can be confident that the seventh auto-regressive coefficient should not be zero. Thus, we choose the $AR(7)$ model. This choice is also corroborated by by performing an LRT between the $AR(6)$ and $AR(7)$ models; this test gives a p-value of 0.0321, confirming that the bigger model is preferable. The acf and pacf of the residuals suggest that no serial correlation remains in the residuals. A plot of the residuals and the qq-plot suggest they are reasonably distributed as white noise. Thus, we have a reasonable model for Y_t, with parameters given in Table 2.4.

Parameter	Value	Standard Error
β_0	0.0014	0.0006
β_1	−0.0050	0.0008
β_2	−0.0024	0.0008
ϕ_1	0.0858	0.0370
ϕ_2	0.0683	0.0371
ϕ_3	0.0549	0.0373
ϕ_4	0.0642	0.0373
ϕ_5	0.0735	0.0373
ϕ_6	0.0831	0.0374
ϕ_7	0.0802	0.0374
a_t standard deviation	0.0069	

Table 2.4: A non-seasonal ARIMA model with seasonal factors is fit to Y_t in Section 2.6.5. In this model, a linear regression is used to deal with seasonality: $Y_t = \beta_0 + \beta_1 \texttt{Saturday}_t + \beta_2 \texttt{Sunday}_t + x_t$, and an ARIMA model is fit to x_t: $(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_4 B^4 - \phi_5 B^5 - \phi_6 B^6 - \phi_7 B^7) x_t = a_t$.

Choosing between models for the residuals x_t of the linear regression

Thus, we have two models to choose between that both deal with seasonality in a deterministic fashion. To check if the model should have a differencing term, we use the Phillips-Perron on ϵ_t, the residuals after the linear model has been fit to deal with seasonality. When this test is carried out, a p-value of 0.01 is found, leading us to reject the null hypothesis that the series has a unit root. Thus the $AR(7)$ model for x_t is chosen.

Number of seasonal levels

As a final consideration, we investigate how many levels the weekly seasonality should have. It is quite plausible that there is a difference between week-days and weekends, but it is less obvious why there would be a difference between Saturdays and Sundays. Thus, we test to check if this extra level is significant by fitting another model with only two levels for week-days and weekends. We carry out an LRT against the model with three levels. When this is done, the LRT is calculated to be 7.18, with a p-value of 0.0074. This indicates that it is best to have three levels for the seasonality: one each for weekdays, Saturdays and Sundays.

Final deterministic seasonality model with ARIMA error term

Thus, the most appropriate model for dealing with seasonality in a deterministic fashion is specified in Table 2.4. This model has three levels for the seasonality: a level for weekdays and separate levels for Saturdays and Sundays. Similar to the model developed in Section 2.6.4, we found the best way to deal with serial correlation was to use autoregressive terms. However, since we have an $AR(7)$ term, a term that links days a week apart, it does suggest that the seasonal levels change as time goes on. We now consider a parallel of this model in a linear regression form.

2.6.6 Regression Model with Seasonal Factors

This model does not utilise ARIMA methodology, but instead uses linear regression. We consider a model of the form:

$$Y_t = \beta_0 + D_t + \sum_{i=1}^{k} \beta_i Y_{t-i} + \varepsilon_t, \tag{2.8}$$

where D_t is a factor with three levels (weekdays, Saturdays and Sundays) and ε_t is an error term. This uses the last k days to calculate a prediction for the level of the present day t, while D_t captures weekly seasonality.

It is necessary to decide on the value of k, the number of past days to use within the regression. To determine k, ANOVA is carried out on the regression model, in such a way that the significance of each additional factor included in the model is tested – see Figure 2.30. We find that only the first seven days, $k <= 7$, contribute significantly to the model. Indeed, the inclusion of Y_{t-7} is only significant at the 10% level; however, we keep Y_{t-7} in the model while we develop it further.

We now consider the factor D_t. We test using ANOVA the initial model with three levels (week-day, Saturday, and Sunday) for D_t against one with just two levels (week-day, weekend). In doing so, it is found that using a three level factor for D_t is a significant improvement to the model at the 5% level, since a p-value of 0.01493 is found. Thus, we keep D_t with three levels.

Finally, we take $k = 6$, since the inclusion of Y_{t-7} is only significant at the 10 % level. The number of levels of D_t is checked again using ANOVA, and three

```
> anova(reg14)
Analysis of Variance Table

Response: Yt
           Df   Sum Sq  Mean Sq F value    Pr(>F)
D3          2 0.004236 0.002118 43.8671 < 2.2e-16 ***
p1          1 0.000528 0.000528 10.9325 0.0009936 ***
p2          1 0.000202 0.000202  4.1737 0.0414342 *
p3          1 0.000130 0.000130  2.6914 0.1013428
p4          1 0.000246 0.000246  5.0848 0.0244477 *
p5          1 0.000351 0.000351  7.2670 0.0071938 **
p6          1 0.000420 0.000420  8.6974 0.0032939 **
p7          1 0.000167 0.000167  3.4546 0.0634988 .
p8          1 0.000051 0.000051  1.0527 0.3052341
p9          1 0.000044 0.000044  0.9053 0.3416856
p10         1 0.000017 0.000017  0.3519 0.5532593
p11         1 0.000053 0.000053  1.1013 0.2943374
p12         1 0.000002 0.000002  0.0426 0.8365213
p13         1 0.000090 0.000090  1.8682 0.1721253
p14         1 0.000026 0.000026  0.5441 0.4609829
Residuals 693 0.033462 0.000048
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Figure 2.30: The results of carrying out ANOVA on the regression model, where the significance of each additional term in relation to including all the previous terms is tested. Terms such as p12, refer to a term in the model corresponding to twelve days previous to the current day i.e. Y_{t-12}

Parameter	Value	Standard Error
β_0	0.0017	0.0003
D_{Saturday}	-0.0067	0.0007
D_{Sunday}	-0.0038	0.0008
β_1	0.0858	0.0377
β_2	0.0527	0.0364
β_3	0.0352	0.0365
β_4	0.0612	0.0364
β_5	0.0660	0.0365
β_6	0.1004	0.0362
σ	0.0069	

Table 2.5: A linear regression model is fit to Y_t in Section 2.6.6; it has the following form: $Y_t = \beta_0 + D_t + \sum_{i=1}^{k} \beta_i Y_{t-i} + \varepsilon_t$.

levels is found significant once more. Thus, we chose the model with $k = 6$, and D_t having three levels. We check the acf and the pacf of the residuals, as we have done previously, and find nothing to suggest that there is serial correlation present in the residuals. A plot of the residuals and their qq-plot suggest they are distributed as white noise. Thus, we have another reasonable model for Y_t which is specified in Table 2.5.

2.6.7 Comparison of Models

In this Section we have fit a number of models to Y_t, the proportion of algorithmic calls after the annual profile A_t has been removed. These models are summarised in Table 2.6. Most of these models allow predictions to be made of future calling proportions, an essential step in an exception reporting system. We now consider which of the models would be most suitable to model Y_t.

In Section 2.6.3 we carry out a time series decomposition on Y_t. Rarely can such a system be used for prediction as usually the estimated trend is non-parametric. However, the decomposition does provide information about Y_t. First, a small increasing trend is still present in the data, with slightly larger perturbations caused by the annual profile A_t not getting into the peaks of call levels near April each year. It has been hoped that by using proportions trending effects would be reduced; here the trend effect has indeed been reduced but it is

Name	Section	Final Fitted Model
Seasonal Decomposition	2.6.3	$Y_t = S_t + T_t + R_t$
Multiplicative Seasonal ARIMA	2.6.4	$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4-\phi_5 B^5-\phi_6 B^6)\nabla_7 Y_t = (1+\Theta_1 B^7)a_t$
Non-Seasonal ARIMA (with Seasonal Factors)	2.6.5	$Y_t = \beta_0 + \beta_1 \texttt{Saturday}_\texttt{t} + \beta_2 \texttt{Sunday}_\texttt{t} + x_t$ $(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4-\phi_5 B^5-\phi_6 B^6-\phi_7 B^7)x_t = a_t$
Linear Regression	2.6.6	$Y_t = D_t + \sum_i^6 \beta_i Y_{t-i} + \epsilon_t$

Table 2.6: The various models that have been fitted to model the proportion of calls relating to vomit, after the annual profile has been removed.

still present. The weekly seasonality (day-of-the-week effect) was confirmed as broadly having three levels: one for week-days and two different levels for Saturday and Sunday. Given more time and data it might be interesting to investigate what causes the difference between Saturdays and Sundays.

Since we then had evidence for Y_t having seasonality, we considered ARIMA models that could address such seasonality. The first such model was a multiplicative seasonal ARIMA model (Section 2.6.4). The seasonal terms in this model appeared to be near redundancy, suggesting that the seasonality would be better dealt with in a deterministic fashion. Mindful of the comments of Chatfield (1978) that ARIMA models can seem very complex, other models may also have the benefit of appearing more straight forward than the conceptually advanced multiplicative seasonal ARIMA model. This can be an important consideration given that exception reporting systems are often used by non-statisticians.

In the next ARIMA model seasonality was modelled deterministically by the use of regressors (Section 2.6.5). A linear regression is carried out to fit the seasonal regressors and an ARIMA model is fit to the resulting residuals. Three levels were found to capture the seasonality well, as would be expected from the time series decomposition. However, the ARIMA model contains a seventh order auto-regressive term, connecting days a week apart. This suggests that while seasonality is best modelled deterministically, it is still undergoing some changes. This is not surprising given our data comes from the very start of the system. The structure of this model is also somewhat more straight forward to explain to a non-statistician.

The linear model developed in Section 2.6.6 is in some sense a parallel of the ARIMA model with seasonal regressors in Section 2.6.5. The linear model has three levels for seasonality and utilises the results of preceding days to form a level for the present day, just as the ARIMA model does. However, they differ in the number of past days they use to form their prediction: the linear model only uses the past six days, where as the ARIMA model uses the past seven days. This reinforces the above observation that seasonality is best modelled deterministically, even though it seems to be changing slightly. This model is arguably the simplest, not relying on knowledge of ARIMA methodology.

If more data were available, we could use part of it as training data to fit the models. The remaining data would then be used to test the different models and

allow comparison between them. We would consider which models gave the 'best' predictions and consider the alarm rates under each model. We could then make an informed decision as to which model would best suit our purpose. However, as we only have two years of data this is not feasible. Thus, we can only choose between the models on the basis of which seems most suited to the data. Since the seasonality seems to be most suitably modelled by a deterministic model, either the non-seasonal ARIMA model in Section 2.6.5 or the linear regression model in Section 2.6.6 seem best. The choice between these two models might be based on the expertise of the people who will be monitoring the system, with the linear model likely being the most straight forward.

2.7 A National Exception Reporting System

We now have all the necessary elements needed for a national exception reporting system to monitor levels of calls to NHS24 relating to people complaining of vomiting. On day t, we can use the information up and to and including that of day t to predict the proportion of algorithmic calls related to vomiting, VP_{pred}, on day $t+1$ and form a confidence interval. Should the observed proportion of algorithmic calls on day $t+1$, VP_{Obs}, be larger than the upper limit of the confidence interval, then an exception would be reported to NHS24. At this point, it is up to NHS24 to decide how they react to the exception. They might dismiss the exception if they know of some obvious reason for it, or they can investigate the log of calls more closely and check for any common traits between those people complaining of vomiting. For instance, if most of the calls come from one geographical area, there might be an outbreak of food poisoning with a side effect of causing people to vomit and thus a localised group of people calling up NHS24 for advice. NHS24 could then call these people back and try to determine the common source for their poisoning.

In more technical detail, the calculations for the exception reporting system, using the linear regression model for Y_{t+1} on day $t+1$, would be as follows:

1. Calculate the prediction $\hat{Y}_{t+1}$ from the fitted regression formula:

$$\hat{Y}_{t+1} = \beta_0 + D_{t+1} + \sum_{i=1}^{k} \beta_i Y_{t-i}$$

2. Calculate the upper $100(1-\alpha)$ confidence limit U_α:
   We expect the prediction $\hat{Y}_{t+1}$ to be distributed normally with mean $\hat{Y}_{t+1}$ and standard deviation σ (in our model $\sigma = 0.0069$). Then, if z_α is the $100(1-\alpha)$-percentile of the normal distribution

$$U_\alpha = \hat{Y}_{t+1} + z_\alpha \sigma$$

3. Then, including the annual profile, A_{t+1}, compare this with observed proportion VP_{Obs}:

$$\text{If } U_\alpha + A_{t+1} > VP_{Obs} \text{ report an exception}$$

4. If an exception is raised, it is reported to NHS24 and they can choose how to respond to it.

We are free to vary the value of α depending on the level of exceptions we wish to have raised; a larger α will lead to more false alarms.

If instead the system used the non-seasonal ARIMA model for Y_{t+q}, the process would be essentially the same but predictions would be found using ARIMA forecasting procedures. Here we have chosen to use a type of an exceedance reporting method, as was used in McCabe (2004), where some upper point of the fitted distribution is compared with an observed value. Other possible alarm methods exist – such as the CUSUM – but we merely use this method here to demonstrate how such a national exception reporting system might work.

2.8 Conclusions

In this chapter we have considered using the data from the telephone health advice line NHS24 to develop a national exception reporting system (NHS24 2008). People can call up NHS24 for health advice twenty-four hours a day, every day of the year. This service is available across the whole of Scotland. As key details of every call are recorded, this call log has the potential to be a very rich data-set for the development of a daily national exception reporting system (see Section 2.3). NHS Direct performs the same function as NHS24 for England and Wales but has been running for a longer time (NHS Direct 2009). The developers of the

NHS Direct exception reporting system have suggested that their system might be able to go beyond detecting widely dispersed 'health events' to detecting local outbreaks of disease if NHS Direct was utilised more (Henning 2004; Doroshenko, Cooper, Smith, Gerard, Chinemana, Verlander, and Nicoll 2005; Cooper 2007; Smith, Cooper, Loveridge, Chinemana, Gerard, and Verlander 2006; Cooper, Verlander, Smith, Charlett, Gerard, Willocks, and O'Brien 2006). Since NHS24 has twice the calling rate of NHS Direct, this condition may be satisfied. Further, the ability to detect local outbreaks may be improved by modelling at the regional level. Our ultimate goal is to explore developing a regional exception reporting system but we start by exploring a national system to help inform our modelling choices when we turn to modelling the regional level.

The amount of data from we had from NHS24 was immense: we had key details of all of the calls from the first two years (2004, 2005) of the system – around a million and a half phone calls. It was necessary to narrow our focus to particular subsets of this data. We chose to focus on those calls that were triaged by the Clinical Assessment System, a set of algorithms that suggest to the nurses at NHS24 the advice they should give to their callers (Cooper, Smith, Baker, Chinemana, Verlander, Gerard, Hollyoak, and Griffiths 2004). Calls diagnosed by the algorithm system were grouped into ten syndromes (Figure 2.2). The purpose and advantages of monitoring such syndromes are considered in Section 2.2; but in summary, they allow for detection of terrorist attacks and earlier detection of 'health events' than more traditional methods of monitoring. This more general monitoring also matches the non-specific diagnosis that the phoneline gives. Some limited research has been conducted on the efficiency in gaining biological samples from callers for more specific diagnosis (Cooper, Smith, Chinemana, Joseph, Loveridge, Sebastionpillai, Gerard, and Zambon 2008). Syndromic surveillance is carried out on the NHS Direct data (Cooper 2007).

We show in Section 2.3 (particularly Figure 2.5) that the number of calls being diagnosed through use of the algorithms is decreasing during the years for which we had data. Later work has shown this trend has continued to the point where focusing on the calls diagnosed by algorithms will not give representative monitoring possibilities (Wilson, Smith, Meyer, Robertson, Baxter, Cooper, and McMenamin 2007). However, grouping the calls in other ways, such as using the 'call reason' field, brings other problems (Kavanagh, Robertson, and McMenamim

2007). At the time of analysis, the algorithmic calls presented the most stable subset of the data to use. Given the data was from the start of the system and unlikely to be completely stable, choosing the most stable elements we could seemed the sensible choice.

At this point, we still had data from ten different syndromes. Again, because of having data from the start of the system, we chose to focus on one syndrome which looked the most stable. In Section 2.5.2 we choose the 'vomiting' syndrome because it is both stable and has large enough counts to make regional modelling feasible. In an exception reporting system, a model of the data is required to allow for predicting future levels of the quantity being monitored. Thus, we then turn to modelling the daily number of calls algorithmically triaged as relating to the syndrome of vomiting.

We first carry out some exploratory modelling on the counts of calls relating to vomit in Section 2.5.3. From this modelling, a number of qualities become apparent about the counts. The number of calls received at the weekend compared to during the week is much higher – about 2.3 times more. The counts generally increase for 2004 and decrease for 2005. This level of trend in such a data-set this small is very hard to model. Thus, we turn to using proportions: we consider the proportion of algorithmic calls that relate to vomiting in Section 2.6. We mention briefly some of problems of using proportions in this section. Proportions are also made use of in NHS Direct's system (Cooper 2007).

On considering the proportions, we find that they are somewhat more stable than the counts. We find that there are two levels of seasonality in the data: annual and weekly (day-of-the-week effect) seasonality. In Section 2.6.1, we consider approaches to deal with the annual seasonality. Since we only have two years of data, standard techniques for extracting the annual pattern are not feasible. Thus, we turn to using a periodic smoother to extract an annual profile (A_t).

The annual profile (A_t) is then subtracted away from the proportion of calls that relate to vomiting to give Y_t. This quantity still has weekly (day-of-the-week effect) seasonality present, so we consider models for it which can take account of this. We expect the proportions on one day to be closely linked with the day before it; that is, we expect there to be significant levels of serial correlation present in Y_t. Given this correlation, we argue that ARIMA models present a natural choice for modelling Y_t, even if such models are not usual in this

context (for a notable exception, see Reis and Mandl (2003)). We try a number of ARIMA models for Y_t, and determine the most suitable one in Section 2.6.5: the weekly seasonality is modelled by seasonal regressors (three levels: week-days, Saturdays and Sundays) and the local trend is captured by a number of auto-regressive terms. A linear model parallel of this ARIMA model is considered in Section 2.6.6; this model has the advantage that it does not utilise ARIMA methodology, so might be preferable if the end user of the exception system has limited statistical experience. Normally, simulations would be carried out under the different models to determine the best choice but the limited data prevents this – see Section 2.6.7 for further discussion of this.

With the annual profile A_t and the model for Y_t, we now have all the modelling elements required for an exception system to monitor calls to NHS24 relating to vomiting. Using an exceedance reporting alarm system, similar to that used in Farrington, Andrews, Beale, and Catchpole (1996), we show how an exception reporting system can be created in Section 2.7. Other alarm choices could be used, such as a CUSUM method, but since our focus is on developing a regional system, we merely wish to present a suggestion for a national system.

The strength of the exception system developed in this Chapter hinges upon the amount and quality of data we have. We have two years of data from the start of the system; such data is unlikely to be uniformly consistent or stable. With only two years of data it is hard to be particularly confident about the annual profile (A_t) that was extracted. Also, the work of Kavanagh et al. (2007) has shown that better results are given by using the call reason field. However, there are some advantages to the system developed in this chapter. First, Kavanagh et al. (2007) do not develop an annual profile in their system, rather using a shifting baseline of data for fitting their models (see McCabe (2004), Burkom (2007) for discussions of this method). Smith et al. (2006) suggest that one of the uses for the data recorded by NHS Direct is reassuring the general public about health issues and emergency planning and exercises. For such activities, having an explicit annual profile as we do here will be useful. Further, Kavanagh et al. (2007) do not address serial correlation, which, as we have argued, is very likely to be present in this data. This may deteriorate their models somewhat. Future work might consider fitting ARIMA models to the national data but with calls categorised by the call reason field and comparing this with the system

developed in Kavanagh et al. (2007).

It is worth noting that the problem of stability that we encountered in this chapter is true for all exception reporting systems. However, this is particularly so for syndromic surveillance systems. As Henning (2004) notes, the data used for syndromic surveillance systems is often not collected primarily for this purpose. Thus, the data can often be affected by artificial pressures not related to natural variations. For instance, as a public body, NHS24 is constantly subjected to review, and their goals updated (NHS24 2006; NHS24 2009). Their function is to provide the best possible service they can to the public of Scotland when that public calls them seeking health advice. As the definition of this 'best service' changes, levels in the recorded data will change too. The advantage of earlier detection in syndromic surveillance comes at the cost of needing to monitor the systems more closely.

Our main purpose in considering a national exception system was to inform choices for developing a regional system. At the regional levels, we are going to have smaller counts and thus smaller proportions. ARIMA models rely on normality, which is reasonable with the size of counts at the national level, but when these counts are divided up between the fourteen health boards normality is likely to break down. Thus, we turn to using GLMs to model the data, noting that in the counts we expect serial correlation and both annual and weekly seasonality. It is quite usual to use GLMs for this purpose (Burkom 2007; McCabe 2004; Farrington, Andrews, Beale, and Catchpole 1996).

Chapter 3

NHS24 GLM Modelling

In Chapter 2, we considered some models that could be used to nationally predict the number of people calling NHS24 complaining of vomiting. As we were dealing with national levels of calls, and so reasonably large numbers, it was reasonable to model them with a normal distribution (McKenzie 2003). However, this approximation is no longer valid when we move to considering regional models; the numbers we are dealing with become smaller as the national counts of calls are split up between the fourteen health boards. The models considered in Chapter 2 all assumed a normal distribution for the counts and so we must to turn to other modelling approaches to deal with with the smaller regional counts. One standard approach is to use generalised linear models (GLMs) and fit Poisson, or related, distributions to them. We consider such models in this chapter, first fitting a GLM to the national counts and then turning to the regional counts. GLMs rely on observations being independent, which is unlikely to be the case in consecutive days of call counts to NHS24: if the call rate was high yesterday it is likely to be high again today. We propose a way of combining GLMs with a Holt-Winters predictor to address any serial correlation in the counts. Using this predictor, we go on to consider a way that predictions between the regions might be linked together. These models can then be brought together to form the basis of a regionally linked exception reporting system. However, as noted previously, we do not have enough data to test the developed system.

3.1 National GLM Rate Model

Typically, a Poisson GLM is used to model counts. However, we show in the following sections how it is possible to use an offset in a Poisson GLM to allow modelling of rates and proportions. This allows us to essentially model the proportion of calls to NHS24 that are related to vomiting. We fit such models to the national counts of calls in Section 3.1.2

3.1.1 Offsets: From Counts to Proportions

As we are modelling counts, it is standard to start by fitting a Poisson distribution. Thus, having specified the response distribution, a link function is required to connect the predictors, X_t, with the mean, μ_t, of the number of calls mentioning 'vomit' on day t. For simplicity, the usual link function for the Poisson, the log-link, was chosen giving:

$$\log \mu_t = \beta X_t.$$

If we introduce $\log T_t$ on the right, where T_t is the total number of calls diagnosed algorithmically on day t, we have:

$$\log \mu_t = \beta X_t + \log T_t,$$

which can be rearranged to give:

$$\log \frac{\mu_t}{T_t} = logVP_t = \beta X_t,$$

which gives a way of modelling the daily rate VP_t, as has been considered in Chapter 2 while fitting a Poisson GLM to the counts of calls. The introduction of a term on the right, which is defined to have a coefficient of one, is sometimes called an 'offset' (see Faraway (2006), McCullagh and Nelder (1983)). Within `R`, quantities can easily be specified as offsets in GLMs.

3.1.2 Initial National Rate Models

Utilising the idea of a rate model, the following Poisson model is fitted to the means μ_t:

$$\log \mu_t = \log T_t + \beta_0 + \beta_1 \log A_t + \beta_2 Ti_t + \sum_{1}^{35} \beta_{i+2} \log(x_{t-i}), \tag{3.1}$$

where: μ_t and T_t are as above; A_t is the annual profile for day t, as found in Section 2.6.1; Ti_t is the day number ($Ti = 1$ on Jan 1st 2004) giving a linear trend; x_s is the observed count on day s; and the β_i are coefficients. We fit $\log T_t$ as an offset. The log of the annual profile (A_t) is used as we expect this to have a multiplicative effect on the number of calls received. The past values are logged so that they are on the same scale as the mean μ_t. Five weeks of past values are included in the model. We hope to dispense with a factor for weekly seasonality by including whole weeks. This number of weeks is chosen as any local trend is likely to be captured within this time.

On fitting this model, the coefficient of the annual profile, β_1, is found to be 0.992 with a standard error of 0.043. This suggests that the $\log A_t$ term should also be treated as an offset, along with $\log T_t$, since the coefficient is very close to one. This approach is suggested in Faraway (2006).

An indication of the fit given by a particular GLM is given by its residual deviance. The model considered here has a residual deviance of 969, with 658 degrees of freedom, suggesting a poorly fitting model. Further, since the residual deviance is much greater than the residual degrees of freedom, it suggests that the Poisson is not a good choice for the response distribution. A likely cause would be over-dispersion, where the variance of the counts is different from their mean; this contrasts with a Poisson fit which assumes the mean and variance are equal. In such cases a quasi-Poisson model can be fitted to the counts. The quasi-Poisson is not a proper distribution; it has the same form as the Poisson distribution, but the variance is defined to differ from the mean by a multiplicative constant called the 'dispersion' parameter. It is a convenient method for obtaining a more accurate and reliable test of the significance of the model parameters (Faraway 2006; McCullagh and Nelder 1983; Hilbe 2007). Estimates of the coefficients remain the same but since the variance is modelled differently, significance tests

of the predictors can be affected. Predictors that were previously significant may become insignificant under the quasi-Poisson model.

To check if a quasi-Poisson GLM is required, the above model is fitted with a quasi-Poisson response. When this is done, the dispersion parameter is found to be 1.47; thus the variance should be 1.47 multiples of the mean. Given this is quite different from 1, it suggests that the quasi-Poisson is a better choice of response distribution. Had the dispersion parameter been near one, there would have been little difference compared to fitting the standard Poisson distribution.

When count data has variance greater than its mean it is said to be 'over-dispersed'. Over-dispersion often happens when the assumption of counts being randomly distributed is violated. For instance, in an epidemiological context, one would expect clustering. Consider, for example, an infection caused by a bacterium (or other underlying cause of infection): the infection has a greater chance of being passed to those that are nearby, potentially creating a cluster of people suffering from vomiting. For a longer discussion on over-dispersion see Hilbe (2007).

Finally, we fit a quasi-Poisson model where the log A_t term is also treated as an offset. We find there is significant serial correlation in the residuals, even though we include many past terms in the model. The acf of the deviance residuals has significant correlations at lags one, five, six and seven. The significant correlation at lag one would have the largest effect on prediction, necessary in exception reporting, since we would most frequently be predicting ahead by one day. Having a significant correlation at lag seven suggests that weekly seasonality is not being modelled suitably. This is surprising since we have included five complete weeks of past observations in the model; we would expect this to be amply sufficient to capture the weekly seasonality. Thus, we alter the national model to include a seasonal factor. Using a seasonal factor may also allow the model to be more parsimonious, as fewer past values may be required. Also, there is no requirement for them to be in whole multiple of weeks since we are no longer using them to capture weekly seasonality.

3.1.3 National Quasi-Poisson Rate Model with Seasonal Factor

To address residual weekly seasonality, we alter the model defined by Equation (3.1) to the following quasi-Poisson model:

$$\log \mu_t = \log T_t + \log A_t + D_t + \beta_0 + \beta_1 Ti + \sum_{1}^{35} \beta_{i+1} \log(x_{t-i}), \tag{3.2}$$

where μ_t, T_t, A_t, Ti_t, x_s and the β_i, are as in Section 3.1.2. Both $\log T_t$ and $\log A_t$ are offsets. The term D_t is a factor with seven levels, one for each day of the week.

We first use ANOVA to check how many past values should be included in the model. F-tests are used in the ANOVA, as suggested in Faraway (2006), since two parameters are being estimated (the mean and the dispersion parameter). Few of the past values contribute significantly to the model; the only ones that do contribute significantly are lags of 6, 7, 21, 28. A number of these are to do with seasonality given they are multiples of seven. We consider a reduced model that just has the first seven lags in it. In this reduced model, lags at four and seven are significant. If we were to keep up to lag four, it would be standard to keep lags one, two and three which are insignificant. Thus, instead, we just consider including the seventh lag, which would allow the seasonality to change. However, in this model, the term at lag seven is not significant, so we include no past observations in our model.

Next, we consider the number of levels in the weekly seasonal factor D_t. We test whether the factor D_t should have seven levels (one for each day of the week), three (week-day, Saturday and Sunday) or two (week-day and weekend). The tests show no significant improvement is given by having more than three levels for D_t (week-day, Saturday and Sunday).

Thus, the model defined by Equation 3.2 becomes:

$$\log \mu_t = \log T_t + \log A_t + D_t + \beta_0 \tag{3.3}$$

where D_t has three levels (week-day, Saturday and Sunday). The coefficients for the reduced model are shown in Table 3.1. Summary statistics of this model can

Model Eqn.	Intercept	Ti	D_{Sat}	D_{Sun}
3.3	-0.0195 (0.0100)	0.00010 (0.00002)	-0.1036 (0.0101)	-0.0602 (0.0099)

Table 3.1: Coefficients for the quasi-Poisson model defined by Equation 3.3, fit to the national counts of calls to NHS24 relating to vomiting. The brackets give the standard errors.

Model Equation	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	Disp Parameter
3.3	1266	1063	203	16	721	718	3	–	1.48

Table 3.2: Summary statistics from the quasi-Poisson rate model defined in Equation 3.3 fit to the national counts of calls to NHS24 relating to vomiting. No AIC statistic can be calculated because the quasi-Poisson is not a proper probability distribution and so no likelihood can be calculated, a required part of the AIC calculation. DoF = Degrees of Freedom.

be found in Table 3.2.

The factor levels for the weekend are both negative, suggesting that the proportion of calls relating to vomiting are lower at the weekend. The reduction on Saturday being around twice the size of the reduction on Sunday matches the pattern that was seen in Section 2.6.3. There is small positive increasing trend over time.

We now turn to considering the residuals. In plots of the residuals, larger residuals occur more frequently than would be expected. Further, there appear to be large residuals at the start and end of each year, along with large residuals during the March and April period. As before, the latter large values correspond to those times in the year when the annual profile, A_t, does not fit well to the annual cycle. A qq-plot of the residuals is reasonably close to a straight line, with a little curvature in the lower half of the distribution but nothing that overly concerns us. The most notable problems are seen when we consider the acf of the deviance residuals. There are significant correlations at lags one, five,

six, seven, eight and ten, and other high order correlations. There are a run of positive correlations, followed by a run of negative correlations, suggesting that trend remains in the residuals. The significant correlations around lag seven suggest there is some seasonality remaining in the residuals. Including the lagged observation x_{t-7} (at time t, data from a week ago) does little to address this seasonality. Including x_{t-1} (at time t, yesterday's observation) does very little to improve the significant correlation at lag one. In summary, there is still significant serial correlation unaddressed in the model.

Given the evidence of poor fit and serial correlation in the residuals, we proceed by fitting a different model. These problems are likely caused in part by using unstable data from the start of the system. The next models we fit include seasonal Holt-Winters components to incorporate a level of local trend. We hope this will deal with the serial correlation and weekly seasonality. Further, we also fit the new model with a negative binomial response. The negative binomial distribution is often used for modelling counts, as it allows for modelling counts with variation different from their means. The negative binomial is a proper probability distribution, as opposed to the quasi-Poisson 'distribution', and so allows more straight forward prediction. We cover some background on these new modelling elements before fitting models that utilise them.

3.2 Modelling Considerations

In this Section we consider the models that are in use at NHS Direct and how they compare with the models we have developed thus far. We propose a different modelling direction using a variable based on Holt-Winters smoothing to address trend. We also consider the negative binomial distribution, which we fit to the counts instead of using the Poisson.

3.2.1 NHS Direct Syndromic Surveillance

Details of the syndromic surveillance used at NHS Direct is taken from Cooper (2007) and Cooper, Smith, Baker, Chinemana, Verlander, Gerard, Hollyoak, and Griffiths (2004). Two systems are used by the Health Protection Agency for monitoring daily call data from 22 NHS Direct sites.

The first system works by calculating upper (99.5%) confidence limits of calls for each syndrome for each site as a proportion of total daily calls for that site. The proportions have a monthly seasonal adjustment to take account of annual seasonal changes in the prevalences of each syndrome (since, for example, we will expect a higher proportion of calls relating to colds and flu during the Winter). An exceedance was noted if any syndrome had a higher proportion of total daily calls than its upper confidence limit.

For syndromes with a higher calling rate, such as cold/flu, cough, fever, difficulty breathing and vomiting, control charts were produced for those NHS direct sites monitoring dense population centres in England (such as London). The counts of calls were modelled by a Poisson GLM. The GLM included a number of terms: the total number of calls received as an offset; an annual seasonal term; linear trend term; public holiday term; day-of-the-week effect if required. As we have seen with modelling the national counts of calls relating to vomiting, the counts can be over dispersed, so scaling is used to address this. It was found that the best way to calculate the upper control limit of the chart was to transform

> "*to approximate normality with zero mean ... and then a back-transformation to the original scale. The resulting expression for the 99.5% upper limit of syndromic calls, used for the control charts, was:*
>
> $$\left(\sinh \frac{z_\alpha/2 + \sqrt{N-0.5}\sinh^{-1}\sqrt{p}}{\sqrt{N-0.5}}\right)^2 (N-0.75) - 3/8$$
>
> *where N is the expected value divided by one less than the scale parameter, p is equal to the scale parameter minus 1, and z_α is the $100 * (1-\alpha)$th centile of the normal distribution. Ad hoc choices of z were used to achieve the desired proportion of purely random exceedances (0.5%).*" (Cooper (2007), p.341-342)

When any calls exceed this upper control limit an exceedance is raised.

When exceedances are raised from either of these systems, a statistician will consult with a medical expert to determine if the exceedance warrants further investigation.

3.2.2 Divergences from the NHS Direct Models

The model we develop in Section 3.1 is similar to the model used in the control chart monitoring methods of the NHS Direct system. The model fit to the NHS Direct data has a Poisson response, with the total number of calls used as an offset (Cooper, Smith, Baker, Chinemana, Verlander, Gerard, Hollyoak, and Griffiths 2004). To deal with annual seasonality they appear to use a monthly factor. This contrasts with our approach of using an annual profile to capture the annual pattern. As we do, a day of the week effect is fitted, but they use different levels for Saturdays and Sundays.

The use of a proportional model, via the use of an offset, was utilised to address 'the gradual year-on-year increase of calls and sudden and local increases in call rates due to local publicity' (Cooper 2007). Thus, their choice is driven by the same reasoning as ours, in looking for stability in the counts. However, for the NHS24 data that we have, this does not appear sufficient to stabilise the counts. We go a step further by investigating the use of past terms to deal with serial correlation. One would hope that this would deal with local trend but it appears to be insufficient. We try using a different type of model that includes a variable that is a prediction from a Holt-Winters model fit to the count data. Holt-Winters is generally regarded as very robust at capturing local trend and so will hopefully deal with the early and non-stable data more effectively than using an offset of total calls. In Section 2.6 we noted the problem of coupling that occurs when modelling proportions; if a reasonably suitable model can be found for the absolute counts, it would suffer none of these difficulties. Holt-Winters modelling is considered further in Section 3.2.3.

In the GLMs we will fit, we will also use a negative binomial response, in contrast to using a Poisson response followed by scaling to address over-dispersion. The negative binomial distribution can directly model over-dispersed counts. Thus, exceedances can be calculated directly and simply, without the need for transformations. We consider the negative binomial distribution further in Section 3.2.4.

3.2.3 Holt-Winters Modelling

Holt-Winters models are a generalisation of simple exponential smoothing. Simple exponential smoothing forms predictions by taking a weighted average of past observations, where the weights decrease exponentially the further they are into the past. This seems intuitively sensible – we expect future observations to be most similar to the recent past and less so to observations separated by greater periods of time. Holt-Winters is a generalisation of simple exponential smoothing that allows for local trend and seasonal factors. For more detail on exponential smoothing, see Appendix B.1. Here we detail the function `HoltWinters` that we will use to fit the Holt-Winters models in `R` and then consider how this can be combined with a GLM to form predictions for use in an exception reporting system.

`HoltWinters` Function

Details here are drawn from `R`'s help page for `HoltWinters`. The `HoltWinters` function allows for both additive and multiplicative Holt-Winters models to be fitted to a time series. We will be fitting models to the log-counts, so the seasonality will become additive. Therefore, we only consider the additive version of the function here. Thus, for a series with additive seasonality of period p, the h-step prediction function for the Holt-Winters model is given by:

$$\hat{Y}[t+h] = a[t] + hb[t] + s[t+1+(h-1) \bmod p],$$

where $a[t]$, $b[t]$ and $s[t]$ are given by:

$$\begin{aligned} a[t] &= \alpha(Y[t] - s[t-p]) + (1-\alpha)(a[t-1] + b[t-1]) \\ b[t] &= \beta(a[t] - a[t-1]) + (1-\beta)b[t-1] \\ s[t] &= \gamma(Y[t] - a[t]) + (1-\gamma)s[t-p] \end{aligned}$$

Unless manually specified, `HoltWinters` calculates α, β and γ by minimising the squared one-step ahead prediction error. In the modelling of counts of calls relating to vomiting, $p = 7$ since we want the Holt-Winters models to capture weekly seasonality. We leave α, β and γ to be determined by the function.

Combining Holt-Winters models with GLMs

In fitting a GLM to the NHS24 counts, we are trying to form a relationship between various covariates and the mean of the counts. Gross and Craig (1974) explore using different forecasting procedures to predict the Poisson means in inventory demand modelling. We propose to combine GLMs and an exponential forecasting procedure to reduce serial correlation and capture local trend when predicting the mean number of calls expected. While in a different context, Gross and Craig (1974) advise the use of exponential smoothing for data whose structure might change, due to exponential smoothing's robustness. From a Holt-Winters model, predictions for future observations can be found. These predictions are then included as a variable in a GLM. This means that both model fitting and predictions for any exception reporting system would have two stages: find the appropriate value from the Holt-Winters model and then include this in the GLM. With modern computing power this extra overhead of calculation is of little concern. Within `R`, predictions can be made from Holt-Winters models by use of the `predict.holtwinters` function. Predicting from this hybrid GLM/Holt-Winters structure is straight forward, since the Holt-Winters forecasts are known at each time point and so the GLM includes the Holt-Winters forecasts as a standard covariate.

3.2.4 Negative Binomial Distribution

In Section 3.1.2 we discuss the phenomenon of over-dispersion which occurs when dealing with counts that have variance markedly different from their mean. In such cases the Poisson distribution is not a suitable fit. One approach to addressing over-dispersion in these models we have seen: one can use a quasi-Poisson model for such data to correct estimates of standard error. However, a quasi-Poisson distribution is not a proper probability distribution. This makes forming predictions from these models much more cumbersome. Another approach to dealing with over-dispersion is to directly model the over-dispersion through a negative binomial distribution. The negative binomial distribution is a proper probability distribution parameterised by two parameters, and so can fit to count data where the mean and variance are different.

Fitting negative binomial models within R can be done easily using the function glm.nb, available from the MASS library (Venables and Ripley 1997). This function uses the following parameterisation of the negative binomial:

$$f_Y(y|\mu,\theta) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)y}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{\theta+y}}, \tag{3.4}$$

where μ and θ are parameters of the distribution, such that:

$$E(Y) = \mu,\ \text{var}(Y) = \mu + \frac{\mu^2}{\theta}.$$

The θ parameter is sometimes known as the 'shape' parameter (see Venables and Ripley (1997)). Estimates for the parameters μ and θ are found by maximum likelihood methods. An estimate for θ is found by holding μ constant. Then, θ is held constant, while μ is estimated. The function glm.nb iteratively alternates between these estimates until convergence.

For further information on the negative binomial distribution, either for fitting models or more generally, see: Hilbe (2007), Venables and Ripley (1997), Lawless (1987), Simon (1962).

3.3 National GLM with Holt-Winters Variable

We begin using these new model elements in fitting another model to the national counts of calls relating to vomiting. We first fit a Holt-Winters model. From this model we form a variable of predictions which is then used in a national negative binomial GLM fit to the same counts.

3.3.1 National Holt-Winters Model

We use the HoltWinters function in R to fit a Holt-Winters model to the log of national calls for 2004 and 2005. Thus, predictions are made on the log-scale. This is a sensible choice, since the GLM we fit in Section 3.3.2 will model the log of the mean as a linear function of the covariates.

The coefficients of this model are shown in Table 3.3. The parameters are found by minimising the one-step ahead prediction error. The value of α, 0.12,

Model	α	β	γ
National	0.11937	–	0.36654

Table 3.3: Coefficients from the Holt-Winters models fit to the log of counts of calls received nationally by NHS24 relating to vomiting during 2004-2005.

is in the range of values that are considered to be most common (0.1–0.3, see Appendix B). This parameter tells us about how the level of the series changes: the larger α is, the fewer past observations that contribute to the prediction of the level. Thus, a larger α suggests more variable data. The same observation applies to γ, in respect of how seasonality changes in the series: a larger γ suggests that the seasonality changes more. We find that the model fitted has no local trend ($\beta = 0$). Predictions from this model are used to form the $pred_{HW}$ variable used in the GLM we fit in the next Section.

3.3.2 Negative Binomial Model Fitting

Initial model

Using the Holt-Winters model, we form a series of predictions which are then incorporated into a negative binomial GLM. We fit this GLM to the count of calls received nationally during 2005. We start with:

$$\log(\mu) = \begin{array}{l} \beta_0 + \beta_1 pred_{HW} + \beta_2 \log(vc_{t-364}) + \beta_3 Ti + Day + Bank_4 \\ +Bank_5 + EMon_4 + EMon_5 \end{array}, \qquad (3.5)$$

where the β_k are coefficients, $pred_{hw}$ is a one-step ahead forecast found by Holt-Winters seasonal exponential smoothing of the logged counts, $\log(vc_{t-364})$ corresponds to the count of calls 364 days previously ($364 = 7 \times 52$, so gives the same day of the week), Ti is the day number from the start of 2005 giving a linear dependence on time, Day is a factor with a different level for each day of the week, $Bank_4$ and $Bank_5$ are factors having two levels reflecting whether the day in 2004 or 2005 was a bank holiday respectively, and $EMon_4$ & $EMon_5$ do similarly for Easter Monday in 2004 and 2005 respectively. We fit this model and then exclude those terms that do not contribute significantly to its fit.

Model	Intercept	$pred_{HW}$	$\log(vc_{t-364})$	Ti
National	1.3903 (0.2483)	0.5760 (0.0520)	0.1755 (0.0380)	-0.0012 (0.0002)

Model	$DayWE$	$Bank_5$	$EMon_5$
National	0.2009 (0.0461)	0.5723 (0.0498)	0.7127 (0.1266)

Table 3.4: Coefficients of the negative binomial model, as defined by Equation 3.6, fit to the national counts of calls received by NHS24 during 2005 relating to vomiting.

Final model

We adapt the initial model (defined by Equation 3.5) to the following:

$$\log(\mu) = \begin{array}{l}\beta_0 + \beta_1 pred_{HW} + \beta_2 \log(vc_{t-364}) + \beta_3 Ti + DayWE + Bank_5 \\ +EMon_5\end{array}, \quad (3.6)$$

where most of the terms are as above, but the Day factor has been replaced by the $DayWE$ factor, which has two levels for week-day and weekends. The coefficients for this model can be found in Table 3.4 and its summary statistics can be found in Table 3.5.

We first consider the coefficients of the national model. It is perhaps surprising that the coefficient of $pred_{HW}$ is not closer to one, given it is a prediction on a log-scale. Further, it is surprising that the $DayWE$ factor is needed, because the seasonality should be captured by the $pred_{HW}$ as this term is based on a seasonal Holt-Winters model. However, the seasonality may not be captured fully from the $pred_{HW}$ term in the GLM because of its coefficient being less than one. The factor levels are as we expect: at weekends, bank holidays and on Easter Monday more calls are received by NHS24 because of doctors' surgeries being closed (note the positive values). There is a small downward trend in the number of calls relating to vomiting received by NHS24 during 2005.

There are a few large residuals corresponding with bank holidays suggesting that the model might be improved by having separate factor levels for each bank holiday. The qq-plot has a little curvature at either end of the distribution, but not so much as to suggest an ill fitting model. A plot of residuals against the predicted log-means suggest the the variance is dealt with appropriately. The only diagnostic plot that really gives us cause for concern is the acf. We find

Model Eqn.	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	θ	θ SE
National	4475	368	4107	92	364	358	6	3162	81.321	10.043

Table 3.5: Summary statistics of the negative binomial model fit to the national counts of calls received by NHS24 during 2005 relating to vomiting, as defined by Equation 3.6. DoF = Degrees of Freedom. SE = Standard Error.

a significant correlation at lag one and a slightly larger correlation at lag three. There are just significant correlations at lags seven and eight. Some isolated higher order correlations are also significant and there are runs of positive and negative correlations. This suggests there is trend remaining in the residuals.

Conclusions

We have found a national negative binomial model for the counts of calls to NHS24, relating to vomiting, during 2005. Generally the model fit is reasonable but there are some problems with trend in the residuals. However, this is no worse than the problems found with the quasi-Poisson model fit in Section 3.1.3. However, we are modelling counts and thus vomiting is independent of the other syndromes: recall the comments in Section 2.6 about modelling the proportions of total counts relating to each syndrome – this *couples* the syndromes together. It is also interesting to note that 92% of the deviance is explained with this model (Table 3.5) compared to 16% with the rate model (Table 3.2). That there is an improvement is no surprise: the negative binomial has the extra parameter θ to address over-dispersion (Hilbe 2007). However, the great improvement in the deviance explained shows how effective modelling with the negative binomial can be. With the different health boards adopting NHS24 at very different paces, it will be hard to find a national model that does not exhibit some problem with trend. We turn to adapting this model to the regional health boards and see if it fares better at the regional levels.

3.4 Regional Models

In this Section we adapt the national model to the regional health board counts. We start by fitting Holt-Winters models to all the regions and comparing them. We then fit negative binomial GLMs to the counts, again including a variable in them based on the Holt-Winters models. Then, we compare these different regional GLMs. The island health boards do not have enough variation in them to be modelled by the negative binomial distribution so we fit binomial models in these regions. Finally, we consider a crude 'network' created by including neighbouring Holt-Winter's predictions in a board's model.

3.4.1 Fits of the Holt-Winters Smoothing

Ideally, we would fit the Holt-Winters' models to all the data we have, particularly since we only have two years of data. However, in a number of the health boards, the number of calls received does not reach a stable level until late in 2004 – see Figure 3.1. In some of the health boards there are relatively large step changes. Due to this, we fit the Holt-Winters models to data from day 300 (October 26$^{\text{th}}$) onwards. Holt-Winters models would of course adapt to these step changes but these would be artificial ones brought about by the system starting rather than natural changes in the level of calls received. Thus, it is not sensible to 'train' the models to these initial step changes. We fit the models using the `HoltWinters` function within `R` on the 'started log counts' ($\log(\text{Count} + 1)$ (Brillman, Burr, Forslund, Joyce, Picard, and Umland 2005)). This change is required as some health boards have no calls on certain days.

The resulting models that are fit are shown in Table 3.6. None of the models utilise local trend but nearly all models utilise exponential smoothing with weekly seasonality. The only exception is Shetland which has a zero for the value of α; since it also has no local trend factor, this means that its model has the same level for the whole series, with a series of seasonal factors that update as time goes on. To understand why, consider the plot of the counts of calls from Shetland in Figure 3.2 and the distribution of its counts in Table 3.7. The number of calls from Shetland is very low, with approximately ninety-five percent of days with either no calls or only one call. For the level of the series, `HoltWinters` calculates 0.2263, which is close to the mean of the counts in the fitted period of 0.2157. The

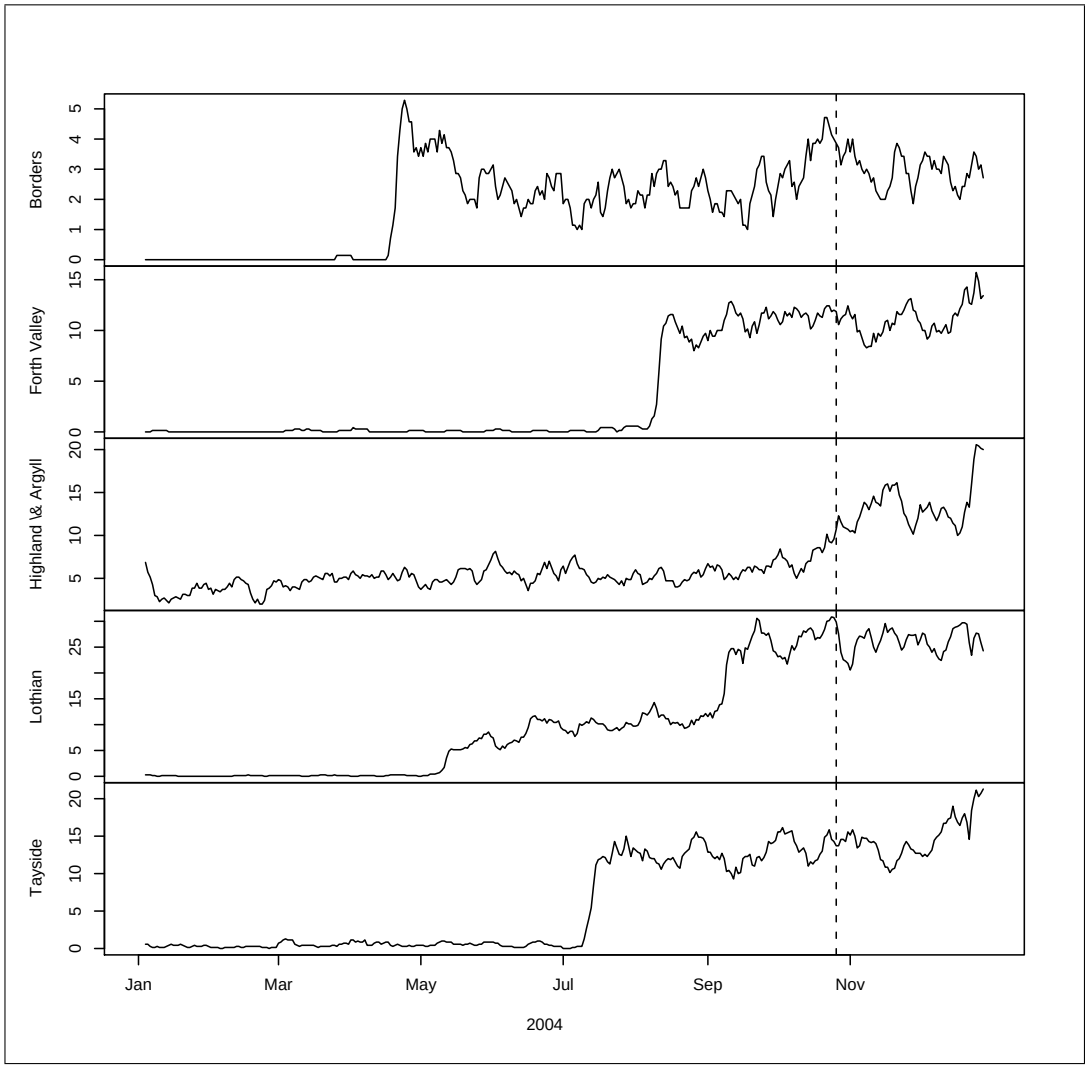


Figure 3.1: The number of calls relating to vomiting during 2004 from health boards where NHS24 utilisation seems to have taken off later. A moving average of order seven has been applied to the data to remove the effect of weekly seasonality. Due to these systems taking a long time to reach a stable level, we start fitting the Holt-Winters models to counts from day 300 onwards (shown by the dotted line, October 26^{th}).

Region	α	β	γ
Ayrshire & Arran	0.06228	–	0.19456
Borders	0.00807	–	0.16104
Dumfries & Galloway	0.01896	–	0.15425
Fife	0.05107	–	0.13255
Forth Valley	0.06020	–	0.18941
Grampian	0.07356	–	0.16407
Greater Glasgow	0.09316	–	0.20466
Highland & Argyll	0.06926	–	0.14596
Lanarkshire	0.09103	–	0.12836
Lothian	0.03522	–	0.26436
Orkney	0.01535	–	0.00164
Shetland	–	–	0.10312
Tayside	0.08538	–	0.18673
Western Isles	0.01646	–	0.11219

Table 3.6: Coefficients from the Holt-Winters models fit to the started-logs ($\log(\text{Count}+1)$) of counts of calls received from each health board relating to vomiting.

Shetland Counts	**Count**						
	0	1	2	3	4		Total
Frequency	314	94	19	4	1		432
%	73	22	4	1	0		100

Table 3.7: The distribution of counts for the number of people that call from Shetland complaining of vomiting, during the period in which the Holt-Winters' models are fitted.

closeness of these two figures is not surprising when we consider how this level is calculated: since α is zero, the initial value for the level determines the level for the series. Initial values for the `HoltWinters` function are found by seasonal decomposition; the initial values for the level are calculated as the mean of the deseasonalised values – for further details see Section B.2. Figure 3.2 suggests there is no change in level for the series and so a constant level seems reasonable.

In the other health boards we generally find that those health boards with the larger populations have larger values of α, the parameter that controls the smoothing of level. The smallest values are found in the southern health boards

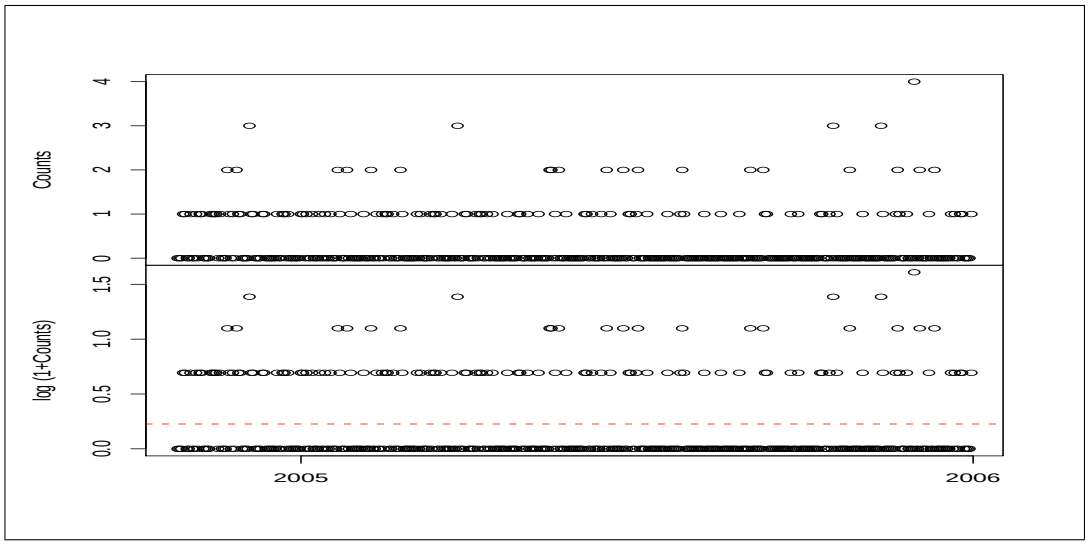


Figure 3.2: The top plot shows the counts of calls from Shetland that relate to vomiting. The bottom plot shows the $log(\text{Count} + 1)$ of calls, the values that the Holt-Winters models are fit to in Section 3.4.1. The time period chosen reflects the values that the `HoltWinters` function fit the Holt-Winters model to; the red dashed line is the level fit by the function.

(Borders and Dumfries & Galloway) and the islands (Orkney and Western isles). Due to the smaller value of α in these places, a greater number of past values contribute to the prediction made at time t. This means that the Holt-Winters models in these regions will adapt more slowly to changes in call rates. However, this is consistent for these regions: the southern and island health boards are sparsely populated meaning that the call rate is unlikely to change quickly. It is perhaps surprising that the value of α for Lothian is not larger: as the health board that contains Edinburgh, the capital of Scotland, it is relatively densely populated, and so we would expect the number of calls to be tied more closely to recent calling rates. We might expect it to be nearer to the value we find for Greater Glasgow, a similarly densely populated health board, with an α of 0.09.

The seasonality smoothing parameter, γ, is generally of a similar magnitude among all the health boards, varying from 0.10 to 0.26. The parameter γ controls how much the seasonal factors are updated: the larger it is, the more the factors are updated – see Section B.1.3 for more detail. We noted above that α was perhaps smaller than we might expect for Lothian; this might be explained by

Orkney Counts	Count				Total
	0	1	2		
Frequency	400	31	1		432
%	93	7	0		100

Table 3.8: The distribution of counts for the number of people that call from Orkney complaining of vomiting, during the period in which the Holt-Winters' models are fitted.

Lothian having the largest value of γ. Some of Lothian's change in level might be captured by the seasonal part of its model. The very small value for Orkney is reasonable when we consider the distribution of its counts, shown in Table 3.8. The counts for Orkney are almost binomial: either there is one call or none; there is only one day on which there was more than one call. Most of the time the series is zero, so there is very little seasonality to adapt to, leading to the very small value for its γ.

Having fitted the Holt-Winters models to each of the health boards, we can now use them to form one-step ahead predictions for each day of 2005. These values are then used as a variable in the negative binomial models that we fit to each health board in the next Section.

3.4.2 Negative Binomial Models

In this Section we fit negative binomial models to the counts of calls from each health board relating to vomiting. The island health boards – Orkney, Shetland and Western Isles – are not suitably distributed for a negative binomial distribution and so we fit a binomial distribution to them in Section 3.4.3.

Fitting Regional Models

For each region we check if the variables that were used in the national model contribute significantly to a regional model. We modify the $\log(vc_{t-364})$ term, the log of calls received 364 days previously, to $\log(vc_{t-364}+1)$ since a number of the boards experience zero counts. Cities and towns in Scotland can also choose to have a number of local holidays (Glasgow Chamber of Commerce 2009). As with national bank holidays, doctors' surgeries tend to be closed on these days and

so we expect more calls to NHS24 during them. The dates of these holidays are determined locally and so vary across Scotland. Thus, we include the $LHol$ factor, which has one level for all local holidays in that region and is zero otherwise. Local holidays were determined by searching for confirmation about particularly large residuals and from general research. For the health boards containing the larger cities, such as Lothian and Greater Glasgow, it is easier to confirm their local holidays; however, this was not possible for all boards.

Having fitted the local covariates to a region, we turn to considering if $pred_{HW}$ predictor terms from neighbouring boards can be used to improve the model for that board. This creates a crude network between the different health boards. Testing for the significance of these predictor terms can be difficult since there is a good deal of correlation and duplication of information amongst them. To choose which, if any, of the neighbours should be included in the model for a board, we use the `step` function in `R` (Venables and Ripley 1997). This function allows for automating model selection using the Akaike Information Criterion (AIC), which for a model is defined as:

$$\text{AIC} = -2 \text{ maximum log likelihood} + 2p,$$

where p is the number of the model's parameters. The `step` function searches through the different possible models, selecting the one that reduces the AIC the most (Faraway 2006). We use the *scope* option of `step` to specify the models we want it to consider. We specify as the smallest model as the one found above, consisting solely of local terms. The largest model considered is one that includes the basic, local, model as well as all of the $pred_{HW}$ terms of its neighbours. Thus, the simplest model that the `step` function can return is the basic local one, the largest one is the basic one with the addition of all of the neighbouring $pred_{HW}$ terms, or a model somewhere between these two. We only consider the $pred_{HW}$ terms of neighbours that include their $pred_{HW}$ term within their own basic models. We use the $pred_{HW}$ term as a proxy for the counts of a region within the model of another region; if the $pred_{HW}$ term for a region does not contribute significantly within its own basic model, it is not likely to represent that region well.

Once the `step` function has returned a model, potentially including the $pred_{HW}$ terms of its neighbours, we consider the significance of its terms. If the inclusion neighbouring $pred_{HW}$ terms makes its own $pred_{HW}$ term insignificant, we choose the basic model as the final model for this region. In such a case we assume that there is a large duplication of information among the $pred_{HW}$ terms and that little additional information is gained through the inclusion of neighbouring $pred_{HW}$ terms. It would seem logical that a region's own $pred_{HW}$ term, in general, will be the best representation of the counts there and so go with that.

Fitted models

The coefficients of the resulting fitted models are shown in Tables 3.9 and 3.10. It is perhaps surprising that the coefficients of $pred_{HW}$ term are not closer to one. Since this variable is a prediction on the log-scale for each health board and so on the same scale as the mean. The coefficient of $pred_{HW}$ in Ayrshire & Arran, 0.50, is reasonably intuitive when coupled with the values of the other coefficients; the sum of this coefficient and the coefficients of $\log(vc_{t-364})$ and $pred_{LN}$ is nearly one. All of these three terms contain values of counts on the log-scale and so their coefficients nearly summing to one gives a weighted average of these three terms. The effects of other terms in the models is likely to explain why the coefficients of the $pred_{HW}$ terms are not closer to one. The $pred_{HW}$ term does not contribute significantly in four health boards: Borders, Dumfries & Galloway and Tayside. The Borders and Dumfries & Galloway health boards have among the smallest values of α in the underlying Holt-Winters models (see Table 3.6). This suggests that their level changes very slowly and that these changes in these boards are better captured over a longer period from the other variables in these models. It is not entirely clear why the term is not significant in Tayside, but it may be because the day-of-the-week effect explains the changes in the mean more.

In those places where the $DayWE$ factor is significant, the levels taken at the weekend are always positive. This is logical: at the weekend doctors' surgeries are closed and so more people will call NHS24. With more people calling NHS24 we will get more calls relating to vomiting. It is perhaps surprising that we need this term in those health boards with the $pred_{HW}$ term, since their underlying Holt-Winters models have a weekly (day-of-the-week) seasonal factor in them. However, the need for $DayWE$ factor might be due to the value of the coeffi-

cient of the $pred_{HW}$ term – since the coefficients of the $pred_{HW}$ terms are not one, the seasonality captured by the underlying Holt-Winters models will not be transferred fully to the GLMs, resulting in the need for this factor. However, we also noted in Chapter 2 that the weekly seasonality of the calls related to vomiting did not appear to be stable. The seasonality captured by this factor might be thought of as capturing the stable element of the weekly seasonality in the counts (it will not change), while the Holt-Winters models capture the changing elements of the seasonality. As we would expect, the largest values of this factor are found in those boards that have no Holt-Winters predictors in them: Borders and Dumfries & Galloway. Ayrshire & Arran probably does not require this factor because the combination of $pred_{HW}$, $\log(vc_{t-364}+1)$ and $pred_{LN}$ terms, with no other terms in the model, are sufficient to capture seasonality.

Other day effects are captured by the: bank holiday factors $Bank_5$ (one level for all the bank holidays in 2005); the Easter Monday factors $EMon_5$ (1 level for Easter Monday during 2005); and the local holiday factor $LHol$ (one level for all the local holidays we have designated in a region). Again, as we would expect, all these factor levels are positive, since more people call NHS24 on these holidays as doctors' surgeries are closed. From calculating the Kendall correlation coefficient between the factor levels of $DayWE$ and each of these different holiday factors, we find there is generally a weak positive correlation between them (Kendall correlation coefficient value of 0.78, 0.50 and 0.81 for $Bank_5$, $EMon_5$ and $LHol$ respectively): the more seasonality that is captured in the GLM by the $DayWE$ factor, the larger the holiday factor levels. Generally, the local holidays and Easter Monday seem to have a larger effect on the number of calls that NHS24 receives.

Few health boards have the $\log(vc_{t-364}+1)$ term in their models. For a lot of boards this will be because of the the calls not reaching equilibrium till quite late in 2004 – see Figure 3.1. We might find with more years of stable data that this term would be significant in more boards. Further, we might expect the $Bank_4$ and $EMon_4$ terms to contribute significantly in more boards. However, since only Greater Glasgow has one of these factors, we do not consider it further here.

The coefficients of the linear trend Ti are all negative, suggesting that the number of calls being diagnosed by clinical algorithm that relate to vomiting are decreasing. This is in agreement with the national model in Section 3.3. This is

Greater Glasgow (GG) – Neighbours: AA, FV, HG, LN, TY; Ignored neighbours : TY; AIC selected: AA, LN;									
$\log(GG) =$	0.7911	$+0.5968P_{HW}$	$+0.1869D$	$+0.4098B_5$	$+0.5187EM_5$	$+0.8855LHol$	$+0.1858pt_t$	$+0.2743B_4$	$-0.0010Ti$
	0.2256	0.0676	0.0557	0.0961	0.1789	0.1354	0.0424	0.1127	0.0002
Forth Valley (FV) – Neighbours: FF, GG, HG, LN, LO, TY; Ignored neighbours : TY; AIC selected: GG;									
$\log(FV) =$	1.5008	$+0.2834P_{HW}$	$+0.6430D$	$+0.9473B_5$	$+1.2431EM_5$	$+1.1074LHol$	$-0.0011Ti$		
	0.2398	0.1032	0.0857	0.1034	0.2448	0.2811	0.0002		
Fife (FF) – Neighbours: FV, LO, TY; Ignored neighbours : TY; AIC selected: LO;									
$\log(FF) =$	0.1216	$+0.5503P_{HW}$	$+0.2014D$	$+0.4710B_5$	$+0.9342EM_5$	$+0.2530P_{LO}$			
	0.2870	0.1228	0.0867	0.1264	0.2909	0.1548			
Tayside (TY) – Neighbours: FF, FV, GG, GR, HG; AIC selected: FF, FV, GG;									
$\log(TY) =$	0.2382	$+0.2424D$	$+0.5687B_5$	$+0.9601EM_5$	$+0.9371LHol$	$-0.0012Ti$	$+0.3424P_{FF}$	$+0.2347P_{FV}$	$+0.2391P_{GG}$
	0.4045	0.1084	0.1249	0.2855	0.3421	0.0003	0.1375	0.1433	0.1425
Highland & Argyll (HG) – Neighbours: FV, GG, OR, SH, TY, WI; Ignored neighbours : OR, SH, TY, WI; AIC selected: GG;									
$\log(HG) =$	1.4969	$+0.3521P_{HW}$	$+0.5914D$	$+0.8256B_5$	$+0.8015EM_5$	$-0.0013Ti$			
	0.3022	0.1187	0.1019	0.1154	0.2996	0.0003			
Grampian (GR) – Neighbours: HG, TY; Ignored neighbours : TY; AIC selected: HG;									
$\log(GR) =$	0.2141	$+0.7074P_{HW}$	$+0.1176D$	$+0.4360B_5$	$+0.1975P_{HG}$				
	0.1325	0.0992	0.0566	0.1063	0.1074				

Table 3.9: The negative binomial models fit to the regional counts of calls to NHS24 from the northerly health boards that relate to vomiting, fit in Section 3.4.2. The numbers below the coefficients are their standard errors. The 'ignored neighbours' are those health boards whose $pred_{HW}$ terms are not in their 'basic' models and so not tested for inclusion as neighbouring values. The 'AIC selected' are those health boards that were selected by the `step` function for inclusion to extend the basic model; however, they are only included if they do not make the $pred_{HW}$ for a region insignificant. Abbreviations: $p_{HW} = pred_{HW}$; p_{AA} is the $pred_{HW}$ term for Ayrshire & Arran; $D = DayWE$; $B_i = Bank_i$; $EM_i = EMon_i$; $pt_t = log(vc_{t-364} + 1)$.

Dumfries & Galloway (DG) – Neighbours: AA, BR, LN; Ignored neighbours : BR; AIC selected: –;										
$\log(DG) =$	1.2954	$+1.0505D$	$+1.0532B_5$	$+1.6555EM_5$	$+1.6188LHol$	$-0.0028Ti$				
	0.0611	0.0592	0.1495	0.2987	0.3364	0.0003				
Borders (BR) – Neighbours: DG, LN, LO; Ignored neighbours : DG; AIC selected: LN;										
$\log(BR) =$	0.0282	$+0.8290D$	$+0.9975B_5$	$+1.3491LHol$	$-0.0008Ti$	$+0.2335P_{LN}$				
	0.4616	0.1505	0.1810	0.4308	0.0005	0.1568				
Ayrshire & Arran (AA) – Neighbours: DG, GG, LN; Ignored neighbours : DG; AIC selected: LN;										
$\log(AA) =$	-0.1476	$+0.4955P_{HW}$	$+0.1616pt_t$	$+0.3306P_{LN}$						
	0.1219	0.1242	0.0487	0.1118						
Lanarkshire (LN) – Neighbours: AA, BR, DG, FV, GG, LO; Ignored neighbours : BR, DG; AIC selected: GG, LO;										
$\log(LN) =$	1.0490	$+0.6114P_{HW}$	$+0.3697D$	$+0.6560B_5$	$+0.8329EM_5$	$+0.8941LHol$	$+0.0521pt_t$	$-0.0012Ti$		
	0.2258	0.0763	0.0729	0.0994	0.2422	0.2644	0.0218	0.0003		
Lothian (LO) – Neighbours: BR, FF, FV, LN; Ignored neighbours : BR; AIC selected: FF, FV, LN;										
$\log(LO) =$	1.5430	$+0.3571P_{HW}$	$+0.4229D$	$+0.4602B_5$	$+1.0216EM_5$	$+1.1430LHol$	$-0.0006Ti$	$+0.2486P_{FF}$	$+0.1484P_{FV}$	$-0.2210P_{LN}$
	0.2908	0.1171	0.0814	0.0843	0.1835	0.1087	0.0002	0.0934	0.0956	0.0838

Table 3.10: The negative binomial models fit to the regional counts of calls to NHS24 from the southerly health boards that relate to vomiting, fit in Section 3.4.2. The numbers below the coefficients are their standard errors. The 'ignored neighbours' are those health boards whose $pred_{HW}$ terms are not in their 'basic' models and so not tested for inclusion as neighbouring values. The 'AIC selected' are those health boards that were selected by the `step` function for inclusion to extend the basic model; however, they are only included if they do not make the $pred_{HW}$ for a region insignificant. Abbreviations: $p_{HW} = pred_{HW}$; p_{AA} is the $pred_{HW}$ term for Ayrshire & Arran; $D = DayWE$; $B_i = Bank_i$; $EM_i = EMon_i$; $pt_t = log(vc_{t-364} + 1)$.

likely to be an artifact of fewer calls being diagnosed by the clinical algorithm and more calls being diagnosed directly by the nurse-practitioners that staff NHS24. Calls in the latter category are not included in our counts. The order of magnitude is similar for most boards but is larger in Dumfries & Galloway because its model contains no $pred_{HW}$ (neither its own or those of its neighbours). However, it should be noted that there is no local linear trend fit in any of the Holt-Winters models - see Section 3.4.1. The Ti trend for Borders only becomes insignificant upon the inclusion of the $pred_{LN}$ term in its model.

We have modelled the relationships between health boards by including in a board's model the Holt-Winters predictors of its neighbours, where they contribute significantly to the model. However, there are a large number of neighbour pairs, so we represent these pairings in Figure 3.3. Take two health boards and label them A and B. If A's Holt-Winters' predictor is included in the model for B, then we put an arrow in Figure 3.3, starting at health board A and ending at health board B. Thus, the Holt-Winters predictor for Lanarkshire $pred_{LN}$ is included in the model for Lanarkshire. We include the relations included in the final models (black) and those relations included by the `step` function but not included in the final models (grey). Recall from earlier that we do not include the neighbouring $pred_{HW}$ terms if they make the $pred_{HW}$ for that region insignificant; see the section on model fitting for a more detailed explanation. Red links indicate boards that are neighbours but were shown to have no links what so ever.

Since the $pred_{HW}$ terms for Borders, Dumfries & Galloway and Tayside were not significant in their basic models, they were not tested for inclusion in other models. However, these boards can still have the $pred_{HW}$ terms of their neighbours in their models. For instance, three boards feed into Tayside and one into Borders. However, there are no links with Dumfries & Galloway. It is worth noting that since Borders and Dumfries & Galloway are on the border with north England, they are 'missing' some of their neighbours as we do not have the data from the health boards in the north of England. Most links seem be between the health boards in the middle of Scotland: Ayrshire & Arran, Borders, Fife, Forth Valley, Greater Glasgow, Lanarkshire and Lothian. This is not surprising since the majority of the Scottish population lives in these regions. Further, these regions also have the the highest densities of individuals and the most commuting

going on between them. These factors would mean that we would expect infection to be transferred most readily between them, making the $pred_{HW}$ terms of neighbours useful for improving the predictions in a region by being a proxy for this transfer. It is perhaps surprising that we do not see more regions with reciprocal links; that is, two neighbouring boards with their $pred_{HW}$ terms in each others' models. Many boards only have a one way relation ships with their neighbours. The two way relationships may indicate contemporaneous relationships, while the one-way may indicate a lagged relationship. However, again we are limited by the quality of the data: we have only fitted these negative binomial models to one year of data, so we should not over-interpret these relationships.

We now consider the relationships that are included in the final models (black arrows in Figure 3.3). Generally, the coefficients of the neighbouring $pred_{HW}$ terms are not significant. This is perhaps not surprising since we expect some correlation between the the neighbouring and local $pred_{HW}$ terms, which will affect tests of significance. Recall that they are chosen through a series of comparisons of AICs, and so we we leave them in. The coefficients of the neighbouring $pred_{HW}$ terms are smaller than the coefficients of the local $pred_{HW}$ terms. This seems sensible: we would expect the local $pred_{HW}$ to be the best 'guide' for the changes in that region, and so would expect it to have the largest value. However, it is important to remember the coefficients also serve to scale the $pred_{HW}$ terms: the $pred_{HW}$ terms give predictions of the log-counts, and so will be larger in those regions that normally get more calls. Thus, in the smaller regions, the $pred_{HW}$ terms of neighbours will have smaller coefficients to scale them down to the appropriate level for that region. As we would expect, barring one term, all of the $pred_{HW}$ terms have positive coefficients. The only exception is the $pred_{LN}$ term in the model for Lothian. It is not entirely clear why this is, but might indicate a negative correlation between the numbers of calls from Lanarkshire and Lothian.

Model Diagnostics and Summary Statistics

In this Section we consider the fits of the negative binomial models fit in the previous Section. Statistics related to the fit of the negative binomial models can be found in Table 3.11.

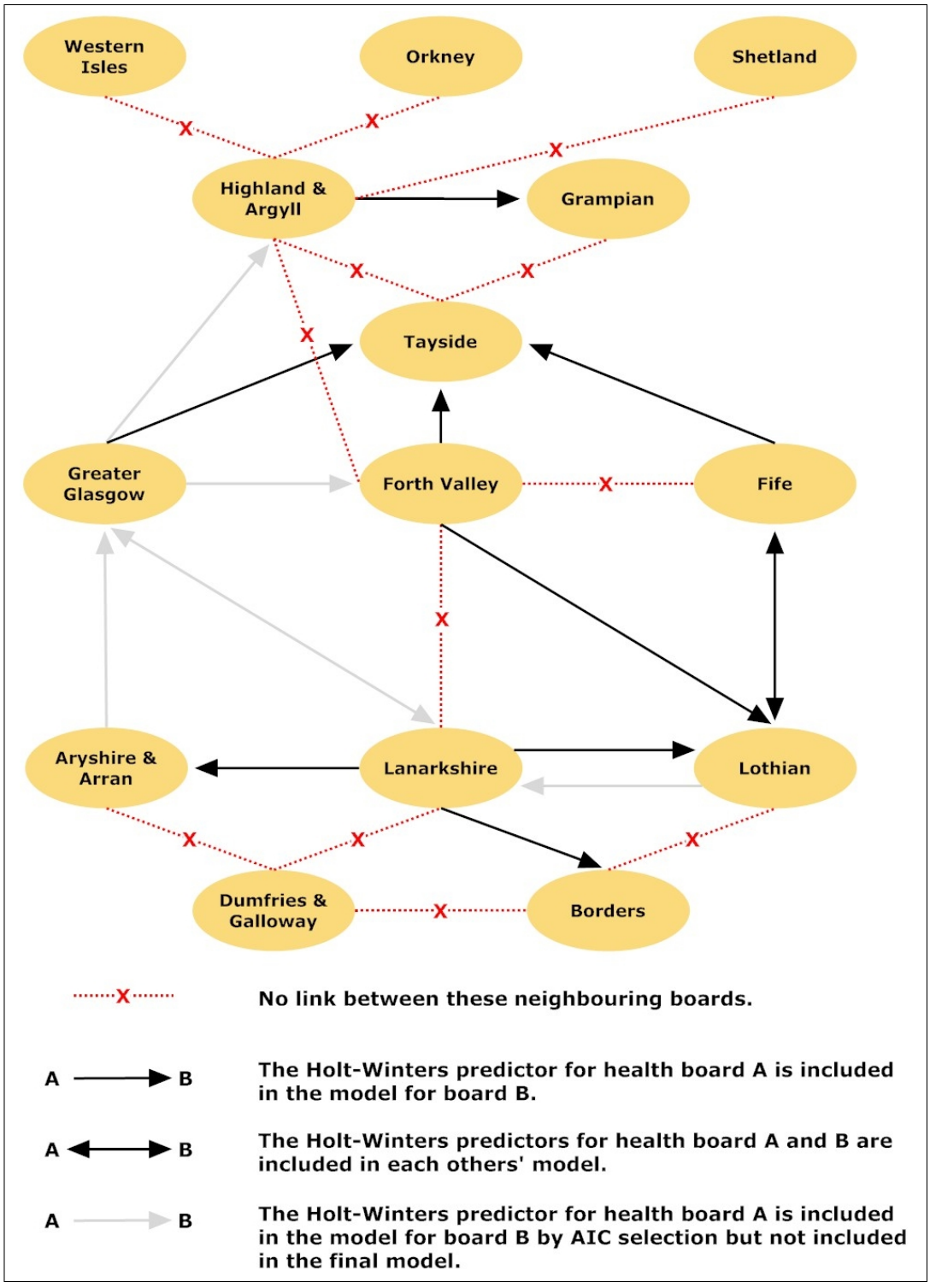


Figure 3.3: This representation of the neighbouring structure of Scotland's health boards shows which Holt-Winters' predictors from other health boards are included in other health board models. Thus, an arrow from from board A to board B, means that the Holt-Winters' predictor from board A is included in the negative binomial model for board B. Thus, Greater Glasgow's negative binomial model includes Ayrshire & Arran's Holt-Winters' predictor.

Region	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	θ	θ SE
Ayrshire & Arran	872	373	500	57	364	361	3	2070	15.546	2.819
Borders	642	390	252	39	364	359	5	1286	62.412	100.674
Dumfries & Galloway	874	405	468	54	364	359	5	1445	49.206	44.522
Fife	824	385	439	53	364	359	5	1940	25.049	7.153
Forth Valley	997	380	617	62	364	358	6	1892	44.369	18.468
Grampian	1222	383	839	69	364	360	4	2122	29.042	6.85
Greater Glasgow	2194	370	1824	83	364	356	8	2378	73.073	18.418
Highland & Argyll	1045	398	647	62	364	359	5	2009	23.141	5.77
Lanarkshire	1497	387	1110	74	364	357	7	2194	29.792	6.563
Lothian	1431	371	1060	74	364	355	9	2176	99.606	44.517
Tayside	1054	383	671	64	364	356	8	1971	22.101	5.505

Table 3.11: Summary statistics from the negative binomial GLMs fit in Section 3.4.2 to the counts of calls relating to vomiting received from each health board DoF = Degrees of Freedom. SE = Standard Error.

Generally, the residuals for each health boards are well behaved with some large values. There is a suggestion that the residuals are larger at the beginning and end of the year. There are some larger residuals on days that we have treated as bank holidays. To address this, future models with more data might consider fitting more than one factor level uniformly to all bank holidays. The residual deviances are of similar orders to the residual degrees of freedom, suggesting reasonable models.

Auto-correlation functions (acfs) of the residuals suggest, in general, that there is no significant serial correlation remaining in the residuals. There are some significant high order correlations but these are at lags that appear to have no links with the underlying system. However, Ayrshire & Arran has a significant correlation at lag 1, which will have most effect on predicting one day ahead. Thus, predictions for Ayrshire & Arran might need to be monitored more closely.

Q-Q plots of the residuals suggest few problems with the fits of these models. For some of the models there is a little curvature at either end of the q-q plots, but nothing that concerns us. Plots of the fitted log-means against the residuals suggest that variance is modelled suitably by the negative binomial distribution.

The value of θ, the shape parameter of the negative binomial, varies quite a lot between the health boards. However, this is not particularly surprising; there is quite a difference in the number of calls we expect to get from Borders (on average, 2.5 calls per day) and Greater Glasgow (30.2 calls per day). If a standard error for a θ is near or larger than its calculated value, it suggests that we either know little about θ or have a poor fitting model. Thus, we are only concerned about Borders and Dumfries & Galloway where this is the case. Besides the island health boards, these are the regions that the fewest calls are received from. Thus, these health boards will have more days with no calls (zero counts) and so might benefit from zero inflated negative binomial models, which we consider further in Section 5.6.1.

Conclusion of negative binomial fitting

In this Section we have fitted negative binomial models to the counts of calls relating to vomiting from twelve of the health boards in Scotland, using data from 2005. The Holt-Winter's variables seems to have captured local trend reasonably well. Various day effects need to be included to account for higher rates of calling when doctor surgeries' were closed. Generally, including calling rates from 364 days previously did not contribute significantly to the models. This is likely because a number of boards did not really start calling NHS24 till part way through 2004. From the diagnostics of these models, they generally seem to be reasonable fits to the data, not suffering the problems that national model had.

3.4.3 Binomial Models

Very few calls are received from the island health boards and so cannot be modelled by a negative binomial model. Instead, we fit a binomial model to them, turning the number of calls received into a binary response: one, if any calls are received, and zero otherwise. We then test for the inclusion of variables that were used in the negative binomial models.

Fitted Regional Binomial Models

The resulting fitted models are shown in Table 3.12. Few variables contribute significantly to the models. Apart from Shetland, there seems to be a weekend

Region	Intercept	$DayWE$	Ti
Orkney	-2.8663 (0.2748)	0.9094 (0.4041)	–
Shetland	-1.0877 (0.1206)	–	–
Western Isles	-2.7913 (1.7908)	0.0167 (0.5947)	-0.0042 (0.0020)

Table 3.12: Coefficients of the binomial models fit in Section 3.4.3 to the health boards where a negative binomial model and Poisson models could not be fit. The binary response is either no calls received, or at least one call received.

effect which we might expect: while it will be hard to get to a doctors' surgery from the island, they will experience even more difficulties during a weekend. The negative trend in the Western Isles suggests that fewer calls are coming from there relating to vomiting. This was also true for a number of the negative models. The Shetland health board simply has an intercept in it, suggesting that the proportion of calls from there is either constant or undergoing such a small change that this modelling does not have the power to detect it.

Model Diagnostics and Summary Statistics

Summary statistics for the these binomial models can be found in Table 3.13. Cross tabulation of days of the week and number calls received suggests that having the $DayWE$ factor is reasonable. From a plot of the number of calls from Western Isles, shown in Figure 3.4, a decreasing trend in the days with calls seems reasonable.

Binomial Modelling Conclusions

In this Section we have fitted three binomial models to the Island health boards of Orkney, Shetland and the Western Isles. Orkney and the Western Isles had a significant week-day/weekend effect in whether a call was received from them or not. The Western Isles showed some evidence for a downward trend over time in the probability of receiving calls on any day. The model for Shetland suggests, as its Holt-Winters model did, that the calling rate from Shetland does not change, since only a constant was fitted to this board.

To fit binomial models to these regions, we have turned the number of calls received into a binary variable: either some calls are received or none are. Due to

Region	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC
Orkney	193	188	5	3	364	363	1	192
Shetland	412	412	0	0	364	364	0	414
Western Isles	473	417	56	12	364	361	3	425

Table 3.13: Summary statistics from the binomial GLMs fit in Section 3.4.3 to the island health boards. DoF = Degrees of Freedom.

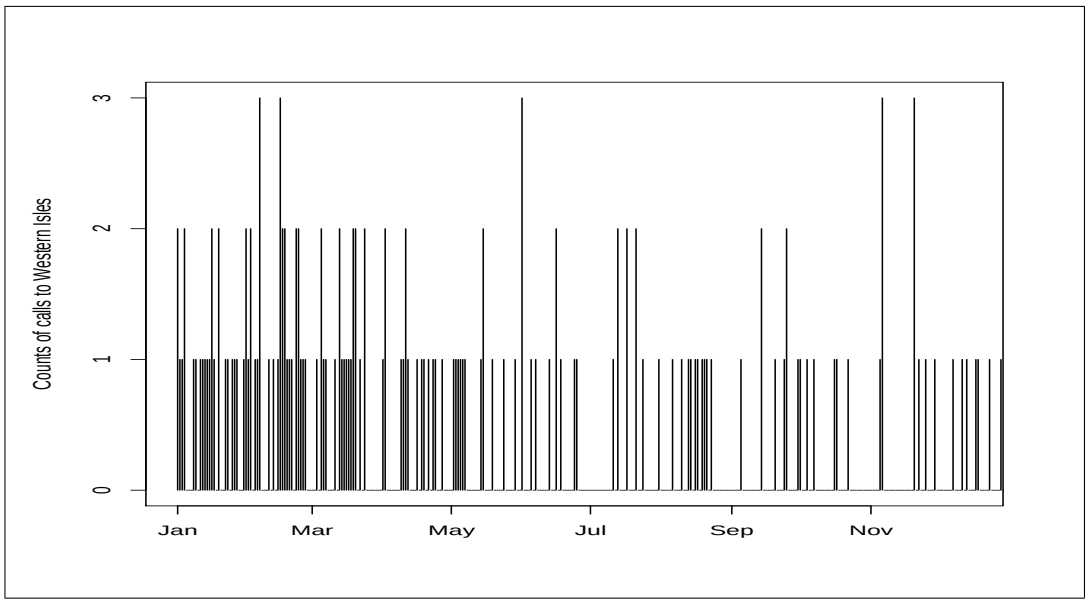


Figure 3.4: The number of calls from the Western Isles during 2005.

this, the exceedance alarm method could not be used with these models. Future work might look at developing an alarm method based on a sequential ratio test (Wald 1945). Such a test could check if the probability of receiving a call has increased, which could be interpreted as an exception. Alternatively, or in addition, it may be best just to define an exception in an island board as receiving more than, say, 2 calls (recall that for these regions it is very unusual to get more than one call in any day).

3.5 Conclusions & Future work

In this Chapter we have developed a number of models that serve as the basis for forming predictions for the number of calls relating to vomiting received by NHS24 both nationally across Scotland and regionally at the health board level. These predictions can then be combined with an alarm method to form a national and regional exception reporting system. Given the considerations of timeliness for detecting exceptions, a version of an exceedance alarm is probably the best choice. Unfortunately, with only two years of data from the very start of the system, there is not enough data to validate these models and investigate how they would work in an exception reporting system. An obvious avenue for future work is obtaining more data from NHS24 and fitting the models to a longer time period. With a sufficient amount of data, it might even be possible to split the data into training and validation/testing periods: we would fit models to the training period and then apply the resulting exception reporting systems to the latter period to see how the system fares. The system could also be extended to other syndromes. With more, hopefully stable, data, utilising past years' values may even be more consistent. Future work might look at different ways of capturing the level in previous years (a simple mean, a weighted average etc.). We might also investigate combining the regional models in some way to monitor the counts nationally.

An important difference between how this system could potentially work and how the exception reporting system at NHS Direct works is in how calls are assigned to a region. At NHS Direct, the data for a region corresponds to all those calls that are received by the call centre that serves a particular region. At times of peak demand, if a call from a region can not be answered by that

region's call centre, it is transferred to another region. Up to ten percent of calls can be transferred in this fashion (Cooper, Smith, Baker, Chinemana, Verlander, Gerard, Hollyoak, and Griffiths 2004). In contrast, the calls to NHS24 have been allocated to their health boards by post codes and so should give a more accurate representation of calls from each region.

Another difference is in what we have modelled. At NHS Direct, the proportion of all calls that relate to a particular syndrome have been modelled through the use of a offset of total calls in their Poisson models. We found that the use of the offset did not stabilise the models particularly well. Instead we model the counts directly using a variable based on a Holt-Winters smoother which captures serial correlation and trend well for most of the regions. Modelling the counts also means we remove the problem of coupling when dealing with proportions, which we discussed in Chapter 2. Future work might look at comparing an exception system based on the models developed here, with the models in place at NHS Direct and the models now in use at NHS24 (Cooper 2007; Kavanagh, Robertson, and McMenamim 2007). These comparisons would also allow for the investigation of the effect in using the negative binomial models as we do here, compared to transforming the Poisson models to take account of over-dispersion, as is done in the other systems.

The models developed here also make more explicit the links between neighbouring health boards in respect of the number of call received. This was done by using some of the Holt-Winters' predictors of neighbouring health boards in each health board's model. Some rudimentary work has shown the inclusion of the neighbouring boards improves the models in the sense of reducing the root-mean-square error of predictions (Robertson, Kavanagh, McKenzie, and Wagner 2008). Future work might look at other ways of linking the boards together. Perhaps using lagged values of the Holt-Winter's predictions in neighbouring boards might improve the models further.

Improvements to the models might be gained by more closely investigating the underlying Holt-Winters models. Currently, the values of the Holt-Winters parameters are found by minimising the one-step ahead prediction errors. However, since the Holt-Winters models are used to form a variable that is included in a GLM, it might be better to find the parameters that minimise the residual deviance in the resulting GLM. Very early work on this has suggested a com-

plex optimisation surface, with optimised values being easily affected by starting points. Given the limited and potentially poor data that we have from the start of the system, we did not investigate this further.

Further improvements would aim to address the effect of large residuals. One way to do this would be to fit the GLMs and down-weight those observations that lead to large residuals. This can be done by fitting a GLM to the data normally. The residuals from this GLM are then inversed and used as weights in re-fitting a GLM to the data. For an example of this, see Farrington et al. (1996).

Chapter 4

National Modelling of *Cryptosporidium* Data

In this Chapter we consider developing an exception reporting system from data held by Health Protection Scotland. Their data comes from a much longer period, approximately twenty years, and is much more stable than the data we had from NHS24. Due to the longer period, we have sufficient data to test any system we develop. Again, our overall goal is to develop a regional exception reporting system. We start this process in this Chapter by considering the background to the data help by Health Protection Scotland, and then going on to develop a national model for *Cryptosporidium*. The models developed here will inform our choices for modelling at the regional, health board, level in later chapters.

4.1 Health Protection Scotland Background & Purpose

Health Protection Scotland (HPS) is a governmental organisation that came into existence on 11th November 2004. HPS was created by the Scottish Executive (the Scottish parliament) 'to strengthen and co-ordinate health protection in Scotland' (Health Protection Scotland (HPS)). Their stated aim is:

> 'To work, in partnership with others, to protect the Scottish public from being exposed to hazards which damage their health, and to limit any impact on health when such exposures cannot be avoided.'
>
> (Health Protection Scotland (HPS))

HPS is the direct successor to the Scottish Centre for Infection and Environmental Health (SCIEH).

The primary role of HPS is to carry out all round surveillance on health related matters in Scotland. This role is far from passive: the surveillance is used to inform and develop environmental and public health policy. HPS also functions as a co-ordinating centre allowing a range of public bodies – such as the Scottish Executive, the emergency services and the NHS health boards – to work more consistently and efficiently together. This is a key role in the event of any large-scale public emergencies occurring in Scotland.

4.2 HPS Data

When treating a patient a doctor will often be presented with a myriad of symptoms. Often these symptoms will point clearly to a single disease. However, if the diagnosis is not clear cut, or as part of standard practice with certain conditions, the treating doctor sends a biological sample from the patient to a diagnostic microbiology laboratory. When these labs identify certain micro-organisms they report them to public bodies. In Scotland, the appropriate body is one of the fourteen regional national health boards – shown in Figure 2.8 – who in turn pass these reports to HPS, and previously its predecessor SCIEH (McCabe 2004). While there are now fourteen health boards (see Section 2.4), this change had not been propagated to the data on which we fit models in this and the following chapters. Thus, we deal with data from fifteen health boards. Bodies equivalent to HPS exist in other countries, such as the Health Protection Agency (HPA) in England and Wales, and the Centers for Disease Control and Prevention (CDC) in the USA. In general, reporting is voluntary (McCabe 2004), but for certain micro-organisms – generally the more dangerous or virulent ones – reporting them is mandatory according to the Public Health (Notification of Infectious Diseases) (Scotland) Regulations 1988 (Parliament 1988).

At HPS these reports are collected together to form a central database. Then from this database weekly counts for the different organisms can be created. These counts form an important way of monitoring for disease outbreaks and epidemics in Scotland. However, there are a very large number of reportable organisms (2,387 in 2004). Monitoring this large number of organisms is not a simple task and, given the range and types of organisms, can require a substantial amount of expert knowledge. This task was made much easier by the development of a national exceedance reporting system in McCabe, Greenhalgh, Gettinby, Holmes, and Cowden (2003) and McCabe (2004), which built upon a system suggested by Farrington, Andrews, Beale, and Catchpole (1996).

One of the more practical issues linked with the data is what constitutes a week. Reports are received by HPS Monday through Friday and so a week is considered to end on a Friday. Then, whichever week starting on a Saturday and ending on a Friday contains January 1st is defined as week one for that year. As McCabe (2004) notes, this has a number of consequences. First, this means that week 1 can often include reports from the previous calendar year. Secondly, this means that if December 31st is a Friday, then that year will have 53 weeks. Since 1988, there have been three such years in the system: 1993, 1999, and 2004, with the next being 2010. Having differing numbers of weeks in some years would cause problems for an exceedance reporting system. Thus, for the historical data, all years are standardised to 52 weeks by taking week 52 and 53, and averaging them and recording that result for week 52. In those years with 53 weeks, during week 52, systems works as normal, using the count data for that week; during week 53, the counts for that week are averaged with those for week 52.

4.2.1 Issues with HPS Data

The HPS data has a number of issues linked with them. McCabe (2004) identifies four particular weaknesses: we briefly re-cap those here and direct the interested reader to McCabe (2004) for greater detail.

Reporting Delays

The delays associated with labs passing details of reports to HPS changed during 2008 and 2009. Previously, a manual fax system was used for the collection

of reports. Now, a more automated electronic system has been adopted, which allows for information to be collected in a timelier fashion. The use of an electronic system also reduces the effect of batching: previously a lab might not pass reports on to HPS if a member of staff was away. On the staff member's return, these reports would be reported in one 'batch'. At worst, this could lead to the suggestion of an organism spontaneously going into outbreak. The data we fit models to come from before the adoption of this electronic system and so might contain batching effects (particularly during holiday periods when the labs are more likely to be closed).

The adoption of an electronic system also changed the date linked with each report. During the period for which we have data, the date linked with each lab report was the date of when the report details were received by HPS. Thus, the counts were likely to be biased and reflect the level of a given organism for a time that has already passed. To reduce the lag between the level noted in the system and the level present in the environment, the date linked with a report has been changed to the date of sample collection from the patient. If this date is not available, then the first available date from the following is used: the date the sample was received by the lab; the date the lab reported the results of the tests back to the doctor; finally, the date on which the report is received by HPS. Reducing the lag between the system level and the actual environmental level may allow for outbreaks to be detected sooner.

With the change in date linked with a report, there is potential for a more obvious delay effect: if details of a report are submitted to HPS a week after the corresponding sample is collected from a patient (time in which, for example, tests may be carried on the sample), past data will need to be revised. It will be important to gauge the magnitude of this effect, to determine if the effect can be ignored or if it should be dealt with by the monitoring system. One way of investigating the effect is to consider the distribution of reporting delays (the time between sample collection and report submission); if most delays are under one week, we may find that this makes little difference to the monitoring system as it deals with weekly totals of reports. If the delay is greater than one week, a mechanism may be required for addressing the delay; an example of such a mechanism is given in Chapter 8.

Inputting Errors

To create counts for each organism, searches will be executed on the databases held by HPS. For instance, when creating the counts for *Campylobacter* a search will select all those records that have the keyword `Campylobacter`; the reports will then be aggregated to calculate the number of reports received for each week. Should any reports pertaining to *Campylobacter* have the organism name mis-spelt, or a variant used (perhaps *C. bacter*) then these reports will not be included in the weekly counts for *Campylobacter*. This would mean that the counts of *Campylobacter* would be under-reported and at worst outbreaks missed. McCabe (2004) notes this was a particular problem for certain organisms, and had to be corrected before his system could be used properly. However, given the system has been in place for some time now, the individuals involved in the reporting system will be aware of this difficulty and are likely to closely monitor the data entered into the system for consistency.

When modelling at the health board level, it will obviously be important to have accurate information on which board a report originates in. To ensure consistency, a simple check where the totals across all boards are combined and compared to the national total can be done. Should the two totals not tally, then we know some reports do not have a board identifier. However, there is no simple check, or fix, for a report incorrectly recording which board it originated in. This may only mean the difference of one report, but given some of the boards have very small populations – for instance the islands in the north of Scotland – the omission or addition of a report to a board could determine whether an exceedance occurs.

Duplicates

Reports passed to HPS pertain to individual *samples* and *not* cases. A doctor may send a lab two samples from the same patient to double check a diagnosis or to determine if a patient is still infected. In either case, a double count would occur for what really accounts to a single infection. This would artificially increase the suggested prevalence of an organism. Some protocols are in place for removing duplicates at HPS, however these are nearly always retrospective. Thus, duplicates are a problem that cannot be eliminated from these counts.

Typing and sub-typing processes

As noted with the issue of data-entry consistency, the recorded organism name determines which, if any, weekly organism count a report is included in. Should the name of the organism change, you can end up with separate and unlinked data-sets. A more subtle problem can occur with sub-typing. Certain organisms can have variants, such as *Salmonella typhimurium* and *Salmonella enteritidis*. These may be initially recorded under *Salmonella* and then recorded under the two sub-types, again leading to unlinked data-sets. However, it is not a simple matter of linking the data-sets, as the counts of *Salmonella typhimurium* are obviously different to those of the broader *Salmonella*. Further complications can be caused by different labs having different standards of sub-typing, such that there will be no standard across the country. For example, a lab could report a case of *Salmonella*, while elsewhere it might be sub-typed as *Salmonella enteritidis*.

There is no simple solution for dealing with new typing and sub-typing of organisms. Generally, if possible, we will have to continue using the old classifications until sufficient data are collected under the new ones to allow the reporting system to be used on the new classifications. Sometimes it will not be possible to combine the new typings in such a way as to recreate the old classifications; in such cases, we will have to wait until enough data are acquired under the new classifications to be able to apply the reporting system as normal. In either case, it is important to convey the ramifications of changing types to users of the system.

The effect of surveillance

Another issue not directly address by McCabe (2004) is the effect that surveillance will have on the counts of organisms. As noted in Section 4.1, the surveillance carried out by HPS is not passive. If the counts can be used to suggest preventative measures, then those measures will be put in place, hopefully reducing the number of reported cases. One example can be seen with the reported cases of *Salmonella enteritidis* infection. A major source of infection was linked with infected hens laying infected eggs containing *Salmonella enteritidis*, which were then consumed by humans. To reduce the risk of infection, a programme of vacci-

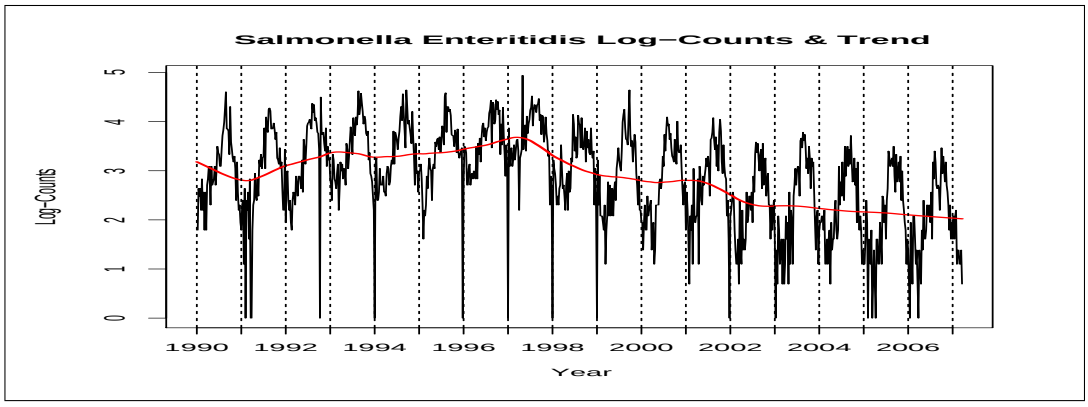


Figure 4.1: The log-counts of *Salmonella enteritidis* in black, with the trend in red, as found by time series seasonal decomposition. For details of the method see Section 4.4.1.

nation was carried on laying hens from 1998 (Health Protection Scotland (HPS) 2004b). This can be clearly seen in the trend of the log-counts, shown in Figure 4.1, found by a time-series seasonal decomposition. From 1998 the trend is distinctly decreasing.

4.2.2 HPS National Exceedance System

The system developed in McCabe (2004) was initially used by SCIEH and continues to be used by HPS. At a conceptual level, for any given week, the system predicts the number of reports it expects for an organism and compares it with the observed number of reports. Then, if the observed reports exceed the expected number by a particular margin, the system indicates that this organism should be further investigated in case of an 'outbreak'. However, the system only indicates *potential* outbreaks (and possibly not all of them); it does not necessarily mean the organism *is* in outbreak.

To calculate the expected number of reports a GLM is used. It uses a seven week window of data from the last five years centred on the week we wish to find an estimate for; thus, the GLM is fitted to 35 values. By choosing this particular data, seasonality does not need to be explicitly modelled, but is captured through the choice of data. Let y_t be the count of the organism at time t. Then, depending on the weekly reporting rate of the organism, one of two different quasi-Poisson

GLMs, with associated dispersion parameter ϕ, are fitted to the mean of the counts μ_t:

Weekly reporting rate	GLM Prediction Model
rate ≤ 5	$\log(\mu_t) = \alpha + \beta * week_t +$ `fact_year`
rate > 5	$\log(\mu_t) = \alpha + \beta * week_t + \gamma * year_t$

where α, β and γ are constants; $week_t$ and $year_t$ are integers indicating which week and year respectively the count at time t relates to; `fact_year` is a factor with a different level for each year the counts are from. Once a GLM is fitted, the model can be used to predict the expected number of counts $\hat{\mu}_0$ for the present week.

Using $\hat{\mu}_0$, a 99% prediction interval is formed, adjusted for skewness by using a $\frac{2}{3}$-power transform. The upper limit of the prediction interval, U is given by:

$$\hat{\mu}_0 \left\{ 1 + \frac{2}{3} z_\alpha \left(\frac{\hat{\tau}}{\hat{\mu}_0} \right) \right\}^{3/2},$$

where z_α is the $100(1-\alpha)$-percentile of the normal distribution, and $\hat{\tau}$ is then given by:

$$\phi + \frac{\text{var}(\hat{\mu}_0)}{\hat{\mu}_0}.$$

An exceedance score is then formed by comparing the upper limit of the prediction interval with the observed value y_0 for this week via:

$$\frac{y_0 - \hat{\mu}_0}{U - \hat{\mu}_0}.$$

Then, if an exceedance score greater than 1 is observed, the system 'triggers' an alarm, informing the user that a higher than expected count has been observed. In such a case, the user can investigate if the given organism is actually in outbreak, or just has higher counts than expected by chance.

Initially, at HPS, $\alpha = 0.10$, while in Farrington et al. (1996), $\alpha = 0.01$. This was extended at a later point with a 'traffic light system' (an example of this approach can be found in Allardice, Wright, Peterson, and Miller (2001)): here, an exceedance when $\alpha = 0.10$ is 'amber' and means the organism should be monitored more closely; an exceedance when $\alpha = 0.01$ is 'red' means the organism should be investigated; otherwise the organism is considered 'green' and safe.

4.2.3 Regional Considerations

We have considered the current national reporting system. We now consider some of the issues relating to developing a regional reporting system.

In reviewing the various methodologies for developing a reporting system, McCabe (2004) considers starting from a spatial dimension, instead of time, such as in Leung, Patel, and McGilchrist (1999) and Raubertas (1989). Outbreaks would then be indicated by detecting clusters of reports. To utilise such a system would require a good deal of geographical information. Unfortunately, this is not available for all reports: only some include the address and post code of the patient that the sample is collected from. Due to this and the administrative structure of HPS, it is more useful to monitor the counts from each health board. If a health board is shown to have a potential outbreak, HPS can advise that health board to monitor the situation more closely.

There are some problems associated with working with the health board level data. The boundaries for the boards are somewhat arbitrary, more likely to be determined by governmental requirements and infrastructure than physical or environmental factors that affect organism spread. For this reason McCabe (2004) considers it to be more fruitful to work at a national level and focus on a method that utilises a time dimension, rather than a spatial one. However, given the national system is in place, the next obvious extension is to consider the health board level counts. Given the time dimension for the counts is very rich – around twenty years of data – we start from the time dimension and then proceed to see how the health boards can be linked together, be this spatially or otherwise. This could potentially allow the spread of infections to be modelled and monitored: for instance, a rise of infections in Lothian might be a precursor to a rise of infections in Glasgow, due to commuters carrying the infection between the two health boards.

McCabe (2004) notes that the reporting rate for different organisms can differ greatly, with many organisms having a very low national reporting rate, and which at a health board level will mean the rates are correspondingly lower. For the modelling considered here, we will only consider those organisms with a higher national reporting rate. This is for two reasons. First, models fitted to the very low counts, possibly with many zero counts, will likely not fit well to

the larger counts. For instance, we saw that the national system fits one of two GLMs depending on the rate of reporting of a particular organism (Section 4.2.2). Secondly, is a reason of practicality. If the rate is very low, then outbreaks are likely to be detected by the national system. Further, given the small number of cases, it will not require much work at HPS to find the location of these potential outbreaks. When the rate is higher, it will be harder to judge potential outbreaks locally and more work to determine the location of the potential outbreak.

Since most outbreaks, particularly the smaller ones, will be localised to a particular area, we can expect the health board level counts to display more serial correlation than at the national level. It should be noted that the national level system does nothing to address serial correlation; McCabe (2004) notes that he found little serial correlation among the counts and so disregarded it. However, serial correlation can have a marked effect on the forecasts made, as standard errors tend to be under-estimated in its presence. The standard errors are used to determine which variables are used in the forecast model through significance tests, so can lead to very different models, and so different forecasts if ignored. Generally, any model will be better for having addressed serial correlation, and so we do our best to incorporate it into the models developed here. For instance, we expect to find many consecutive weeks of zeros: by modelling this, serial correlation may be reduced.

In summary, modelling at the regional level presents different issues to the national level. The system developed by McCabe (2004) is a very general purpose tool, having to deal with around three thousand organisms. As any regional system developed is likely to be applied to far fewer organisms we can afford to develop a system in greater detail and try to deal with such things as serial correlation and potential inter-board relationships.

4.2.4 Exploratory Analysis: Organism Choice

To better understand the structure of the counts at a health board level, it was decided to take one organism and initially apply similar techniques as were utilised with the NHS24 data, using Generalised Linear Models (GLMs) as the basic modelling tool. HPS suggested six organisms with reasonably large rates as suitable ones to explore: Methicillin Resistant *Staphylococcus Aureus* (MR*SA*), *Norovirus*,

Salmonella Enteritidis, *Salmonella Typhimurium*, *Campylobacter*, *Cryptosporidium*. To choose between them we consider the log-counts of these organisms, from 1990 to the beginning of 2007, as they have two advantages. Medical counts are, as standard, assumed to be Poisson distributed and so modelled with a Poisson GLM, which typically uses a log-link function. By looking at the log-counts we see a form of the data that covariates will be fitted to in a GLM. The log function also serves as a variance stabilising function. As the counts are Poisson distributed, as the counts increase so does their associated mean, and by definition their variance; by applying the log function the multiplicative structure becomes additive, which is easier to interpret.

We decided against modelling MR*SA* and *Norovirus*, since they experience a step change in level, as seen in Figure 4.2 The trends in this section were estimated by using time-series seasonal decomposition; the method is explained in Section 4.4.1 As MR*SA* is primarily a Hospital Acquired Infection (HAI), its change in level is probably attributable to the growing prevalence that HAIs have had in the media's consciousness of recent years. As the public becomes more concerned about HAIs, authorities will need to monitor them more closely, leading to an increased reporting rate that does not necessarily indicate a greater natural prevalence. The change in reported cases of *Norovirus* is likely to have some link with a greater natural prevalence but is affected by inconsistent reporting standards between public health labs, meaning that 'accurate interpretations of changes in reporting rates over time or between NHSB [health boards] is impossible' (Smith-Palmer and Cowden 2003). When it comes to modelling these organisms, it may be necessary to disregard the data before the change in level; we could model MR*SA* from 2001 and *Norovirus* from 2003. The national system tries to deal with such changes in level by fitting models to data from only the last five years. However, for this exploratory analysis, it seems sensible to pick an organism with no such step change.

Salmonella Typhimurium was rejected, as its counts are becoming increasingly small. The counts can be seen in Figure 4.3. Towards the end of the data, particularly for 2005 and 2006, nearly all of the counts are in single digits; this will mean that at the health board level, many counts will be zero, and so not that interesting to model.

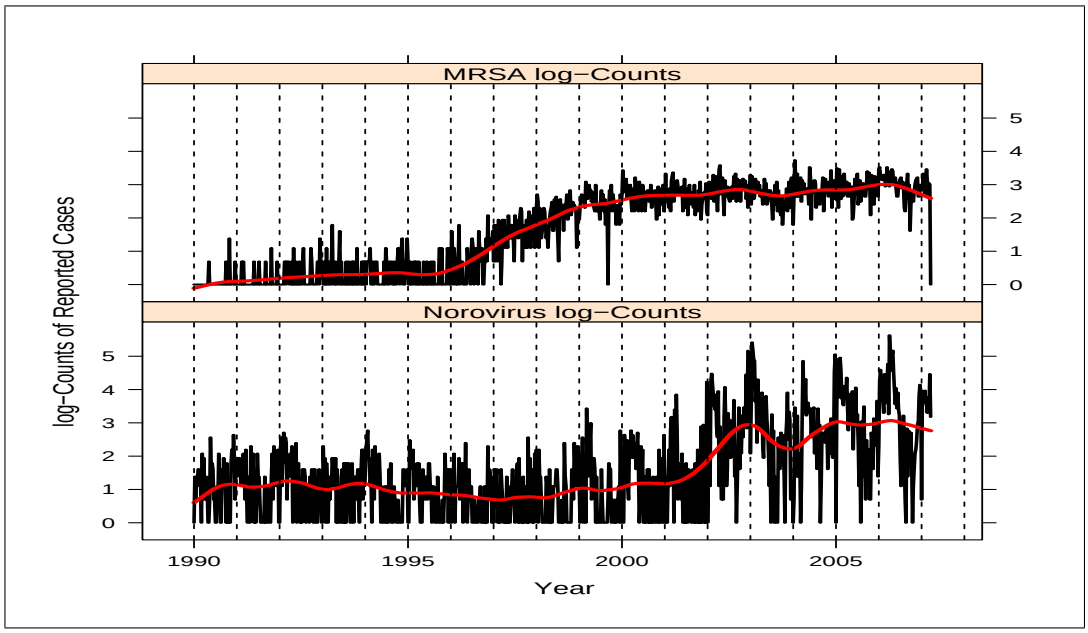


Figure 4.2: The log-counts of MR*SA* and *Norovirus* in black, with estimated trend in red as found by time-series decomposition.

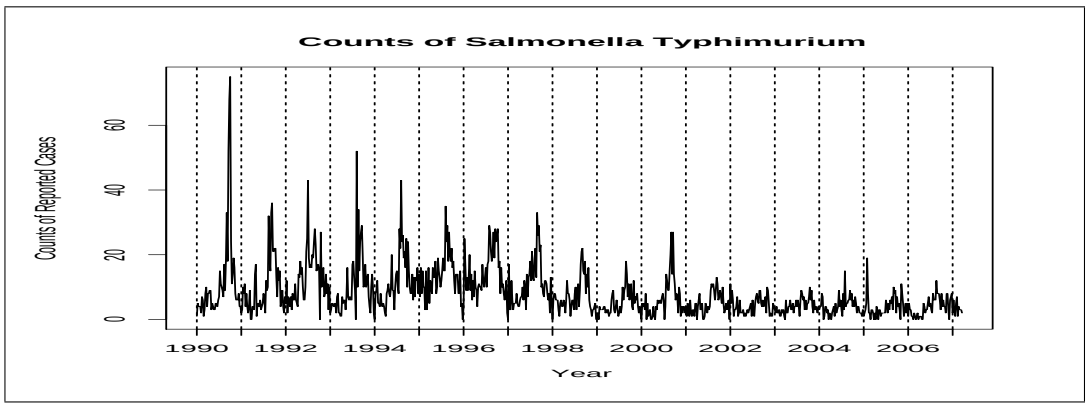


Figure 4.3: The counts of cases reported to HPS of *Salmonella Typhimurium*.

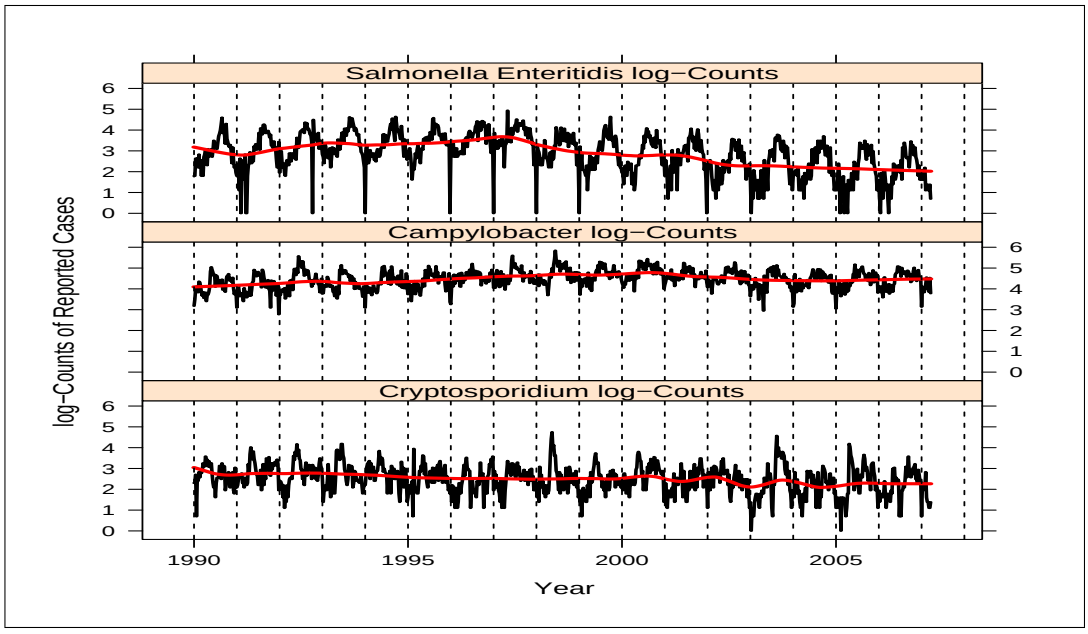


Figure 4.4: The log-counts of *Salmonella Enteritidis*, *Campylobacter* and *Cryptosporidium* in black, with estimated trend in red as found by time-series decomposition.

Of the remaining organisms, the trend in the log-counts of *Salmonella Enteritidis* is most variable, as can be seen Figure 4.4. Relatively, the counts of *Campylobacter* and *Cryptosporidium* are quite stable, so we chose between those. These two organisms do differ in average weekly reporting rates quite markedly: 93 and 14 for *Campylobacter* and *Cryptosporidium* respectively. We chose to try exploring the *Cryptosporidium* counts, since they are lower, but not too low as with *Salmonella Typhimurium*: we expect any models developed from *Cryptosporidium* will scale up appropriately for those organisms with higher rates.

The decision in this Section between the organisms has been mostly motivated by the trends in their reporting rate. This seems a reasonable step in this exploratory analysis, but is something we will need to address in any final system developed. One approach might be to disregard some past data, as is done with the national system. However, we leave such considerations for later chapters.

4.2.5 Development Process of Regional Exception Reporting System for *Cryptosporidium*

Having decided to focus our efforts on developing a regional exception reporting system for *Cryptosporidium*, we adopt a top down approach. We first consider the biological background for *Cryptosporidium* in Section 4.3, to help us make more informed modelling decisions. We then consider suitable models for the national counts of *Cryptosporidium* in Section 4.4. These models will help us develop the regional models in the next chapter. The regional models can then be used for the prediction element of the exception reporting system.

4.3 *Cryptosporidium*: Biological Background

The first reported cases of human infection by *Cryptosporidium* are recorded in 1976 (Ungar 1990; Nime, Burek, Page, Holsher, and Yardley 1976). Since then, the prevalence and threat posed by *Cryptosporidium* has continued to increase (Guerrant 1997). The threat posed is exacerbated by the ever increasing numbers of immuno-compromised humans, particularly in the AIDS stricken African continent.

Cryptosporidium is a parasite that is smaller than a red blood cell and has a number of sub-types or *genotypes*. The primary human pathogens are *Cryptosporidium hominus* and *Cryptosporidium parvum* (Pollock, Young, Smith, and Ramsay 2008; Peng, Xiao, Freeman, Arrowood, Escalante, Weltman, Ong, MacKenzie, Lal, and Beard 1997). Both cause Cryptosporidiosis, an infection whose main symptom is watery diarrhoea, along with dehydration, weight loss, fever, nausea and vomiting (Centers for Disease Control and Prevention 2007). These symptoms normally present around a week after infection and can last from one to two weeks, with most people recovering fully (Association of Medical Microbiologists 1997). However, for those who are already immuno-compromised, such as those suffering with AIDS, *Cryptosporidium* can be much more serious, sometimes fatally so (Caccò, Thompson, McLauchlin, and Smith 2005). Guerrant (1997) is a good source for a more detailed medical description of the infection.

Like a number of parasites, *Cryptosporidium* can affect a wide range of animals; in particular, *Cryptosporidium parvum* has been shown to transfer to hu-

mans from other animals (Hunter and Thompson 2005). In each host the parasite goes through an entire life cycle. It lives in the intestine of an afflicted animal and passes out 'oocysts' in the stools of its host. An 'oocyst' is a thick walled structure that the parasite develops within and serves as a vehicle to carry the parasite to other animals through infected stools. Infection is thought to be spread through infected animal waste entering the water cycle (Smith, Robertson, and Ongerth 1995); however the magnitude of this effect is unclear (Robertson 2009).

Between humans, infection is often passed because of poor food hygiene and under-cooking meat (Laberge and Giffiths 1996). Usually, these infections are related to *Cryptosporidium hominus*. Infections can spread quickly, as even after symptoms have abated, individuals remain infectious for several weeks. In the UK, the infection is most common among children aged between one and five years (Health Protection Agency). Generally *Cryptosporidium hominus* is restricted to humans but there have been reported cases of infection of animals (Xiao and Fayer 2008; Smith, Nichols, Mallon, MacLeod, Tait, Reilly, Gray, Reid, and Wastling 2005).

Controlling the spread of *Cryptosporidium* is difficult primarily because of its strongly durable oocyst form (Association of Medical Microbiologists 1997; Guerrant 1997). In this form it can survive, but not multiply, for extended periods (particularly in moist conditions) (Meinhardt, Casemore, and Miller 1996). For example, *Cryptosporidium* can survive the chlorination present in swimming pools for several days, making them common sources of infection (Centers for Disease Control and Prevention 2007; Fayer, Morgan, and Upton 2000). When oocysts enter water supplies they are very hard to eradicate (Drinking Water Inspectorate 2001; Goldstein, Juranek, Ravenholt, Hightower, Martin, Mesnik, Griffiths, Bryant, Reich, and Herwaldt 1996). The infection rate is also high because it takes very few oocysts to infect a new host (Chappell, Okhuysen, Sterling, Wang, Jakubowskui, and DuPont 1999; Okhuysen, Chappell, Crabb, Sterling, and HL 1999; Association of Medical Microbiologists 1997; DuPont, Chappell, Sterling, Okhuysen, Rose, and Jakubowskui 1995); however, there is disagreement over the precise level required for infection (Scottish Parliament Information Centre (SPICe) 2002; Messner, Chappell, and Okhuysen 2001; Fayer, Morgan, and Upton 2000). In human water supplies, it is often hard using standard techniques to reduce the levels of oocysts to those that will not cause in-

fections (Drinking Water Inspectorate 2001). An 'integrated multiple barrier approach' has been shown to be effective (Pollock et al. 2008). Newer water purification techniques, utilising *Cryptosporidium*'s ultraviolet light and ozonation sensitivities, are being investigated (Rochelle et al. 2004; Korich et al. 1990; Awwa Research Foundation 2008). The resilience of *Cryptosporidium* has led to it becoming one of the most common causes of waterborne infections (Centers for Disease Control and Prevention 2007).

There have been a number of large outbreaks of *Cryptosporidium* infection within the UK (Chartered Institute of Enviromental Health 2007). Within Scotland, two of the most noticeable outbreaks have been linked with contaminated water supplies which provide water to Glasgow and Edinburgh (Drinking Water Quality Regulator 2003; Scottish Parliament Information Centre (SPICe) 2002). There was also a large outbreak in the water supply to Aberdeen (Mukerjee 2002). The link between water supplies and infections are further explored in Pollock, Young, Smith, and Ramsay (2008). Water supplies can often become infected during periods of heavy rainfall, when there can be insufficient capacity in treatment plants to treat all the water, or surface water enters the water supply unusually (Health Protection Scotland (HPS) 2002d; Scottish Parliament Information Centre (SPICe) 2002). Thus, when modelling the counts of *Cryptosporidium* it may be worth considering the effect of rainfall. Smaller, and often more local outbreaks, tend to be associated with three particular sources: people visiting the countryside and getting infections from sheep and other cattle, particularly at wildlife centres (considered in Sayers, Dillion, and Connolly (1996), Miron and Kenes (1991), with an example outbreak in Health Protection Scotland (HPS) (2005)); tourists returning home often carry infections, particularly if they have returned from parts of the world with poorer sanitary conditions (considered in Smerdon, Nichols, Chalmers, Heine, and Reacher (2003), with an example outbreak in Health Protection Scotland (HPS) (2003b)); and swimming pools (considered in Robertson, Sinclair, Forbes, Veitch, Kirk, Cunliffe, Willis, and Fairley (2002), Fayer, Morgan, and Upton (2000), with an example outbreak in Health Protection Scotland (HPS) (2002a)).

We note above that there are two genotypes of *Cryptosporidium*, with different infection routes. Due to the different routes, the genotypes have different levels at different times of the year. Ideally, we would consider the genotypes separately.

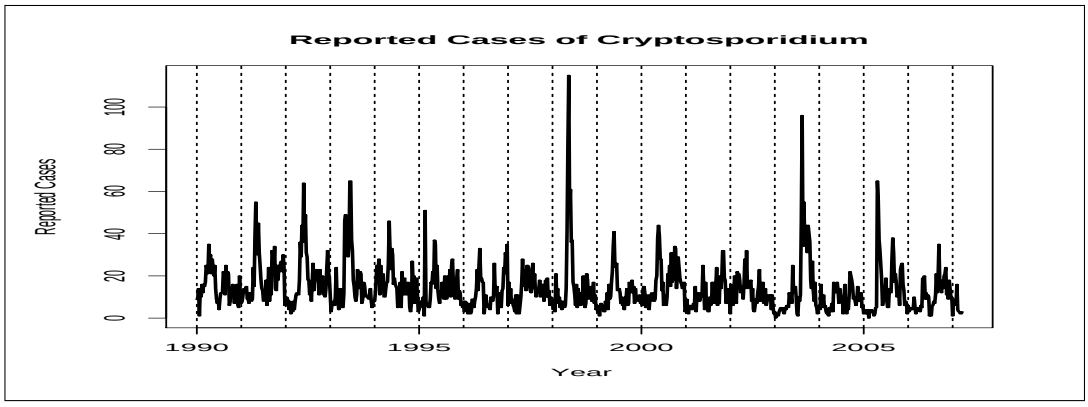


Figure 4.5: The reported cases of *Cryptosporidium* received by HPS from 1990 to early in 2007.

However, identifying which genotype a lab sample belongs to is very difficult due to the structure of the oocysts (Fayer et al. 2000). For this reason there is no subtyping of *Cryptosporidium* by HPS, so their weekly reports of *Cryptosporidium* include both types anonymously (Pollock, Ternent, Mellor, Smith, Ramsay, and Innocent 2009).

4.4 National Modelling

Before fitting any regional models we first consider the *Cryptosporidium* counts at the national level to gain a better understanding of the data. For consistency, we continue to work with data from the same years as were used in Section 4.2.4, where the counts from 1990 to the twelfth week in 2007 were used. The counts are shown in Figure 4.5. From a visual inspection, there is a suggestion of a decrease in reporting level. There are also some relatively very high counts, such as in 1998, 2003 and 2005, which might be linked to seasonal peaks or outbreaks.

4.4.1 Time-series Decomposition

As a first step in analysing the counts we carry out a time-series decomposition on them. We do this on the log-counts as we expect the seasonality to be multiplicative. Organism infection is generally multiplicative, as when one host is infected they can infect several others, and those others can go on to infect several others

beyond themselves and so on. Thus, the seasonal factors themselves will scale in a multiplicative way and so by using the log-counts the seasonality becomes additive. We first calculate the seasonal factors. A crudely de-seasonalised version of the log-counts is found by running a moving average of fifty-two over them and then subtracting it from the original log-counts. Using this de-trended series, we calculate the mean value for each set of weeks in turn (all the first weeks, the second weeks and so on), creating a seasonal factor for each week in the year. The resulting seasonal factors are shown in Figure 4.7.

Cryptosporidium seems to be at its most prolific during moist conditions, which perhaps explains why it is higher during the end of Spring through to early Summer, and Autumn through to the beginning of Winter. Further, a number of outbreaks have been linked with heavy rainfalls (for instance Health Protection Scotland (HPS) (2002d)). Given these considerations, we go on to investigate including variables that measure rainfall and temperature in our models – see Section 4.4.3. The seasonal levels seem to change in a reasonably smooth fashion, suggesting that seasonality might be modelled acceptably by trigonometric terms. This would be preferable to using a different factor level for each week in the year as this would require fifty-two different values. *Cryptosporidium* is somewhat unusual in having four turning points in its seasonal pattern but this has been noted elsewhere (Scottish Centre for Infection and Enviromental Health (SCIEH) 1998). This means it may require more harmonics than some other organisms.

From the work of Pollock, Ternent, Mellor, Smith, Ramsay, and Innocent (2009), the seasonality of *Cryptosporidium* seems to be partly explained by its different genotypes. Consider Figure 4.6 taken from Pollock et al. (2009), with permission (Pollock 2009), which shows the monthly totals for a two year period of *Cryptosporidium hominis* (human only genotype) and *Cryptosporidium parvum* (human and animal genotype). The peak in May of *C. parvum* explains the peak around week twenty in Figure 4.7. Pollock et al. (2009) attribute the rise of *C. parvum* to lambing and calving at this time year, which is supported by others (McLauchlin, Amar, Pedraza-Diaz, and Nichols 2000). Lambs have a greater susceptibility to *Cryptosporidium*, and so as they become infected environmental levels of *Cryptosporidium* increase (Santín, Trout, and Fayer 2007). The more diffuse second peak in Figure 4.7 is driven by *C. hominis* which is bimodal with peaks in August and October. These peaks may be explained by holiday periods,

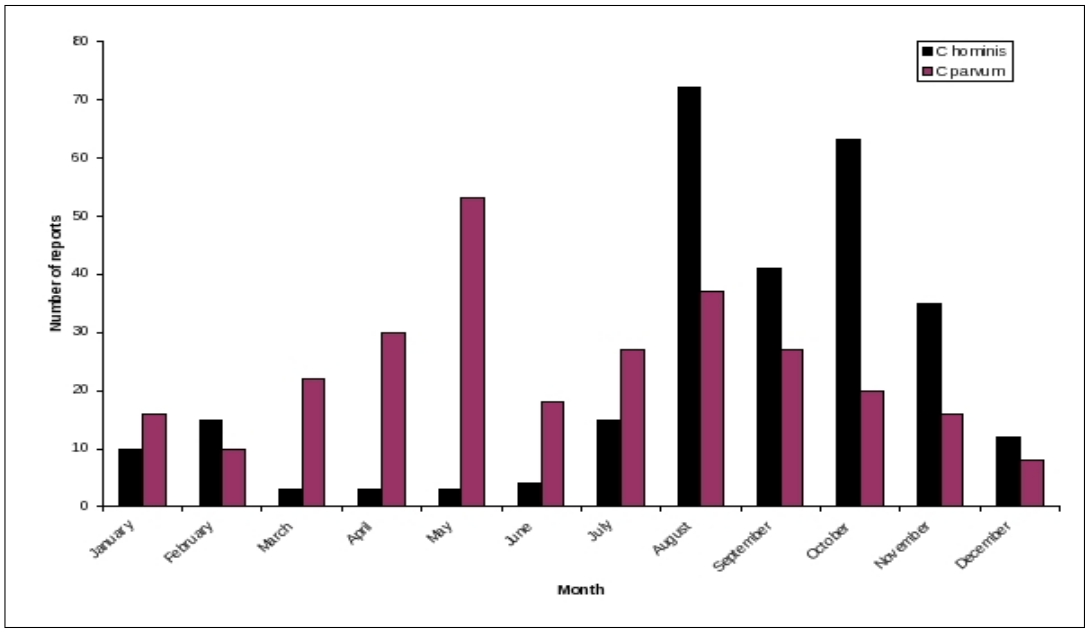


Figure 4.6: Month of the year and counts of reports by species of *Cryptosporidium*, from June 2005 to June 2007: *Cryptosporidium hominis* is shown in black and *Cryptosporidium parvum* in purple. Taken with permission from Pollock, Ternent, Mellor, Smith, Ramsay and Innocent (2009).

when people will engage in more recreational activities and foreign travel (Pollock, Ternent, Mellor, Smith, Ramsay, and Innocent 2009). Both of these increase the risk of infection as individuals have a greater risk of being exposed to infected people.

We then estimate the trend in the log-counts by subtracting the seasonal factors from the log-counts and using a smoother on the resulting residuals. To calculate the trend as observed in Figure 4.4, Friedman's SuperSmoother (the `supsmu` function in `R`) is used (Friedman 1984). The estimated trend suggests a slow decrease over time in the log-counts. It seems that the trend may be modelled by a linear trend term, which we can test by fitting a straight line term in models fitted to the log-counts. Between 2000 and the end of 2006 there are some fluctuations from this 'straight' line. There is some evidence that the early fluctuations may be linked the outbreak of Foot and Mouth Disease (FMD) in Scotland in the southern health boards Borders and Dumfries & Galloway during 2001. Strachan, Ogden, Smith-Palmer, and Jones (2003) compare the rate

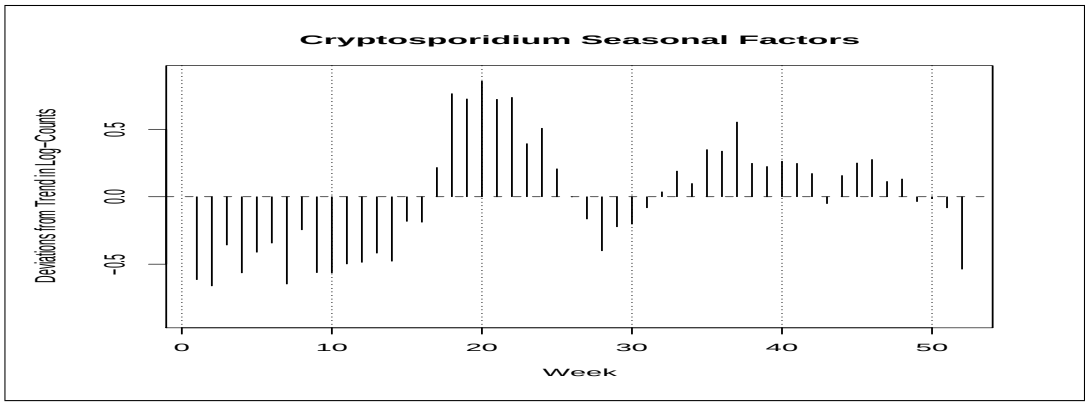


Figure 4.7: The seasonal factors as estimated from the log-counts of *Cryptosporidium* by time-series decomposition. Of the diseases considered here, it is the only one to exhibit two 'peaks' and two 'troughs' in its seasonal pattern.

of *Cryptosporidium* in these two afflicted boards with the rate in the rest of the country during the infectious period. They find that the rates of infection in Borders and Dumfries & Galloway are significantly lower than the rates in the rest of the country during the period of infection in 2001. This is attributed to the measures which were taken by the UK government to deal with the FMD outbreak: in the afflicted health boards in Scotland, approximately one million cattle were slaughtered and strict restrictions placed on the movements of cattle. These actions seem to have had the result of reducing the levels of *Cryptosporidium parvum* in the environment and thus reducing the cases of infections in humans (Strachan et al. 2003). The results in Strachan et al. (2003) suggest that the rates recover during 2002. We may find that fitting a factor for the infectious period and the recovery period may improve our models. For more information on the FMD outbreak in 2001, see Department for Food, Environment and Rural Affairs (2008). There was a small outbreak of FMD in 2007, but it did not affect Scotland (Department for Food, Environment and Rural Affairs 2009).

4.4.2 GLM Modelling

We start by fitting a negative-binomial GLM to the *Cryptosporidium* national counts, since, as with the NHS24 data, we find the counts to be over-dispersed.

We fit the following GLM to the mean μ_t of the counts at week t:

$$\log(\mu_t) = \beta_0 + \beta_1 Ti_t + trig(10)_t, \tag{4.1}$$

where: β_0, β_1 are coefficients; Ti is the week number; $trig(10)_t$ are the first 10 harmonics of sine and cosine at week t with period 52. Thus $trig(2)_t$ would denote, given β_0 and β_1 have already been used:

$$\beta_2 \sin\left(\frac{2\pi t}{52}\right) + \beta_3 \cos\left(\frac{2\pi t}{52}\right) + \beta_4 \sin\left(\frac{4\pi t}{52}\right) + \beta_5 \cos\left(\frac{4\pi t}{52}\right).$$

Here, Ti models a linear trend and the $trig$ terms are used to model the seasonal pattern in the counts. We test the elements of the model for statistical significance and find that only the first four harmonics of the trigonometric terms are necessary. Thus, we fit a new GLM to reflect this:

$$\log(\mu_t) = \beta_0 + \beta_1 Ti_t + trig(4)_t, \tag{4.2}$$

where the terms are as defined previously. The estimated coefficients for this model can be found in Table 4.1 and θ, the dispersion parameter of the negative binomial, is 4.87. This is quite a large value for θ, suggesting a great deal more variation than we would expect from standard Poisson counts. From the constant, β_0, we can see that $\exp(2.889) \approx 18$ reports are expected each week, when $Ti = 0$, before any effects of trend or seasonality are considered. This is slightly higher than the average count of 14 that was found in Section 4.2.4. As was speculated, there appears to be a decreasing linear trend in the number of reports received, shown by the negative sign of β_1. Since $\exp(52 \times -0.0005461) = \exp(-0.028) = 0.972$ (using a greater level of precision for β_1), we expect 2.28% fewer reports each year. Using a deterministic linear trend as we do here must be done carefully. We cannot extrapolate very far into the future since counts could become unrealistically small.

The seasonal pattern fit by the GLM is shown in Figure 4.8, superimposed on the seasonal factors found in Section 4.4.1 by time-series decomposition. Seasonality seems to be modelled relatively well during weeks thirteen to forty, but the cosinusoids seem to adapt poorly to the rest of the year. As noted previously, *Cryptosporidium* is one of the few organisms with a seasonal pattern with

Model	Intercept	Ti_t	$\sin\left(\frac{2\pi t}{52}\right)$
Equation 4.2	2.888 (0.040)	-0.000546 (0.000078)	-0.163 (0.028)

Model	$\cos\left(\frac{2\pi t}{52}\right)$	$\sin\left(\frac{4\pi t}{52}\right)$	$\cos\left(\frac{4\pi t}{52}\right)$	$\sin\left(\frac{6\pi t}{52}\right)$
Equation 4.2	-0.268 (0.029)	-0.299 (0.028)	-0.060 (0.029)	0.179 (0.029)

Model	$\cos\left(\frac{6\pi t}{52}\right)$	$\sin\left(\frac{8\pi t}{52}\right)$	$\cos\left(\frac{8\pi t}{52}\right)$
Equation 4.2	0.108 (0.029)	-0.132 (0.029)	-0.161 (0.029)

Table 4.1: The coefficients, and corresponding standard errors in brackets, calculated for the GLM defined by Equation 4.2 (trend and first four harmonics for seasonality).

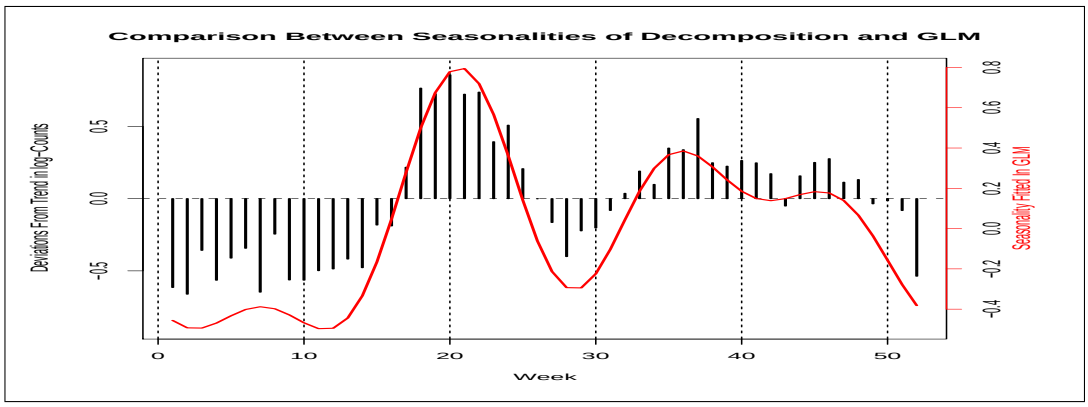


Figure 4.8: The seasonal factors as estimated from the log-counts of *Cryptosporidium* by time-series decomposition, shown in black, and in red the seasonal values fitted in the GLM defined by Equation 4.2 (trend and first four harmonics for seasonality).

four turning points, so we expect other organisms to be modelled better with fewer cosinusoids. For instance, the seasonal patterns of *Salmonella Enteritidis* and *Salmonella Typhimurium* shown in Figure 4.9 are very smooth and probably would be modelled very well by one harmonic. However, the cosinusoid terms give the most parsimonious way of modelling seasonality, so we continue to use them, particularly as they will be most easily extendable to other organisms. It is notable that all three of the organisms considered here have large negative seasonal factors for weeks one and fifty-two. This is likely to be caused by delayed reporting due to the holidays at the beginning and end of the year.

The auto-correlation function (acf) of the deviance residuals is shown in Figure

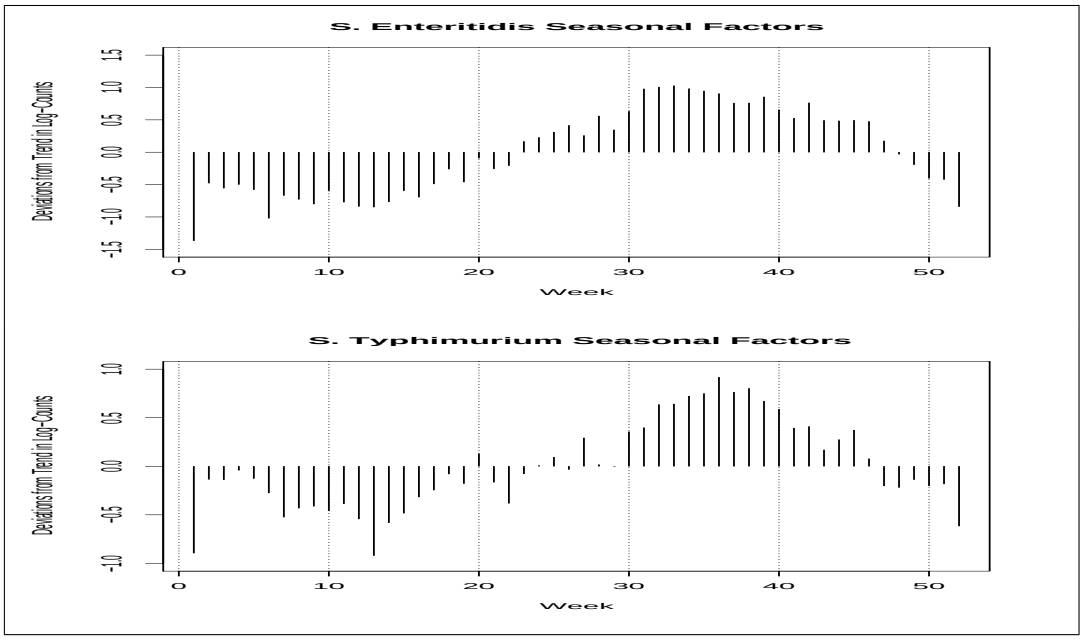


Figure 4.9: The seasonal factors found from a seasonal decomposition carried out on *Salmonella Enteritidis* and *Salmonella Typhimurium*. The general smoothness of their seasonal pattern suggest they will be modelled well by trigonometric terms.

4.10 and indicates there is strong significant residual serial correlation. The low order correlations exhibit strong slowly decreasing positive correlation, suggesting the presence of trend that is not modelled appropriately by the GLM. We focus on removing the low order correlations and hope that this will deal with most of the residual trend.

One of the ways to address serial correlation in such a model is by including lagged values of the quantity being modelling (Brandt, Williams, Fordham, and Pollins 2000). This serves as a proxy to the serial correlation rather than modelling it explicitly. We alter the model defined by Equation 4.2 to the following:

$$\begin{aligned}\log(\mu_t) \quad = \quad & \beta_0 + \beta_1 Ti_t + trig(4)_t + \beta_{10}\log(x_{t-1}) + \beta_{11}\log(x_{t-2}) \\ & + \beta_{12}\log(x_{t-3}) + \beta_{13}\log(x_{t-4}) + \beta_{14}\log(x_{t-5})\end{aligned},$$

where x_t are the counts of *Cryptosporidium* for week t. We use the log of the past values so that they are on the same scale as the mean μ_t. More lagged values could be included but we use five as a starting point. When this model is fitted we find that only the first two lagged counts are statistically significant. Note

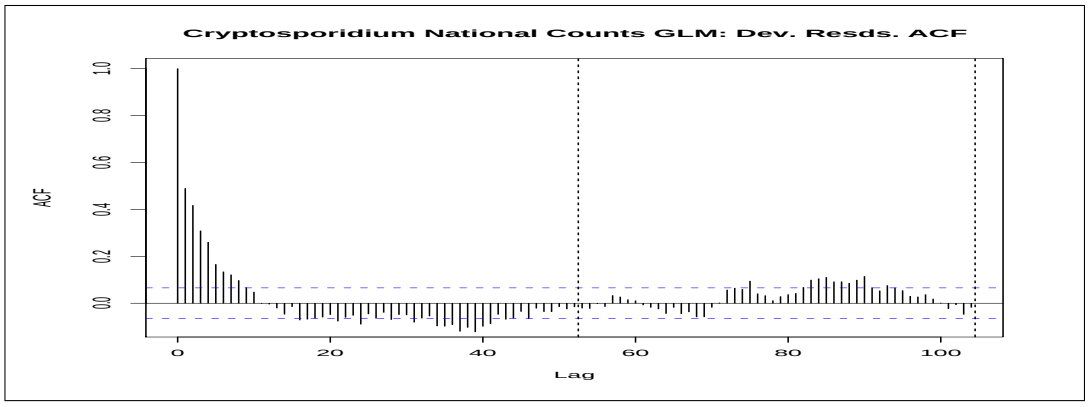


Figure 4.10: The acf of the deviance residuals from the GLM defined by Equation 4.2 (trend and first four harmonics for seasonality) for the first two years of lags.

that x_{t-1} is the variable that explains the most deviance as we would expect.

Thus, we fit a model with just x_{t-1} and x_{t-2}:

$$\log(\mu_t) \quad = \quad \beta_0 + \beta_1 Ti_t + trig(4)_t + \beta_{10}\log(x_{t-1}) + \beta_{11}\log(x_{t-2}). \qquad (4.3)$$

The acf of the deviance residuals for this model can be found in Figure 4.11. The level of residual correlation is now much less, with the correlation at low order lags being much lower (cf Figure 4.10). Thus, we have one approach to dealing with the residual serial correlation.

Another approach is to use a form of the smoothed counts as a variable. We apply the Holt-Winters' filter to the log-counts and use the resulting one step-ahead forecasts as a variable in a GLM, as we did with the NHS24 modelling in Chapter 3. Again, the log-counts are used, as this means the predictions are then on the log-scale, matching the scale of the counts through the log-link function. In essence, the Holt-Winters' smoother allows for exponential smoothing to be carried out on a time series, but also includes elements to incorporate seasonality and a linear trend. A greater explanation of Holt-Winters' smoothing can be found in Appendix B. The smoothing parameters for the HW_t term can be found in Table 4.2. The parameter for local level (α) is quite high at 0.37 (normally expected to be between 0.1 and 0.3 – see Appendix B), which is consistent with there being quite a lot of serial correlation in the counts. Seasonality is fitted but without any local linear trend.

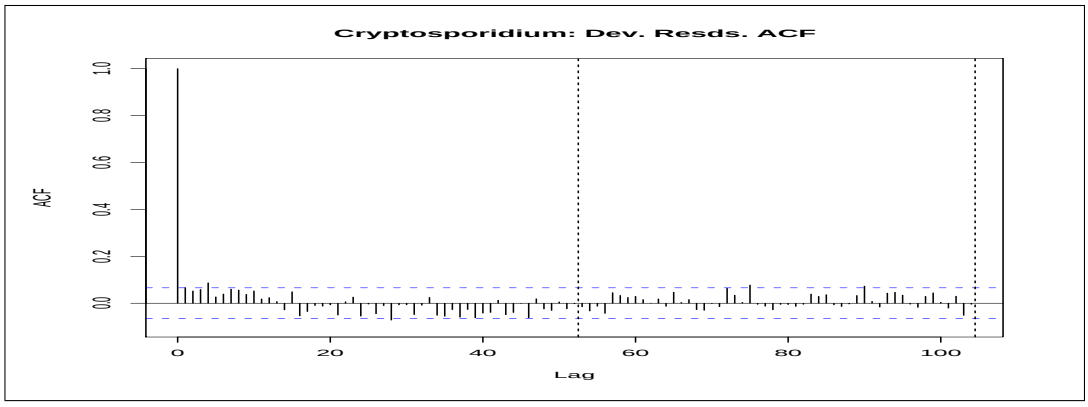


Figure 4.11: The acf of the deviance residuals from the GLM defined by Equation 4.3 (trend, first four harmonics for seasonality and two past observations) for the first two years of lags.

Parameter	Value
Local level – α	0.369
Local trend – β	0
Seasonal – γ	0.214

Table 4.2: The parameters of the Exponential Smoothing (Holt-Winters' form) that is used to create the HW_t term used in the model defined by Equation 4.4 (trend, first four harmonics for seasonality and Holt-Winters one-step ahead prediction). These parameters were left free to vary and found by minimising $\sum e_t^2$; see Appendix B.

Thus, the GLM with trend and seasonality defined in Equation 4.2 is then extended to:

$$\log(\mu_t) \quad = \quad \beta_0 + \beta_1 Ti_t + trig(4)_t + \beta_{10} HW_t, \tag{4.4}$$

where HW_t is the one-step ahead Holt-Winters prediction of the log-counts at time t, and the other terms are as before. In fitting seasonal elements of the Holt-Winters smoother, the function `HoltWinters` loses one year of data, so the GLM is fitted to counts from 1991 onwards, instead of 1990 as was done previously.

When we fit the model defined by Equation 4.4, we find that the HW_t term is statistically significant. The acf of the deviance residuals can be found in Figure 4.12. There appears to be evidence of correlations at lags one and two.

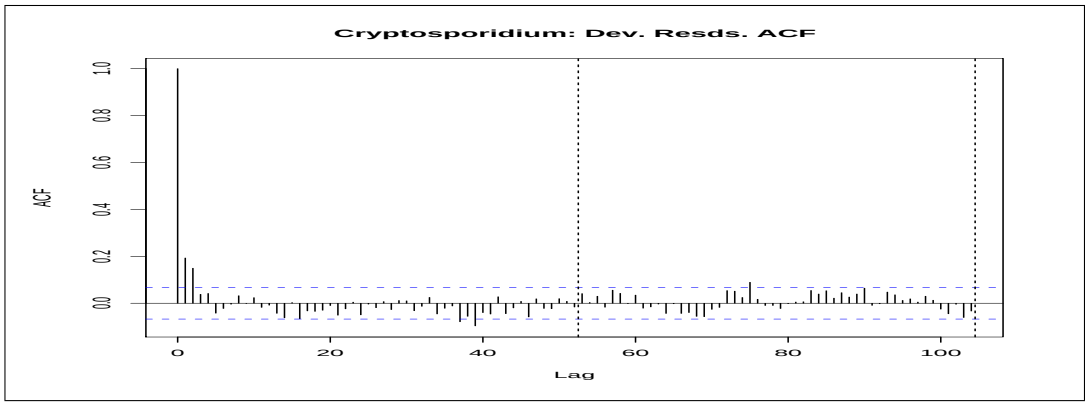


Figure 4.12: The acf of the deviance residuals from the GLM defined by Equation 4.4 (trend, first four harmonics for seasonality and Holt-Winters one-step ahead prediction) for the first two years of lags.

To allow for easier comparison between these two approaches of removing serial correlation, we fit both GLMs defined by Equations 4.3 and 4.4 to data from 1991 onwards. The coefficients calculated for these models can be seen in Table 4.5, with the model fits summarised in Table 4.4. The coefficient of HW is the largest, barring the constant, by 2 orders of magnitude and so can dominate the other terms. As the HW_t term is a one-step ahead predictor, we might expect its coefficient to be nearer to 1, as opposed to 0.57; this value may be due to the seasonal or trend elements of the Holt-Winters' smoother performing poorly; we investigate this more closely in Section 5.4. In both models there is downward trend of similar magnitude.

The latter model (Equation 4.4) seems to be a very slightly better fit than the other, as it has a smaller residual deviance (830 to 835 respectively – see Table 4.4). However, since the models are not nested, it is best to use the AIC statistic to compare them. Recall from the definition we gave in Section 3.4.2, that a smaller AIC suggests a better fitting model. Thus, the model that includes past observations (defined by Equation 4.3) is a slightly better fit than the model that includes the HW_t term (defined by Equation 4.4). However, in both these measures, the models are quite close to each other. Also, we are only comparing the GLMs with these two measures; they take no account of the extra level of modelling that is required when the exponential smoother is used. The model using past values seems to remove serial correlation more effectively than the one

Model	Intercept	Ti_t	$\sin\left(\frac{2\pi t}{52}\right)$	$\cos\left(\frac{2\pi t}{52}\right)$
Equation 4.3	1.288 (0.105)	-0.000237 (0.000069)	-0.045 (0.024)	-0.132 (0.024)
Equation 4.4	1.455 (0.118)	-0.000270 (0.000073)	-0.105 (0.024)	-0.169 (0.025)

Model	$\sin\left(\frac{4\pi t}{52}\right)$	$\cos\left(\frac{4\pi t}{52}\right)$	$\sin\left(\frac{6\pi t}{52}\right)$	$\cos\left(\frac{6\pi t}{52}\right)$	$\sin\left(\frac{8\pi t}{52}\right)$
Equation 4.3	-0.125 (0.025)	-0.082 (0.023)	0.072 (0.024)	0.105 (0.023)	-0.013 (0.023)
Equation 4.4	-0.163 (0.026)	-0.014 (0.024)	0.099 (0.025)	0.025 (0.024)	-0.089 (0.024)

Model	$\cos\left(\frac{8\pi t}{52}\right)$	$\log(x_{t-1})$	$\log(x_{t-2})$	HW_t
Equation 4.3	-0.148 (0.023)	0.344 (0.034)	0.227 (0.033)	–
Equation 4.4	-0.069 (0.025)	–	–	0.515 (0.040)

Table 4.3: The coefficients fitted to the various terms of models defined by Equations 4.3 (trend, first four harmonics for seasonality and two past observations) and 4.4 (trend, first four harmonics for seasonality and Holt-Winters one-step ahead prediction) fitted to the counts of *Cryptosporidium* from 1991.

using the HW_t term (cf Figures 4.11 and 4.12). The model utilising Holt-Winters is somewhat more parsimonious in the GLM (one variable, to the two variables of past values), although, this should be balanced with the trade-off that a year of data that is lost to fitting the seasonal elements of Holt-Winters. However, given there is so much data, the latter point is less of a concern. Further, it is likely that the Holt-Winters model could be further refined if we analysed the parameters of the smoother more closely. The Holt-Winter's model also has the advantage of allowing straight forward predictions more than one week ahead; in the case of the model using past observations, a recursive calculation would have to be made.

Given both modelling approaches seem to be feasible, we will use both of them in Sections 5.3 and 5.4 and then compare them in Section 5.5. Further work could be done to refine the national model by looking at the residuals and perhaps incorporating past outbreaks or the effects of Foot and Mouth Disease. However, given our focus is on the regional modelling we move directly to that, after considering the effect of including weather variables.

4.4.3 Weather Considerations

Many organisms multiply at a greater rate in warmer temperatures and so we might expect higher levels of them during the summer months. This might ex-

Model	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	θ	θ SE
Eqn. 4.3	1867	835	1032	55	843	832	11	5378	7.543	0.559
Eqn. 4.4	1664	830	834	50	843	833	10	5465	6.350	0.440

Table 4.4: The summary statistics from the models defined by Equations 4.3 (trend, first four harmonics for seasonality and two past observations) and 4.4 (trend, first four harmonics for seasonality and Holt-Winters one-step ahead prediction) fitted to the counts of *Cryptosporidium* from 1991.

plain the structure of the seasonal patterns of *Salmonella*, shown in Figure 4.7. Such patterns are often explained well by seasonal factors or trigonometric terms. However, some organisms may be tied more tightly to the short-term changing environmental conditions. By including variables that measure things such as rainfall and temperature, we can model these short term environmental effects directly. It also provides of a way of modelling weather effects that are one off: say, an unusually heavy rainfall one year, which would not be modelled well by seasonal factors or trigonometric terms. The weather indicators might also serve as a proxy for human activity that might affect reported cases of particular organisms. For instance, if the weather is warm, people are more likely to be engaged in outside activities which could lead to more infections of certain organisms.

Given the biology of *Cryptosporidium*, it seems that a measure of rainfall will be the most likely candidate for affecting the reported cases of it. For instance, a particularly heavy rainfall may carry more oocysts into the water supply leading to more infections.

Unfortunately, gaining long term weather information is not a straight forward task. The obvious place to go for this information in the UK is the MET Office (2008b). They will provide six months of daily data for free, but this is only from 2000 (MET Office 2008c). For longer periods there is a significant charge. Further, when enquires were made about obtaining data at a regional level from

weather stations, it transpired that the data is not stored in a digital form. For these reasons, we cannot include general weather variables.

However, the Met Office does provide free access to two particular data-sets. The first of these is HadUKP series, a data-set that has recorded daily precipitation in the UK in some form since 1776 (Alexander and Jones 2001). From this data-set it is possible to get daily measurements of rainfall in Scotland but only at the national level. We form an appropriate variable below and see if it contributes significantly to the national models. Secondly, the Met Office provides free access to the thirty year averages for rain fall, temperature, etcetera. We see if these averages can replace the trigonometric terms for modelling the seasonal pattern.

Including the HadUKP Rainfall Data in the National Models

Having obtained the daily rainfall data from MET Office (2008c) for Scotland, we calculate weekly totals, where the weeks correspond to those used within the HPS system. We denote the total rainfall for week t by r_t, where the total is measured in centimetres. We extend the national models by including r_t and lags of it:

$$\begin{aligned}\log(\mu_t) &= \beta_0 + \beta_1 r_t + \beta_2 r_{t-1} + \beta_3 r_{t-2} + \beta_4 r_{t-3} + \beta_5 r_{t-4} \\ &\quad + \beta_6 Ti_t + trig(4)_t + \beta_{15}\log(x_{t-1}) + \beta_{16}\log(x_{t-2}) \qquad (4.5)\\ \log(\mu_t) &= \beta_0 + \beta_1 r_t + \beta_2 r_{t-1} + \beta_3 r_{t-2} + \beta_4 r_{t-3} + \beta_5 r_{t-4} \\ &\quad + \beta_6 Ti_t + trig(4)_t + \beta_{15} HW_t \qquad (4.6)\end{aligned}.$$

The coefficients of the models are shown in Table 4.5. We find that r_t and r_{t-2} do not contribute significantly to either national model. Since r_t is not significant, it suggests that it takes over a week for the level of rain at time t to affect the levels of *Cryptosporidium*. As r_{t-2} is not significant, it is reasonable just to extend the national models by the inclusion of r_{t-1}:

$$\begin{aligned}\log(\mu_t) &= \beta_0 + \beta_1 r_{t-1} + \beta_2 Ti_t + trig(4)_t + \beta_{11}\log(x_{t-1}) + \beta_{12}\log(x_{t-2}) \qquad (4.7)\\ \log(\mu_t) &= \beta_0 + \beta_1 r_{t-1} + \beta_2 Ti_t + trig(4)_t + \beta_{11} HW_t. \qquad (4.8)\end{aligned}$$

Model	r_t	r_{t-1}	r_{t-2}	r_{t-3}	r_{t-4}
Equation 4.5	-0.016 (0.009)	0.023 (0.010)	-0.011 (0.010)	0.026 (0.010)	0.023 (0.009)
Equation 4.6	-0.016 (0.010)	0.020 (0.010)	-0.012 (0.010)	0.025 (0.010)	0.029 (0.010)

Model	Intercept	Ti_t	$\sin\left(\frac{2\pi t}{52}\right)$	$\cos\left(\frac{2\pi t}{52}\right)$
Equation 4.3	1.165 (0.114)	-0.000228 (0.000068)	-0.055 (0.024)	-0.178 (0.028)
Equation 4.4	1.328 (0.126)	-0.000259 (0.000073)	-0.116 (0.024)	-0.216 (0.030)

Model	$\sin\left(\frac{4\pi t}{52}\right)$	$\cos\left(\frac{4\pi t}{52}\right)$	$\sin\left(\frac{6\pi t}{52}\right)$	$\cos\left(\frac{6\pi t}{52}\right)$	$\sin\left(\frac{8\pi t}{52}\right)$
Equation 4.3	-0.124 (0.024)	-0.086 (0.023)	0.077 (0.023)	0.106 (0.022)	-0.016 (0.023)
Equation 4.4	-0.161 (0.026)	-0.020 (0.024)	0.105 (0.024)	0.027 (0.024)	-0.090 (0.024)

Model	$\cos\left(\frac{8\pi t}{52}\right)$	$\log(x_{t-1})$	$\log(x_{t-2})$	HW_t
Equation 4.3	-0.147 (0.022)	0.344 (0.034)	0.223 (0.033)	–
Equation 4.4	-0.068 (0.025)	–	–	0.511 (0.039)

Table 4.5: The coefficients fitted to the various terms of models defined by Equations 4.5 (trend, first four harmonics for seasonality, two past observations and rain totals) and 4.6 (trend, first four harmonics for seasonality, Holt-Winters one-step ahead prediction and rain totals) fitted to the counts of *Cryptosporidium* from 1991. The total amount of precipitation for week t is denoted r_t.

The coefficients for these models can be found in Table 4.6. However, in these models r_{t-1} is not significant and so should be left out of the models. This would reduce the models to those that were developed in Section 4.4.2. Since higher lags of rainfall were significant, there may be some way of including the HadUKP rainfall data in such a way that it does contribute significantly to the national models. For instance, future work might look at using a rainfall total from a period of longer than a week.

Thirty Year Averages

In the national models developed (defined by Equations 4.3 and 4.4) seasonality is modelled primarily by eight trigonometric terms. If seasonality could be modelled in fewer terms this would make the models more parsimonious. One way of doing this may be to use the thirty year averages of weather to form seasonality variables. Unlike most data held by the MET office, this data is freely available for most weather stations in Scotland and so we would not have a problem forming regional values (MET Office 2008a). The most recent averages have

Model	Intercept	r_{t-1}	Ti_t	$\sin\left(\frac{2\pi t}{52}\right)$
Equation 4.7	1.234 (0.108)	0.017 (0.009)	-0.000234 (0.000069)	-0.042 (0.024)
Equation 4.8	1.413 (0.121)	0.014 (0.010)	-0.000268 (0.000073)	-0.103 (0.024)

Model	$\cos\left(\frac{2\pi t}{52}\right)$	$\sin\left(\frac{4\pi t}{52}\right)$	$\cos\left(\frac{4\pi t}{52}\right)$	$\sin\left(\frac{6\pi t}{52}\right)$	$\cos\left(\frac{6\pi t}{52}\right)$
Equation 4.7	-0.148 (0.025)	-0.123 (0.025)	-0.083 (0.023)	0.073 (0.024)	0.106 (0.022)
Equation 4.8	-0.182 (0.026)	-0.161 (0.026)	-0.015 (0.024)	0.100 (0.025)	0.027 (0.024)

Model	$\sin\left(\frac{8\pi t}{52}\right)$	$\cos\left(\frac{8\pi t}{52}\right)$	$\log(x_{t-1})$	$\log(x_{t-2})$	HW_t
Equation 4.7	-0.013 (0.023)	-0.149 (0.023)	0.348 (0.034)	0.225 (0.033)	–
Equation 4.8	-0.089 (0.024)	-0.069 (0.025)	–	–	0.516 (0.040)

Table 4.6: The coefficients fitted to the various terms of models defined by Equations 4.7 (trend, first four harmonics for seasonality, two past observations and rain total lagged by one week) and 4.8 (trend, first four harmonics for seasonality, Holt-Winters one-step ahead prediction and rain total lagged by one week) fitted to the counts of *Cryptosporidium* from 1991.

been calculated for 1971-2000, and records the monthly: minimum and maximum temperatures; days of air frost; hours of sunshine; total rainfall; and number of days with more than 1mm of rainfall.

Taking the monthly averages for Scotland MET Office (2008d), we interpolate between the months to form weekly averages. To illustrate this process consider total rainfall. We first divide the total monthly rainfall by the number of days in each month, as the longer months would be expected to have more rain. This gives the average daily rainfall for each month. This step is only done with those measures it makes sense for; for instance we do not average the maximum temperature, though others have done this. The monthly daily rainfall average, R_t, is then taken as a response variable, and we fit a linear model to the data:

$$\begin{aligned} R_t &= \beta_t + \beta_1 \sin(ft) + \beta_2 \cos(ft) + \beta_3 \sin(2ft) + \beta_4 \cos(2ft) + \beta_5 \sin(3ft) \\ &\quad + \beta_6 \cos(3ft), \end{aligned}$$

where $f = \frac{2\pi}{364}$, the β_i are coefficients, and t is the middle day for each month that R_t is recorded for. In this modelling we have shortened the year to 364 days by dispensing with a day in January (note the period of 364 in the trigonometric terms). This means that the number of days in the year is a multiple of fifty-two. Note that we are fitting a model to twelve pieces of data that has seven

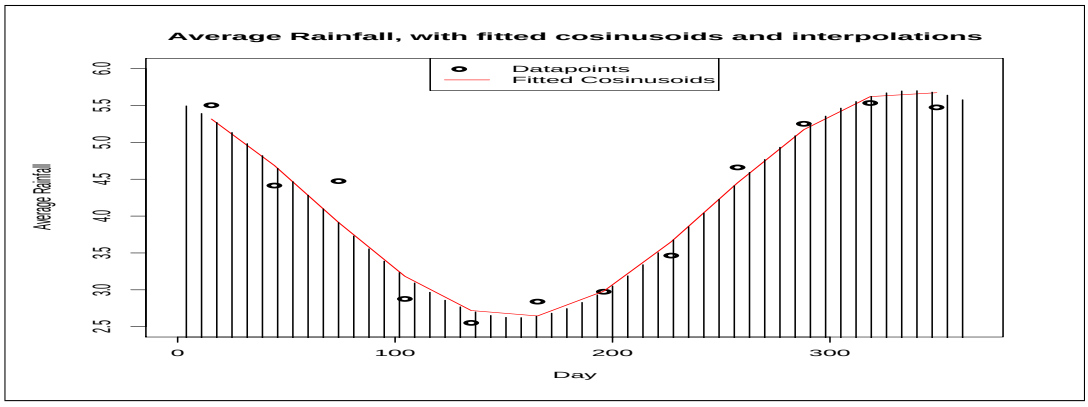


Figure 4.13: The points in black represent the monthly daily average of rainfall calculated from the 30 year (1971-2000) monthly averages of total rainfall. The line in red gives the fit of the linear model defined by Equation 4.9 to these values. The vertical lines in black represent the interpolations from the model corresponding to 52 weekly values to match with the HPS time periods.

parameters. We disregard those harmonics that are not significant. In the above case, only the first harmonic is significant, so the model simplifies to the following:

$$R_t \quad = \quad \beta_t + \beta_1 \sin(ft) + \beta_2 \cos(ft)\,. \tag{4.9}$$

The coefficients of this model are then used to calculate the values for R_t for 52 values, giving the corresponding weekly levels. This process is most clearly seen in a diagram as in Figure 4.13. We then repeat this for the other measures in the thirty year monthly averages, and find that the number of harmonics required for interpolation varies between one and two, as summarised in Table 4.7.

Unfortunately, we know that none of these measures will be able to explain the seasonality in *Cryptosporidium*, since we know from the national models that four harmonics are required to model the seasonal pattern. Further, there is correlation between these different climatic measures: if there are more days of rain, we would expect more rainfall; less trivially, when it is cold, we tend to expect more rain. Due to the correlation between the measures, it is hard to statistically test which is the best measure to be included in the models. This is exacerbated by having to interpolate between the monthly readings to form weekly estimates. Thus, the approach of using the thirty year averages to form a seasonal pattern cannot be used with *Cryptosporidium*; it may work well with some of the other

Weather Measure	Required harmonics
Maximum Recorded Temperature	2
Minimum Recorded Temperature	2
Days of Air Frost	2
of Sunshine	1
Rainfall	1
Days of Rain	1

Table 4.7: The different measures that are recorded in the 30 year meteorological averages, with the number of harmonics that are used in their interpolation, as explained in Section 4.4.3.

organisms considered in Section 4.2.4. Ideally, the averages would be calculated regionally from the raw weekly values for the different climatic variables in each region. Again, the practicalities of obtaining this data from the MET Office might make this investigation impractical.

Weather Data Conclusions

At present there is not sufficient data from the Met Office to investigate further using weather measurements within the national model. Further, due to the bi-modal seasonality, the weather variables are not well suited to modelling the seasonality of *Cryptosporidium*. If in the future the cases of *Cryptosporidium* can be sub-typed into *Cryptosporidium parvum* and *Cryptosporidium hominis*, the seasonality maybe simpler to model, possibly by using weather variables.

4.5 Conclusions

In this Chapter we have considered the context in which we aim to develop a regional exception reporting system. We will be using data routinely collected by Health Protection Scotland (HPS), an organisation tasked with monitoring the health of the Scottish population (Section 4.1). In Section 4.2, we describe how their data collects together reports from micro-biology laboratories across the country. Doctors routinely send samples from their patients to these labs to confirm diagnoses of infections by particular organisms. For certain organisms, typically the rarer and/or more dangerous ones, positive tests must be reported

to HPS. These reports are then aggregated to provide weekly counts of cases of the different organisms. The current exception reporting system used to monitor these weekly counts is detailed in Section 4.2.2. This system is a general purpose one, well suited to modelling a wide range of organisms at the national level. However, we consider some of the problems of applying this system to the regional level in Section 4.2.3. Notably, regional systems are likely to have to deal with large numbers of zero counts and substantial serial correlation. The current national system was not developed with these factors in mind.

To begin the process of developing a regional exception reporting system, we decide to focus our attention on one of six organisms suggested by HPS (Section 4.2.4). A great many of the organisms monitored by HPS occur very infrequently; it would not be very interesting to consider developing an exception reporting system for these. We choose to focus our modelling on *Cryptosporidium*, as its counts are relatively stable. We consider the biological background of *Cryptosporidium* in Section 4.3. A regional exception reporting system for this organism could be of wide benefit, since it is becoming one of the leading sources of water-borne infection around the world.

We adopt a top-down strategy in the modelling of *Cryptosporidium*: we start by modelling its national counts. We begin with a time series decomposition in Section 4.4.1. We find that there is some trend in the counts and that the seasonality is bi-modal. From our considerations in Section 4.3, we know that this bi-modality is likely caused by the counts of *Cryptosporidium* including cases of two particular sub-types: *Cryptosporidium parvum* and *Cryptosporidium hominis*. The first of these primarily accounts for infections originating with animals and the latter for infections originating in humans. Due to these different primary sources of infection, the different sub-types peak at different times in the year.

We then fit a negative binomial GLM to the national counts in Section 4.4.2. We initially fit a model with a linear trend and trigonometric terms to model the seasonality. The trigonometric terms do not model the seasonality perfectly but are much more parsimonious than the alternative of seasonal factors. Besides, because of the bi-modal seasonality of *Cryptosporidium*, we expect modelling its seasonality to be difficult; the trigonometric terms are likely to be well suited to capturing the seasonality of the other organisms. We find that there is significant serial correlation in the national counts, so we try two different approaches

to address this. We first try including lagged observations within the model (as done elsewhere: Brandt, Williams, Fordham, and Pollins (2000)). The second approach utilises including a Holt-Winters predictor as was done with the NHS24 modelling in Chapter 3. Both approaches remove most of the serial correlation but the first seems to be more effective. Either of these models could be combined with an alarm method to create a national exception reporting system for *Cryptosporidium*. This system would be different to the one considered in Section 4.2.2 by dealing with serial correlation. Since both approaches produce reasonable national models, we will use the structure of them as the basis for the regional models we go on to develop in Chapter 4.

We briefly investigate whether weather variables can be used to improve the national models. Unfortunately, it was not possible to gain access to all the data that would have been required to do this thoroughly. Nothing definitive was determined, but there are some suggestions that it might be a fruitful direction for future work.

Thus, this Chapter has served to lay the ground work for developing models of regional counts of *Cryptosporidium*, as we do in the next Chapter. These regional models will then serve as the basis for a regional exception reporting system.

Chapter 5

Regional *Cryptosporidium* Modelling

In this Chapter we fit GLMs to the regional counts of *Cryptosporidium*. These models can then be used as the basis for a regional exception reporting system to regionally monitor *Cryptosporidium* across Scotland. We begin by investigating the differences between the regions in Section 5.1 to help inform our modelling. Since past outbreaks in the data-set can negatively affect the models we fit, we consider outbreaks that may be present in the data, and how we might deal with them, in Section 5.2. We then fit two sets of GLMs to the regional counts: in Section 5.3, GLMs which include past observations in their models; in Section 5.4, GLMs which include a term based on Holt-Winters smoothing. These two sets of models are then compared in Section 5.5. Finally, we draw together lessons from the modelling and suggest directions for future work in Section 5.6.

5.1 Regional Considerations

When modelling at the regional level there are more nuances to be aware of than when modelling at the national level. We have noted previously that outbreaks tend to happen at a regional level; indeed, this is our main reason for developing the regional exception reporting system. This means that we are likely to come across outbreaks which may affect the regional models. In Section 5.2, we detail the 'outbreaks' that will be tested for statistical significance in models that we

will develop. Here, we consider some issues linked with dealing with the smaller counts at the regional level.

Counts at the regional level are not uniformly smaller. Each health board has different populations and geographies which will affect the number of reported cases of *Cryptosporidium*. To convey some of these differences we find the population and area of each health board from the 2001 Census (General Register Office for Scotland 2001). The Scottish Neighbourhood Statistics (2009) are also another good source of such information. The Census contains information on the number of people and area within each postcode sector. By using a database, the postcode sectors can easily be aggregated into their corresponding health boards, leading to the results in Table 5.1. Recall, as was noted early in Chapter 4, that we include Argyll & Clyde in our considerations even though it has now been closed. The data to which we fit regional models has not been updated to reflect this closure and so we still have fifteen health board regions to consider (see Figure 2.8 for further detail).

The population and area values are graphically shown in Figures 5.1 and 5.2 respectively. We note that Glasgow and Lothian have the largest populations, since they contain the two largest cities in Scotland (Glasgow and, Scotland's capital, Edinburgh). The island regions to the north of Scotland – Western Isles, Shetland and Orkney – have very small populations. As one would expect, those places with larger populations have higher weekly rates of reporting – see Table 5.1. In terms of area, Highland covers by far the most and Glasgow the least. Glasgow covers this small area due to the very large population contained within it and the associated heavy administrative burden.

Using the population and area statistics, the density of population in each health board can be calculated. The density of the different health boards is depicted in Figure 5.3, and recorded in Table 5.1: Glasgow has the highest density since it is the city with the largest population among the cities in Scotland. We can compare the density of population with the rate of infection (total number of cases of *Cryptosporidium* divided by population), again recorded in Table 5.1. One would expect those places with higher densities to have a higher rate of infection, since people are on average in closer proximity to each other, allowing infection to spread more quickly and easily. To investigate potential links between rate and density, we plot log-density against rate of infection in Figure 5.4. There

is a weak positive correlation between log-density and rate, with Spearman's statistic being 0.25 for the given data. This suggests, to a limited degree, that as population density increases so does infection rate. However, there is a suggestion of another contributing factor which might explain why health boards such as Highland and Grampian have higher rates than we might expect. This other factor might be linked with the prevalence of cattle in these health boards. Given the biology of *Cryptosporidium* (see Section 4.3), having large numbers of cattle can lead to a greater prevalence of *Cryptosporidium* in rural areas. Thus, the amount of farm land as well as density might influence the rate. In densely populated areas there may be fewer sources of infection, but this infection can be transmitted more easily (i.e. Glasgow); in rural areas there may be more sources of infection, but infection is spread less easily between humans (Highland). Grampian has a mix of large areas of farmland and densely populated areas (for instance Aberdeen, one of the largest cities in Scotland), and so, perhaps, explains why it has the largest rate of infection. However, this is speculation; we would require more detailed data on each health board to confirm this theory. Some time after this analysis was carried out, the work of Pollock, Ternent, Mellor, Smith, Ramsay, and Innocent (2009) came to our attention. They help explain some of the pattern: in rural areas, levels of *Cryptosporidium parvum* will be higher with the greater number of cattle present; in urban areas, levels of *Cryptosporidium hominis* will be higher among greater concentrations of human populations. Thus, it is not a surprise that we expect the rate of infection to be highest in those boards that have both cities and farming areas. Of course, the different regions will also have different weather patterns which will also affect the rate. Further consideration of this is outside our primary focus of developing a reporting system and so we move on.

The weekly rates of reported cases for the period under consideration are shown in Table 5.2. Excluding the island health boards, Borders has the lowest average rate of 0.33 reports per week while Lothian has the highest at 2.45. The time series for these two health board regions are shown in Figure 5.5. With the smaller counts at the regional level, there are more weeks with no reported cases. Indeed, most weeks have *no* cases, as seen in Figure 5.6. The regions with smaller populations have a greater number of weeks with no reports. Often in the health boards with lower weekly rates of reporting, there are runs of weeks with

Health board	Population	%	Area (Hectares)	%	Density	Rate per 10,000	Weekly Reporting Rate
Argyll & Clyde	426,431	8	796,595	10	0.54	1.15	0.94
Ayrshire & Arran	368,149	7	336,721	4	1.09	0.65	0.46
Borders	105,950	2	445,966	6	0.24	0.76	0.15
Dumfries & Galloway	147,625	3	643,333	8	0.23	1.15	0.33
Fife	346,580	7	131,454	2	2.64	0.46	0.31
Forth Valley	282,900	6	280,014	4	1.01	1.13	0.62
Glasgow	841,847	17	55,263	1	15.23	1.95	0.96
Grampian	540,227	11	881,940	11	0.61	3.04	3.15
Highland	208,340	4	2,530,343	32	0.08	1.92	0.77
Lanarkshire	574,365	11	247,601	3	2.32	1.03	1.13
Lothian	779,223	15	175,305	2	4.44	1.04	1.56
Orkney	19,245	<1	98,881	1	0.19	0.52	0.02
Shetland	21,988	<1	143,836	2	0.15	0.00	0.00
Tayside	372,639	7	732,031	9	0.51	0.75	0.54
Western Isles	26,502	<1	299,886	4	0.09	0.00	0.00
Scotland	5,062,011		7,799,170		0.65	1.12	10.94

Table 5.1: The population and area for each health board as calculated from the *2001 Census: Standard Area Statistics*. The density is calculated by dividing the population by area giving the number of people per hectare. Rate gives the number of reported cases of *Cryptosporidium* in 2001 for a given health board, divided by the board's population, multiplied by 10,000. Weekly reporting rates gives the total number of reports received during 2001 divided by fifty-two for each health board.

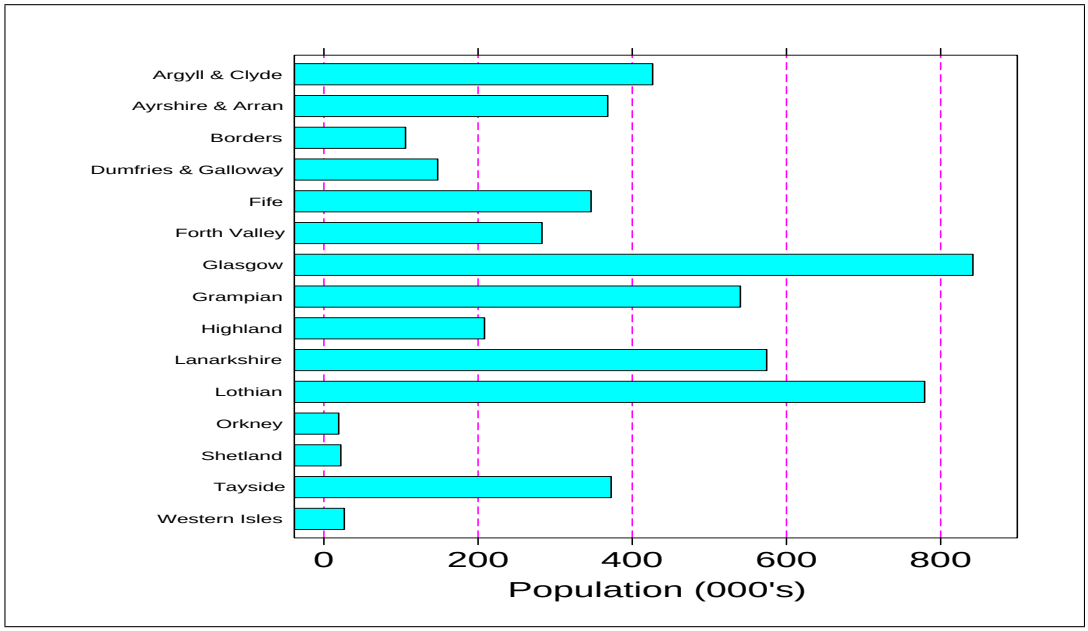


Figure 5.1: The population within each health board, according to the *2001 Census: Standard Area Statistics*.

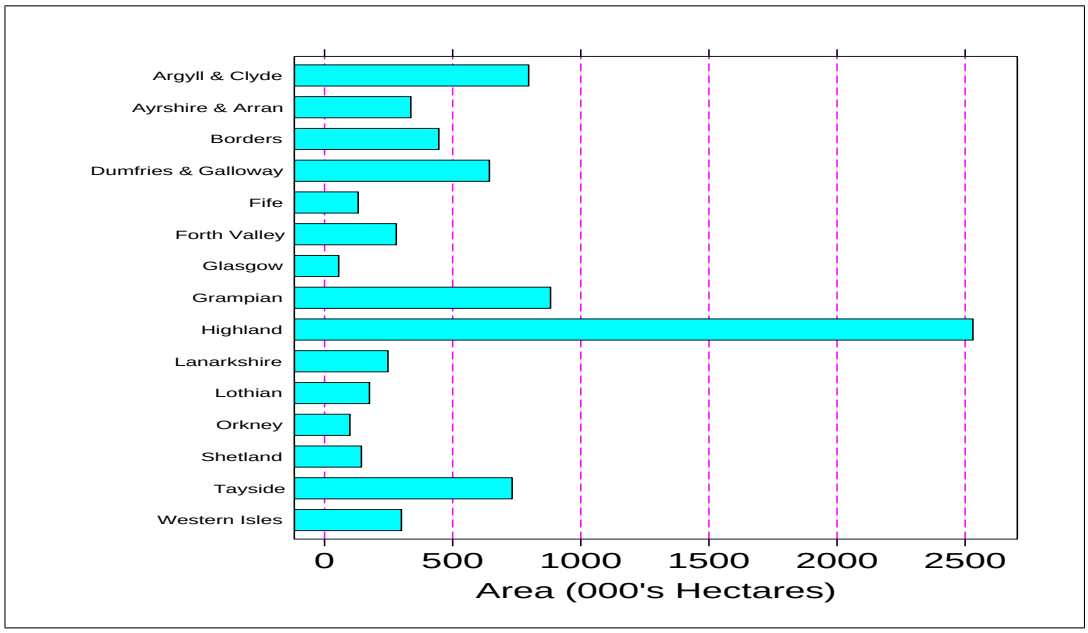


Figure 5.2: The area within each health board according to the *2001 Census: Standard Area Statistics*.

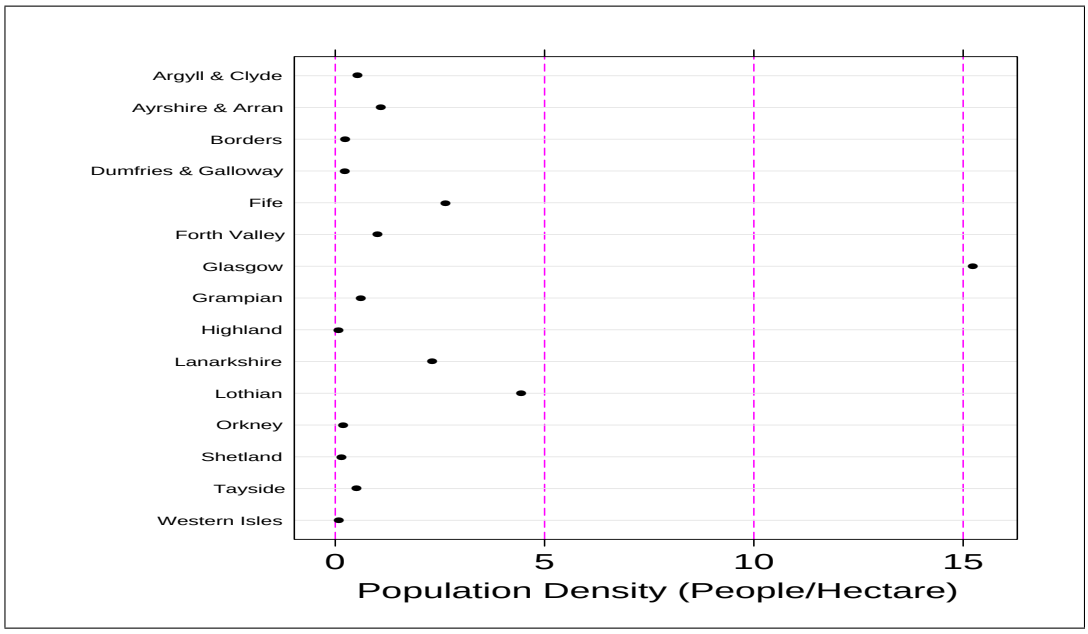


Figure 5.3: The density of populations within each health board according to the *2001 Census: Standard Area Statistics*.

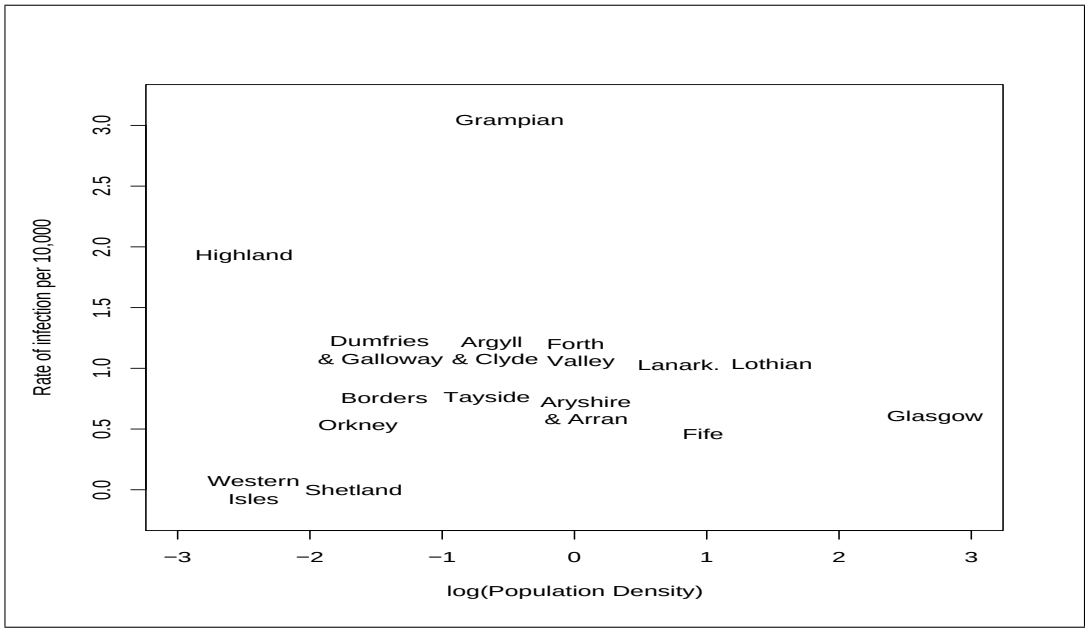


Figure 5.4: The rate of *Cryptosporidium* infection per 10,000 population in 2001 against log-density. Spearman's statistic for these variables is 0.248.

Health board	Weekly Rate of Reported Cases
Argyll & Clyde	1.39
Ayrshire & Arran	0.56
Borders	0.33
Dumfries & Galloway	0.77
Fife	0.44
Forth Valley	0.53
Glasgow	1.53
Grampian	2.37
Highland	0.71
Lanarkshire	1.92
Lothian	2.45
Orkney	0.04
Shetland	0.04
Tayside	0.99
Western Isles	0.03
Scotland	14.10

Table 5.2: The weekly rate of reported cases of *Cryptosporidium* from 1988 to the forty-fifth week in 2007.

no reports; this can be seen quite clearly in a time plot of reported cases from Borders, shown in Figure 5.5. In case this observation can be used to improve our regional models, we create a variable that measures the run length of zero counts – $zLen$. This will be an integer that counts the number of weeks since a week with one or more cases. Similarly, it is possible there will be clustering of non-zero reporting weeks and so we create a $pLen$ variable, a integer variable that counts the number of past weeks since a zero count. These variables are most clearly understood from Table 5.3, where values of $zLen$ and $pLen$ are shown against a short series of counts from Glasgow.

5.2 Regional Outbreaks

When an outbreak happens at a regional level, we expect the counts in that region to be higher than expected. For instance, consider Tayside in Figure 5.7: in 2005 there is a week with a count of fifty reported cases, in contrast to the weekly reporting rate of one case per week – this is very likely to be an

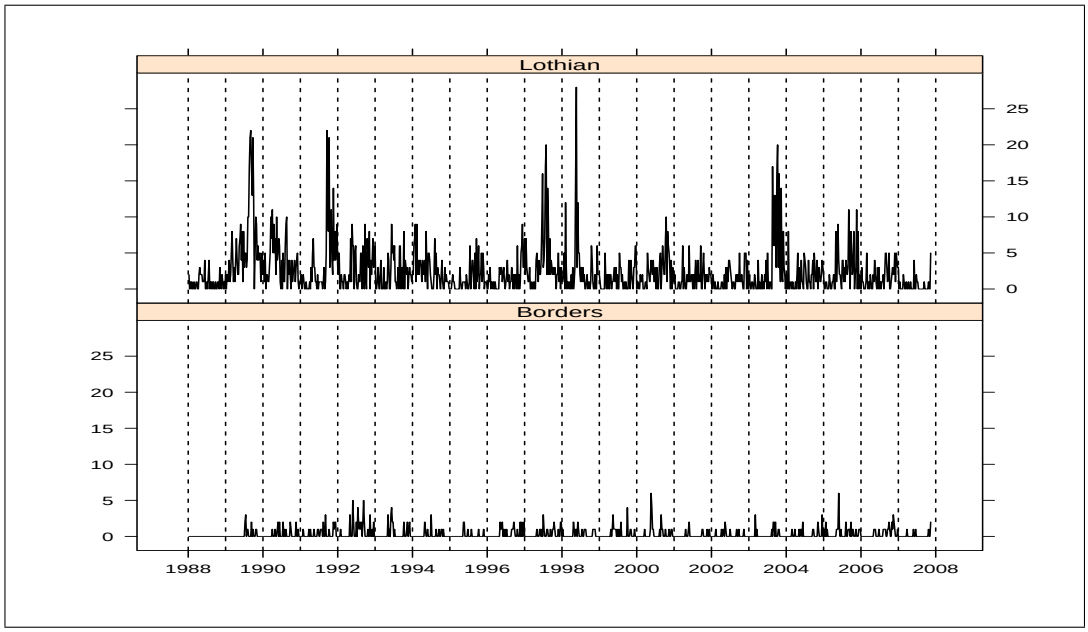


Figure 5.5: The reported cases of *Cryptosporidium* for Lothian and Borders.

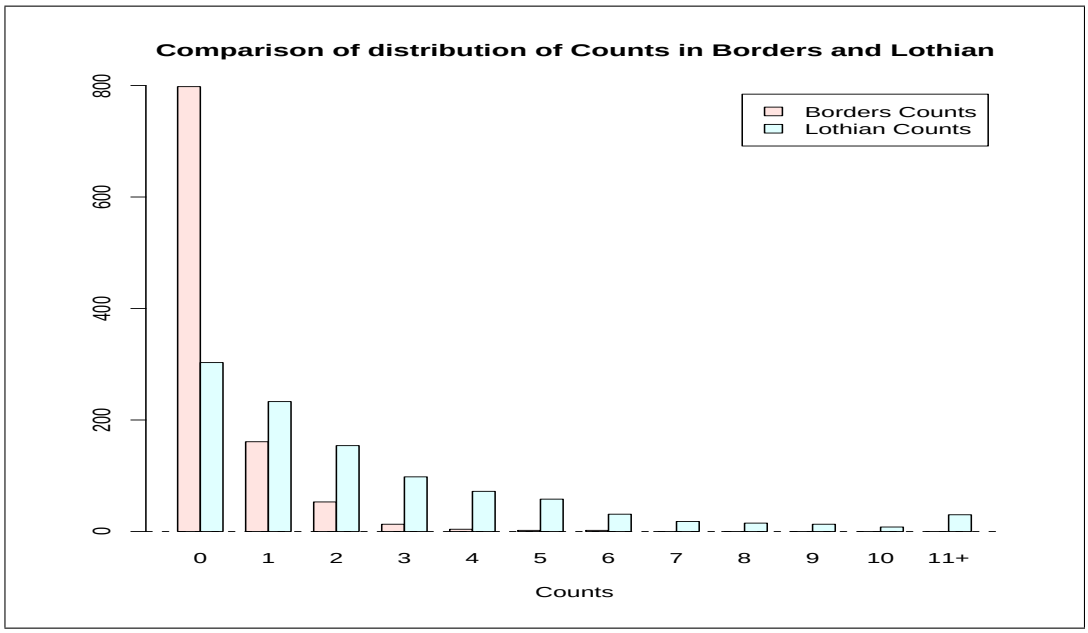


Figure 5.6: The frequency of different counts per week of reported cases of *Cryptosporidium* for Borders and Lothian.

Year	Week	Reported Cases	$zLen$	$pLen$
2005	28	1	0	5
2005	29	0	0	6
2005	30	0	1	0
2005	31	0	2	0
2005	32	0	3	0
2005	33	0	4	0
2005	34	1	5	0
2005	35	5	0	1
2005	36	4	0	2
2005	37	4	0	3
2005	38	1	0	4
2005	39	0	0	5
2005	40	3	1	0
2005	41	0	0	1

Table 5.3: Example *Cryptosporidium* counts from Glasgow, with the corresponding values for $zLen$ and $pLen$ which measure the run lengths of zero and non-zero reporting periods.

outbreak. While outbreaks are rarely so obvious, ignoring them could lead to the development of poor models. Thus, a number of outbreaks, both identified by us and ones identified in HPS literature, will be tested for statistical significance in our regional models. For those weeks that were considered to be part of an outbreak, a factor is fitted to them, with a factor level for each week. The details of the 'outbreaks' that will be tested for statistical significance can be found in Table 5.4. This list is far from exhaustive but with twenty years of counts for twelve health board areas, it is unfeasible to identify all outbreaks. When the models are used for prediction and an identified exceedance is validated as a legitimate outbreak, the appropriate model should be updated with a factor to reflect this outbreak.

An alternative method for dealing with outbreaks is to fit a GLM as normal; the resulting residuals can then be used to refit the GLM, weighting observations by the inverse of the original residuals. This means that observations with larger residuals will contribute less to the model fit, which should result in fitted models that are less affected by outbreaks.

Region	Year	Weeks	Evidence for Outbreak
Fife	2006	18-19	–
Glasgow	2000	16-21	Health Protection Scotland (HPS) (2000b)
	2003	33	Health Protection Scotland (HPS) (2003a)
	2006	16	Health Protection Scotland (HPS) (2002c)
Grampian	1992	10	–
	1995	8	–
	2001	3	Mukerjee (2002)
Lanarkshire	1998	18-21	–
	2003	33	–
Lothian	2002	46-48	Health Protection Scotland (HPS) (2002a)
Tayside	1994	45-46	Health Protection Scotland (HPS) (1994)
	2000	29	Health Protection Scotland (HPS) (2000a)
	2002	34-37	Health Protection Scotland (HPS) (2002b)
	2005	16-20	Health Protection Scotland (HPS) (2005)

Table 5.4: Potential outbreaks that we test for significance in the regional models.

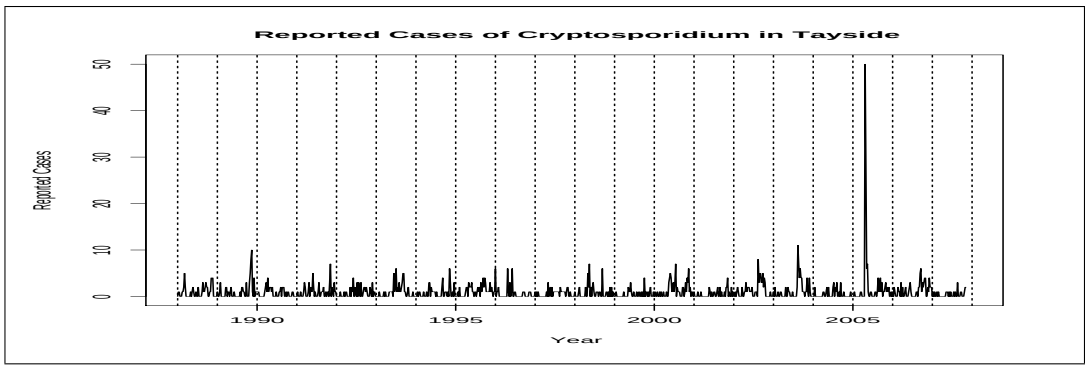


Figure 5.7: The reported cases of *Cryptosporidium* in Tayside, with a very likely outbreak cases of *Cryptosporidium* in 2005.

Each week, HPS publishes a weekly digest which 'contains current news and articles as well as surveillance reports' (Health Protection Scotland (HPS) 2008a). Most of the the recent reports can be searched electronically (Health Protection Scotland (HPS) 2008b; Health Protection Scotland (HPS) 2008a). This allows us to search for identified outbreaks of *Cryptosporidium* within the time range of counts we are considering.

5.3 Regional Modelling with Past Terms

At the regional level, we only consider the health boards on the mainland of Scotland, to the exclusion of Orkney, Shetland and Western Isle health boards. As these are islands, they have very small and inconsistent counts as shown in Figure 5.8. For these health boards, setting an arbitrary number as an exceedance level would likely be sufficient as a detection system. In this Section, we consider the development of regional models for the remaining health board regions and utilise past observations to deal with serial correlation. In the modelling carried out here, data from 1988 till week 45 in 2007 will be used.

5.3.1 GLM Fitted Models

For each region, we test eight seasonal harmonics, a linear trend, $zLen$, $pLen$, and up to five past observations for statistical significance. We chose to use up to five past observations since an individual is rarely infectious for more than three weeks and so we would expect most serial correlation to captured sufficiently within this window – see Section 4.3. The resulting models can be found in Table 5.5, with the factor levels fitted to the outbreaks in Table 5.6 and summary statistics for the model fits in Table 5.7.

Comments on Fitted Models

To compare the seasonality fit to the twelve regions is quite a complex task. To ease the process, we apply hierarchical cluster analysis between the seasonal patterns that result from the trigonometric harmonics fitted to each region. This analysis groups those regions together which have similar seasonal patterns. Hierarchical clustering applied to a series of n-objects produces a series of nested

Region	Intercept	$\sin\left(\frac{2\pi}{52}\right)$	$\cos\left(\frac{2\pi}{52}\right)$	$\sin\left(\frac{4\pi}{52}\right)$	$\cos\left(\frac{4\pi}{52}\right)$	$\sin\left(\frac{6\pi}{52}\right)$	$\cos\left(\frac{6\pi}{52}\right)$
Argyll & Clyde	0.011 (0.129)	-0.021 (0.057)	-0.070 (0.060)	-0.173 (0.058)	-0.092 (0.059)	0.081 (0.058)	0.151 (0.058)
Ayrshire & Arran	-0.737 (0.097)	-0.109 (0.084)	-0.349 (0.090)	-0.350 (0.087)	-0.224 (0.085)	0.157 (0.085)	0.196 (0.085)
Borders	-1.431 (0.087)	-0.536 (0.108)	-0.624 (0.107)	-0.629 (0.104)	0.030 (0.098)	–	–
Dumfries & Galloway	0.136 (0.095)	0.045 (0.067)	-0.564 (0.077)	-0.416 (0.070)	-0.251 (0.073)	0.084 (0.070)	0.147 (0.070)
Fife	-1.140 (0.082)	-0.333 (0.088)	-0.416 (0.094)	-0.168 (0.088)	-0.207 (0.088)	–	–
Forth Valley	-0.944 (0.121)	-0.354 (0.092)	-0.292 (0.092)	-0.349 (0.090)	-0.247 (0.096)	0.103 (0.090)	0.174 (0.090)
Glasgow	-0.237 (0.114)	-0.059 (0.056)	-0.213 (0.057)	-0.239 (0.056)	-0.090 (0.055)	0.209 (0.055)	0.163 (0.056)
Grampian	-0.076 (0.115)	-0.049 (0.050)	-0.135 (0.051)	-0.115 (0.050)	-0.130 (0.051)	0.188 (0.050)	0.175 (0.051)
Highland	-0.435 (0.066)	-0.399 (0.080)	-0.518 (0.075)	-0.472 (0.075)	-0.014 (0.078)	0.211 (0.075)	0.139 (0.073)
Lanarkshire	0.143 (0.125)	-0.056 (0.050)	-0.065 (0.050)	-0.200 (0.050)	-0.126 (0.051)	0.075 (0.051)	0.195 (0.050)
Lothian	0.185 (0.126)	-0.257 (0.052)	-0.148 (0.048)	-0.171 (0.049)	-0.019 (0.050)	0.035 (0.049)	0.115 (0.048)
Tayside	-0.607 (0.066)	-0.360 (0.066)	-0.244 (0.064)	-0.263 (0.063)	-0.066 (0.065)	0.051 (0.063)	0.166 (0.063)

Region	$\sin\left(\frac{8\pi}{52}\right)$	$\cos\left(\frac{8\pi}{52}\right)$	Linear Trend	$zLen$	$pLen$	$FMDInfect$
Argyll & Clyde	-0.133 (0.059)	-0.227 (0.059)	-0.000842 (0.000158)	–	–	–
Ayrshire & Arran	–	–	–	-0.054 (0.018)	–	–
Borders	–	–	–	–	–	–
Dumfries & Galloway	-0.065 (0.068)	-0.147 (0.068)	-0.000954 (0.000169)	-0.065 (0.017)	–	-0.675 (0.341)
Fife	–	–	–	–	–	–
Forth Valley	0.039 (0.089)	-0.281 (0.088)	–	-0.036 (0.017)	–	–
Glasgow	–	–	-0.000348 (0.000137)	–	–	–
Grampian	-0.086 (0.051)	-0.309 (0.052)	-0.000401 (0.000130)	–	-0.014 (0.007)	–
Highland	-0.166 (0.071)	-0.153 (0.071)	–	-0.044 (0.021)	–	–
Lanarkshire	-0.013 (0.050)	-0.193 (0.050)	-0.000760 (0.000136)	–	-0.040 (0.013)	–
Lothian	-0.042 (0.049)	-0.117 (0.048)	-0.000520 (0.000129)	0.144 (0.046)	–	–
Tayside	-0.058 (0.062)	-0.186 (0.061)	–	–	–	–

Region	$\log(x_{t-1}+1)$	$\log(x_{t-2}+1)$	$\log(x_{t-3}+1)$	$\log(x_{t-4}+1)$	$\log(x_{t-5}+1)$
Argyll & Clyde	0.289 (0.065)	0.292 (0.065)	0.168 (0.066)	0.180 (0.066)	–
Aryshire & Arran	0.130 (2.118)	–	–	–	–
Borders	0.536 (0.144)	–	–	–	–
Dumfries & Galloway	–	–	–	–	–
Fife	0.400 (0.126)	0.302 (0.128)	–	–	–
Forth Valley	0.122 (0.128)	0.287 (0.121)	0.375 (0.119)	–	–
Glasgow	0.353 (0.059)	0.333 (0.061)	0.078 (0.062)	-0.004 (0.061)	0.186 (0.060)
Grampian	0.129 (0.052)	0.295 (0.050)	0.285 (0.050)	0.263 (0.050)	0.114 (0.051)
Highland	–	–	–	–	–
Lanarkshire	0.351 (0.056)	0.247 (0.054)	0.217 (0.053)	0.085 (0.053)	0.114 (0.053)
Lothian	0.340 (0.055)	0.262 (0.049)	0.210 (0.048)	–	–
Tayside	0.377 (0.074)	0.312 (0.072)	–	–	–

Table 5.5: The GLMs fitted in Section 5.3.1 to the *Cryptosporidium* data in the different regions, utilising past observations to deal with serial correlation. The tables give the coefficients of the variables and their associated standard errors in brackets.

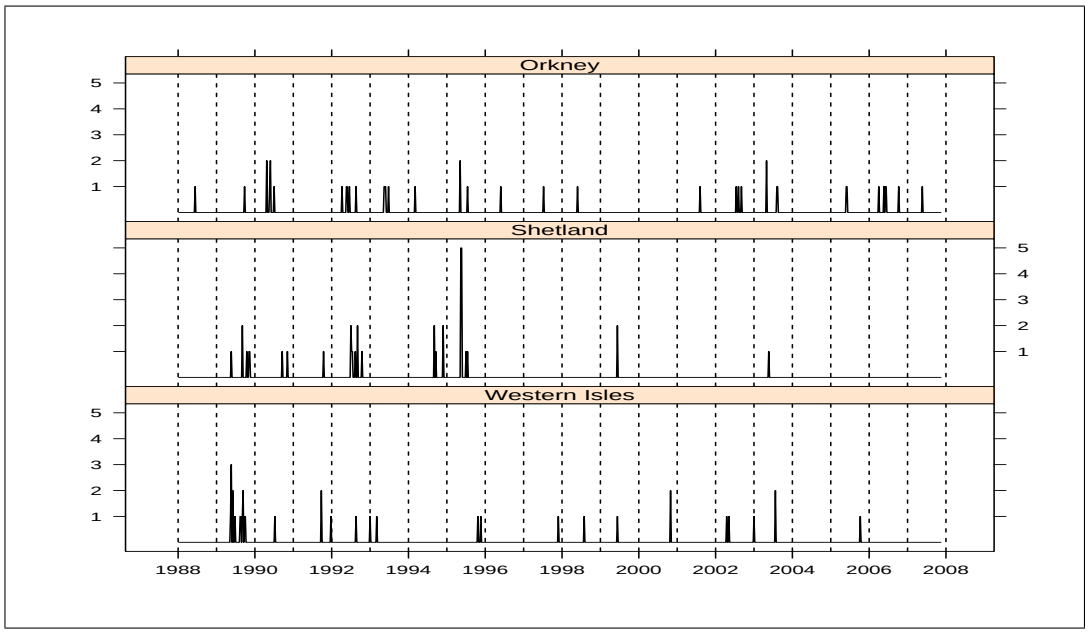


Figure 5.8: The counts of reported cases of *Cryptosporidium* from Orkney, Shetland and Western Isle health boards – small island health boards to the north and west of Scotland.

Region	Year	Start Week	Week 1	Week 2	Week 3	Week 4	Week 5
Fife	2006	18	3.373 (1.090)	–	–	–	–
Glasgow	2003	33	1.912 (0.852)	–	–	–	–
Grampian	1995	8	3.590 (0.824)	–	–	–	–
Lanarkshire	1998	18	1.977 (0.769)	–	–	–	–
	2003	33	2.506 (0.772)	–	–	–	–
Tayside	2000	29	1.883 (0.809)	–	–	–	–
	2005	16	4.300 (0.726)	2.070 (0.779)	–	–	–

Table 5.6: The factor level values for outbreaks found significant, fitted to regions when using past observations within the GLMs fitted in Section 5.3.1.

Model	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	θ	θ SE
Argyll & Clyde	1261	980	281	22	980	967	13	2997	1.365	0.140
Ayrshire & Arran	864	762	102	12	980	972	8	1811	0.881	0.143
Borders	773	659	114	15	980	975	5	1419	1.027	0.229
Dumfries & Galloway	1147	894	253	22	980	969	11	2138	1.889	0.350
Fife	831	737	94	11	980	973	7	1683	0.932	0.167
Forth Valley	866	731	135	16	980	968	12	1836	0.750	0.107
Glasgow	1418	1030	388	27	980	967	13	3093	1.727	0.197
Grampian	1500	1079	421	28	980	964	16	3730	1.566	0.143
Highland	1049	877	173	16	980	971	9	2169	1.422	0.225
Lanarkshire	1457	1046	411	28	980	963	17	3417	1.874	0.190
Lothian	1435	1074	361	25	980	967	13	3857	1.583	0.136
Tayside	1340	960	380	28	980	967	13	2448	2.042	0.338

Table 5.7: Summary statistics from the negative binomial GLMs utilising past observations fit in Section 5.3.1. DoF = Degrees of Freedom. SE = Standard Error.

partitions. At the lowest level of the partitioning, there are n partitions each containing one object. An 'agglomeration algorithm' is then used to decide which two objects are 'closest' together; the two closest objects are then grouped into one partition. We are then left with $n-1$ partitions and the agglomeration algorithm combines the next closest two partitions. This process repeats until the original n partitions are nested within in a single partition that contains all of the original objects. The user of hierarchical clustering is then left to decide at which level of the partition merging they will form their clusters from. This can be most clearly seen from a *dendrogram*, which is a diagrammatic representation of the merging of partitions that occurs with each successive application of the agglomeration algorithm. For an example of a dendrogram see the top of Figure 5.9: at the bottom of the dendrogram we have the n objects which are then combined together until they are contained within a single group at the top of the dendrogram. The user decides the clusters by making a horizontal cut across the dendrogram; thus in the dendrogram in Figure 5.9, three clusters will be formed with the chosen cut. The height on the dendrogram measures how 'far' objects are apart – thus, in the dendrogram, LO and TY are closer together than FV and HG. For information on cluster analysis, see Everitt and Dunn (2001). The

choice of agglomeration algorithm can have quite an effect on the clusters found. We use the 'complete linkage' agglomeration algorithm, which takes the distance between two partitions as the maximum distance between any element of the first partition and any element of the second partition. Using this measure, the two partitions that are the closest together are combined into one partition at each application of the algorithm. Other measures of 'inter-cluster dissimilarity' exist, such as 'single linkage'. However, complete linkage is reasonably robust, so we use that method here.

We use the `hclust` procedure in `R` with complete linkage to perform our hierarchical clustering. When looking at seasonality one can looked at the scaled or unscaled seasonal patterns implied by the trigonometric terms. The different views stress different elements of the seasonality: the former focuses on the pattern through the year, while the latter focuses on the magnitude of the seasonal effects.

We define the seasonal value $Seas_{HT}$ for health board H for week T (taking integers between 1 and 52) by:

$$Seas_{HT} = \sum_{i=1}^{k_H} \beta_{Hi} \sin\left(\frac{2\pi i T}{52}\right) + \sum_{j=1}^{k_H} \beta_{Hj} \cos\left(\frac{2\pi j T}{52}\right), \tag{5.1}$$

where the β_{Hi} and β_{Hj} are the coefficients of the i^{th} and j^{th} harmonics of the sine and cosine terms respectively for healthboard H, contained in Table 5.7. The term k_H gives the number of harmonics used in the model for health board H.We then define the vector $Seas_H$ by:

$$Seas_H = (Seas_{H,1}, Seas_{H,2}, \ldots, Seas_{H,52}). \tag{5.2}$$

This means that $Seas_H$ contains the unscaled seasonal pattern (on the predictor scale) for a year (52 weeks) fitted by the trigonometric terms in Table 5.7. We start applying cluster analysis to this set of twelve seasonal vectors (one for each health board). The corresponding dendrogram is shown in the top of Figure 5.9. We make a cut at the height of 2.6 which results in the four groups shown at the bottom of Figure 5.9. The largest groupings seemed to be explained primarily by a split between the East (Fife, Forth Valley, Lothian, and Tayside) and the West (Ayrshire & Arran, Argyll & Clyde, Glasgow, Grampian and Lanark-

shire) in Central Scotland. The only inconsistent region in these two groupings is Grampian, which, while in the East, is included in the West grouping. This leaves two remaining groupings: Highland and Borders; Dumfries & Galloway. These groupings have the largest seasonal amplitude which may be linked with these regions having higher levels of farming activity than other regions. All regions have a peak between weeks 19 and 21; other peaks and troughs in the year vary widely across the different regions. The groupings seemed linked with how big the peaks are in the Summer compared to the Spring peak: for those regions in the East, the Summer peaks are around the same size if not bigger than the Spring peak (Fife is notable as having a much larger peak in the Summer than the Spring peak); in the other groups the Summer peaks are smaller than the Spring peak.

Next, we apply cluster analysis to the scaled patterns. Define the scaled seasonal pattern $SSeas_H$ for health board H by:

$$SSeas_H = \frac{1}{\max Seas_H - \min Seas_H} \left(Seas_H - \min Seas_H\right), \tag{5.3}$$

where min and max give the smallest and largest elements of $Seas_H$ respectively. Thus, $SSeas_H$ is a vector of the seasonal pattern $Seas_H$ standardised to range between zero and one. We carry out cluster analysis on the set of twelve scaled seasonal vectors. The resulting dendrogram and groupings are shown in Figure 5.10. We cut the dendrogram at a height of 1.4. Again, four groupings are suggested: Argyll & Clyde, Lanarkshire and Grampian; Aryshire & Arran, Dumfries & Galloway, and Glasgow; Forth Valley, Highland, Lothian and Tayside; Borders and Fife. As noted above, the scaled seasonal patterns focus on the similarities in the patterns of the seasonal cycle; thus, Borders and Fife are likely grouped together because they are the only two regions with only two harmonics fitted to them. Similarly, Aryshire & Arran and Glasgow are together, which only have three harmonics fitted to them. However, it is perhaps inconsistent that Dumfries & Galloway, a region with four harmonics fitted, is included within in this latter group. Again, a main distinguishing characteristic appears to be how high the Summer peak is in comparison to the Spring peak (between weeks 19-22): in the first two groups the Summer peak is lower than the Spring peak; in the latter two groups the Summer peak is higher, sometimes exceeding the Spring peak.

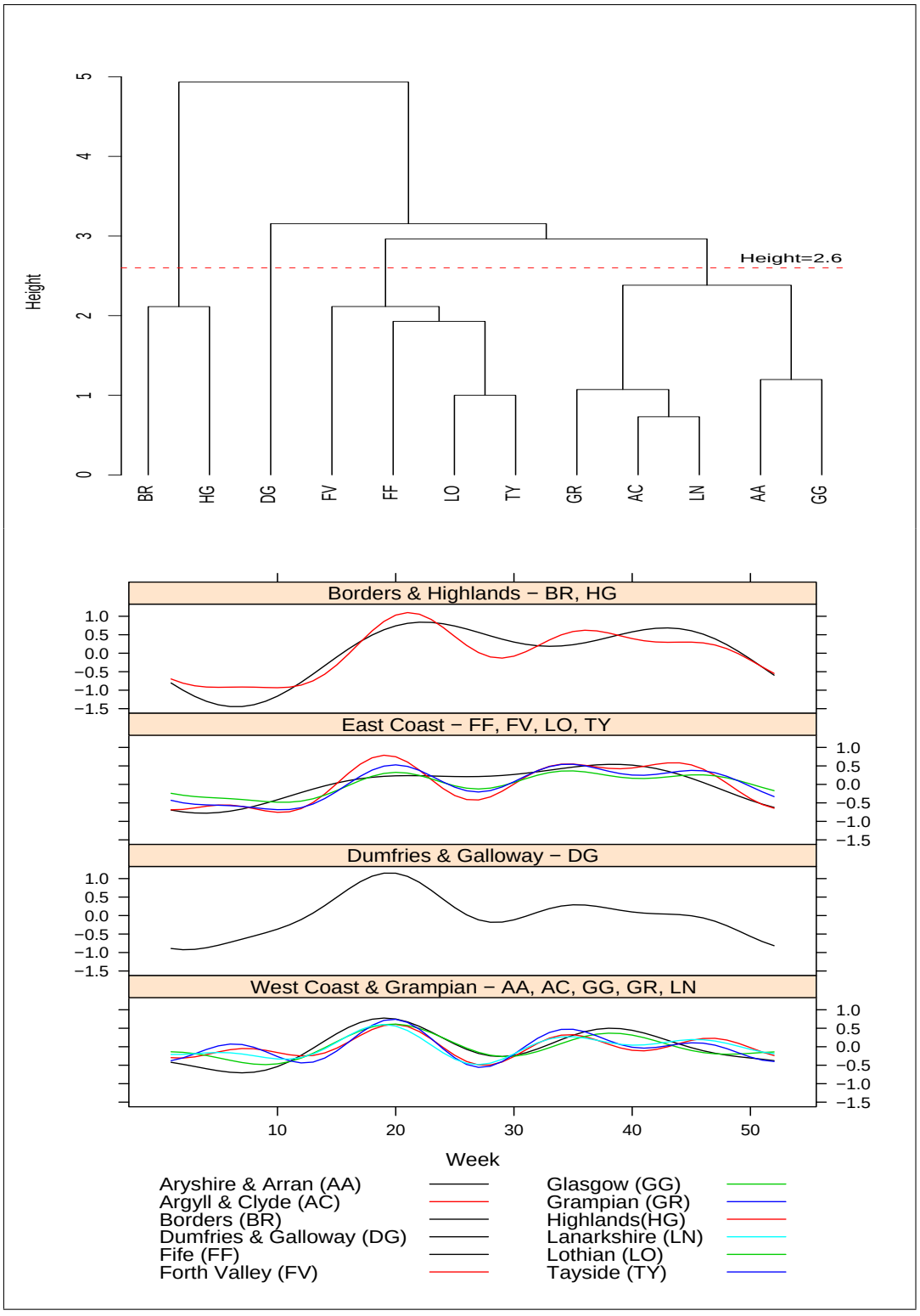


Figure 5.9: Hierarchical clustering applied to the unscaled seasonal patterns fit by the trigonometric terms, on the predictor scale, from the models fit in Section 5.3.1. The top plot is a dendrogram corresponding to the clustering. The red dashed line on the dendrogram represents where we have 'cut' it to form our clusters of regions with similar seasonality. The bottom plot shows the seasonality of each region grouped by the clustering suggested within the dendrogram.

Both sets of seasonal groupings are consistent with the reported findings in Pollock et al. (2009), where they record that there are more cases of *Cryptosporidium hominis* in the East than the West. We noted in Section 4.4.1 that *C. hominis* drives the Summer peak, so it is reasonable that one of the key differences between these groups is the Summer peak. They also found a difference between rates of infection in the North and South, which our groupings reflect to a degree (Borders and Dumfries & Galloway, in the South, are in different groups to the one containing the most northerly board, Highland).

We can also look at the seasonal patterns exponentiated, allowing us to see the seasonal cycles on the scale of the counts. Unfortunately, given the nature of the exponential function, the cluster analysis focuses on the differences between the peaks, which become more pronounced when exponentiated and practically ignores the differences between the seasonal troughs. However, no difference is suggested in the groupings of the unscaled seasonal patterns.

Six regions have a linear trend fitted within their models: Argyll & Clyde, Dumfries & Galloway, Glasgow, Grampian, Lanarkshire, and Lothian. In all these cases the trend is negative and small, but still significant, with the same orders of magnitude. This suggests that the number of reported cases of *Cryptosporidium* is decreasing, as was suggested by the national model in Section 4.4. These six regions are among the seven boards with the largest reporting levels; thus, the other regions might also have a decreasing trend that the modelling here does not have the power to detect.

Only five regions have the $zLen$ term (the term that measures the run length of zeros): Aryshire & Arran, Dumfries & Galloway, Forth Valley, Highland and Lothian. In all the regions, except for Lothian, the $zLen$ term is negative; these regions have the smallest average weekly reporting rates apart from Borders and Fife. This suggests that in these regions the GLM over predicts during a run of zeros. This is perhaps to be expected, given the covariates we are using, as zero will never be predicted by the mean. The positive value for Lothian might be because it is highly populated region, second only to Glasgow (see Figure 5.1): a zero for Lothian indicates that the mean is likely to be positive soon, in contrast to the other regions where a zero is more likely to indicate a subsequent zero.

Two of the twelve regions include the $pLen$ term in their models (the term that measures the run length of positive counts): Grampian and Lanarkshire. In

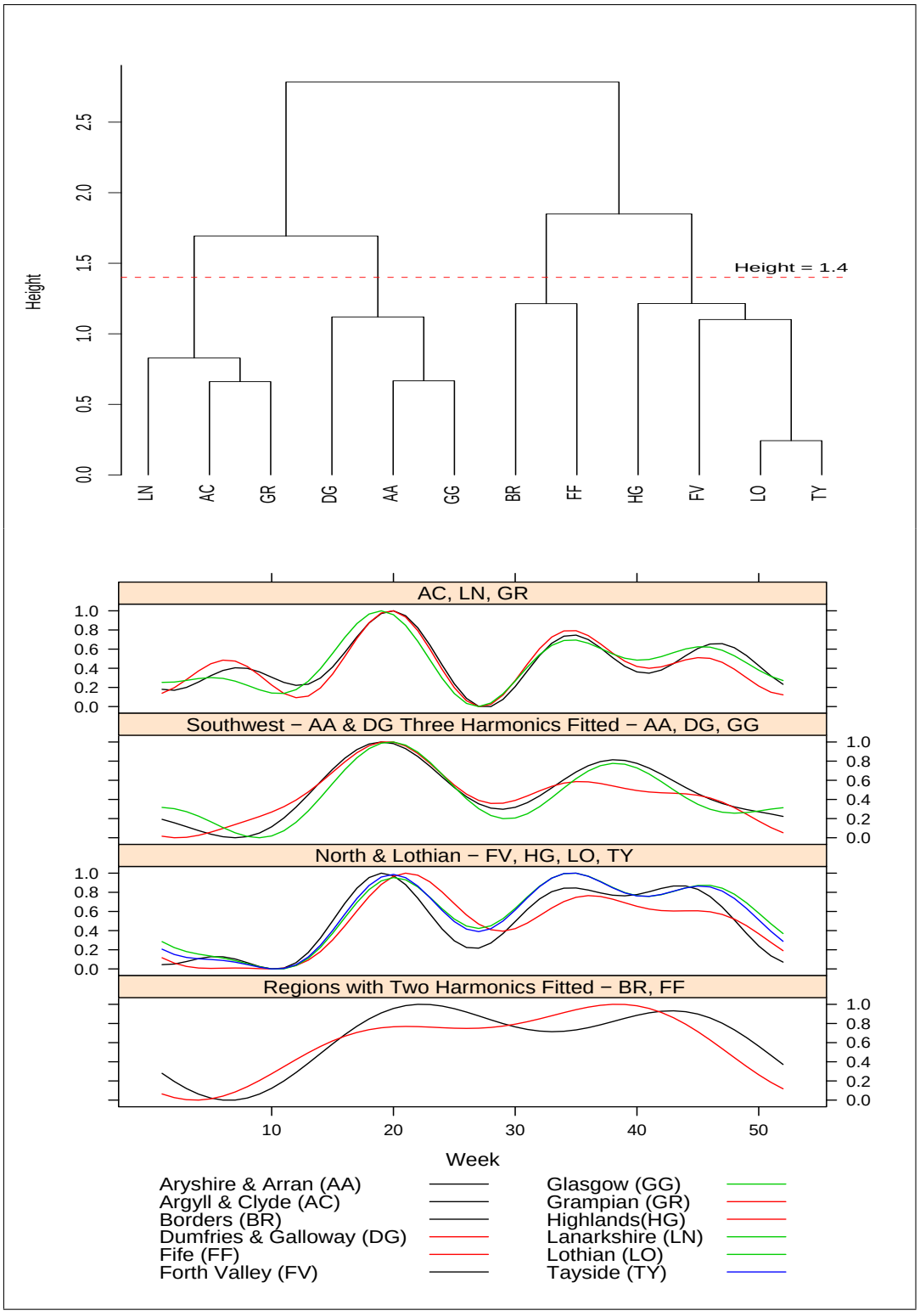


Figure 5.10: Hierarchical clustering applied to the scaled seasonal patterns fit by the trigonometric terms, on the predictor scale, from the models fit in Section 5.3.1. The seasonal patterns have been scaled to go between zero and one before the clustering has been applied. The top plot is a dendrogram corresponding to the clustering. The red dashed line on the dendrogram represents where we have 'cut' it to form our clusters of regions with similar seasonality. The bottom plot shows the seasonality of each region grouped by their clustering given by the dendrogram.

both cases, *pLen* is negative and small in magnitude. This might indicate that these models over-predict and this term serves as a corrective to that. These models also have the maximum number of past values and trend fitted, perhaps indicating that trend is not dealt with adequately in these regions.

We note in Section 4.4.1 that Borders and Dumfries & Galloway have been shown to have a significantly lower rate of infection of *Cryptosporidium* due to the outbreak of Foot and Mouth Disease (FMD) in 2001. To check if this affects our models we include a factor ($FMDInfect$) for those weeks in 2001 when FMD was considered active in these boards (Strachan, Ogden, Smith-Palmer, and Jones (2003) state weeks 13-38); thus:

$$FMDInfect = \begin{cases} 1 & \text{Weeks } 13-38, 2001 \\ 0 & \text{otherwise} \end{cases} .$$

This factor was only significant for Dumfries & Galloway. In this health board, the coefficient was negative (-0.675) indicating that reported cases of *Cryptosporidium* were lower during this time as we would expect. Strachan et al. (2003) also suggest that the infection rate climbs back up during 2002. We included a factor to reflect this, in case it affected the models, but the factor was found to be non-significant.

Up to five past values are included in the models for each region. To allow for easy comparison, their coefficients are plotted in Figure 5.11. Two regions utilise no past values: Dumfries & Galloway and Highland. This perhaps is sensible when we consider that these regions have the smallest densities of populations: it is harder for infection to spread between humans and so there is likely to be less correlation between successive reports of *Cryptosporidium*. In seven of the other regions, the pattern is generally as we would expect in the coefficients of the past values: there is either only a single past value included in the model or the coefficients get smaller as the order of the past value increases. This is intuitive as we expect the reports of cases for this week to correlate most closely with the week just gone, less so with the week before, and even less with the week before that and so on. The pattern for Glasgow is a little odd, with the first three past values' coefficients decreasing in size as we would expect, the fourth past value being very small and the fifth being bigger than the third. This might be

an artifact of the relatively very high population density in Glasgow (see Figure 5.3). It is speculative but it might be possible that once a source of infection has ended in one place in Glasgow, that 'stream' of infection might have gone around the population to come back to the original source in around five weeks. In Grampian, the second coefficient is bigger than the first but then the coefficients decay as we would expect. This might be linked with the combination of rural and dense city regions contained in Grampian, giving it its central position in Figure 5.4; this is discussed in Section 5.1. No obvious explanation seems apparent for the increasing size in the coefficients of the past values fitted in the Forth Valley.

Details of the outbreaks fitted are shown in Table 5.6. It seems the effect of most outbreaks are felt generally for a week, or two at most. This might be to do with the biology of *Cryptosporidium* and the time it takes for a sample to be tested. Of course, the effect on prediction of an outbreak will be decreased by the use of past values in our models.

Model Diagnostics

We now check the acf of the deviance residuals, to see if the residuals are distributed like white noise, indicating a good model fit. In most of the regions, this is the case: the low order correlations are not significant (or only just), and the higher order correlations that are significant are isolated at 'meaningless' lags (i.e. none of these acfs have a significant correlation at lag 52, which might indicate some residual seasonality). Tayside has a significant correlation at lag five, which might be removed by fitting up to five past values in Tayside's model; however, the extra past values were not found to be significant, so they were left out of the model. Glasgow has a correlation at lag 52 which is just significant. Since more harmonics were not found significant, and since the the correlation at lag 52 is just significant, we do not worry about it further. In Grampian, nearly all the correlations in the acf are positive, with a possible suggestion of a linear decay, suggesting a trend in the residuals These artifacts are likely due to the very low counts that have occurred in Grampian, during 2006 and 2007, compared to the previous years. The acf of Lothian's deviance residuals have small significant correlations at lags 4, 5, 9 and 10, suggesting some residual serial correlation, possibly for a similar reason to Grampian. These correlations might be removed by including x_{t-4} and x_{t-5} in the model; however, we do not include those here,

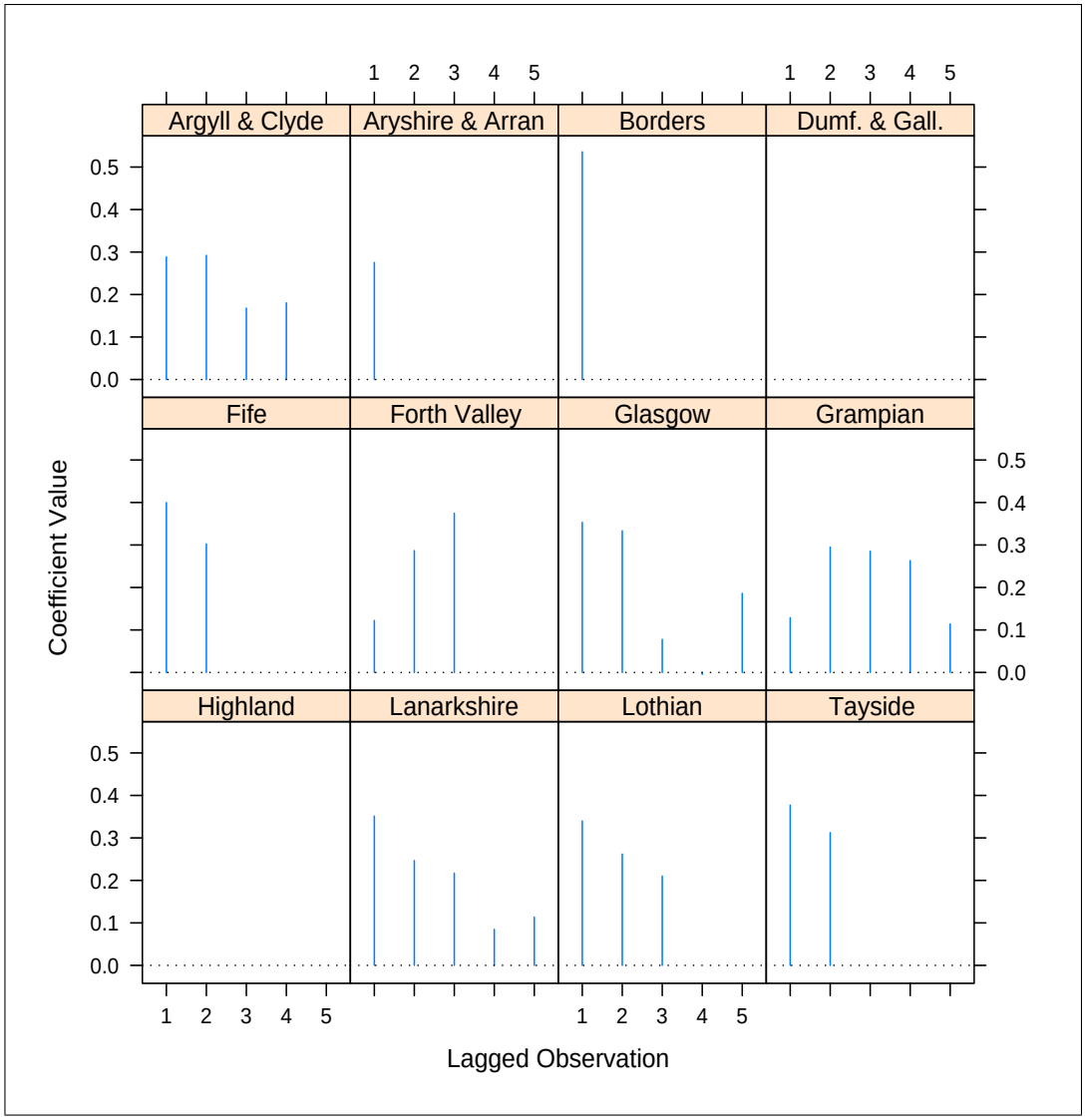


Figure 5.11: The values of the coefficients of the past observations from the models fitted in Section 5.3.1.

since they did not contribute significantly to the region's model. For Tayside, there is a significant correlation at lag five, but otherwise the residuals are well behaved. In Argyll and Clyde, there are significant correlations at lags 6, 10, and 52. The latter correlation suggests that the model for this region does not deal sufficiently well with seasonality in this health board. This might be improved by including a higher order of harmonics. However, in most cases, serial correlation has been dealt with reasonably adequately in these models.

Unfortunately, plotting the deviance residuals suggests problems with the model fits. Consider the deviance residuals from the model fitted to Highland in Figure 5.12. There is quite a distinct pattern in the negative residuals, suggesting that there are times when the model systematically over predicts the counts (since the residual deviance, r_D is defined as $r_D = \text{sign}(y - \hat{\mu})\sqrt{d_i}$ (Faraway 2006)). This also accounts for the 'kink' that can be seen in the qq-plot of the deviance residuals, shown in Figure 5.13. On closer inspection this pattern occurs when there are long runs of zeros, as shown in Figure 5.14: the zero counts cause the seasonal pattern to be reflected in the negative residuals. Thus, for those regions with higher average levels of reporting, such as Glasgow, such patterns are much less distinct, since there are fewer runs of zeros. In general, such a pattern would suggest quite serious problems with the model fit. For instance, it might suggest that seasonality had not been dealt with. However, the effect here is just a product of the particular data we are modelling and not indicative of deficiencies in the models: long periods of zero reports are to be expected with the low reporting rates and there is no easy way to escape the pattern that results, which is reflected in the negative residuals. We discuss some other models that might deal with this problem in Section 5.6. These factors suggest that residuals that are based on the difference between the fitted means and the data are not going to be useful here, since any problems are going to be obscured in pattern of the negative residuals. One approach to dealing with this is to consider other types of residuals; here, we consider 'modal residuals' which look at the difference between the modes of the fitted distributions and the observed count i.e. our predictor is the mode rather than the mean of the conditional distribution.

We define the *modal residual*, r_M, to be the difference between the reported count y and the mode $\hat{m}$ of the distribution fitted by the model i.e. $r_M = y - \hat{m}$. A general formula for finding the mode of the negative binomial distribution can

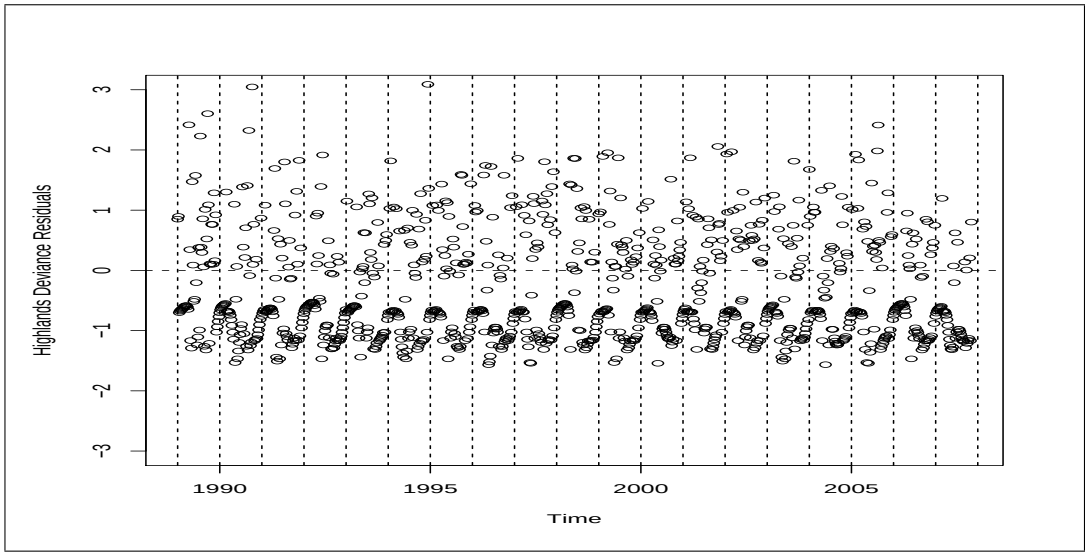


Figure 5.12: The deviance residuals from the model fitted to the Highlands that uses past observations (fitted in Section 5.3.1). There is clearly a pattern in the negative residuals.

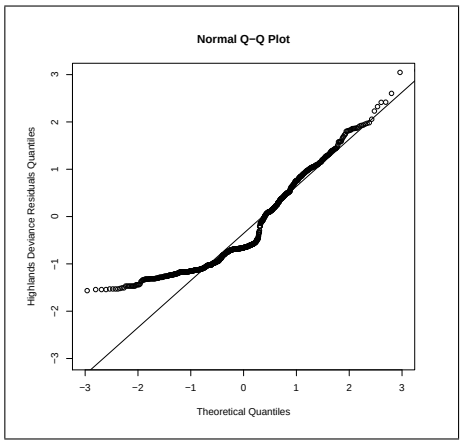


Figure 5.13: The qq-plot of the deviance residuals from the model fitted to the Highlands health board (fitted in Section 5.3.1). Since the deviance residuals are not normally distributed, we would not expect them to follow the straight line perfectly; however, the 'kink' and the change in residuals around the zero quantile suggests something odd, reflecting the pattern shown in Figure 5.12.

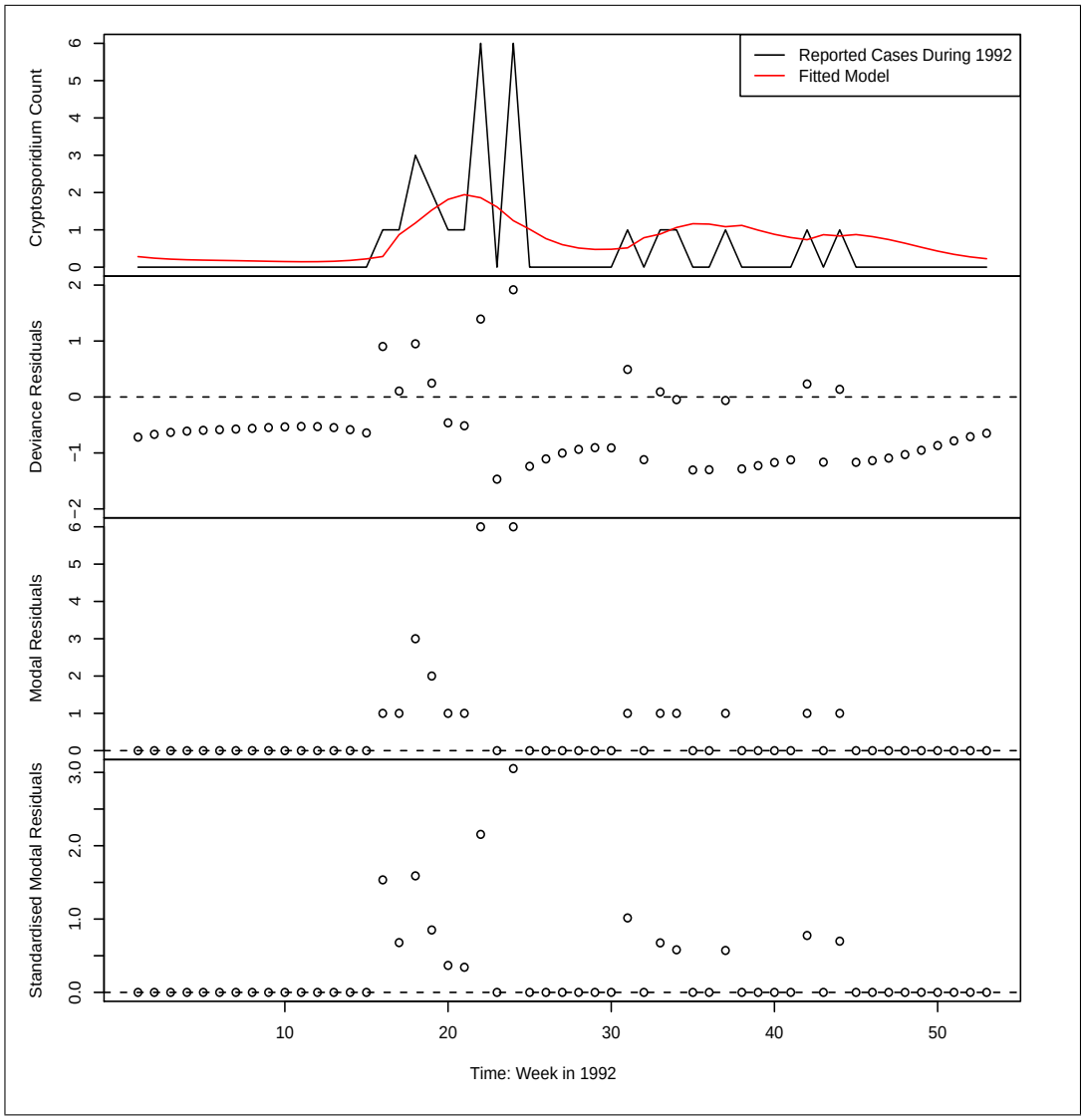


Figure 5.14: From the top: the reported cases of *Cryptosporidium* in the Highlands during 1992 and the values fitted by the model developed in Section 5.3.1; the deviance residuals; the modal residuals; the standardised modal residuals. When there is a long run of no reported cases, the deviance residuals have a smooth pattern to them, primarily corresponding to the seasonality fitted in the model to Highland. This pattern is removed in both forms of the modal residuals.

be found by considering the difference between successive terms of the probability distribution and considering sign changes, as we do in Appendix C. Letting

$$\beta = \mu\left(1 - \frac{1}{\theta}\right),$$

we know from the Appendix that then if $\beta < 0$, then mode is at zero, otherwise the mode is at $\lfloor \beta \rfloor$. The modal residuals for Highland's model during 1992 and 1993 are shown in Figure 5.14. It is not an error that the modal residuals look similar to the original counts: the counts for Highland are mostly quite small (often zero) and so the modes of the resulting distributions are nearly always zero, resulting in the modal residuals looking similar to the pattern of reports received during this period. However, the main thing to note is that since the patterned effect has been removed from the residuals, we can more easily see any real deficiencies in the models. Negative modal residuals are not impossible but much rarer; they tend to only happen in the regions with larger weekly reporting rates. In such regions, we can have a week at time t with quite a high count. Then, due to the the inclusion of past observations within the models here, the large count at time t will make the mode at $t + 1$ large. Sometimes, the mode at $t + 1$ will then be large enough to make the modal residual at $t + 1$ negative. However, generally, the counts are most frequently small and so modes are most often zero or one, giving non-negative modal residuals. Of course, in those regions where the dispersion parameter θ is smaller than one, the mode will always be zero and so we will always get non-negative modal residuals.

It is common practice, and more useful, to have standardised residuals, making them more comparable amongst themselves. The easiest way to do this is to bootstrap at each week, where we have a fitted distribution from the GLM. From the fitted distribution we take a random sample Z_i of size n and calculate the root mean square standard error associated with the mode $\hat{m}_t$ at week t:

$$\sqrt{\frac{\sum_{i=1}^{n}(Z_i - \hat{m}_t)^2}{n}}. \tag{5.4}$$

The modal residuals are then divided by this error to give the *standardised modal residuals*. Letting $n = 100,000$ is sufficient to generally reduce the maximum differences between successive runs of standardised residuals to less than 0.05;

larger n does make successive runs more consistent, but does not make much reduction to the maximum difference observed. An example of the standardised modal residuals can be found in Figure 5.14.

Before commenting on the standardised modal residuals, we note that the mode could also be used for prediction, as opposed to using the mean to parameterise the negative binomial and, for example, finding the 95% quantile. Instead, we could bootstrap at each week, finding the empirical distribution of differences between the mode and the values of a bootstrap sample. From this empirical distribution, we could calculate a 95% quantile that gives a distance such that 95% of the differences are smaller than this value; this distance could then be added to the mode to give an exceedance limit. Then, in any week, where the reports that week exceed this value, then an automatic warning would be produced. However, for a reasonable bootstrap size, this exceedance limit is going to be about the same as the value found by using the mean. Thus, to save calculation, we may as well use the mean.

We can now look at the standardised modal residuals for any patterns that would indicate a poor model fit, without being distracted by patterns in the negative deviance residuals. From the standardised modal residuals no evidence of bad fits is apparent. In Dumfries & Galloway there are perhaps a few non-contiguous years in the 1990s, with the suggestion of a pattern in a few weeks. If we were particularly worried about these, then we could fit a model to Dumfries & Galloway for the last ten years, excluding the period for which the years have a pattern in their residuals. In general, these models seem an acceptable fit to the data. The dispersion parameter of the negative binomial, θ, varies from 0.75 to 2.04 (Table 5.7) which seem reasonable.

5.4 Exponential Smoothing Regional Modelling

In this section we model the regional counts of *Cryptosporidium* using a variable based on exponential smoothing to try to deal with local trend. This is the same as was done in the NHS24 regional modelling; within that chapter, we discuss the motivation in Section 3.2.3 and discuss how to combine the Holt-Winters variable with the GLM in Section 3.4.1. For more information about exponential smoothing see Appendix B. We consider which form of exponential smoothing

is best for the modelling here in Section 5.4.1. Then, GLMs are fitted to the different regions utilising the created variable in Section 5.4.2.

5.4.1 ES Parameter Selection

We apply the Holt-Winters smoothing to the log of the counts. This makes the seasonality additive and so we use the additive form of the `HoltWinters` function within `R`. This function gives us the flexibility to choose between different versions of ES (Simple Exponential Smoothing (SES), ES with trend, ES with no trend & seasonality, and ES with trend & seasonality). Chatfield and Yar (1988) provide a number of guidelines on choosing the appropriate form of ES. While these are sensible, we propose a slightly different approach given our purpose for using ES: we will use the form of ES that explains the most deviance in the GLMs. Thus, to each health board, we fit four different negative binomial GLMs:

$$\log(\mu_t) = \beta_0 + \beta_1 SES_t, \tag{5.5}$$
$$\log(\mu_t) = \beta_0 + \beta_1 EST_t, \tag{5.6}$$
$$\log(\mu_t) = \beta_0 + \beta_1 SESSEAS_t, \tag{5.7}$$
$$\log(\mu_t) = \beta_0 + \beta_1 HW_t, \tag{5.8}$$

where SES_t, EST_t, $SESSEAS$, HW_t are the filtered series given by `HoltWinters` for carrying out SES, ES with linear trend, Holt-Winters with seasonality and no trend, or full Holt-Winters respectively. Values of the smoothing parameters under each form of exponential smoothing are found by minimising $\sum e_t^2$ – see Section B.1. The results of fitting these different models are shown in Table 5.8, where the residual deviance after fitting each type of smoothing is shown. Thus, for each region, whichever model has the least residual deviance suggests that that method of ES is best.

From fitting these models a number of results become apparent. It seems that ES does not benefit from being seasonal in most cases, with only two areas having lower residual deviance when a seasonal form of ES is used. Similarly, ES does not seem to benefit from having a linear trend (EST). Some of the more complex forms of ES reduce to simpler forms of ES when the simpler form is more optimal. For example, consider the Argyll & Clyde area: when the model

Region	Null Deviance	Residual Deviance from model using SES	EST	SESSEAS	HW	'Best' ES technique
Argyll & Clyde	1168	982	SES	992	SESSEAS	SES
Aryshire & Arran	794	748	751	752	SESSEAS	SES
Borders	688	664	SES	666	SESSEAS	SES
Dumfries & Galloway	999	885	SES	888	SESSEAS	SES
Fife	739	721	726	725	726	SES
Forth Valley	767	727	731	730	SESSEAS	SES
Glasgow	1259	1016	SES	1014	SESSEAS	SESSEAS
Grampian	1247	1066	SES	1070	SESSEAS	SES
Highlands	911	878	SES	882	SESSEAS	SES
Lanarkshire	1248	1038	1041	1049	SESSEAS	SES
Lothian	1333	1078	1082	1080	SESSEAS	SES
Tayside	993	916	919	907	908	SESSEAS

Table 5.8: The reduction in deviance resulting from fitting GLMs defined by Equations 5.5, 5.6, 5.7 and 5.8, with each having one variable corresponding to different types of Exponential Smoothing. Those entries that have the name of another ES technique in them correspond to those more complex techniques that have reduced to simpler ones e.g. HW with no trend reduces to SESSEAS. The "*best* ES technique" then, in this context, is the one which leaves least residual deviance in the resulting GLMs.

with SES is compared to the model with EST, there is practically no difference in residual deviance (the difference starts with the second decimal place). On closer inspection, for SES $\alpha = 0.1994641$, while with EST we find that $\alpha = 0.1994681$ and $\beta = 0$. Typically, when such 'simplification' occurs, there is agreement in parameters to three or more decimal places. In ten out of the twelve cases HW reduces to SESSEAS. EST reduces to SES in half of its uses.

The most surprising result is probably that SES does better than seasonal forms of ES in most cases; elsewhere, we have seen good evidence for *Cryptosporidium* being seasonal. To investigate why it is only in certain regions that a seasonal form of ES does better, we carry out a time series decomposition on certain regions, using the techniques applied in Section 4.4.1. We choose two regions (Argyll & Clyde and Lanarkshire) where SES has the largest improvement over other forms of ES, and the two regions (Glasgow and Tayside) where a seasonal form fares better. In these regions we calculate the seasonal factors from the log-counts, shown in Figure 5.15. By doing so, we can compare the differences in seasonality, and perhaps explain why the latter group does better with seasonal forms of ES. The main difference between the two groups seems to

be that the seasonal factors change more consistently and smoothly in the regions where a seasonal form of ES does better. The seasonal factors calculated here are similar to the seasonal factors in SESSEAS: the factors for any one week are mostly independent of the other factors for other weeks. This contrasts with the use of trigonometric harmonics, where the counts for all weeks contribute to the seasonal level for all other weeks. Since we can often get zero counts when a count might be expected, this 'borrowing effect' – using information from other weeks to contribute to the seasonal level for a particular week – probably contributes to a more realistic seasonal pattern. Thus, it is possible that seasonal forms of ES do not do as well as SES because of using seasonal factors. In the GLMs we go on to fit, we find that trigonometric harmonics are significant, suggesting that the data is seasonal but that seasonality is not modelled well with seasonal factors.

For simplicity of interpretation, we will choose one method of ES for all regions. In light of the evidence of Table 5.8, the best form of ES will be SES – simple exponential smoothing with no trend or seasonality. For those regions where a seasonal version of ES fares better, their seasonality will be incorporated in the GLMs through the use of trigonometric harmonics. In some areas, where EST does not simplify to SES, there is a potential of some trend; as with seasonality, we will try to incorporate this into our GLMs by using a linear trend term.

5.4.2 Regional GLMs with SES Terms

As we noted previously, with the inclusion of ES in the GLMs, model fitting becomes a two stage process. We first consider the exponential smoothing, and then proceed to the GLM fitting.

Exponential Smoothing

We use SES – simple exponential smoothing with no trend and no seasonality – in all the regions. The parameters found by the `Holtwinters` function (with β and γ manually specified as zero) are shown in Table 5.9. The α parameters are in the range (0.08,0.27), with the median value being 0.19. The range of values seem quite sensible given the comments we made in Section B.1.

There is a suggestion that α is larger in those areas of larger population:

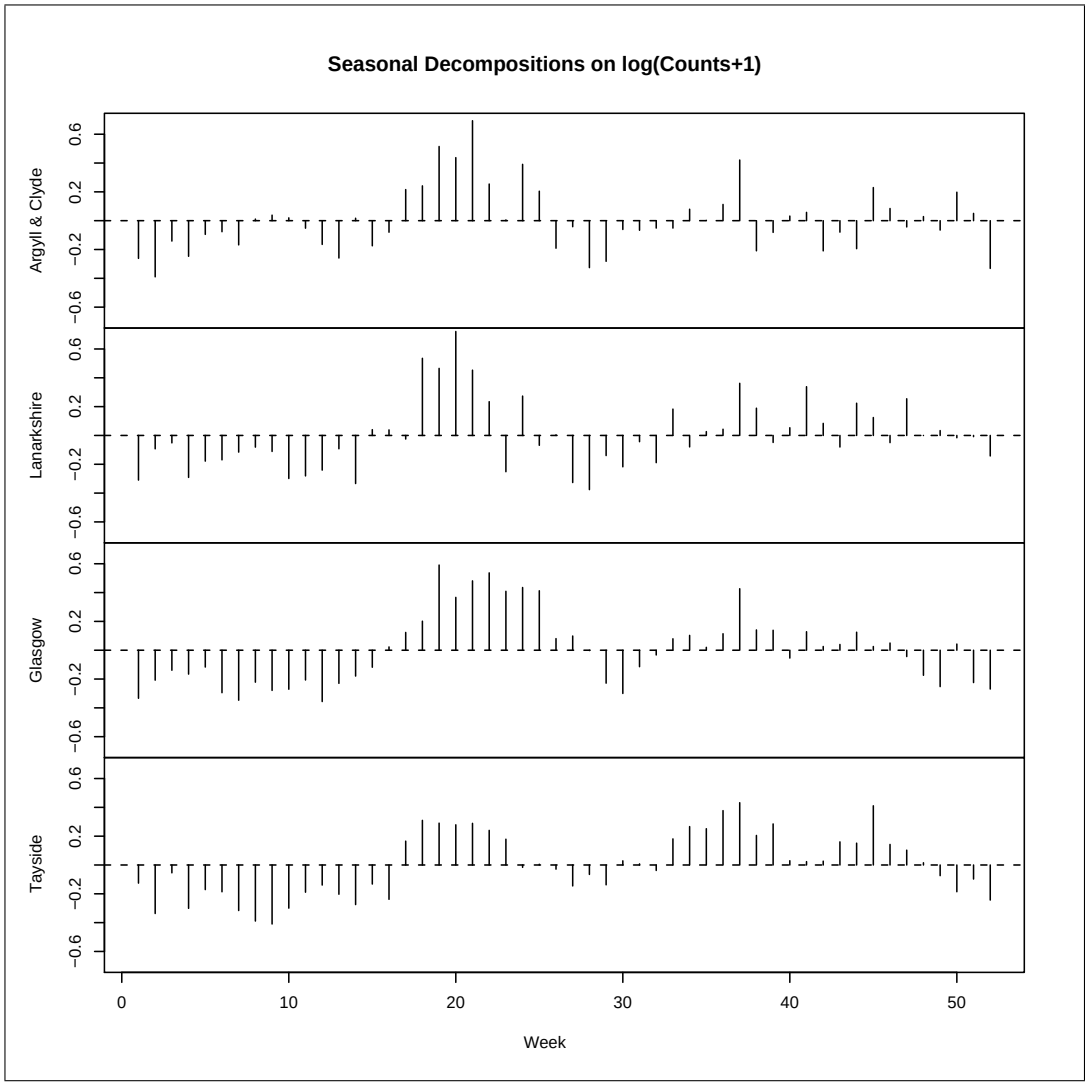


Figure 5.15: The seasonal factors (deviation from the trend) from a seasonal decomposition carried out on $\log(\text{Counts}+1)$ of (from the top) Argyll & Clyde, Lanarkshire, Glasgow and Tayside. The method used is detailed in Section 4.4.1. In Section 5.4.1 we see that the 'best' type of exponential smoothing for the top two regions is simple exponential smoothing, while in the bottom two, exponential smoothing with seasonal factors fares better. This may be because changes in the seasonal pattern are smoother in these regions.

Region	α
Argyll & Clyde	0.1995
Aryshire & Arran	0.1693
Borders	0.1112
Dumfries & Galloway	0.1984
Fife	0.0823
Forth Valley	0.1207
Glasgow	0.2677
Grampian	0.1808
Highlands	0.1373
Lanarkshire	0.2314
Lothian	0.2205
Tayside	0.2497

Table 5.9: The smoothing parameters found by `HoltWinters` for each region when SES is used in Section 5.4.2 on the regional counts of *Cryptosporidium*.

Kendall's τ statistic is 0.55 between α and regional populations (taken from the General Register Office for Scotland (2001)). A larger α means that observations are tied more strongly to preceding weeks. With more people in a region, infections will be able to spread more easily between people, resulting in greater levels of correlation between reporting levels of successive weeks. In regions with smaller populations, successive reports are less likely to be linked, and more likely to be independent cases, leading to less correlation between successive weeks. In less populated regions, seasonality may also have a greater effect on reporting levels.

GLMs: Comments on Fitted Models

The negative-binomial GLMs fit to each region can be found in Table 5.10. All variables were tested for inclusion in the GLMs and rejected if they did not contribute significantly (at the 5% level) to the model. The summary statistics for the models can be found in Table 5.12.

We first consider the SES coefficients. Since these are predictions on the log-scale, we might initially assume that their coefficient would be near one. However, we saw in the NHS24 modelling that this was often not the case because of the interaction with the other terms in the model – see Section 3.3.2. In two regions

Region	Intercept	SES	$\sin\left(\frac{2\pi}{52}\right)$	$\cos\left(\frac{2\pi}{52}\right)$	$\sin\left(\frac{4\pi}{52}\right)$	$\cos\left(\frac{4\pi}{52}\right)$	$\sin\left(\frac{6\pi}{52}\right)$
Argyll & Clyde	-0.284 (0.163)	1.237 (0.136)	0.012 (0.061)	-0.041 (0.064)	-0.200 (0.061)	-0.121 (0.064)	0.113 (0.062)
Ayrshire & Arran	-1.038 (0.161)	1.130 (0.371)	-0.060 (0.085)	-0.311 (0.093)	-0.316 (0.087)	-0.275 (0.090)	0.146 (0.087)
Borders	-1.666 (0.138)	1.693 (0.514)	-0.428 (0.118)	-0.672 (0.107)	-0.680 (0.104)	-0.030 (0.101)	–
Dumfries & Galloway	-0.466 (0.199)	1.217 (0.309)	0.131 (0.071)	-0.435 (0.084)	-0.365 (0.073)	-0.326 (0.075)	0.072 (0.071)
Fife	-1.446 (0.152)	1.906 (0.505)	-0.288 (0.092)	-0.515 (0.095)	-0.172 (0.089)	-0.268 (0.088)	–
Forth Valley	-1.386 (0.134)	1.919 (0.363)	-0.334 (0.100)	-0.387 (0.100)	-0.370 (0.094)	-0.278 (0.102)	0.136 (0.095)
Glasgow	-0.676 (0.091)	1.472 (0.146)	0.003 (0.056)	-0.194 (0.058)	-0.242 (0.055)	-0.124 (0.057)	0.217 (0.055)
Grampian	-0.355 (0.127)	1.377 (0.101)	-0.018 (0.050)	-0.174 (0.050)	-0.159 (0.049)	-0.149 (0.051)	0.244 (0.050)
Highland	-0.787 (0.131)	0.681 (0.301)	-0.358 (0.088)	-0.504 (0.076)	-0.478 (0.075)	-0.026 (0.080)	0.219 (0.075)
Lanarkshire	-0.095 (0.143)	1.222 (0.125)	-0.032 (0.050)	-0.059 (0.050)	-0.223 (0.050)	-0.150 (0.052)	0.109 (0.050)
Lothian	-0.073 (0.156)	1.031 (0.098)	-0.209 (0.056)	-0.209 (0.051)	-0.196 (0.052)	-0.064 (0.053)	0.062 (0.052)
Tayside	-0.783 (0.085)	1.055 (0.136)	-0.284 (0.069)	-0.280 (0.064)	-0.252 (0.063)	-0.118 (0.066)	0.056 (0.064)

Region	$\cos\left(\frac{6\pi}{52}\right)$	$\sin\left(\frac{8\pi}{52}\right)$	$\cos\left(\frac{8\pi}{52}\right)$	$zLen$	$pLen$	linTrend
Argyll & Clyde	0.173 (0.062)	-0.191 (0.062)	-0.240 (0.062)	–	–	-0.000661 (0.000177)
Ayrshire & Arran	0.211 (0.086)	-0.031 (0.085)	-0.174 (0.085)	-0.038 (0.019)	–	–
Borders	–	–	–	–	–	–
Dumfries & Galloway	0.193 (0.071)	-0.059 (0.068)	-0.180 (0.068)	-0.040 (0.018)	-0.097 (0.034)	-0.000723 (0.000190)
Fife	–	–	–	–	–	–
Forth Valley	0.180 (0.096)	-0.002 (0.094)	-0.326 (0.094)	–	–	–
Glasgow	0.185 (0.055)	-0.035 (0.054)	-0.126 (0.054)	–	-0.051 (0.019)	–
Grampian	0.150 (0.050)	-0.146 (0.050)	-0.257 (0.050)	–	-0.022 (0.007)	-0.000340 (0.000130)
Highland	0.162 (0.074)	-0.153 (0.072)	-0.177 (0.072)	–	–	–
Lanarkshire	0.217 (0.050)	-0.048 (0.050)	-0.213 (0.050)	–	-0.039 (0.014)	-0.000624 (0.000141)
Lothian	0.153 (0.051)	-0.075 (0.051)	-0.119 (0.051)	0.105 (0.046)	–	-0.000388 (0.000142)
Tayside	0.201 (0.063)	-0.077 (0.062)	-0.226 (0.062)	–	–	–

Table 5.10: The GLMs fitted in Section 5.4.2 to the *Cryptosporidium* data in the different regions, ustilising SES to deal with serial correlation. The tables give the coefficients of the variables and their associated standard errors in brackets.

Region	Year	Start Week	Week 1	Week 2	Week 3	Week 4	Week 5
Fife	2006	18	3.264 (1.109)	–	–	–	–
Grampian	1992	10	1.819 (0.825)	–	–	–	–
	1995	8	3.629 (0.820)	–	–	–	–
Lanarkshire	1998	18	1.876 (0.776)	–	–	–	–
	2003	33	2.559 (0.776)	–	–	–	–
Tayside	1994	45	1.894 (0.828)	–	–	–	–
	2000	29	1.690 (0.819)	–	–	–	–
	2005	16	4.235 (0.735)	2.470 (0.749)	–	–	–

Table 5.11: The factor level values for outbreaks found significant, fitted to regions when using SES within the GLMs fitted in Section 5.4.2.

Model	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	θ	θ SE
Argyll & Clyde	1261	980	281	22	980	970	10	2991	1.366	0.141
Ayrshire & Arran	872	760	112	13	980	970	10	1806	0.909	0.149
Borders	762	653	109	14	980	975	5	1422	0.971	0.208
Dumfries & Galloway	1161	892	269	23	980	968	12	2128	1.988	0.376
Fife	821	732	89	11	980	974	6	1684	0.895	0.157
Forth Valley	854	728	125	15	980	971	9	1837	0.72	0.101
Glasgow	1398	1025	374	27	980	970	10	3094	1.667	0.185
Grampian	1505	1071	434	29	980	967	13	3713	1.578	0.143
Highland	1052	878	174	17	980	971	9	2169	1.434	0.228
Lanarkshire	1442	1046	396	27	980	967	13	3420	1.828	0.184
Lothian	1427	1071	356	25	980	969	11	3856	1.565	0.133
Tayside	1329	955	374	28	980	967	13	2450	1.987	0.321

Table 5.12: Summary statistics from the negative binomial GLMs with SES fit in Section 5.4.2. DoF = Degrees of Freedom. SE = Standard Error.

we find that the coefficient is near one: Lothian and Tayside. However, more generally, the coefficient is usually quite a bit larger than one, with the exception of Highlands, where it is smaller than one. We check for correlations between the SES coefficients and board population, the value of alpha in ES that creates SES_t, the intercept, θ and average regional counts, but only small negative ones were found, suggesting no interesting relationships.

We use hierarchical clustering once more to suggest groups from the fitted seasonal patterns, as was done in Section 5.3.1. We apply cluster analysis to the vectors of unscaled seasonal patterns, as defined in Equation (5.2) but based on the coefficients contained in Table 5.10. The corresponding dendrogram is shown at the top of Figure 5.16. We cut the dendrogram at 2.9 resulting in three groupings: the first group contains mostly coastal regions on the edge of Scotland – Argyll & Clyde, Ayrshire & Arran, Dumfries & Galloway, Grampian, Glasgow, Lanarkshire, Lothian, and Tayside; Borders is on its own; North Central areas (Forth Valley, Highland) and Fife. As before, there is a reasonable agreement on the largest peak at around week 20. The latter two groupings have the largest amplitudes, and have the smallest values of α, the ES smoothing parameter – see Table 5.9: this may indicate that in these regions seasonality plays a more important role than in the first grouping, with their higher dependence on past values. In the latter two groupings, the trough following the Spring peak at around week 20 is shallower than with the first two groupings.

Next, we apply the cluster analysis to the scaled seasonal patterns contained in the vectors defined by Equation (5.3), but based on the coefficients contained in Table 5.10. The resulting dendrogram and groupings are shown in Figure 5.17. We cut the dendrogram at the height of 1.6, resulting in three groups: the West and Grampian – Argyll & Clyde, Ayrshire & Arran, Dumfries & Galloway, Grampian, Glasgow and Lanarkshire; a corridor from the East to the North – Forth Valley, Highland, Lothian and Tayside; those regions where two harmonics were fitted – Borders and Fife. Again, as we would expect, there is agreement on the spring peak around week 20 being the largest peak. The later peaks and troughs in the year tend to determine the groupings: in the regions in the West, the trough following the Spring peak is lower than the equivalent trough in the East and North group. In the East and North group, the Summer peak is more comparable to the Spring peak, than with the West and Grampian peak. Borders

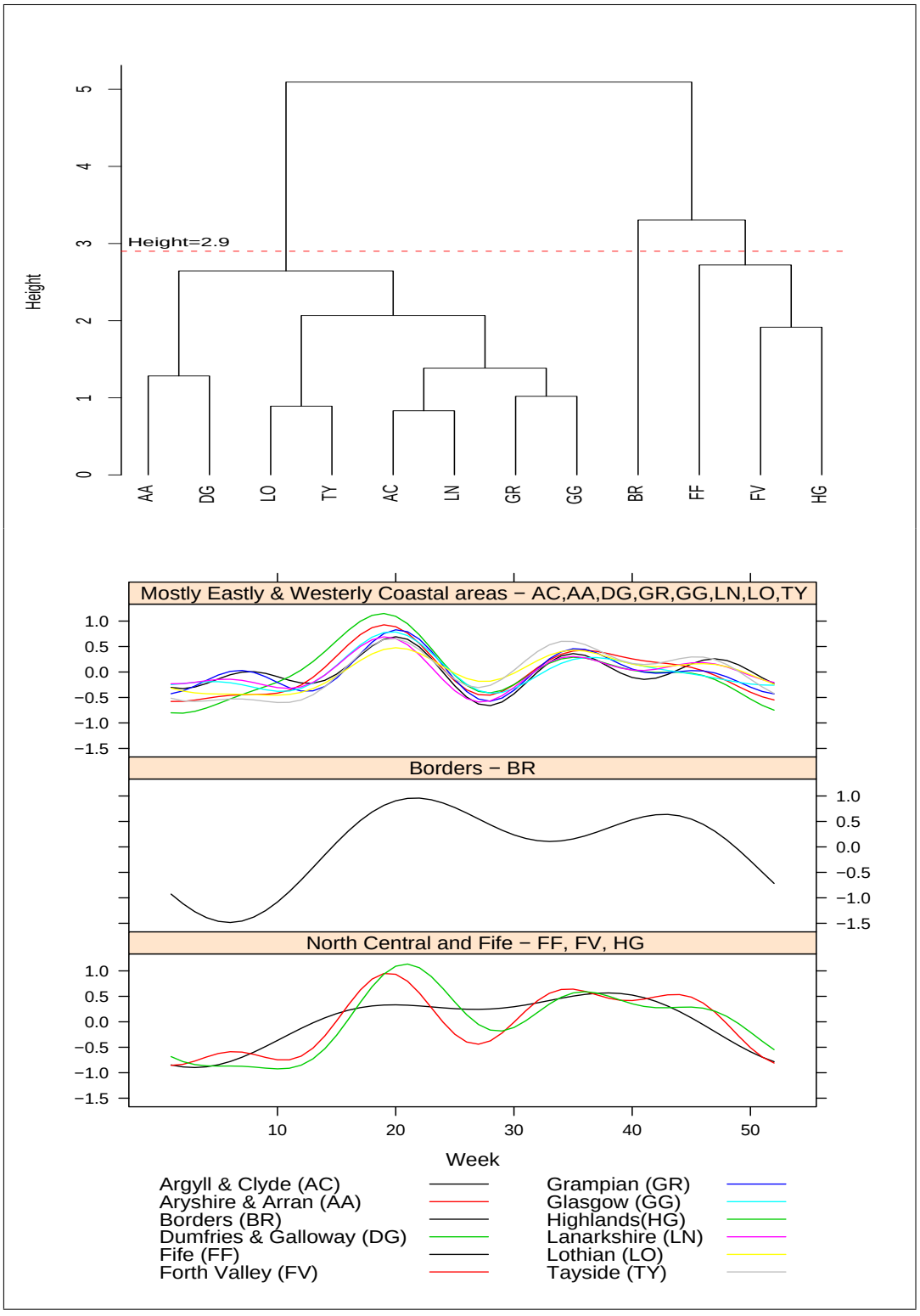


Figure 5.16: Hierarchical clustering applied to the unscaled seasonal patterns fit by the trigonometric terms, on the predictor scale, from the models fit in Section 5.4.2. The top plot is a dendrogram corresponding to the clustering. The red dashed line on the dendrogram represents where we have 'cut' it to form our clusters of regions with similar seasonality. The bottom plot shows the seasonality of each region grouped by the clustering suggested within the dendrogram.

and Fife again have a reduced number of harmonics fitted to them, resulting in them seeming quite different to the other regions. These patterns are again consistent with the regional variations suggested in Pollock et al. (2009), as we comment in Section 5.3.1.

Only three regions have a *zLen* term (that measures the run-lengths of zeros): Aryshire & Arran, Dumfries & Galloway, and Lothian. In the first two regions the *zLen* coefficient is reasonably small and negative, suggesting that the models over predict during runs of zeros in these regions, allowing this term to serve as a corrective. The models are likely to over predict during runs of zeros in these regions, since they have among the smallest average weekly counts. In Lothian, the coefficient is an (absolute) order of magnitude larger, but positive. This is likely because of the higher weekly reporting rate there – see the comment in Section 5.3.1.

Four regions include a *pLen* term (that measures run-lengths of positive numbers): Dumfries & Galloway, Grampian, Glasgow and Lanarkshire. All the coefficients are of a similar small order of magnitude and negative. The negative value suggests that in these regions the models over predict slightly, and these terms work as a corrective to that.

In these models the 2001 Foot and Mouth disease outbreak was found to have no significant effect. Thus, no factors were included in the models for Borders and Dumfries & Galloway to reflect the outbreak.

Five regions have a linear trend term fitted: Argyll & Clyde, Dumfries & Galloway, Grampian, Lanarkshire, and Lothian. In all of these regions the trend is very small and negative, but still significant. This suggests that in these regions the cases of *Cryptosporidium* are decreasing slowly over time. These regions have the largest average weekly counts, possibly meaning the other regions have a decreasing trend that the modelling here does not have the power to detect.

For details of the outbreaks that were included as factors (one factor level for each week of the perceived outbreak) see Table 5.11. Most outbreaks seemed to only affect the first week of the outbreak.

The values fitted to θ, the dispersion parameter for the negative binomial, vary between 0.7 and 2, which seem reasonable values.

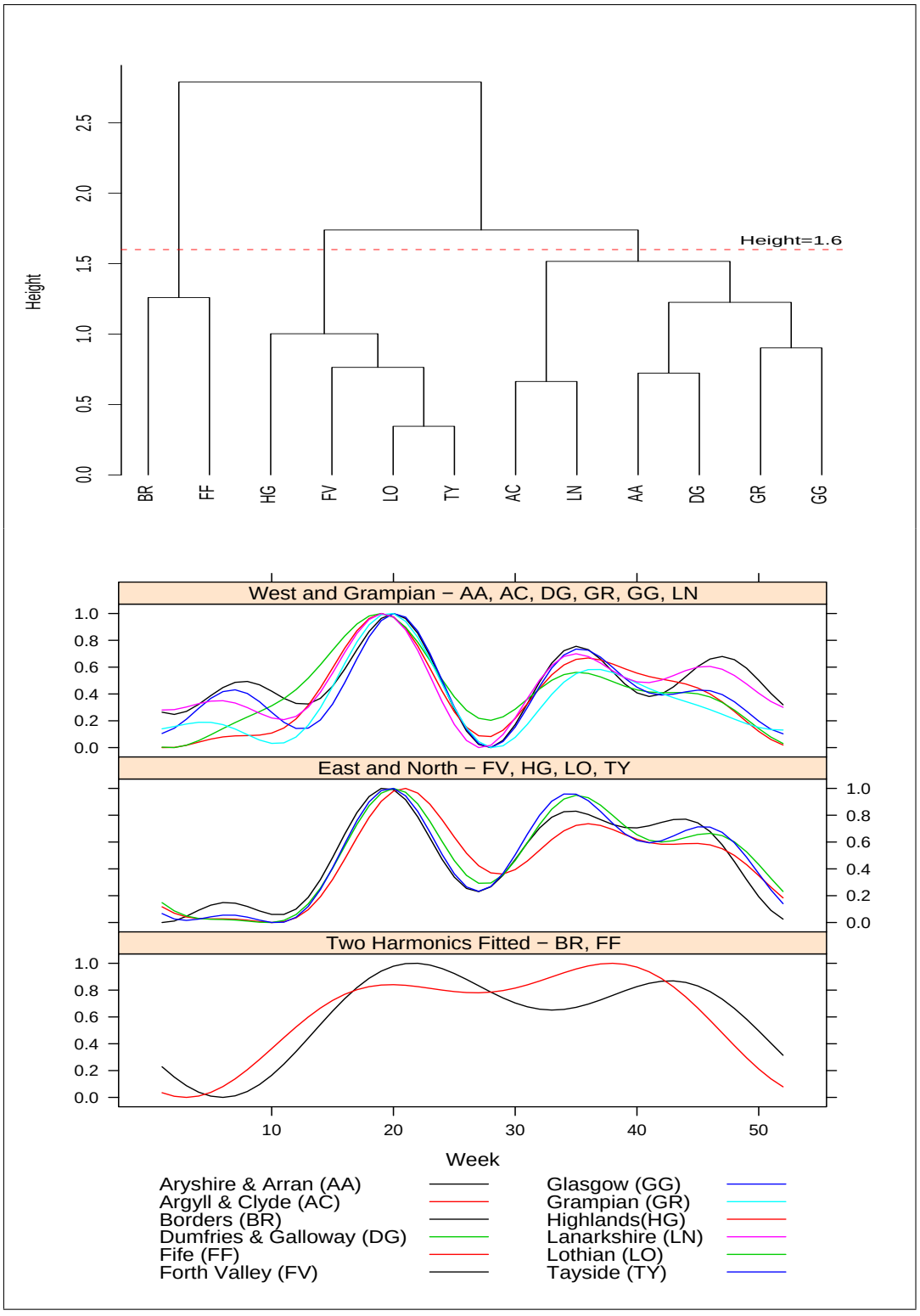


Figure 5.17: Hierarchical clustering applied to the scaled seasonal patterns fit by the trigonometric terms, on the predictor scale, from the models fit in Section 5.4.2. The seasonal patterns have been scaled to go between zero and one before the clustering has been applied. The top plot is a dendrogram corresponding to the clustering. The red dashed line on the dendrogram represents where we have 'cut' it to form our clusters of regions with similar seasonality. The bottom plot shows the seasonality of each region grouped by their clustering given by the dendrogram.

GLMs: Model Diagnostics

We check the acfs of the deviance residuals to check for any residual serial correlation. In eight of the twelve regions, there is nothing to concern us. In Grampian, there is a just significant correlation at lag 4, but it is isolated, so we do not worry about it further; otherwise the acf seems reasonable. For Glasgow, there is a significant correlation at lag 2, but there is no easy way to address this (we could include past values, but that will prevent us from comparing the two different modelling approaches). There is also a significant correlation at lag 52, but it is only just significant. We also note that the acf for Tayside indicates no problems with seasonality. Together, these two factors suggests that modelling seasonality in the GLMs and not the ES term does not seem to particularly harm the models – see Section 5.4.1 for a discussion on this. Borders has a significant correlation at lag one. This is perhaps not surprising when we note from Figure 5.11 that Borders has the largest coefficient for $\log(x_{t-1}+1)$; it is twenty percent larger than the equivalent coefficients in the other regions. The implied high level of correlation between successive weeks is unlikely to be captured well by any ES technique. Argyll and Clyde has a relatively large significant correlation at lag ten, but it is isolated and at a lag of no known importance, so we ignore it. Again, there is a positive correlation at lag 52, suggesting some residual seasonality, as we found in the modelling using past values in Section 5.3.1. If this becomes an issue we can fit a model on a restricted set of the data.

We next turn to considering the deviance residuals. As in Section 5.3.1, a pattern exists in the negative deviance residuals that does not actually indicate poor fitting models; thus we turn to considering the modal residuals defined in the same Section 5.3.1. We use the same size of bootstrap sample, n=100,000, to standardise the modal residuals. From the standardised modal residuals, we find that nine of the regions have no noteworthy features. However, Borders, Forth Valley and Grampian have some large residuals. In the case of Borders, the large residuals seem to be linked with a particularly large seasonal peak. If the large residuals cause concern, a factor can be included in the appropriate models to address them, as was done to deal with outbreaks (see Section 5.2). In the Forth Valley residuals, there are suggestions of patterns in 1996 and 1998: if these are a concern the model could be fit to a subset of the data.

5.5 Regional Modelling Comparisons

In this Section we compare the two sets of regional models: the set fit in Section 5.3.1 which uses past observations in the GLMs to deal with serial correlation and the set fit in Section 5.4 that include a single term based on exponential smoothing to do this. We will refer to these sets of models as 'past observation models' and 'SES models' respectively. We defer to Chapter 6 comparing the predictions given by these systems, alongside the predictions from the system developed in McCabe (2004) (see Section 4.2.2).

In both sets of models, seasonality was modelled through the use of trigonometric harmonics. In the SES models, the number of significant harmonics is most consistent at four. However, in both sets of models, Borders and Fife take the fewest harmonics and have the largest seasonal amplitudes. This may suggest that seasonality plays a more important role in these two regions compared to the others; this is supported by the fact that the number of past observations included in these regions are among the lowest (see Figure 5.11), and the parameters of SES fit to these regions are among the smallest (see Table 5.4.2). When we consider the seasonal patterns (see Figures 5.10 and 5.17), there is a general split between the East and West (sometimes including the North). In both sets of models, Grampian is somewhat of an exception, since its seasonal pattern is more similar to the West grouping and yet it lies on the East coast of Scotland. This might be linked with its particular mix of farming and urban areas (this was touched on in Section 5.1). Generally, the seasonal pattern for Borders stands out on its own because of the low numbers of harmonics fitted to it. In all regions, the largest peak in the seasonal cycle happens around week twenty during the Spring (see Section 4.3 for a discussion of the possible reasons for this). The seasonal patterns seem to be separated primarily by differing timings for peaks and troughs during the Summer and Autumn periods.

Both sets of regional models fit a very small linear downward trend for five regions: Argyll & Clyde, Dumfries & Galloway, Grampian, Lanarkshire, Lothian. In the SES models, a linear trend is also fit to Glasgow. These are the regions with the highest populations and so give the highest weekly counts. As noted previously, this may suggest there is a downward trend in the other regions that is too small to detect. In general, reporting levels seem to be decreasing.

The *zLen* variable was used in the regional models to record the run-length of successive zero counts. For both modelling approaches the term is included in Ayrshire & Arran and Dumfries & Galloway GLMs. In the past observation models, the term is also included in the Forth Valley and Highland GLMs. Generally, for both sets of models, the *zLen* coefficient is small and negative. This suggests that in these regions the models over predict during counts of zeros and this term works as a corrective to that. We saw in Section 5.3.1 that some of this over prediction is caused by the seasonality fit during periods of zero counts. The greater use of *zLen* terms in the past observation models may indicate that the SES models adapt better to low counts. For both sets of models, Lothian is somewhat of an oddity with respect the *zLen* term: Lothian is the only region with a positive coefficient for *zLen* and is also an order of magnitude larger than the other *zLen* coefficients. This may be because the longest run of zeros in Lothian is eight, while in the other regions there tend to be run lengths in the low teens. Note also that Lothian has the highest weekly reporting rate (see Table 5.2). This means that *zLen* may serve a different purpose in Lothian to the other regions: in Lothian, when we observe a zero, we expect the counts to increase soon, while in the other regions, when we observe a zero we expect to continue seeing more zeros.

The *pLen* variable measures the run-length of successive non-zero counts. Both modelling approaches fit a *pLen* term to Grampian and Lanarkshire. The SES approach also fits a *pLen* term to Dumfries & Galloway and Glasgow. In all cases, the coefficients of the terms are negative with the same orders of magnitude. This suggests that *pLen* serves to correct over prediction during non-zero counts. It may indicate that the SES method is slightly more prone to over predicting non-zero values.

The dispersion parameters for both sets of models are quite similar. Both sets vary from about 0.7 to about 2. Three boards with dispersion parameters less than one have the largest over-dispersion (Ayrshire & Arran, Fife and Forth Valley). The over-dispersion seen in the other boards is less. In all regions the standard error of the dispersion parameter is at least three to four times smaller than its value. This indicates that we do not have problems with the fits of the models as we saw with some regions in the NHS24 regional modelling (see Section 3.4.2).

The two different methods of dealing with serial correlation have different advantages. By using past observations the model fitting and the subsequent predicting is a straight forward one step process. However, the models are less parsimonious, with up to five terms of past observations included in the GLMs (more past terms might be used in certain of the regions, further exacerbating the point). The SES GLMs are more parsimonious since one term, SES, is only ever included in the models to deal with serial correlation. However, model fitting and prediction is a two stage process. First, the SES prediction must be found and then fed into the GLM for a prediction as a second stage. However, both of these stages can still be carried out easily and quickly. We can also contrast the two different approaches by considering the model summaries shown in Tables 5.7 and 5.12. For nine of the regions, the SES GLMs have the smallest residual deviance, one region has the least residual deviance using the past observation and two regions are about equal. This would suggest that the SES GLMs are better. We can also consider the AIC of the GLMs. There does not seem a lot of differences between the two approaches: six regions have a lower AIC with past observations, five regions have a lower AIC with SES and in one region there is nearly no difference (Highland). The biggest improvement in AIC happens in favour of SES. The SES models also seem to have slightly fewer artifacts in their residuals. Before comparing the predictive abilities of the systems more closely in Chapter 6, we would favour the SES models since: they are more parsimonious; they reduce the residual deviance the most; and where they are better in terms of AIC, they tend to make a bigger improvement in AIC, than the past observation models make when they are better in terms of AIC.

In these models we have omitted any element that might model the regions interacting with each other; there is no reason that an infection of *Cryptosporidium* would respect the artificial administrative boundaries that form the health boards. Thus, we may expect the counts in Lothian to have some relationship with the counts in its neighbouring health boards Borders, Fife and Lanarkshire. We investigate these relationships in the following Section and see why they cannot be modelled well within the GLMs developed here.

5.5.1 Interactions in Neighbouring Regional Counts

In Sections 5.3.1 and 5.4.2, we fit the best regional models that can be fitted to the counts of *Cryptosporidium* with the given variables. Seasonality, local trend, run-lengths of zero and non-zero counts, and local outbreaks have been incorporated into these models. We ruled out the use of climatic measures, such as rainfall and temperature, because of practicalities (see Section 4.4.3). There are not many other obvious variables that we can try to include in the fitted regional models. However, a remaining refinement might be given by including relationships between the counts in the different health boards. One way of determining if there are any such relationships is by considering the Cross-Correlation Function (ccf) of the residuals of the fitted models. Since the SES models (Section 5.4.2) are more parsimonious, we will use the residuals of these models in the present section.

A ccf function shows the correlations between the observations of two time series at different lags. In doing so, it allows one to check for potential interdependencies between the two time series. Thus, we can take all pairs of neighbouring health board models in turn and analyse the ccf of their residuals to investigate any potential interdependencies between them. Some examples will help illustrate this process. The Highland and Grampian health boards are neighbours and their ccf can be found in Figure 5.18. We denote the deviance residuals of the models at time t for Highland and Grampian as HGR_t and GRR_t respectively. Then, at lag k, the correlation between HGR_{t+k} and GRR_t is calculated. Bands that represent 95% confidence intervals for correlations due to two series of unrelated white noise can also be calculated and are shown in Figure 5.18. Thus, if correlations fall outside this bounded range there is evidence to suggest that there is some structure between the series. However, most would likely consider the ccf in Figure 5.18 to suggest little structure between the two series of residuals. There are significant correlations at lags -22, 5 and 12. However, these correlations are isolated and only just statistically significant or at lags of no special significance in the problem. Further, as 95% confidence intervals are used, it is not unexpected to get around two or three correlations that are statistically significant, even though there are no real relationships.

It should be noted that the ccf can be affected by structures, such as trend, in either of the two time series it is calculated from. This may lead to the suggestion of invalid relationships between the two time series. To avoid complications such as this, Diggle (1989) recommends 'pre-whitening' the two time series being compared by filtering out any trend or other structures. For instance, this could be done by fitting an ARIMA model. Then, by dealing with residuals of the ARIMA models, we are dealing with 'whitened' data; any structure suggested by the ccf of the these series relates only to relationships between the two series, and not any contained within either one. This is essentially what we do here by using the residuals of the SES models fitted to the regional counts of *Cryptosporidium*: we have removed as much structure as possible from within each series, so that any correlations indicated by the ccf of the residuals are likely to validly suggest relationships *between* regions. Thus, whenever we have used a ccf, we have also considered the acfs of the individual series of residuals to check for any artifacts that might be carried over to the ccf, which might suggest invalid relationships. For example, in Figure 5.18, the acfs of residuals for Highland and Grampian are given. The residuals for Highland have no features which would lead us to suspect they are not distributed as white noise. Apart from a just statistically significant and isolated correlation at lag four, the Grampian residuals are also reasonably close to white noise.

Other patterns in the ccf graphs can occur. An illustrative example can be found in Figure 5.19, where the ccf of the residuals for the models of Borders and Lanarkshire are shown. The ccf function has statistically significant correlations at lags -1 and -6. The lag at -6 is reasonably high order, isolated and only just statistically significant and thus, we focus on the much larger correlation at lag -1. This suggests there is a correlation between the counts in week $t-1$ in Borders and week t in Lanarkshire. Thus, a way of improving the model for Lanarkshire may be to include in its model the reported counts of infections in Borders lagged by one, or some similar measure lagged by one week.

Unfortunately, this improvement cannot be applied everywhere. Next, consider the ccf of the residuals for Lanarkshire and Dumfries & Galloway models, shown in Figure 5.20. In the ccf for these two series, there is a significant positive correlation at lag zero. This means in general that as the counts in Lanarkshire change, the counts in Dumfries & Galloway can be expected to fluctuate con-

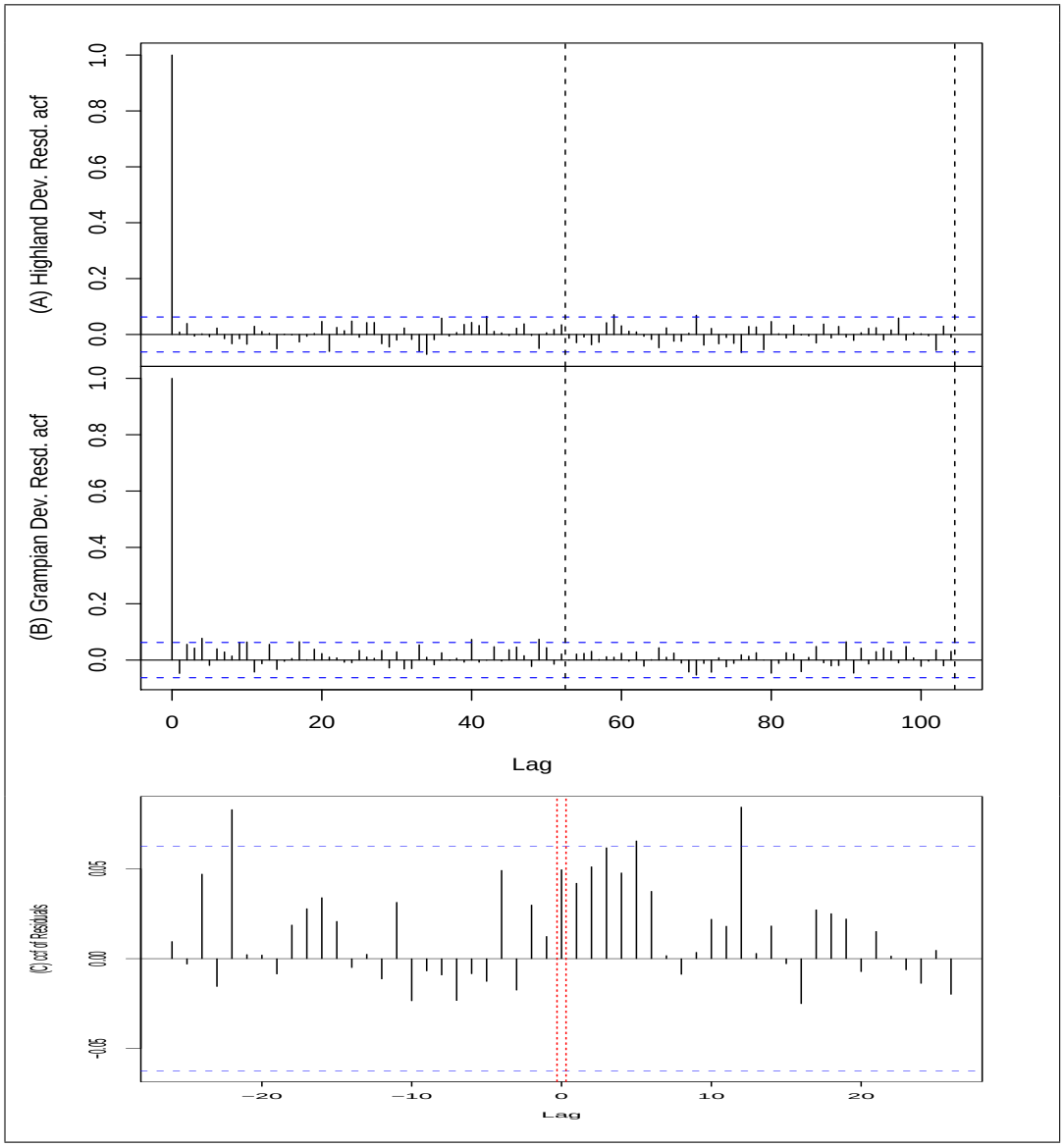


Figure 5.18: Graphs (A) and (B) show the acf of the deviance residuals from the regional models fitted to Highlands and Grampian health boards respectively in Section 5.4.2. Graph (C) shows the cross-correlation function of the above residuals.

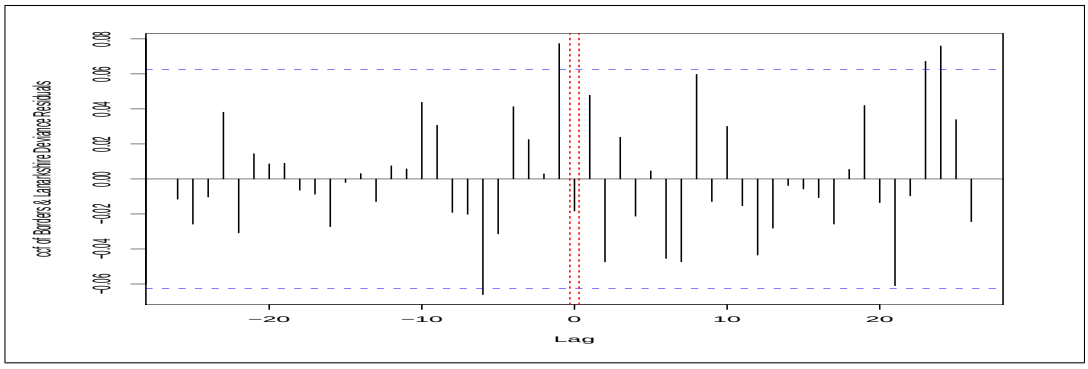


Figure 5.19: The cross-correlation function of the deviance residuals from the regional models fit to Borders and Lanarkshire in Section 5.4.2.

temporaneously in a similar way, as the correlation is positive. To model this contemporaneous relationship well would likely require a very different modelling paradigm than is adopted in Sections 5.3.1 and 5.4.2. In these Sections we fit a GLM to each region separately. To model a contemporaneous relationship would require the models for both boards to be fit simultaneously, possibly with a bi-variate distribution. Of course, this is just considering the modelling of two neighbouring boards; to fit a model that can incorporate contemporaneous relationships between all health boards would likely require all regional models to be fitted simultaneously. We consider the implications of this in Section 5.6.

A similar complication happens between Argyll & Clyde and Glasgow. The ccf of their residuals is shown in Figure 5.21. The ccf of the residuals of these two boards is quite complex: there is a large positive correlation at lag zero, and other significant correlations at low order lags, both positive and negative. These suggest that the counts both fluctuate contemporaneously and also interact with each other over a longer period of time. This greater complexity in the relationships probably reflects the relatively large population densities in these two boards and the effects of people commuting between them. Again, this type of relationship cannot be modelled well via the current approach, and so suggests a different approach is required if we wish to model regional interactions.

The ccfs for all pairs of neighbouring health boards are summarised in Table 5.14. The ccfs provide an indication of the typical movement of *Cryptosporidium* infections through Scotland. For instance, the ccf in Figure 5.19 suggests that

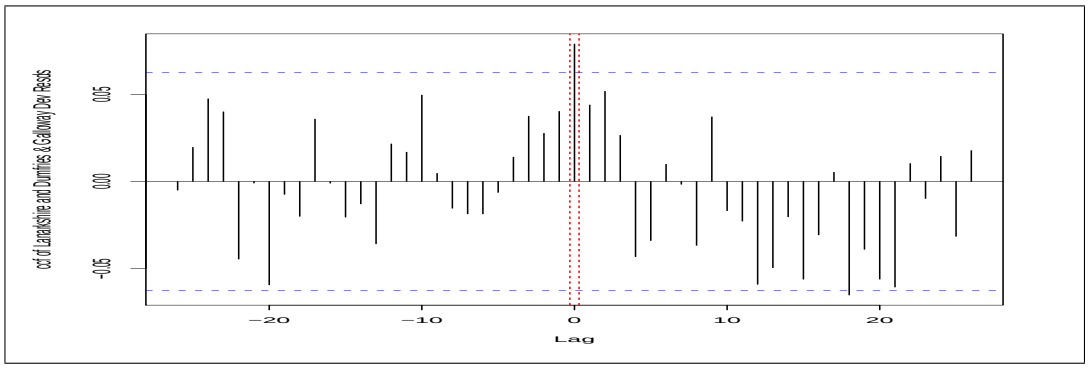


Figure 5.20: The cross-correlation function of the deviance residuals from the regional models fit to Lanarkshire and Dumfries & Galloway in Section 5.4.2.

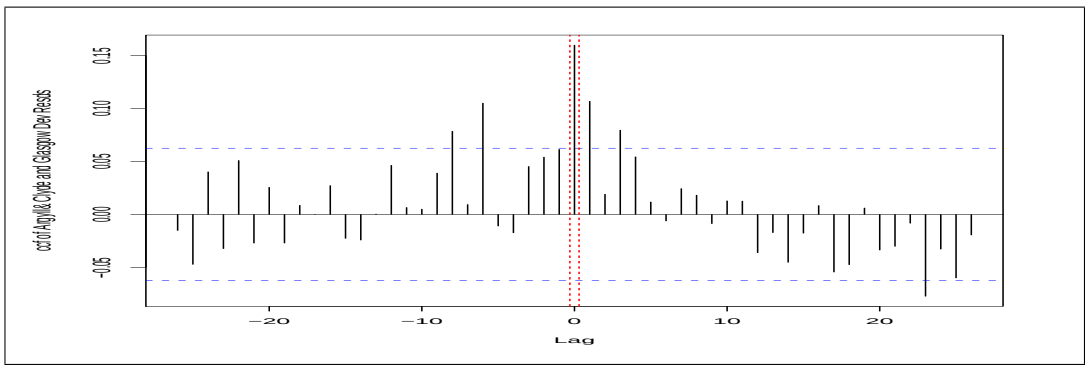


Figure 5.21: The cross-correlation function of the deviance residuals from the regional models fit to Argyll & Clyde and Glasgow in Section 5.4.2.

the level of infections in Borders move into Lanarkshire in the week following. In some neighbouring boards, there is no link (Figure 5.18), in others there is a positive contemporaneous correlation in counts (Figure 5.20) and the remainder exhibit a more complex neighbouring interaction in the spread of *Cryptosporidium*, sometimes incorporating a contemporaneous element as well (Figure 5.21).

A pictorial representation of the spread of *Cryptosporidium* can be found in Figure 5.22, and the geographical placement of the health boards can be seen on the right of Figure 2.8. Green lines on the diagram indicate those boards that have a contemporaneous relationship between them: infection in one board will affect the neighbouring boards in under a week. It seems that infection spreads most quickly through the East-West band (Argyll & Clyde, Forth Valley, Fife, Glasgow, Lanarkshire and Lothian). This is likely linked with these boards having the largest populations. Infection seems to spread more slowly in a northerly direction (Borders, Lanarkshire, Forth Valley, Tayside and Grampian). These factors may be explained by the higher levels of commuting that happen between the population centres in the East and West, compared to the lower levels of commuting North to South. Some of the geographically larger boards have no links with other large neigbouring boards (Highlands and Grampian, and Dumfries & Galloway and Borders). The is probably explained by the low densities of population in these regions, who will carry relatively few infections across regional borders.

Health board 1	Health board 2	Significant Correlations	Neighbouring Relations				Notes
			None	Contemporaneous	One way	Interaction	
Highlands	Grampian	-22, 5, 12	✓				
Highlands	Tayside	-9, 0, 8, 13		✓			
Highlands	Argyll & Clyde	-6, 7, 26	✓				
Tayside	Fife	-24, 0, 3		✓			
Tayside	Forth Valley	-12, 3, 14, 16			TY←FV		Slow movement.
Lothian	Lanarkshire	-15, -2, 0, 14		✓		✓	
Borders	Lanarkshire	-1, 23, 24			BR→LN		
Borders	Dumfries & Galloway	-11, -9, -5, 15	✓				
Glasgow	Lanarkshire	-24, -16, -4, 0, 1		✓		✓	
Glasgow	Aryshire & Arran	-26, -20, 14, 17	✓				
Aryshire & Arran	Argyll & Clyde	-23, -21, -2, 2, 6				✓	
Grampian	Tayside	-22, -7	✓				
Aryshire & Arran	Lanarkshire		✓				
Aryshire & Arran	Dumfries & Galloway	-14	✓				
Lanarkshire	Dumfries & Galloway	0, 18		✓			
Tayside	Argyll & Clyde	-18, -17, -8, 13	✓				
Argyll & Clyde	Forth Valley	-9, -8, -2, 0, 2, 4, 12		✓		✓	
Argyll & Clyde	Glasgow	-8, -6, 0, 1, 4		✓	AC←GG		
Forth Valley	Fife	-22, -20, -1, 5, 9		✓	FV→FF		
Forth Valley	Lothian	-20, -13, -12, -1, 0, 1, 5, 23		✓		✓	
Forth Valley	Lanarkshire	-1, 4, 19				✓	
Forth Valley	Glasgow	-7, 0, 1, 3, 14, 16		✓	FV←GG		
Lothian	Borders	5			LO←BR		Slow movement
Lothian	Fife	15, -2, 0, 2, 26		✓		✓	

Table 5.14: This table gives the statistically significant correlations from the cross correlation function applied to the residuals of all neighbouring pairs of SES models (fit in Section 5.4.2). Particular correlations suggest particular interactions between the health boards; see Section 5.5.1 for an explanation of them.

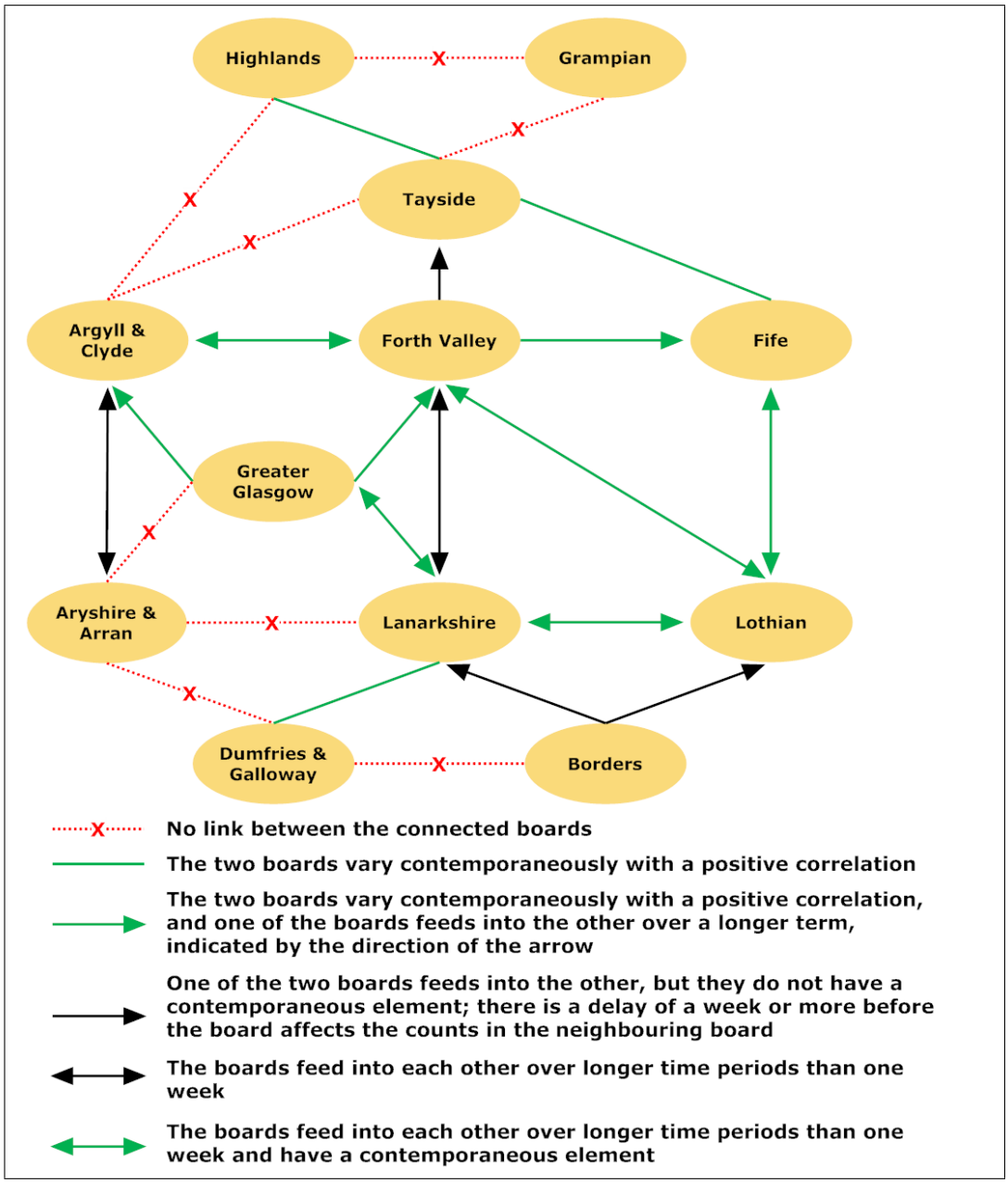


Figure 5.22: A pictorial representation of potential infection spread of *Cryptosporidium*, as suggested by the ccfs of residuals of neighbouring health boards, summarised in Table 5.14.

5.6 Conclusions

In this Chapter we have developed two modelling approaches that could be used in the regional monitoring of *Cryptosporidium*. Modelling at the regional level brings with it particular issues considered in Section 5.1: there we consider the differences between health boards and their different weekly reporting rates. Then in Section 5.3.1, we fit regional models with past observations in the GLMs to deal with serial correlation. The residuals of these GLMs initially suggest that the models are a poor fit; however, on closer investigation, the patterns in the residuals turn out to be a by-product of the seasonality and low (often zero) counts and are not an indication of poor fitting models. To better investigate the fit of these models, we introduce the idea of 'modal residuals' which look at the difference between the mode of the fitted distributions and the observed counts. This provides a way of removing the pattern from the residuals, making it easier to see any true short comings in the models. Alternatively, one could model the zero counts directly; we consider this further in Section 5.6.1.

We then move on to regional modelling that uses a variable based on Exponential Smoothing (ES) to address serial correlation (Section 5.4). Somewhat unusually, we find that simple exponential smoothing seems the best form of ES to use, even though the counts are seasonal in nature. This is likely because of the way that the `HoltWinters` function in `R` deals with seasonality – see Section 5.4.1. Future work might look at modelling the seasonality using trigonometric terms fitted by using discounted least squares. A further alternative might be to fit the models using Generalised Additive Models (GAMs), which are used in Chapter 7. In such models, smoothing splines are used to model the data, which allow for more complex patterns to be fitted than are generally possible with standard GLMs (Zuur, Ieno, Walker, Saveliev, and Smith 2009). For example, non-parametric polynomials can be fitted to the data (Wood 2006b). This would be particularly useful for *Cryptosporidium* with its bi-modal seasonality. We make use of GAMs in Chapter 7 to deal with fitting a more complex seasonal pattern than is possible with straight forward application of trigonometric terms within a GLM.

The two sets of regional models are compared in Section 5.5. We find that both sets of regional models give evidence for the counts of *Cryptosporidium*

decreasing in the health boards with larger populations. Seasonality is in broad agreement between the two sets of models: there is a split in seasonality between the East and West coast. Grampian is somewhat unusual in being on the East coast but having a seasonal pattern that is closer to that of health boards on the West coast. The Borders health board seems to have the largest seasonal amplitude. These results agree with those reported in Pollock et al. (2009). Both sets of models seem to do comparatively well addressing serial correlation, but the models utilising ES give more parsimonious GLMs.

One element not considered in the regional models is the relationship between the counts of the different health boards. We investigate such relationships in Section 5.5.1 by looking at the cross-correlation function of the residuals of neighbouring regional models. We find that some relationships are delayed: for example, the counts in Borders might give an indication of the counts expected in Lanarkshire in the week following. Other relationships are contemporaneous: for example, the counts for Lanarkshire and Dumfries & Galloway seem to have some relation between their counts during the same week. The latter relationship cannot be modelled by independent GLMs and so suggests a different modelling paradigm will be required if we wish to model all types of inter-board relationships. One possible approach is presented in Shaddick and Wakefield (2002), where a Bayesian model is used to simultaneously model air pollution over time from several locations simultaneously. We consider this further in Section 5.6.2.

Thus, we have two sets of tailored regional models that seem reasonable candidates for forming predictions for the reported cases of *Cryptosporidium* and thus calculating exceedance values. If HPS receives more reports than this exceedance value for a particular region, then a warning would be flagged up and it would be known that this region may have an outbreak. This would mean that HPS would have less work to do in locating a potential outbreak and greater power to detect smaller outbreaks. These two factors should allow for earlier detection of potential outbreaks. This work formed the basis of a presentation given at the Royal Statistical Society's Young Statisticians' Meeting (Wagner, McKenzie, and Robertson 2009).

Having tailored systems is reasonable if we wish to merely monitor *Cryptosporidium*. However, if we wish to apply these monitoring systems to other organisms, it will be best if we can fit the same basic model to all regions. This

basic model could then, more or less, be applied to the other organisms. It is more desirable for general exception reporting systems to use the same underlying model, as it removes the necessity of users becoming familiar with numerous specific models. This in turn increases the usability of the system. Hence, in the next Chapter, we develop two generic models for monitoring *Cryptosporidium*, one that utilises past observations and another that utilises a variable based on Holt-Winters smoothing. These models will be developed from the modelling carried out in this Chapter. We then take these two generic models and combine them with an exceedance alarm method to form two exception reporting systems. Finally, we apply these two systems, along with the system developed in McCabe (2004), to the regional counts of *Cryptosporidium* from 2006. This will allow us to compare and contrast the three systems.

5.6.1 ZIPs: Directly Modelling Zero Counts

We found in the regional modelling that many counts are zero with a clustering effect among them. We dealt with this in various ways, such as the use of the $zLen$ variable as a proxy for the correlation we expect in runs of zeros. This is a reasonable approach given we intend to use these models for predictions. However, the zeros still have an effect, particularly in the residuals, necessitating the need for the use of modal residuals (Section 5.3.1). Future work might look at developing models that address the high counts of zeros directly. Here we consider one potential approach further.

First, it is useful to consider the types of zeros that can occur with the HPS data. Zuur et al. (2009) divide zeros into two types: 'true' zeros and 'false' ones. 'True' zeros are ones that naturally occur in the process being modelled. Thus, in the context of *Cryptosporidium*, a 'true' zero would indicate that *Cryptosporidium* currently has a low prevalence in the general environment at a particular time. A 'false' zero could be caused by a number of different sources that prevent the underlying process be observed. Thus, within our context here, false zeros could be caused by a lab failing to report a positive test result for *Cryptosporidium*. However, we have no way to distinguish between these types of zeros and so must use sophisticated techniques to deal with the different types. The labels used to distinguish between these types of zeros differs; Ridout, Demétrio, and Hinde

(1998) label them 'structural' (true) and 'sampling' (false) zeros. In any case, we observe more zeros than would be expected with either the Poisson or negative binomial distributions. To deal with such situations, these distributions have been extended to produce Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB) distributions (Zuur et al. 2009). The over-dispersion within the *Cryptosporidium* data is very likely to be caused by the 'excess' zeros and so, if we are able to use a zero inflated distribution, we may find that it is not necessary to use the more general negative binomial over the Poisson distribution. Thus, we present the ZIP distribution here, but the general ideas apply straight forwardly to extending the negative binomial (Zuur et al. 2009).

We use the notation from Ridout et al. (1998) to introduce a ZIP. Inflated models have often been described as mixture models, since, as presented above, their zeros come from two different sources (Zuur et al. 2009). Observing a false zero is a binomial process: we either observe a false zero with probability ω or a zero from the Poisson process with probability $1-\omega$. Otherwise, the counts are distributed as usual from a Poisson distribution with parameter λ. Combining these results together gives the ZIP distribution:

$$Pr(Y=y) = \begin{cases} \omega + (1-\omega)\exp(-\lambda), & y=0 \\ (1-\omega)\exp(-\lambda)\lambda^y/y!, & y>0. \end{cases}$$

The parameters ω and λ can then be related to sets of covariates X and Z, with parameters β and γ (Lambert 1992):

$$\log(\lambda) = X\beta \quad \text{and} \quad log\left(\frac{\omega}{1-\omega}\right) = Z\gamma.$$

The EM-algorithm can then be applied to calculate estimates of β and γ (Lambert 1992; Ridout, Demétrio, and Hinde 1998). The two sets of covariates can be completely separate or have elements in common; if they are the same, the model can be simplified (Lambert 1992). It is quite possible that the covariates that govern ω will be different to those that govern λ, or possibly just exhibit different relationships. For instance, with the *Cryptosporidium* counts, we might well expect the seasonality to be different in the two components: in β, the seasonality would model the seasonal prevalence of *Cryptosporidium* in the environment; in

γ, the seasonality would model when false zeros are more likely, such as at the start and end of the year when labs are more likely to be closed.

Zero inflated distributions have been used for some time, both on ecological data (for example Agarwal et al. (2002), Welsh et al. (1996), Gupta et al. (1996)) and in other contexts such as manufacturing (for example Xie et al. (2001), Freund et al. (1999), Ridout et al. (1998), Lambert (1992)). However, there have been few extended considerations of these distributions (Zuur et al. 2009). Some mention is given in Cameron and Trivedi (1998), Hardin and Hilbe (2007), with a strong practical introduction in Zuur et al. (2009) and some further detail in Hilbe (2007). Software directly supporting these distributions is very limited. In fact, Zuur et al. (2009) note that it is only recently that the different software implementations have started giving similar results. One commercial package is LIMDEP, available at http://www.limdep.com (Ridout et al. 1998); however it is very expensive. Standard installations of `R` do not come with native support for inflated distributions but packages that do provide this include `VGAM` (Yee 2007) and `pscl` (Jackman 2008). A comparison of the different inflated distributions and examples using `pscl` are given in Zeileis, Kleiber, and Jackman (2008).

To our knowledge, no exception reporting system has made use of zero inflated models. These models present a natural choice for use in exception reporting systems as these systems come to be applied more widely to regional data that contains many zeros. As the software becomes more standard and reliable, this should prove a rich and rewarding avenue for future work.

5.6.2 Modelling Inter-board Relationships

In Section 5.5.1, we saw a number of inter-board relationships that were contemporaneous. To fit such relationships really requires that the models for all health boards are fit simultaneously. This is not possible using the modelling approach adopted in this Chapter. However, a potential approach may be found by adapting the one used in Shaddick and Wakefield (2002).

Shaddick and Wakefield (2002) present a spatio-temporal approach for modelling the levels of four pollutants from eight monitoring sites, over three years, simultaneously. They do this by using a hierarchical dynamic linear model. They let Y_{spt} denote the observed level of pollutant p at spatial location s on day t and

assume:

$$Y_{spt} = X_{spt}\beta + \theta_{pt} + m_s + \upsilon_{spt},$$

where: υ represents measurement errors which are treated as independently and identically distributed (iid) $N(0, \sigma^2_{sp})$; m_s is used to model the spatial effect of being at site s; θ_{pt} is used to give temporal and pollutant dependence; β is a $q \times 1$ vector of coefficients; and X_{spt} is a $1 \times q$ vector of regressors that can change spatially and temporally, and can be pollutant specific. This gives the model for the observed data.

Underlying this model there are another two models. The first of these model the spatial-pollutant relationships. The collection of random effects $m_p = (m_{p1}, \ldots, m_{pS})'$, $p = 1, \ldots, P$ arise from the multivariate normal distribution (MVN):

$$m_p \sim MVN(\mathbf{0}_s, \sigma^2_{pm}\Sigma_{pm}), \tag{5.9}$$

where: $\mathbf{0}_s$ is an $S \times 1$ vector of zeros; σ^2_{pm} is the between-site variance for pollutant p; and Σ_{pm} is the $S \times S$ correlation matrix. Elements of Σ_{pm} give the correlation between sites. Different functions can be used to calculate appropriate values for these correlations. Shaddick and Wakefield (2002) calculate the correlation using a function of the distance between sites. This model is stationary.

Next, the temporal-pollutant model:

$$\theta_{pt} = \theta_{p,t-1} + w_{pt}, \tag{5.10}$$

for $p = 1, \ldots, P$, where: $w_t = (w_{1t}, \ldots, w_{Pt})'$ are iid MVN with zero mean and variance-covariance matrix Σ_P. Within this matrix are the variances σ^2_{wp}, allowing each pollutant to have different amounts of temporal dependence. There are $P(P-1)/2$ covariance terms which can be used to model the relationships between pollutants. This model structure produces a first order smoothing model, similar to an AR(1) structure.

Shaddick and Wakefield (2002) then proceed to fit this model by using Monte Carlo Markov Chain (MCMC) methods within the WinBUGS software (Spiegelhalter, Thomas, Best, and Lunn 2005). Hyperpriors are required for the initialisation of the MCMC methods – see Shaddick and Wakefield (2002) for details of these.

This model contains all the key elements of the models that we have fit in this Chapter. If we just wish to model *Cryptosporidium*, then the above model from Shaddick and Wakefield (2002) can be greatly simplified to modelling just one 'pollutant'. Then, Y_{s1t} would denote the count of reported cases from health board s on week t. The temporal-pollutant element of the model would deal with some of the serial correlation within the counts. If this was not sufficient to address all the serial correlation, elements such as past observations or a variable based on Holt-Winters smoothing could be included in β. Alternatively, Equation 5.10 could be adapted to give an equivalent higher order of AR. Inter-board board relationships would be modelled within Equation 5.9, through appropriate covariance values in Σ_{pm}.

More generally, one could obviously extend this approach to modelling more than one organism simultaneously. For instance, we might model all organisms that relate to respiratory infection together. However, this systems presents one of the largest deviations from the current system and would likely require significant development. MCMC methods can be very computationally intensive, even with contemporary computing power. Due to this, this approach may not be feasible. However, this system does provide an elegant way of dealing with inter-board relationships. Investigations would also be required to determine how easily such models can be interpreted.

Chapter 6

Comparison of Regional *Cryptosporidium* Exception Systems

In Chapter 5 we develop two sets of models for forecasting the levels of *Cryptosporidium* at the regional health board level. The first set use past-observations in their models (Section 5.3) and the second set incorporate a Holt-Winters term (Section 5.4). In Chapter 4 we describe the system developed in McCabe (2004), that is currently used to monitor the national counts of *Cryptosporidium* (Section 4.2.2). Currently, the only complete option HPS has ready to monitor regional counts would be to use this national system on the regional data. To compare with this system, we develop two sets of uniform models (same variables in all models) in this Chapter, based on those developed in the previous one. These models are fit to data from 1988 to the end of 2005 and then used to produce forecasts for 2006. These forecasts can then be used within an exceedance alerting method to form exceptions for 2006. We compare the results of these systems, with those that result from applying the system developed in McCabe (2004) to the regional counts of *Cryptosporidium*. We find some evidence suggesting that the uniform models developed here may be more suited to regional exception reporting monitoring and consider the changes required for extending these systems beyond Cryptosporidium. However, it should be remembered throughout the comparisons made, that we are comparing systems that have been tailored

specifically to *Cryptosporidium*, against the very general system developed in McCabe (2004); we expect some of the noted improvements to be due to this tailoring.

6.1 Uniform modelling

6.1.1 Choice of Data

We fit the uniform models to data from 1988 to the end of 2005. Even though we have data till the forty-fifth week in 2007, we will produce forecasts for 2006. This results in forecasting for a whole calender year starting in January, which seems most natural. For consistency with the modelling in the previous Chapter, the simple exponential fitting will be applied from 1988 to the end of 2005 and the GLMs will be fitted to data from 1989 to the end of 2005.

6.1.2 Variable Choice

We now develop two sets of uniform regional models: that is, within in each set, the same set of variables are used in the model for each health board. One set will include past observations and the other will include a variable based on simple exponential smoothing (SES). In the tailored models developed in Chapter 5, a number of different variables were used in each region. In each model we only included those variables that contributed significantly to the model. When developing uniform models, we choose variables that benefit the largest number of boards, and/or the largest proportion of the Scottish populations. The choice with some variables is straight forward. For seasonality, nearly all regions have four trigonometric harmonics fitted to them in Chapter 5 and so we include four harmonics in each regional model. The $zLen$ and $pLen$ variables – those that measure the run lengths of zeros and positive numbers respectively – are only significant in a small number of the regional models, so both variables are excluded. The choice with the linear trend is not so immediately obvious. Coefficients of linear trend in the tailored models are very small – see Tables 5.5 and 5.10. However, the linear trend term takes some of largest values when compared with the other variables, since it takes integer values that correspond to the week number as counted from the start of the series in 1988, to its end nearly twenty

years later; thus it takes integer values between 1 and 1040. This contrasts with the trigonometric terms which take values between -1 and 1, and the other terms which typically take small positive values. Thus, even though the linear trend terms have very small coefficients, they have the potential to have a noticeable effect on the calculated means since they take such large values at points. Given the linear trend term is included within the models fitted to regions that contain the largest populations – see Table 5.1 – we include it in our models here. When the tailored models were fitted, we noted that there may be a linear trend in the other regions that the models did not have the power to detect, due to the decreasing gradients being very small. Thus, including a linear trend seems unlikely to degrade these models. However, including such a deterministic linear trend in the system will necessitate the fit of the models being reviewed periodically. For simplicity, we do not consider the inclusion of past outbreaks in the uniform models. This means that the uniform models have no provision for dealing with past outbreaks, in contrast to the ERS system which does (see Section 5.2).

So far, we have decided on variables that both sets of models will have in common; both sets of models will include four trigonometric harmonics and a linear trend term for all regions. One set of models will include the variable based on SES smoothing. There is nothing remaining to decide for these models. For the other models that include past observations, it remains to decide how many past observations should be included in each regional model. The number included in the tailored past observation models varied between zero and five (the maximum number considered) – see Figure 5.11. In the tailored modelling, three boards included five past observations in their models: Glasgow, Grampian, and Lanarkshire. Since these three boards contain approximately forty percent of the Scottish population (Table 5.1), so choose to include five past observations in the uniform past observation models.

6.1.3 Fitted models

Thus, we have decided on the variables that will be included in our uniform models. The coefficients of the models that utilise past observations, along side four trigonometric harmonics and a linear trend term, can be found in Table 6.1, with the summary statistics of these models given in Table 6.2. The parameters

of the smoothing that forms the SES variables for use in the other set of models are shown in Table 6.3. Coefficients of the models that include this variable alongside the four harmonics and linear trend are shown in Table 6.4, with their summary statistics given in Table 6.5. Given we looked in depth at the similarities and differences between the regional models in Chapter 5, we do not do so here. However, it is interesting to note that the SES models have a smaller AIC than the corresponding past observation models in nine of the twelve regions, suggesting better fitting models.

6.2 Systems Comparison Methodology

6.2.1 Choice of data

We fit the uniform models to data from 1988 to the end of 2005. By the nature of the system in McCabe (2004), it will only use data from 2001 – 2005, even though we provide it with data from 1988.

6.2.2 Producing Forecasts

To produce forecasts for the past observation models is straight forward. We fit the models to data from 1988 to the end of 2005. These fitted models are then used to form predictions for the fifty-two weeks of 2006; this is done easily from within `R` by using the `predict.glm` function. This function takes new values of variables in a GLM and calculates the mean that results from these values (Faraway 2006). Thus, we calculate the value of the variables used in the past observations models for each week in 2006 and then pass them to `predict.glm` function to find the corresponding forecasts for each week.

As noted previously in Section 5.4, fitting the SES models is a slightly more complex two-stage process. First, we must find the SES one-step ahead predictions for 2006. To do this, we first fit SES models to data from 1988 to the end of 2005 by using the `HoltWinters` function. From these models, estimates of the smoothing parameters α are found by minimising the squared one-step ahead prediction errors; we denote these estimates $\hat{\alpha}$. Recall that in the Holt-Winters models considered in Section 5.4.1, we found that simple exponential smoothing (with the one parameter α) was the best form of exponential smoothing for

Region	Intercept	$\sin\left(\frac{2\pi}{52}\right)$	$\cos\left(\frac{2\pi}{52}\right)$	$\sin\left(\frac{4\pi}{52}\right)$	$\cos\left(\frac{4\pi}{52}\right)$	$\sin\left(\frac{6\pi}{52}\right)$
Arygll & Clyde	-0.057 (0.136)	0.012 (0.059)	-0.068 (0.061)	-0.181 (0.059)	-0.072 (0.061)	0.080 (0.060)
Ayrshire & Arran	-0.702 (0.170)	-0.071 (0.090)	-0.292 (0.098)	-0.299 (0.093)	-0.224 (0.096)	0.128 (0.093)
Borders	-1.388 (0.175)	-0.530 (0.129)	-0.630 (0.121)	-0.611 (0.122)	0.075 (0.120)	0.041 (0.115)
Dumfries & Galloway	-0.439 (0.150)	0.020 (0.069)	-0.483 (0.087)	-0.323 (0.077)	-0.299 (0.076)	0.053 (0.074)
Fife	-1.230 (0.158)	-0.294 (0.095)	-0.429 (0.101)	-0.175 (0.097)	-0.257 (0.096)	-0.026 (0.095)
Forth Valley	-1.065 (0.158)	-0.316 (0.095)	-0.291 (0.095)	-0.307 (0.093)	-0.235 (0.098)	0.130 (0.092)
Glasgow	-0.170 (0.125)	-0.027 (0.058)	-0.267 (0.062)	-0.227 (0.058)	-0.086 (0.060)	0.215 (0.058)
Grampian	0.109 (0.114)	-0.016 (0.051)	-0.113 (0.052)	-0.082 (0.050)	-0.154 (0.052)	0.177 (0.051)
Highland	-0.568 (0.137)	-0.406 (0.084)	-0.505 (0.081)	-0.429 (0.078)	-0.016 (0.081)	0.212 (0.078)
Lanarkshire	0.142 (0.133)	-0.061 (0.053)	-0.103 (0.054)	-0.185 (0.054)	-0.133 (0.055)	0.061 (0.054)
Lothian	0.195 (0.125)	-0.221 (0.055)	-0.179 (0.050)	-0.150 (0.051)	-0.066 (0.051)	0.041 (0.051)
Tayside	-0.697 (0.127)	-0.255 (0.078)	-0.365 (0.074)	-0.362 (0.073)	-0.125 (0.075)	0.010 (0.072)

Region	$\cos\left(\frac{6\pi}{52}\right)$	$\sin\left(\frac{8\pi}{52}\right)$	$\cos\left(\frac{8\pi}{52}\right)$	Linear Trend
Arygll & Clyde	0.148 (0.059)	-0.123 (0.061)	-0.233 (0.061)	-0.000659 (0.000175)
Ayrshire & Arran	0.202 (0.093)	0.007 (0.092)	-0.192 (0.092)	-0.000785 (0.000259)
Borders	0.274 (0.114)	-0.132 (0.108)	-0.169 (0.107)	-0.000224 (0.000268)
Dumfries & Galloway	0.161 (0.073)	-0.024 (0.071)	-0.149 (0.071)	-0.000682 (0.000200)
Fife	0.152 (0.094)	-0.037 (0.092)	-0.113 (0.092)	-0.000021 (0.000242)
Forth Valley	0.174 (0.092)	0.033 (0.092)	-0.289 (0.092)	-0.000095 (0.000241)
Glasgow	0.140 (0.059)	-0.051 (0.057)	-0.111 (0.057)	-0.000367 (0.000161)
Grampian	0.082 (0.051)	-0.123 (0.052)	-0.315 (0.052)	-0.000171 (0.000138)
Highland	0.152 (0.076)	-0.134 (0.075)	-0.179 (0.075)	-0.000010 (0.000194)
Lanarkshire	0.200 (0.054)	-0.029 (0.054)	-0.235 (0.054)	-0.000654 (0.000159)
Lothian	0.143 (0.050)	-0.025 (0.050)	-0.115 (0.050)	-0.000339 (0.000144)
Tayside	0.309 (0.073)	0.105 (0.072)	-0.233 (0.072)	0.000207 (0.000189)

Region	$\log(x_{t-1}+1)$	$\log(x_{t-2}+1)$	$\log(x_{t-3}+1)$	$\log(x_{t-4}+1)$	$\log(x_{t-5}+1)$
Arygll & Clyde	0.268 (0.066)	0.280 (0.066)	0.152 (0.068)	0.152 (0.067)	0.084 (0.067)
Ayrshire & Arran	0.327 (0.128)	0.167 (0.131)	0.164 (0.132)	0.102 (0.135)	0.068 (0.137)
Borders	0.351 (0.154)	0.174 (0.159)	0.110 (0.164)	-0.179 (0.175)	0.287 (0.166)
Dumfries & Galloway	0.180 (0.091)	0.043 (0.092)	0.183 (0.092)	0.224 (0.093)	0.111 (0.094)
Fife	0.301 (0.134)	0.265 (0.135)	-0.013 (0.141)	0.263 (0.137)	0.242 (0.138)
Forth Valley	0.176 (0.121)	0.301 (0.120)	0.407 (0.119)	0.237 (0.123)	0.009 (0.127)
Glasgow	0.347 (0.062)	0.293 (0.063)	0.050 (0.064)	0.012 (0.064)	0.166 (0.063)
Grampian	0.052 (0.051)	0.241 (0.050)	0.229 (0.050)	0.232 (0.050)	0.054 (0.051)
Highland	0.027 (0.094)	0.149 (0.093)	0.065 (0.095)	0.069 (0.096)	-0.047 (0.098)
Lanarkshire	0.307 (0.055)	0.226 (0.056)	0.155 (0.056)	0.090 (0.056)	0.081 (0.056)
Lothian	0.224 (0.050)	0.214 (0.050)	0.174 (0.050)	0.096 (0.050)	0.084 (0.050)
Tayside	0.404 (0.084)	0.178 (0.087)	0.113 (0.089)	-0.151 (0.091)	0.209 (0.089)

Table 6.1: The GLMs fitted in Section 6.1 to the *Cryptosporidium* data in the different regions, utilising past observations to deal with serial correlation. The tables give the coefficients of the variables and their associated standard errors in brackets.

Region	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	θ	θ SE
Argyll & Clyde	1129	904	225	20	883	869	14	2827	1.377	0.145
Ayrshire & Arran	765	668	98	13	883	869	14	1616	0.825	0.138
Borders	717	596	121	17	883	869	14	1294	1.097	0.264
Dumfries & Galloway	1041	820	221	21	883	869	14	2009	1.920	0.357
Fife	752	668	84	11	883	869	14	1529	1.012	0.195
Forth Valley	790	674	116	15	883	869	14	1716	0.754	0.110
Glasgow	1271	928	343	27	883	869	14	2824	1.718	0.201
Grampian	1229	982	247	20	883	869	14	3587	1.556	0.137
Highland	962	805	158	16	883	869	14	2022	1.459	0.239
Lanarkshire	1239	939	300	24	883	869	14	3199	1.603	0.154
Lothian	1284	974	310	24	883	869	14	3579	1.606	0.142
Tayside	1025	811	214	21	883	869	14	2284	1.187	0.141

Table 6.2: Summary statistics from the negative binomial GLMs utilising past observations fit in Section 6.1. DoF = Degrees of Freedom. SE = Standard Error.

Region	α
Argyll & Clyde	0.2001
Ayrshire & Arran	0.1641
Borders	0.0382
Dumfries & Galloway	0.2053
Fife	0.0847
Forth Valley	0.1065
Glasgow	0.2567
Grampian	0.1793
Highland	0.1264
Lanarkshire	0.2358
Lothian	0.2235
Tayside	0.2527

Table 6.3: The smoothing parameters found by the `HoltWinters` function for each region when SES is applied in Section 6.1 to the regional counts of *Cryptosporidium*.

Region	Intercept	SES	$\sin\left(\frac{2\pi}{52}\right)$	$\cos\left(\frac{2\pi}{52}\right)$	$\sin\left(\frac{4\pi}{52}\right)$	$\cos\left(\frac{4\pi}{52}\right)$
Argyll & Clyde	-0.295 (0.154)	1.191 (0.129)	0.043 (0.059)	-0.042 (0.062)	-0.211 (0.058)	-0.090 (0.061)
Ayrshire & Arran	-0.876 (0.202)	1.268 (0.361)	-0.037 (0.091)	-0.288 (0.098)	-0.324 (0.091)	-0.252 (0.096)
Borders	-1.711 (0.247)	2.363 (0.858)	-0.500 (0.131)	-0.723 (0.117)	-0.651 (0.121)	0.091 (0.118)
Dumfries & Galloway	-0.580 (0.170)	0.973 (0.220)	0.055 (0.070)	-0.461 (0.088)	-0.339 (0.076)	-0.323 (0.076)
Fife	-1.444 (0.206)	1.790 (0.527)	-0.288 (0.099)	-0.511 (0.101)	-0.199 (0.096)	-0.302 (0.097)
Forth Valley	-1.279 (0.194)	1.789 (0.383)	-0.325 (0.097)	-0.381 (0.097)	-0.373 (0.091)	-0.275 (0.099)
Glasgow	-0.311 (0.135)	1.033 (0.110)	0.026 (0.059)	-0.260 (0.062)	-0.263 (0.057)	-0.107 (0.060)
Grampian	-0.036 (0.126)	0.972 (0.093)	0.008 (0.052)	-0.144 (0.052)	-0.123 (0.050)	-0.158 (0.052)
Highland	-0.603 (0.176)	0.367 (0.343)	-0.392 (0.092)	-0.527 (0.078)	-0.453 (0.076)	-0.019 (0.082)
Lanarkshire	-0.058 (0.148)	1.036 (0.104)	-0.038 (0.053)	-0.094 (0.054)	-0.202 (0.053)	-0.153 (0.055)
Lothian	0.050 (0.138)	0.917 (0.086)	-0.204 (0.056)	-0.221 (0.050)	-0.174 (0.051)	-0.084 (0.051)
Tayside	-0.795 (0.134)	0.979 (0.163)	-0.203 (0.080)	-0.391 (0.075)	-0.376 (0.073)	-0.164 (0.075)

Region	$\sin\left(\frac{6\pi}{52}\right)$	$\cos\left(\frac{6\pi}{52}\right)$	$\sin\left(\frac{8\pi}{52}\right)$	$\cos\left(\frac{8\pi}{52}\right)$	linTrend
Argyll & Clyde	0.108 (0.059)	0.160 (0.059)	-0.173 (0.059)	-0.233 (0.059)	-0.000524 (0.000179)
Ayrshire & Arran	0.150 (0.092)	0.230 (0.092)	-0.013 (0.091)	-0.216 (0.091)	-0.000680 (0.000268)
Borders	0.073 (0.114)	0.278 (0.113)	-0.156 (0.106)	-0.172 (0.106)	-0.000199 (0.000272)
Dumfries & Galloway	0.070 (0.073)	0.162 (0.072)	-0.038 (0.070)	-0.143 (0.070)	-0.000595 (0.000206)
Fife	0.010 (0.095)	0.158 (0.095)	-0.054 (0.092)	-0.117 (0.092)	0.000067 (0.000244)
Forth Valley	0.156 (0.093)	0.175 (0.093)	-0.039 (0.091)	-0.308 (0.091)	-0.000035 (0.000243)
Glasgow	0.254 (0.057)	0.155 (0.058)	-0.069 (0.057)	-0.113 (0.057)	-0.000298 (0.000164)
Grampian	0.229 (0.051)	0.057 (0.051)	-0.185 (0.051)	-0.273 (0.051)	-0.000187 (0.000139)
Highland	0.231 (0.077)	0.155 (0.076)	-0.155 (0.074)	-0.179 (0.074)	-0.000017 (0.000195)
Lanarkshire	0.089 (0.054)	0.223 (0.053)	-0.060 (0.053)	-0.253 (0.053)	-0.000535 (0.000164)
Lothian	0.072 (0.051)	0.157 (0.050)	-0.044 (0.050)	-0.103 (0.050)	-0.000280 (0.000148)
Tayside	0.011 (0.072)	0.332 (0.073)	0.100 (0.072)	-0.270 (0.071)	0.000215 (0.000191)

Table 6.4: Coefficients of the GLMs fitted in Section 6.1 to the *Cryptosporidium* data in the different regions, utilising SES to deal with serial correlation. The tables give the coefficients of the variables and their associated standard errors in brackets.

Region	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	θ	θ SE
Argyll & Clyde	1126	904	223	20	883	873	10	2821	1.371	0.144
Ayrshire & Arran	763	667	96	13	883	873	10	1610	0.816	0.135
Borders	714	598	116	16	883	873	10	1291	1.081	0.258
Dumfries & Galloway	1039	819	220	21	883	873	10	2001	1.911	0.354
Fife	738	662	75	10	883	873	10	1528	0.944	0.174
Forth Valley	774	670	105	14	883	873	10	1716	0.713	0.101
Glasgow	1254	925	329	26	883	873	10	2824	1.661	0.191
Grampian	1206	976	231	19	883	873	10	3590	1.498	0.129
Highland	957	803	154	16	883	873	10	2017	1.431	0.232
Lanarkshire	1226	939	287	23	883	873	10	3200	1.567	0.149
Lothian	1269	971	298	23	883	873	10	3579	1.568	0.136
Tayside	1001	807	194	19	883	873	10	2290	1.115	0.128

Table 6.5: Summary statistics from the negative binomial GLMs with SES fit in Section 6.1. DoF = Degrees of Freedom. SE = Standard Error.

our purpose. To force the `HoltWinters` function to carry out simple exponential smoothing, we manually specify the parameters β and γ as zero. We then fit a SES models to data from 1988 to the end of 2006, manually specifying the smoothing parameter α as $\hat{\alpha}$. From these models we can extract the values of SES (the variable in the SES models that are the simple exponential smoothing one-step ahead predictions) that would be calculated had the system been run in real time during 2006. Normally, the predictions could be found more easily by using the `predict.HoltWinters` function. The other variables in the SES GLMs can be calculated directly for 2006. We then have all the information required to proceed as with the past observation models: we use `predict.glm` to calculate the corresponding weekly forecasts from the appropriate variable values for each week.

To obtain forecasts from the system developed in McCabe (2004), we were given a copy of its coded implementation in use at HPS. We shall refer to this system as the Exceedance Reporting System (ERS). This implementation has been written in `R`. We apply the system to the regional counts of *Cryptosporidium*. The system is reasonably straight forward to use, requiring a separate text file for each set of counts that the system is to be applied to; for an example of such a file see Figure 6.1. We create a suitable text file for each health board (excluding the

Cryptosporidium Argyll & Clyde	2	1988	1
Cryptosporidium Argyll & Clyde	2	1988	2
Cryptosporidium Argyll & Clyde	2	1988	3
Cryptosporidium Argyll & Clyde	0	1988	4
Cryptosporidium Argyll & Clyde	0	1988	5
Cryptosporidium Argyll & Clyde	0	1988	6
Cryptosporidium Argyll & Clyde	0	1988	7
Cryptosporidium Argyll & Clyde	0	1988	8
Cryptosporidium Argyll & Clyde	0	1988	9
Cryptosporidium Argyll & Clyde	0	1988	10

Figure 6.1: A small section of a suitable data-file that can be used by the Exceedance Reporting System (ERS), which is described in Section 4.2.2. The first two columns specify the disease and subtype that the counts pertain to; we use the sub-type to record which region the counts come from. The third column gives the number of reported cases of *Cryptosporidium* reported for that year and week combination, recorded in the fourth and fifth columns respectively.

island health boards) and then apply the ERS to obtain forecasts and exceptions. Results are output in further text files, which we re-read back into `R` to allow easy comparison between the forecasts of this system and the uniform models. No substantive changes were made to the system other than to adapt its output to ease analysis.

6.2.3 Producing Exceptions

Using the uniform models, we apply an exceedance alerting method to the forecasts of each system. This creates exceptions for 2006 and allows comparison with the exceptions given by the ERS. McCabe (2004) found that this method gave the best results for the national monitoring of organisms. If either set of the uniform models were adopted by HPS, we would expect them to 'slot into' the present system, replacing the forecast model developed in McCabe (2004) when dealing with *Cryptosporidium*. Thus, the exceedance alerting method would be applied to them and so we compare the three sets of models using this alarm method. In Section 6.4 we consider further the implications of the type of alerting method used.

As was adopted in McCabe (2004), we extend the exceedance reporting method to have multiple stages, similar to the monitoring system used in Allardice, Wright, Peterson, and Miller (2001). Recall the general principle of an exceedance reporting system: we find an expected distribution of counts for the present week; from this, an upper quantile that corresponds to a particular p-value is found and if the observed counts *exceed* this quantile an exception is raised; thus, such exceptions can also be referred to as exceedances. We use p-values of 0.01 and 0.1, so that if an observed count is larger than the quantile corresponding to 0.01, a 'red' exception is raised; if the observed count is smaller than this quantile, but larger than the quantile corresponding to a p-value of 0.1, an 'amber' exception is raised; otherwise the organism is considered in the 'green' state and no exception is reported.

With the uniform models developed here, finding the quantiles is straight forward: these models give the forecasted means for 2006, along with values of θ (the negative binomial dispersion parameter). These values are then used as parameters in the negative binomial quantile function `qnbinom`, to calculate the values used for comparison to determine if a reported count should raise an exception. However, a complication arises because we are generally dealing with very low counts; thus, the fitted distributions have most of their probability in the very small integers (obviously, since we are dealing with counts, they are either zero or a positive integer). For instance, consider a negative binomial distribution X, with mean 0.6 and dispersion parameter 0.25. These values are chosen for clarity of demonstration but are not atypical for the regional models. Part of the probability function $f(x)$ ($f(x) = P(X = x)$) is shown in Figure 6.2. Here we see that most of the probability is 'contained' between zero and eight, with higher values having very small probabilities. Now, consider the cumulative distribution function $F(x)$, which is defined by $F(x) = P(X \leq x)$. When we use the quantile function `qnbinom` to find the amber exceedance limit, we seek the largest x such that $F(x) = 0.9$ (since we are using a p-value of 0.1 for this level); this is shown visually in Figure 6.3. With this particular distribution, two is shown to be the required value. However, the p-value linked with two is 0.134 ($P(X \geq 2) = 1 - P(X < 2) = 1 - f(1) - f(2) = 1 - 0.736 - 0.130 = 0.134$). More generally, this means that by using `qnbinom` directly we will have more exceptions than we would expect for our chosen reporting levels: here, for example, we seek

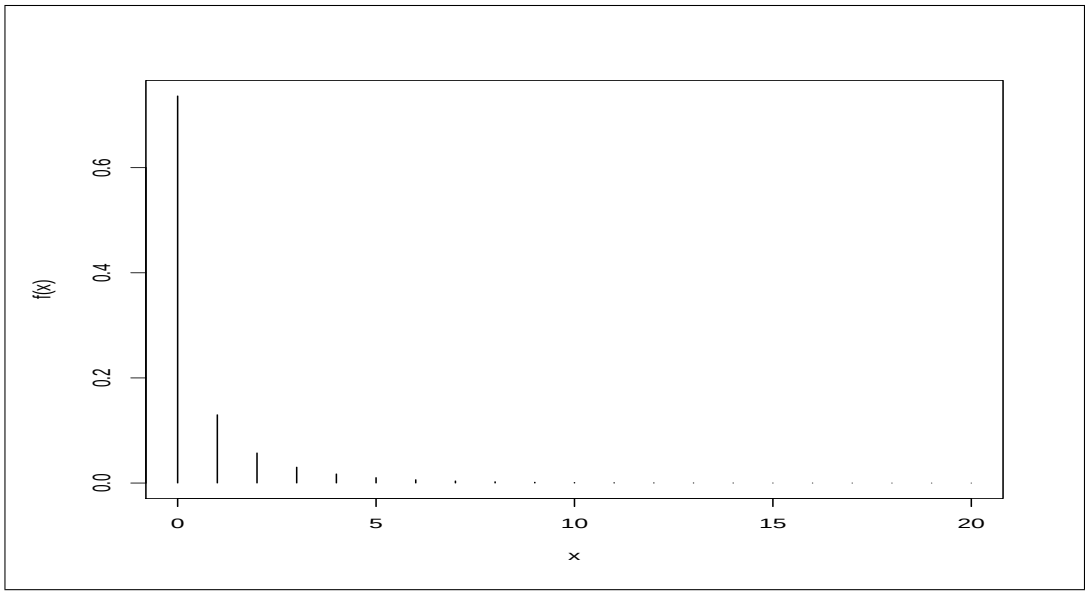


Figure 6.2: The probability densities for a negative binomial distribution X with mean 0.6 and dispersion parameter (θ) 0.75. Since we are dealing with very small counts at the regional levels, much of the probability is contained in the small integers.

a reporting level related to p-value of 0.1, but rather find one related to a p-value of 0.13. We choose to add one to the values returned by `qnbinom`, so that we always have corresponding p-values smaller than those sought. We would expect this to reduce the false reporting rate. This more conservative regional reporting rate is likely sensible, since we have increased the number of time series to be monitored (one national time series, now with an additional twelve regional ones (excluding the islands)). As McCabe (2004) notes, if an exception rate is too high, users will start to disregard an exception system. In general, with the potential of many more regional series to be monitored, and thus more exceptions to be investigated, reducing the number of exceptions will hopefully reduce user fatigue towards the system.

The output of the ERS system automatically indicates any exceptions. We alter the code to return quantiles corresponding to 'amber' and 'red' reporting levels. However, the underlying reporting procedures remain unchanged. Also note that McCabe (2004) includes a heuristic in his system, where *any* exception is ignored if there are fewer than five cases in the previous four weeks.

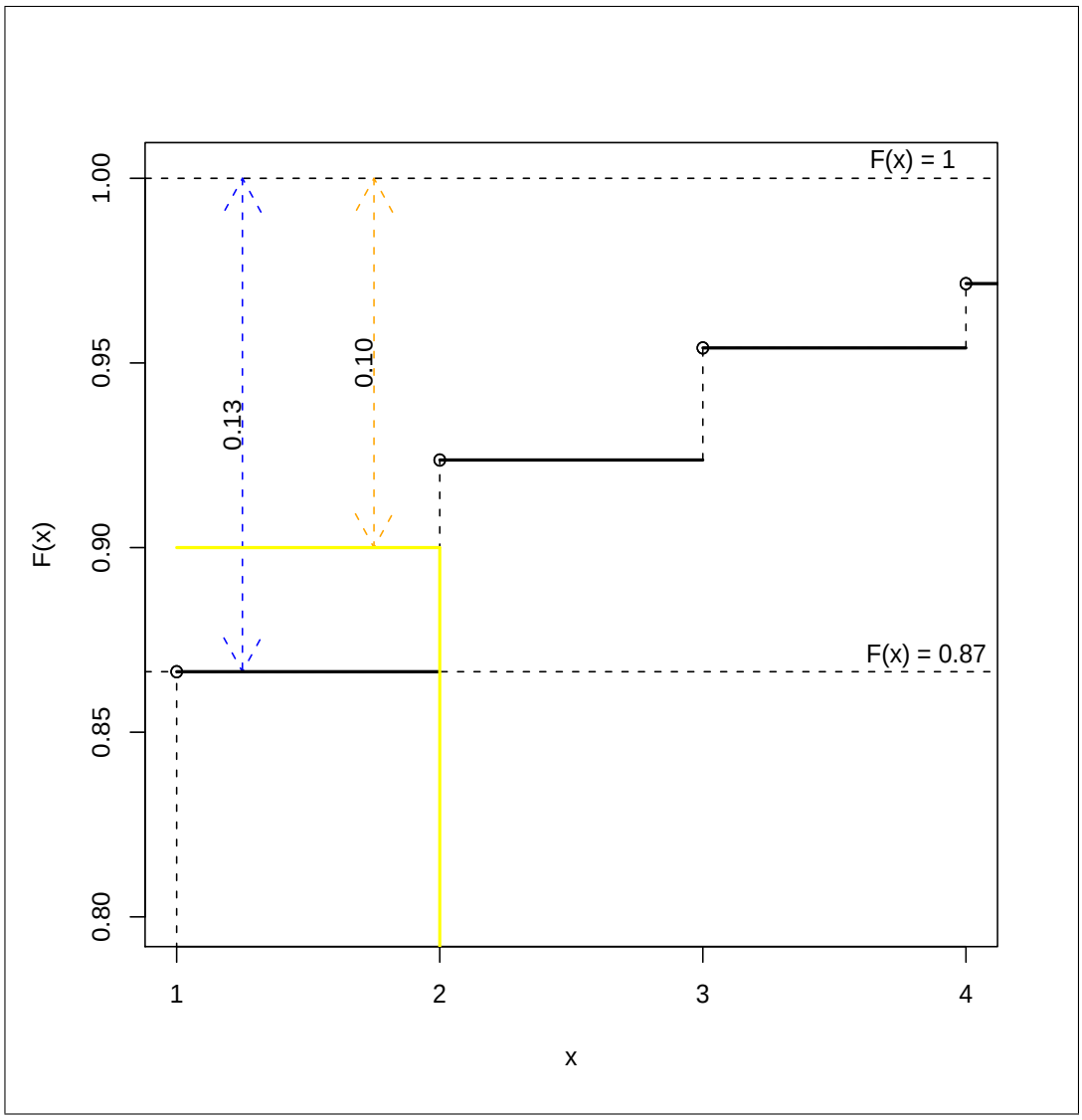


Figure 6.3: A graph of part of the cumulative distribution function for a negative binomial distribution X with mean 0.6 and dispersion parameter (θ) 0.75. The yellow line gives a graphical representation of what happens when we use the quantile function `qnbinom` to calculate the quantile for 'amber' alerts: with a p-value of 0.1, we find the suitable quantile is 2 (orange arrow). However, the p-value for 2 ($P(X \geq 2)$) is 0.13 (blue arrow).

6.3 Results from Comparisons of the Systems

Our primary purpose in comparing these systems is to look for differences in how they function as exception reporting systems. Thus, in Section 6.3.2, we compare the exceptions raised under each system. However, an essential step of an exception reporting system is forecasting, so we first focus on the forecasting abilities of each system in the next Section.

6.3.1 Forecast Errors

Forecast error is generally defined as the difference between a forecast and the observed value. Thus, for each health board, we calculate forecast errors (one for each week of 2006) from each of the three systems. To compare these sets of errors we proceed as in Burkom (2007): for each set of forecast errors, within each region, we find the median absolute value and the median absolute percentage error. These values are summarised in Table 6.6. It is important to remember that we are dealing with very small counts at the health board level; many regions do not have any weeks with more than four reports and many weeks have no reported cases. This means that many percentage errors cannot be calculated because the corresponding observed value is zero and so calculating the percentage would involve dividing by zero.

The median absolute forecast errors suggest a number of conclusions. The SES models and past observation models tend to give very similar forecast errors, suggesting they give similar forecasts. Both the ERS system and the SES models give the smallest forecast errors in six of the health boards. When the ERS gives the smallest forecast error in a region, the error tends to be close to that of the error in the other two systems. In Glasgow and Lothian, the SES and the past observation error is a third smaller than the ERS. These are the most densely populated health boards and so we might expect them to have the highest levels of serial correlation; an infection of *Cryptosporidium* is more likely to be passed between members of their densely packed populations and so stay 'active' for a longer time over consecutive weeks, increasing serial correlation. Thus, the SES and past observation models may give these relatively large improvements since they have elements to model serial correlation unlike the ERS. Ayrshire and Arran is the only health board where the ERS does (relatively) much better than

the other systems; its predictions and observed counts are shown in Figure 6.6. The observed values for 2006 seem quite 'spiky' – they jump up to one or two and then back to zero quickly. Since the past observation and SES models have terms in them for serial correlation, they will adapt more slowly to these spikes, causing larger forecast errors. There is a weak suggestion that the ERS tends to give the smallest errors in those boards to the South, West and North, while the SES models do better in the South-East and Glasgow. If we recall Figure 5.22, we note that there seems to be most interaction in the counts of *Cryptosporidium* between those boards in the Southeast and Glasgow, possibly related to higher levels of commuting between these boards. This may increase serial correlation and so be better modelled by the SES and past observation models.

There are not many percentage error values to compare in Table 6.6 because of the problem noted above about dividing by zero. At first glance these percentage errors can seem very large; however, this is primarily because we are dealing with such small counts. A forecast error of only one can often give a percentage error of fifty percent or more. In the four boards that we can compare, we find that the past observation and SES models give smaller percentage errors. Once more, our two systems appear to have similar results.

Another measure used to compare errors is the root mean square (rms) error of the forecasts:

$$\sqrt{\frac{\sum_{t=1}^{52} (f_t - o_t)^2}{52}},$$

where f_t and o_t are the forecast and observed counts respectively for week t. A graph comparing the different rms errors of the models for each board can be found in Figure 6.4. Squaring the errors means that larger errors will be penalised more than smaller errors. Generally, the past observation and SES models always do better than the ERS by this metric. In some cases the ERS is quite a bit worse than the other systems – particularly in Tayside. Again, the past observation and SES models tend to give similar results. Fife is the only board where the ERS does better than the other models, but the difference is only slight. By the rms error metric, the past observation and SES models do better in Ayrshire and Arran.

Combining these comparisons – the median analysis and then the rms error comparison – suggest that the past observation and SES models are somewhat

Forecast Error Measure	Argyll & Clyde	Ayrshire & Arran	Borders	Dumfries & Galloway	Fife	Forth Valley	Grampian	Glasgow	Highland	Lanarkshire	Lothian	Tayside
Median Absolute Deviation												
ERS System	0.53	0.01	0.30	0.43	0.63	0.27	1.59	0.99	0.39	0.74	1.62	1.07
Past Observations Models	0.62	0.36	0.33	0.41	0.57	0.32	1.12	1.12	0.51	0.74	1.11	0.89
SES Models	0.66	0.36	0.37	0.42	0.59	0.31	1.03	1.05	0.52	0.70	1.04	0.91
Median Absolute % Error												
ERS System	–	–	–	–	–	–	–	98.97	–	99.44	84.36	99.29
Past Observation Models	–	–	–	–	–	–	–	70.79	–	78.70	52.85	68.69
SES Models	–	–	–	–	–	–	–	72.58	–	74.61	50.69	73.30

Table 6.6: Comparison of forecasts from the three different systems. The highlighted cells indicate which method gives the smallest error for a particular health board. Many of the percentage errors can not be calculated because the corresponding observed value is zero and so calculating the percentage would involve dividing by zero.

poorer than the ERS at predicting values close to zero. This is sensible when we consider that the seasonality in the past observation and SES models: they model seasonality by the use of trigonometric terms and so the resulting means from these models can never get to zero. The ERS only has a linear trend and a year factor in its model (see Section 4.2.2) and so can more easily get closer to zero. Since any small predicted mean will lead to a mode at zero, this does not present much of a problem. If our primary focus were on forecasting, it may make more sense to consider modal forecast error. However, the ERS does not allow for easy calculation of a mode. The rms error graph also suggests that the past observation and SES models perform better at predicting non-zero values. These factors combine to suggest that the past observation and SES models give the best forecasts, with little difference between these two systems. These are similar to results in Burkom, Murphy, and Shmueli (2007), Burkom (2007). The lack of much difference is reasonable: both sets of models address the same modelling elements, if in slightly different ways.

6.3.2 Exception Comparisons: Using the Systems for Exception Reporting

We now turn to considering the number of exceptions raised under each system when an exceedance alerting method is used. The systems are taken in pairs and for each health board we record the number of 'amber' and 'red' exceptions. In each comparison pair we also record how many weeks have exceptions under both systems and how many weeks only have exceptions under either one of them. These values are recorded in Tables 6.7 (ERS and past observation system), 6.8 (ERS and SES system) and 6.9 (past value and SES system). We also show the results of the three systems in Figures 6.5 to 6.16: each Figure shows the three systems applied to the counts for 2006 for a particular healthboard, with observed, predicted and exceedance values for each system.

We first consider the ERS and past observation system (Table 6.7). The past observation system generally raises fewer or equal numbers of exceptions in all but two boards; in these two boards, the past observation system only raises one further exception. Very few, only three, 'red' exceptions are raised by the past observation system. This contrasts with the ERS where quite a large number of

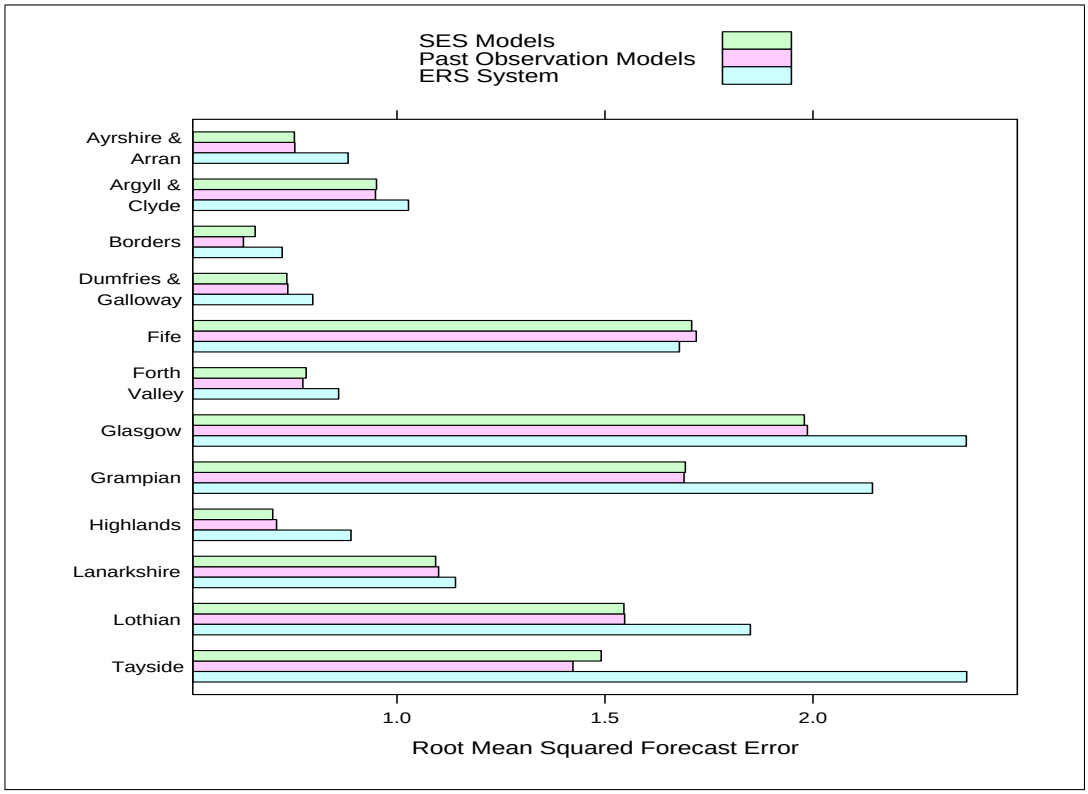


Figure 6.4: The root mean squared forecast error for each of the systems within each of the health boards for 2006.

'red' exceptions are raised; indeed, in three boards more 'red' exceptions are raised than 'amber' exceptions. This latter quality seems odd and is likely to increase user fatigue towards the system; it seems more intuitive that there should be a good deal fewer 'red' than 'amber' exceptions. The ERS system would raise many more exceptions if it did not use the heuristic of ignoring exceptions when there have been fewer than five reports in the previous four weeks. This is particularly true for the first fifteen or so weeks of 2006, as shown in Figures 6.6, 6.9, 6.14, and 6.16. However, this heuristic fails when it comes to Glasgow because of Glasgow's higher average reporting rate: see Figure 6.11. In the first nine weeks of 2006 there are six 'red' exceptions. However, because of the more consistent seasonal modelling in the past observation system, no exceptions are raised during this period. Visually inspecting the reported cases for this period suggests that the treatment of seasonality in the past observation models is more consistent; the counts are not particularly large for Glasgow. This demonstrates some of the problems of using set numerical level rules rather than directly modelling elements. Consecutive exceptions are also less likely with the past observation system because of the use of past observations in the models to remove serial correlation: when the count is high one week it is expected to be high the next week. This is perhaps seen most clearly with Glasgow in Figure 6.11, where the ERS has five consecutive exceptions towards the end of 2006. Users of the past observation system would need to be clear on the effect this would have; if an 'amber' exception is raised, the system is less likely to escalate to a 'red' exception than with the ERS. Thus, 'amber' exceptions from the past observation models should be investigated more thoroughly; however, this greater level of effort is offset by their being fewer exceptions under these models. By implication, this means that 'red' exceptions also increase in epidemiological significance.

Next, we turn to comparing the ERS and the SES models (Table 6.8). Many of the above comments also apply between these two sets of models. The SES models raise fewer exceptions than the ERS and of the exceptions that it does raise, a very small number of them are 'red'. The SES models also deal better with the beginning of the year due to how seasonality is dealt with in these models.

It remains to compare the past observation and SES models (Table 6.9). The total number of exceptions raised by both of these systems is nearly equal at thirty-four with the past observations models and thirty-six with the SES mod-

els. The number of 'red' exceptions raised is very small, with three raised by past observations and two by SES models. Nine boards had equal numbers of exceptions and the majority of these exceptions occurred during the same weeks.

When we consulted the HPS weekly reports, we found no outbreaks reported for 2006. However, in Fife during week forty-two, there are twelve reported cases of *Cryptosporidium* (see Figure 6.9). This is about twenty-five times larger than the average weekly count of cases for Fife and so is strongly suggestive of an outbreak, or at least an exception that should be investigated further. It is reassuring that all three systems report a 'red' exception for this week. Unfortunately, there are no other particularly obvious high counts which may raise exceptions and allow for further comparison between the systems.

We noted in Section 6.2.3 that the way we calculate the p-values in the past observation and SES models makes them more conservative than the ERS, and thus they produce fewer exceptions. However, a number of the 'extra' exceptions in the ERS are caused by the heuristic rule (ignoring any exception if there have been fewer than 5 cases reported in the previous four weeks) failing in certain cases. In the past observation and SES models, this heuristic is not required because of the more precise modelling of seasonality by trigonometric terms, which are better suited to capturing the seasonal pattern of *Cryptosporidium* throughout the year. This can be clearly seen in Figures 6.5 to 6.16, where the envelopes formed by the exceedance limits are generally much more smooth and consistent under our models compares to ERS. Further, the use of modelling elements to remove serial correlation reduces the likelihood of consecutive runs of exceptions; thus, when an exception occurs, it is important to consider it more carefully than under the ERS. As noted previously, this more conservative approach is justified when monitoring the greater number of regional time series. There are few differences to inform the choice between the past observation and SES models. However, the SES models may be preferred due to their more automatic design (the past observation models require a choice to be made about the number of past observations to included in their models).

	Argyll & Clyde	Ayrshire & Arran	Borders	Dumfries & Galloway	Fife	Forth Valley	Glasgow	Grampian	Highland	Lanarkshire	Lothian	Tayside
ERS												
Amber	2	1	2	1	2	2	5	1	2	3	4	6
Red	0	2	0	0	3	0	13	0	0	2	0	6
Total	2	3	2	1	5	2	18	1	2	5	4	12
Past Observation												
Amber	2	6	2	1	4	2	6	1	0	1	1	5
Red	0	0	0	0	1	0	1	1	0	0	0	0
Total	2	6	2	1	5	2	7	2	0	1	1	5
Comparison												
Agree	2	1	1	0	3	1	6	1	0	1	1	3
ERS only	0	2	1	1	2	1	12	0	2	4	3	9
PO only	0	5	1	1	2	1	1	1	0	0	0	2

Table 6.7: Exceedances comparisons for ERS and past observation models.

	Argyll & Clyde	Ayrshire & Arran	Borders	Dumfries & Galloway	Fife	Forth Valley	Glasgow	Grampian	Highland	Lanarkshire	Lothian	Tayside
ERS												
Amber	2	1	2	1	2	2	5	1	2	3	4	6
Red	0	2	0	0	3	0	13	0	0	2	0	6
Total	2	3	2	1	5	2	18	1	2	5	4	12
SES												
Amber	2	6	3	1	4	2	6	1	0	1	1	7
Red	0	0	0	0	1	0	0	1	0	0	0	0
Total	2	6	3	1	5	2	6	2	0	1	1	7
Comparison												
Agree	2	1	1	0	3	1	5	1	0	1	1	5
ERS only	0	2	1	1	2	1	13	0	2	4	3	7
SES only	0	5	2	1	2	1	1	1	0	0	0	2

Table 6.8: Exceedances comparisons for ERS and SES models.

	Argyll & Clyde	Ayrshire & Arran	Borders	Dumfries & Galloway	Fife	Forth Valley	Glasgow	Grampian	Highland	Lanarkshire	Lothian	Tayside
Past Obsevation												
Amber	2	6	2	1	4	2	6	1	0	1	1	5
Red	0	0	0	0	1	0	1	1	0	0	0	0
Total	2	6	2	1	5	2	7	2	0	1	1	5
SES												
Amber	2	6	3	1	4	2	6	1	0	1	1	7
Red	0	0	0	0	1	0	0	1	0	0	0	0
Total	2	6	3	1	5	2	6	2	0	1	1	7
Comparison												
Agree	2	6	2	1	5	2	6	2	0	1	1	5
PO only	0	0	0	0	0	0	1	0	0	0	0	0
SES only	0	0	1	0	0	0	0	0	0	0	0	2

Table 6.9: Exceedances comparisons for past observation and SES models.

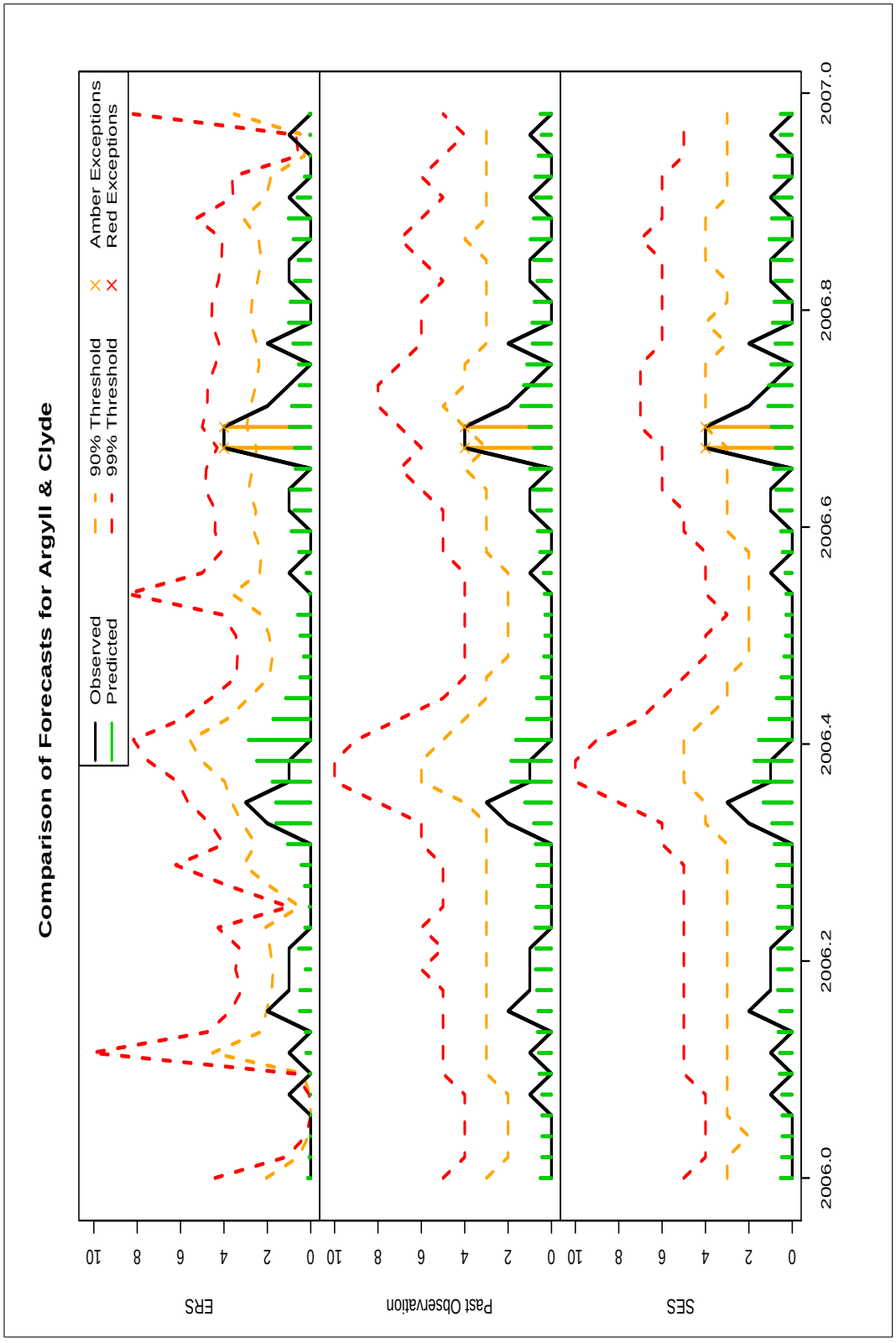


Figure 6.5: The three different forecast systems for 2006, with predictions and exceedance levels for Argyll & Clyde.

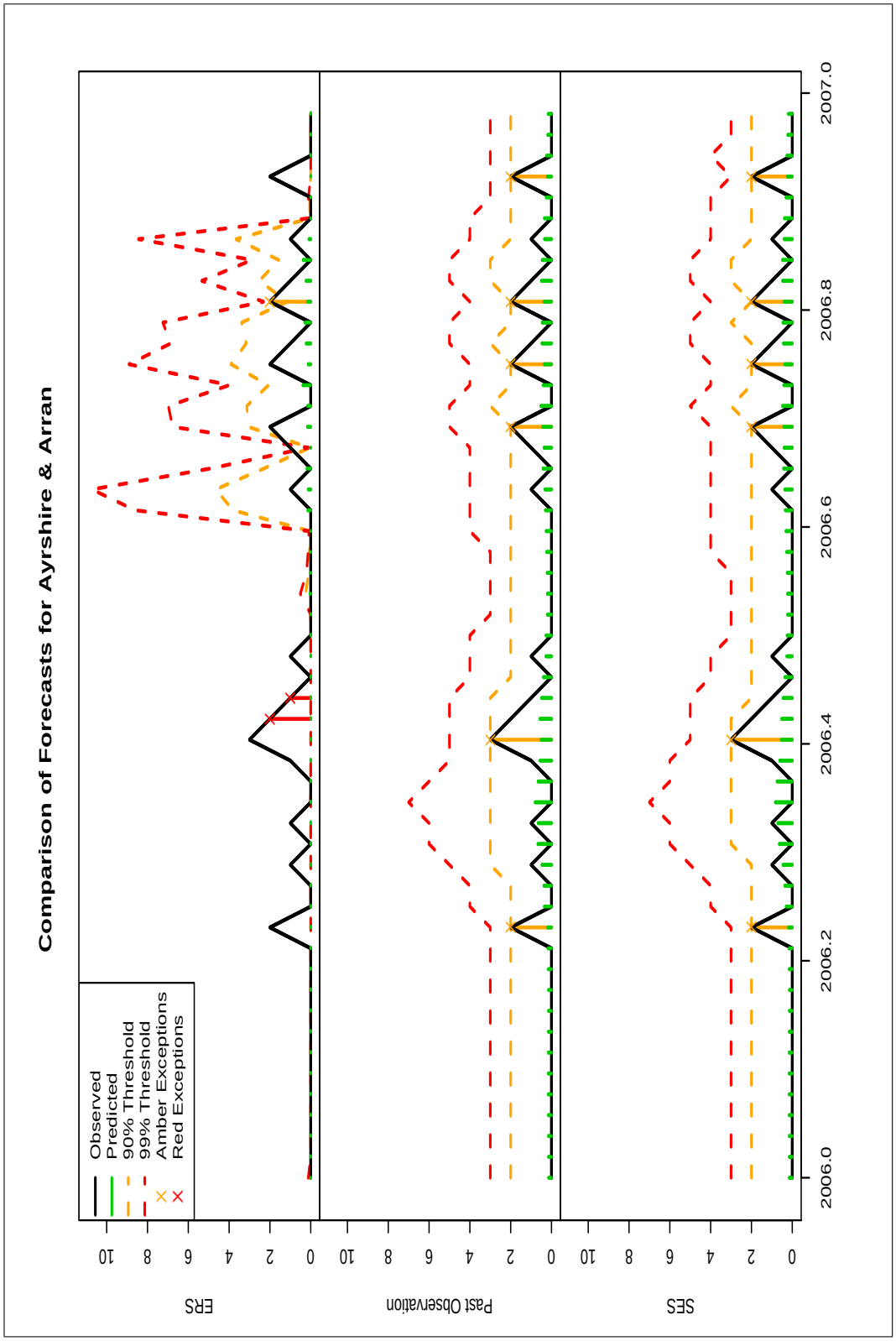


Figure 6.6: The three different forecast systems for 2006, with predictions and exceedance levels for Aryshire & Arran.

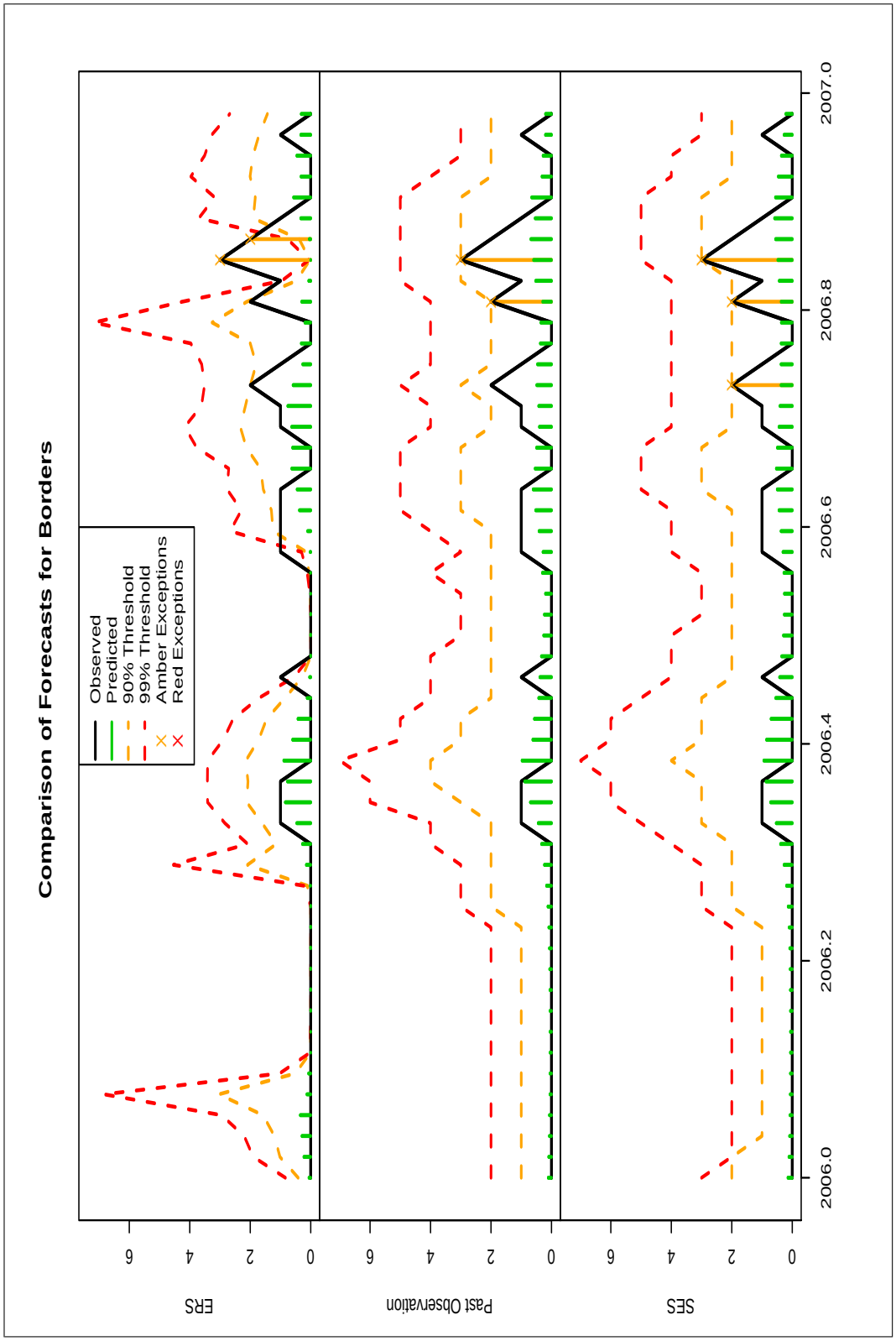


Figure 6.7: The three different forecast systems for 2006, with predictions and exceedance levels for Borders.

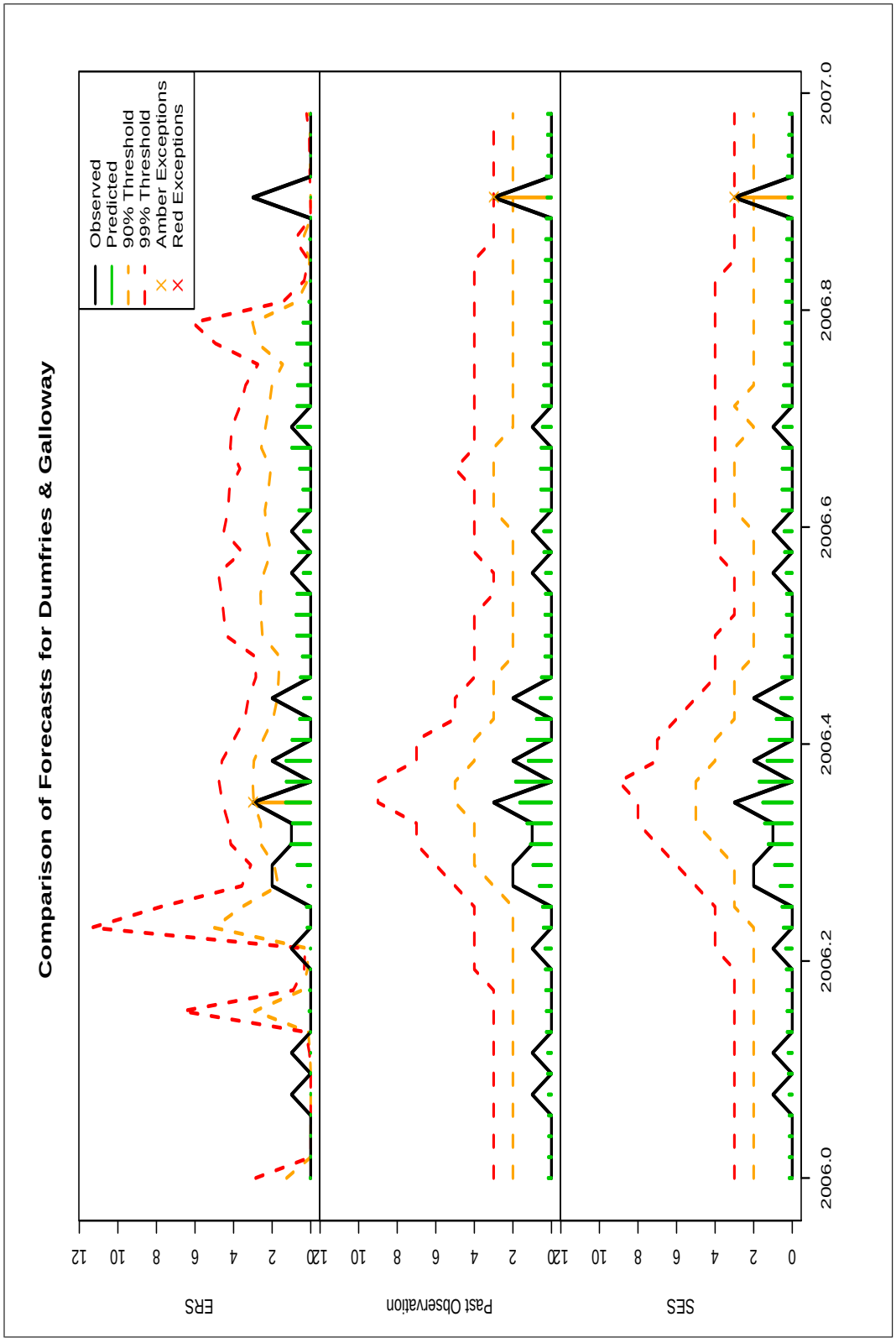


Figure 6.8: The three different forecast systems for 2006, with predictions and exceedance levels for Dumfries & Galloway.

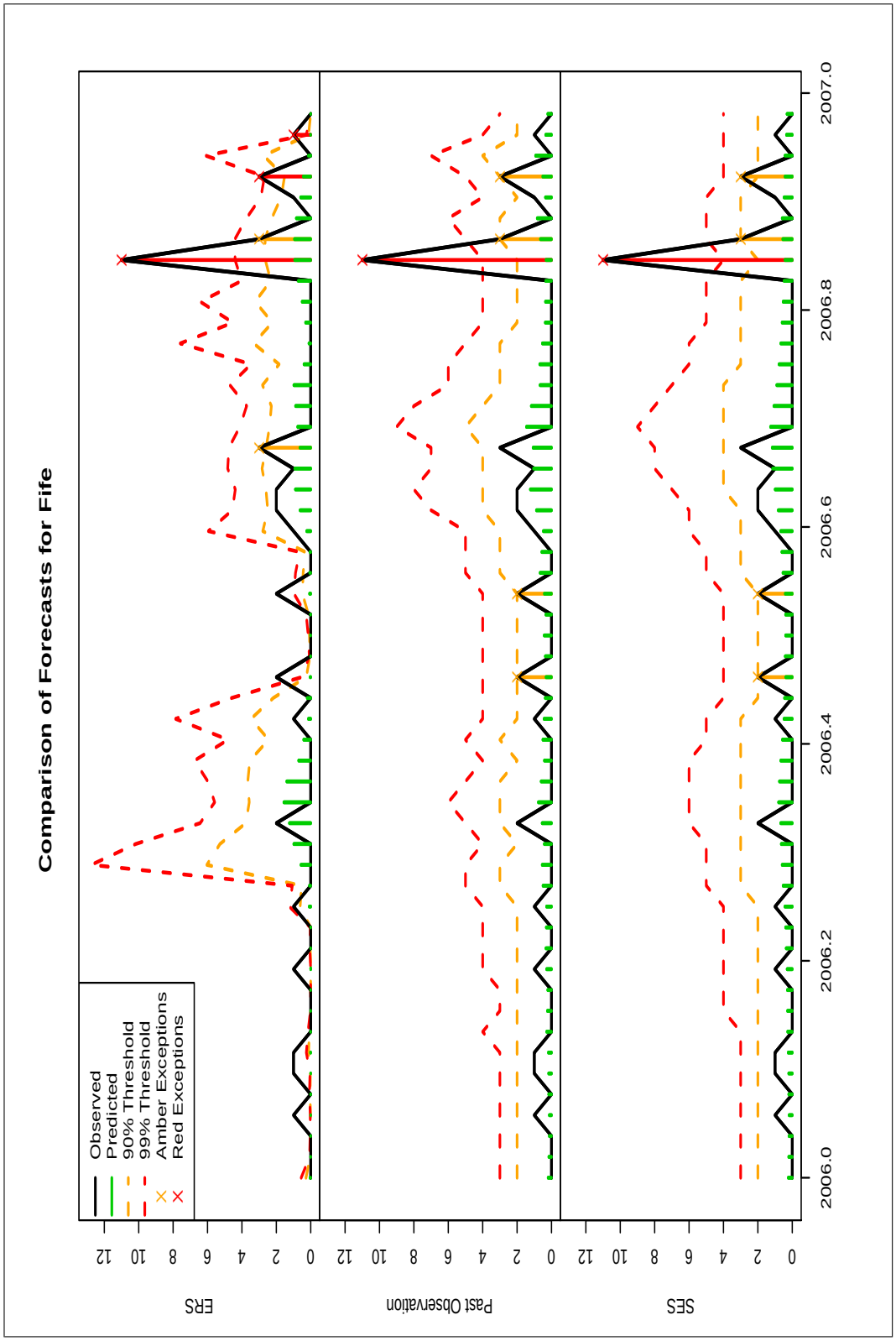


Figure 6.9: The three different forecast systems for 2006, with predictions and exceedance levels for Fife.

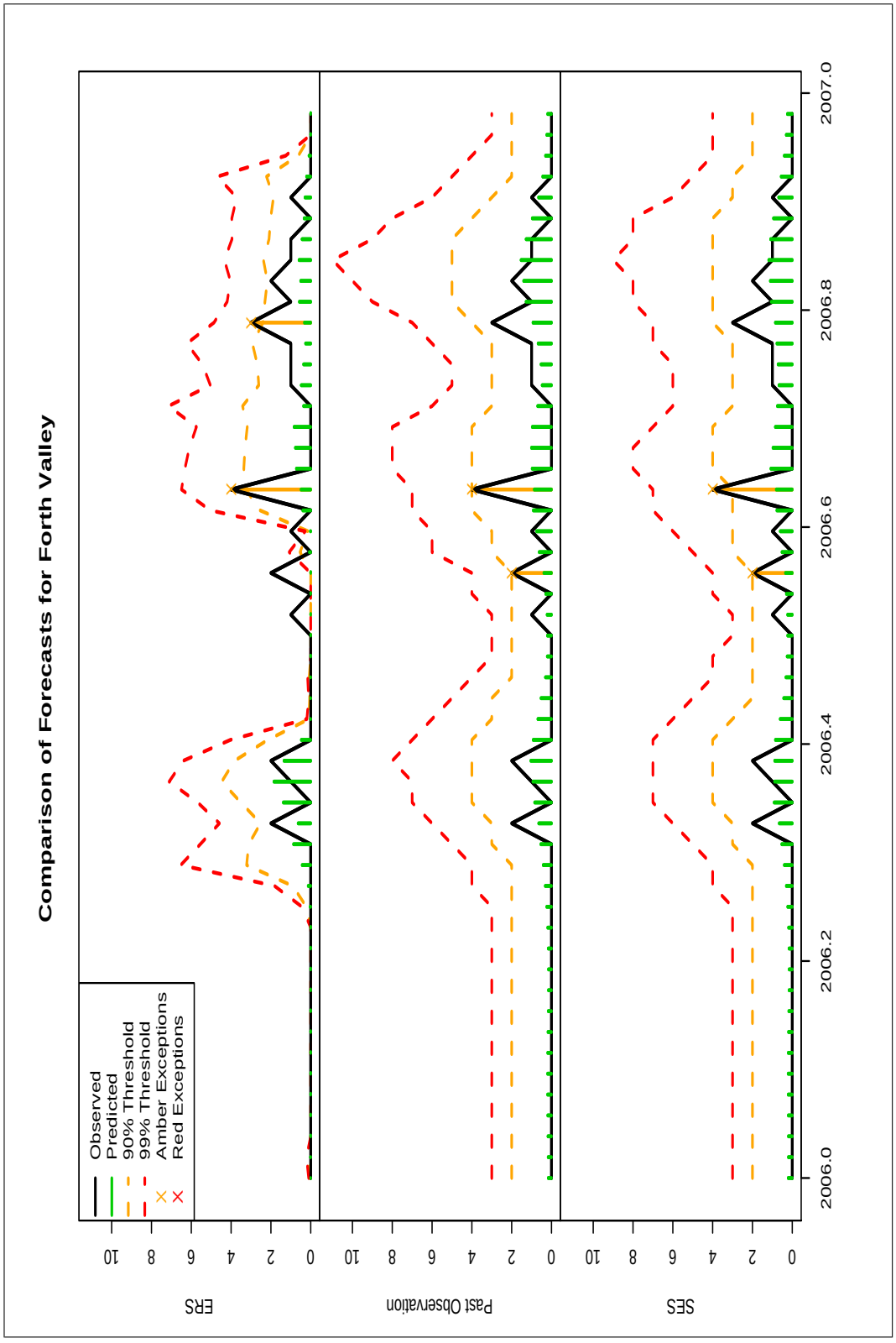


Figure 6.10: The three different forecast systems for 2006, with predictions and exceedance levels for Forth Valley.

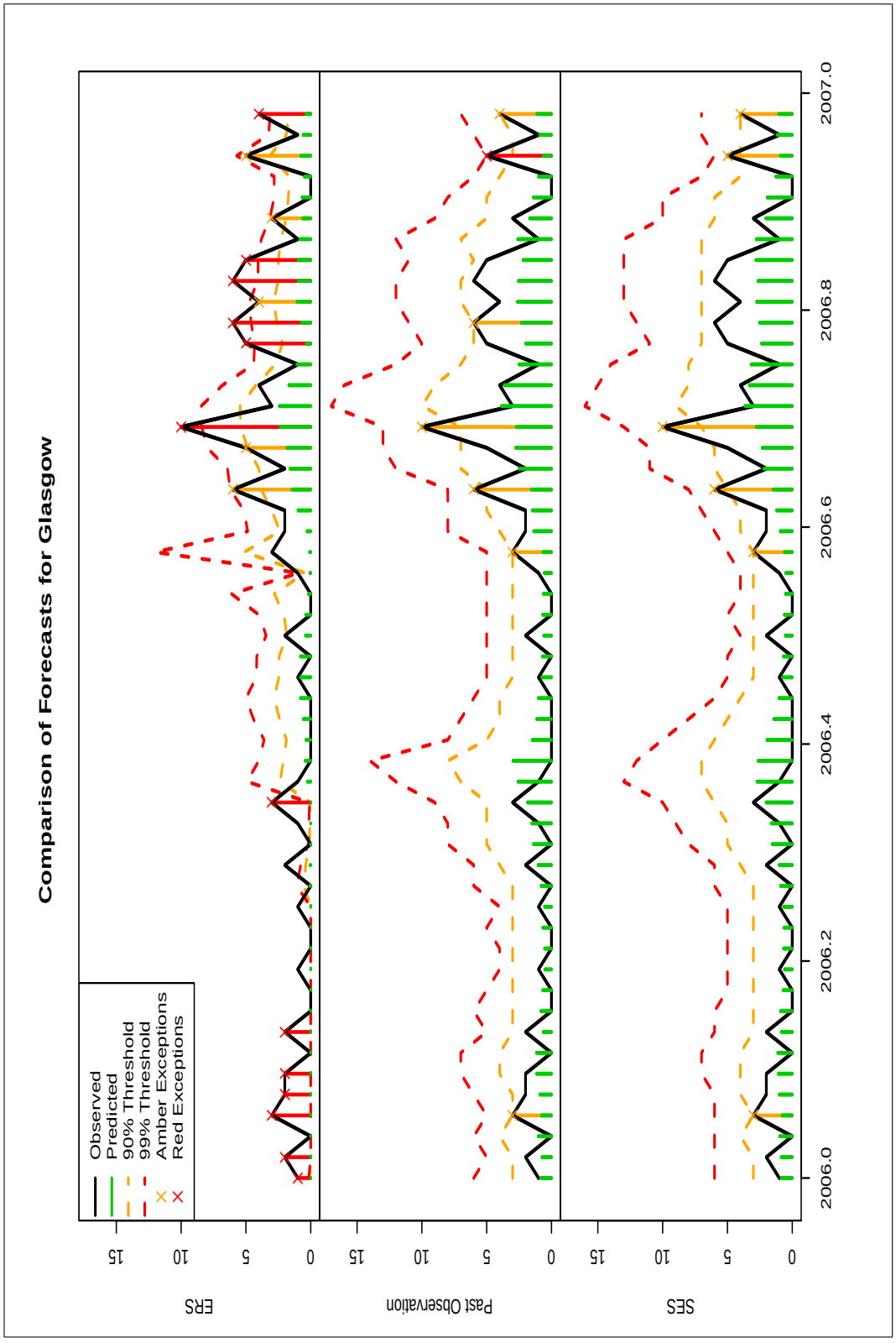


Figure 6.11: The three different forecast systems for 2006, with predictions and exceedance levels for Glasgow.

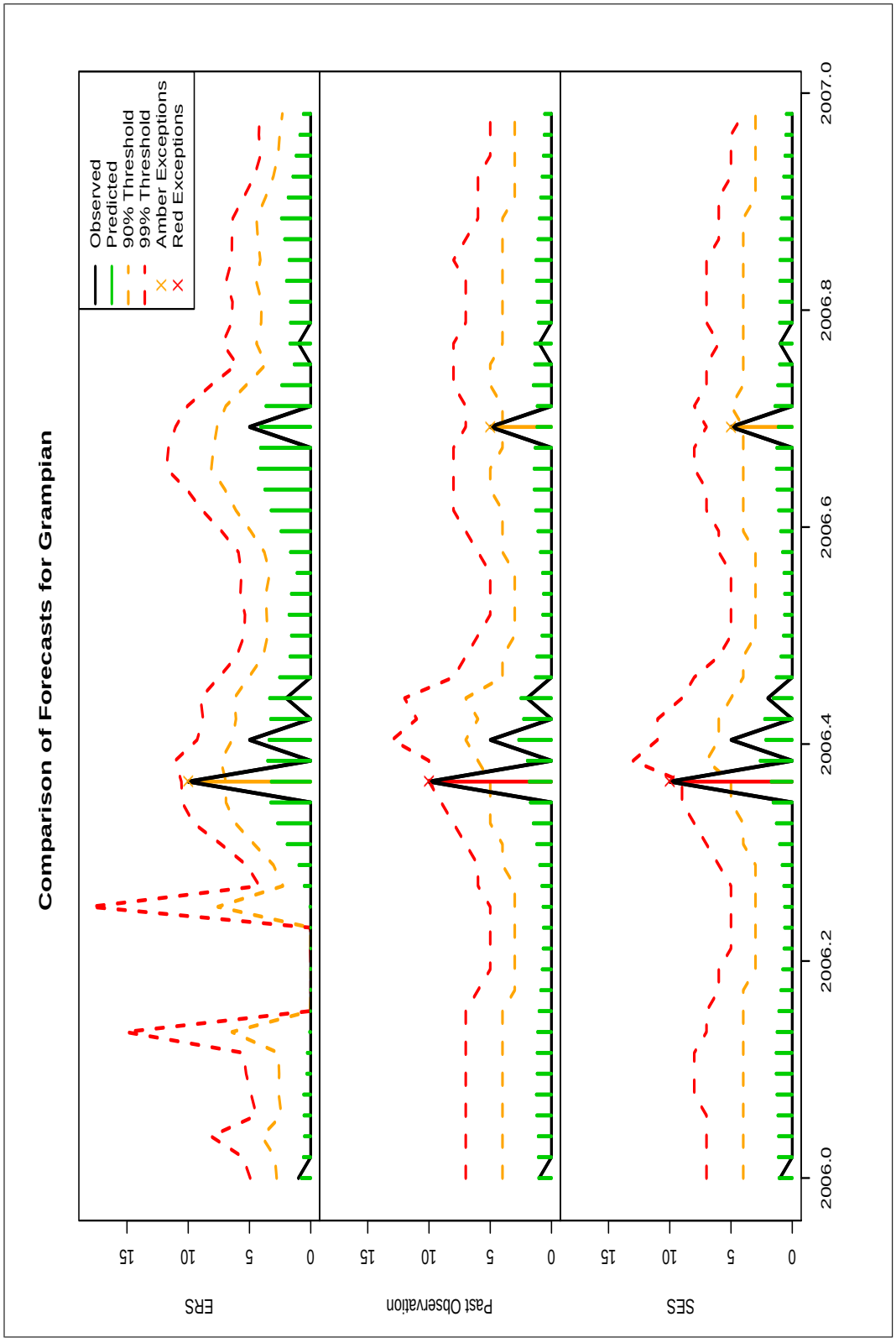


Figure 6.12: The three different forecast systems for 2006, with predictions and exceedance levels for Grampian.

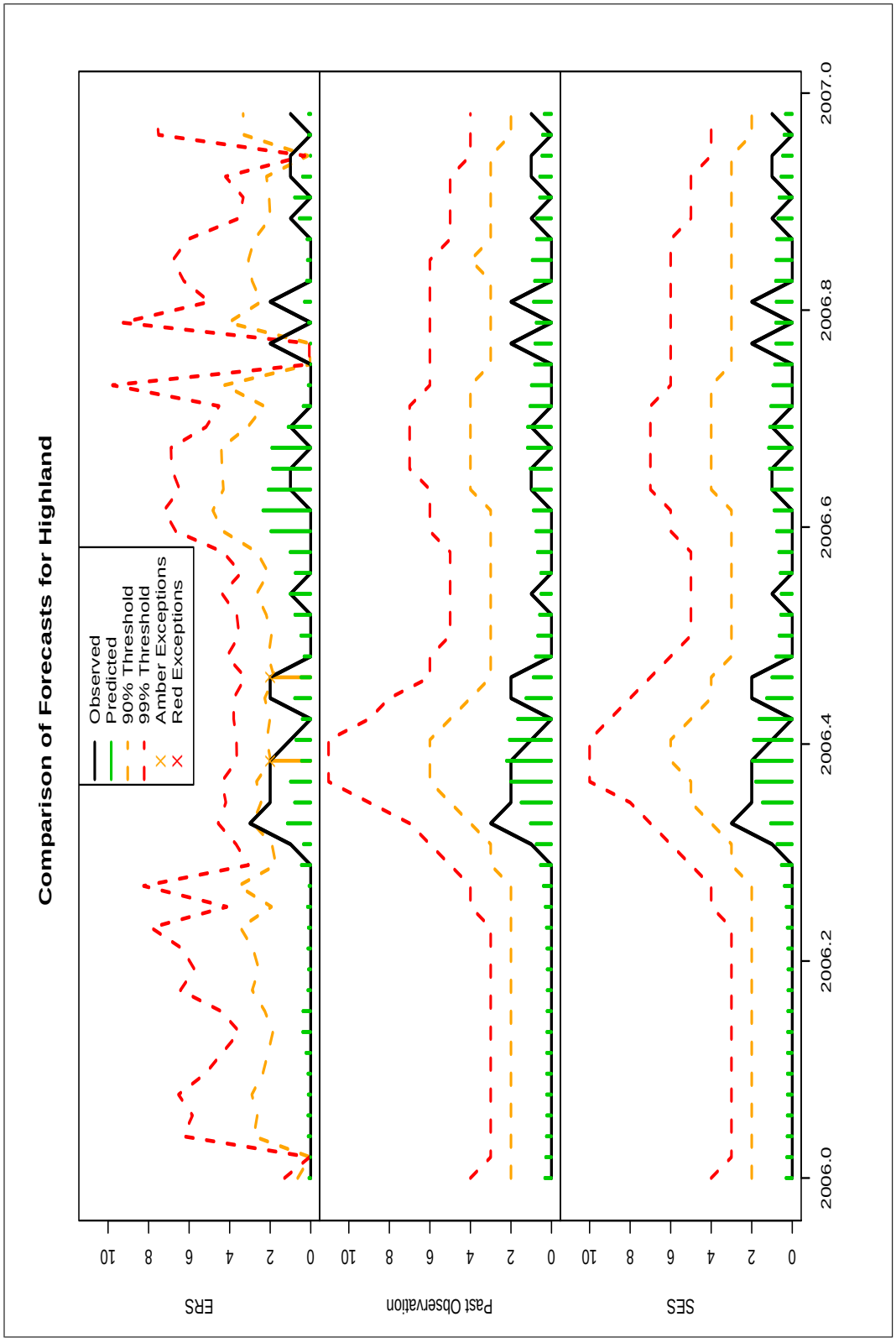


Figure 6.13: The three different forecast systems for 2006, with predictions and exceedance levels for Highland.

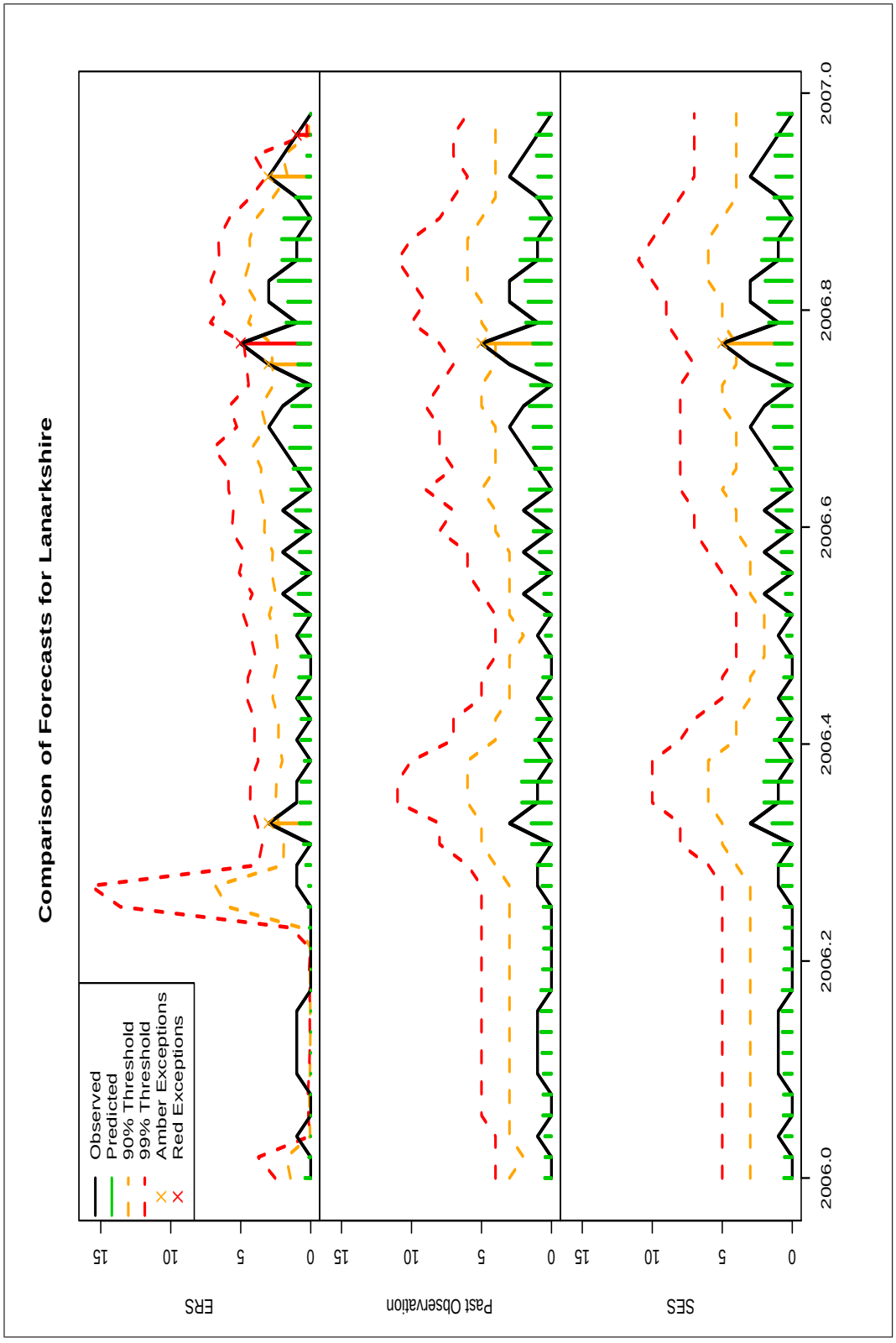


Figure 6.14: The three different forecast systems for 2006, with predictions and exceedance levels for Lanarkshire.

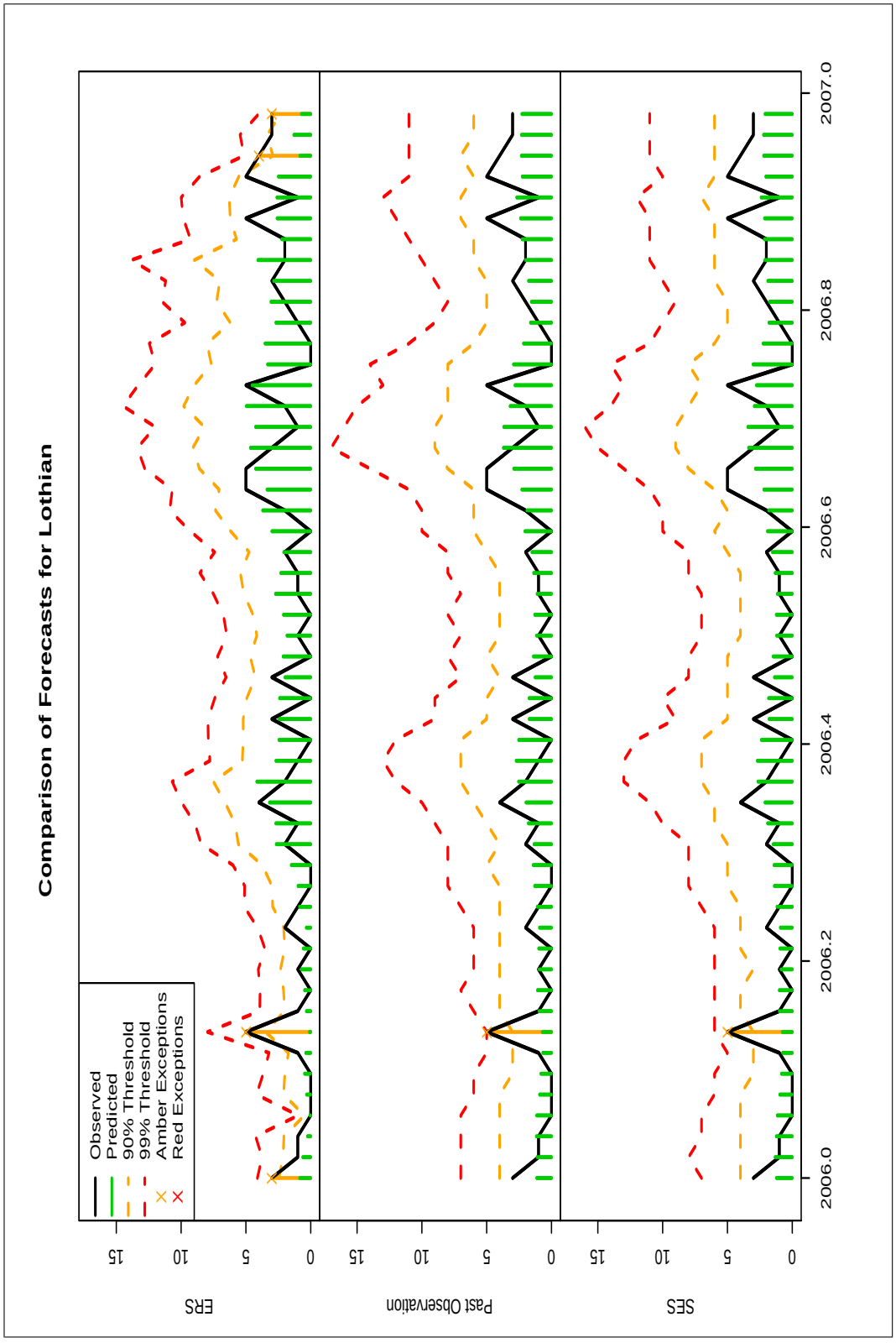


Figure 6.15: The three different forecast systems for 2006, with predictions and exceedance levels for Lothian.

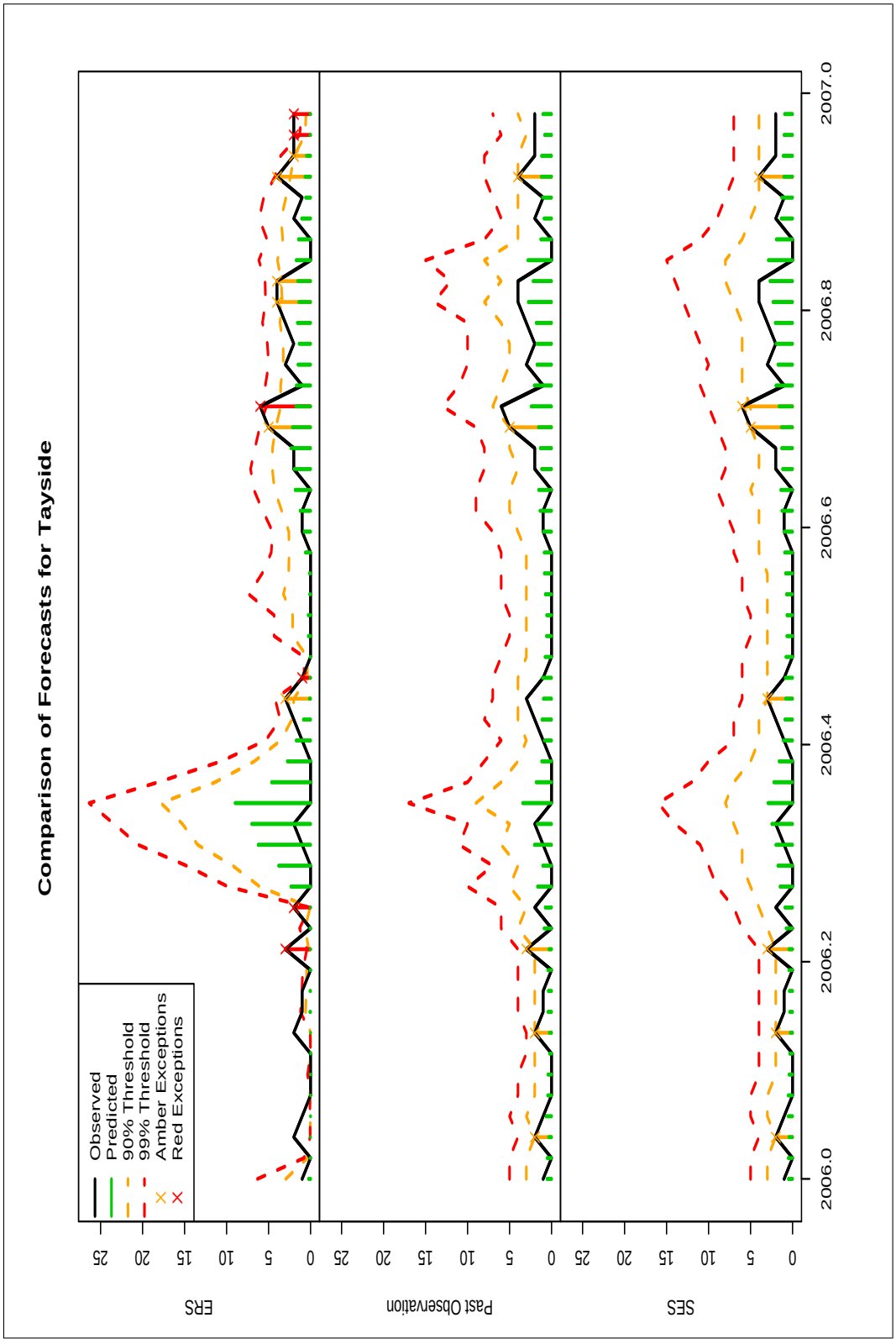


Figure 6.16: The three different forecast systems for 2006, with predictions and exceedance levels for Tayside. The large threshold 'envelope' in the ERS system is caused by an outbreak (fifty cases) that occurred during 2005 around the same time of year.

6.4 Conclusions & Future Work

Before our concluding comments about the comparisons between the uniform models and the ERS system, we reiterate an important caveat noted at the beginning of this Chapter: we are comparing systems that have been tailored specifically to *Cryptosporidium* to the very general ERS system, which was designed to be applicable for the exceedingly wide range of organisms monitored by HPS. Thus, it is not surprising that we find the systems we have developed, seem to be better suited to monitoring *Cryptosporidium*. Also, note that the ERS system has procedures for dealing with past outbreaks, while the other systems just ignore them (see Section 6.1.2). It is better to consider the comparisons as highlighting possible benefits of adopting elements of the models developed in this thesis in regional reporting systems.

For *Cryptosporidium*, the uniform models developed here seem more suited to forecasting reported cases rather than the ERS. In the ERS, seasonality is dealt with through the choice of base-line data; to produce a forecast for week t it will use the weeks corresponding with $t-3$ to $t+3$ in the previous five years. By using weeks around week t the fitted regression model is hoped to have a level that is about right for that time in the year. Unfortunately, at the regional level there are many zeros in the data which can obscure the seasonal pattern. By using trigonometric terms, as in the past observation and SES models, it is easier to capture a more representative seasonal pattern. First, we fit the seasonal pattern from a much longer period of data – eighteen years. Also, with the trigonometric modelling, all weeks contribute to fitting the seasonal pattern, as opposed to a specific subset of weeks from previous years. This mitigates the effects of having many zero-counts obscuring the underlying seasonal pattern. However, as demonstrated in Burkom (2007), having a non-adaptive regression model tends to give among the poorer forecast errors; we resolve this through the inclusion of terms in the past observation and SES models to address serial correlation. This results in forecasts that are generally comparable to the ERS, and often better, reducing the forecast error noticeably, particularly in densely populated areas (see Section 6.3.1).

The exception comparison using an exceedance alerting method in Section 6.3.2 also suggests that the past observation and SES models are more suited to

monitoring *Cryptosporidium*. This comparison may not be entirely fair given the way we have chosen to find the exceedance limits in the past observation and SES models; note the comments about them being more conservative in Section 6.2.3. However, we note that this is a reasonable step given the small counts we have at the regional health board level. We observe, perhaps unsurprisingly, that more exceptions are raised by the ERS than by the other systems. For a number of health boards the ERS also gives many more 'red' exceptions than 'amber' exceptions, which seems counter intuitive. A number of these exceptions are caused by the ERS model performing poorly at the start of the year when counts are very low. In the regions with low average counts, this problem is resolved through the use of the heuristic where exceptions are ignored if there are fewer than five counts in the preceding four weeks before the forecast. However, this heuristic fails when applied to boards with larger average counts, such as Glasgow. This heuristic is not required with the past observation and SES models, as the seasonality is captured sufficiently well by their trigonometric terms to forecast suitably low predictions at the start of the year.

A point particularly worthy of note is the transformation that is required by the ERS to normalise the data for producing its exceedance score – see Section 4.2.2. A two-thirds power transform is applied across all organisms, as McCabe (2004) found it to be the best general transform for achieving normality. The application of one transform uniformly across all organisms is dubious, but the application to small counts is even more so. As counts become smaller, as we expect at the regional level, they will become increasingly skew and so harder to normalise. It is much more appropriate to model then with the negative binomial distribution, as we do with the past observation and SES models. By doing so, we fit a proper distribution to each times series, rather than fitting a quasi-Poisson distribution and then having to apply a transform to achieve normality for probability calculations as in the ERS. Thus, each time series will have its own specific better fitting distribution. Even if the past observation and SES models are not applied in their entirety, we would strongly encourage future work to investigate the adoption of the negative binomial distribution into the ERS. This point is reinforced when we consider that the majority of organisms monitored by the ERS having very small counts (McCabe 2004). As we noted in Chapter 5, further improvements may also be found by using ZIPs – see Section 5.6.1.

These factors lead us to recommend the SES models as the best form for regionally monitoring *Cryptosporidium*. The ERS is a good general purpose monitoring system but has deficiencies which are not best suited to regional counts. The past observation models are similar in their quality of forecasts to the SES models, but the SES models are the most general with the potential for easy automatic adaption to other organisms. Contemporary computing power is such that the extra step of model fitting required in the SES models for producing the exponential smoothing forecasts is of very little concern. The combination of exponential smoothing and a regression model allow for the combination of the advantages of both modelling tools. Serial correlation can be dealt with by the smoother. The regression model allows for a suitable distribution to be fitted easily and other variables to contribute towards prediction. For instance, holiday effects could be easily included in the regression model, which cannot be done easily with a simple exponential smoothing prediction (a noted problem (Burkom 2007)). A more complex model could be developed to address these elements simultaneously, but these separate elements are robust and more easily explainable to non-experts.

A good first direction for future work would be to apply the SES models to the other organisms that HPS asked us to consider modelling regionally (Section 4.2.4). Some of these organisms under go large step changes in reporting level and so we would need to investigate the effect this would have on the models. It is possible that the SES term would adapt sufficiently well, allowing for more years of information to be included in the model to inform the seasonal pattern. Alternatively, models could be fitted to data after the step-change. It is likely that fewer trigonometric harmonics would be required for modelling seasonality in these other organisms; recall that it is only because we are modelling two strains of *Cryptosporidium* that its counts exhibit bi-modal seasonality and so requires four trigonometric harmonics (see Section 4.3). In the case of MR*SA*, we may find that it is not very seasonal due to the organism primarily being a hospital acquired infection. However, including terms to address serial correlation may well be very fruitful with MR*SA*, since there are likely to be lots of outbreaks of related cases causing a good deal of serial correlation. Through modelling these organisms it will be interesting to see if the best form of exponential smoothing is always simple exponential smoothing, or if a more complex form is better.

The extension of the SES models to other organisms would also permit a fairer comparison with the ERS system to be carried out. An even more equitable comparison would result if a common form of the SES model was applied to all the organisms considered. For instance, a good starting form for the general model might be to include the following variables: an SES term (see Section 5.4.2 for a reminder of this term's purpose); two trigonometric harmonics to capture seasonality (most seasonal patterns are likely to be uni-modal, in contrast that of *Cryptosporidium*); and a linear trend term. We would hope, as was found with *Cryptosporidium*, that such a system would better model regional seasonality and serial correlation than the ERS system, hopefully leading to better surveillance qualities. With *Cryptosporidium*, we found evidence of few outbreaks and those that were present seemed to have little effect on the models developed. As we consider other organisms, we are likely to come across more outbreaks. It will be important to investigate how such outbreaks affect the models. We may find that it is better to have some automatic approach for dealing with outbreaks, such as the one adopted by the ERS system (see Section 5.2).

It may be interesting to investigate the choice of alerting method. We have chosen to apply exceedance alerting in this chapter since it is the the method in place at HPS and the one that the ERS is programmed to apply. As Burkom (2007) notes, exceedance alerting is best suited to monitoring data for 'spiked' increases. Of the outbreaks we have come across reported in the HPS weekly bulletins, this seems a reasonable choice. As McCabe (2004) notes, it is also one of the more timely methods, important for combating point sources of infection. However, it may be interesting to investigate the use of CUSUM charts. Since the SES and the past observation models are adaptive, it is quite conceivable that they would miss slow increases in reporting rates. The use of CUSUM charts, possibly in addition to an exceedance method, would reduce this risk as they are better suited for picking up small changes (Burkom 2007). The CUSUM charts might have to be monitored by more statistically expert individuals, but since they would be monitoring for smaller and slower changes, they are unlikely to need to be monitored as closely or as frequently as the exceedance alerts. Beyond this, future work could look at addressing the relationships between health boards. Modelling the different organisms at the health board level would provide opportunities to investigate if other organisms exhibit strong and/or similar

correlations between regions. If they do, forecasts may be improved by modelling these relationships as suggested in Section 5.6.

We may also consider future work which departs further from these considerations. It might be interesting to investigate if modelling a slightly different quantity gives informative results for surveillance. Instead of looking at just the identified cases, of say, *Cryptosporidium*, we might consider this as a proportion of all lab tests (solely for *Cryptosporidium* or in general). This may be more representative of the natural prevalence of the organism, since, if more tests for an organism are being performed, we expect more reported cases. This assumes this data can be obtained by HPS; it may present an unreasonable burden on the microbiology labs to expect them to report all lab tests, including the negative ones. A further, separate, approach may be to investigate the use of modal forecasts for surveillance.

Chapter 7

Monitoring Mortality: Modelling Totals of Recorded Deaths

During 2009 a new strain of influenza emerged which came to be known as swine influenza A (H1N1), or simply 'swine flu' (though naming conventions differ slightly (World Health Organization 2009b)). This new strain is thought to have developed from viruses found among swine, resulting in its given name (Smith et al. 2009). The sub-type of swine flu is not a new one among previous human flu infections. However, the structure of the new virus is sufficiently different from previous strains that there were strong concerns that large proportions of the human population would not be resistant to infection (particularly among the young) (Centers for Disease Control and Prevention 2009). These fears were realised when, after the first reported cases in April 2009, swine flu was declared to be a global pandemic in mid-June 2009 by the World Health Organization (Centers for Disease Control and Prevention 2009; BBC 2009). Generally, swine flu causes mild symptoms in most reasonably healthy adults (World Health Organization 2009a). However, swine flu has concerned the medical community and the general public at large because of its very contagious nature and the limited efficacy of vaccines in preventing infection by this new strain (World Health Organization 2009a; World Health Organization 2009c). With this threat to public health and general public concern, HPS responded by increasing surveillance of flu related conditions and other appropriate information streams within Scotland.

This intensifying response by HPS is in line with health agencies around the world. Systems more usually applied to seasonal flu were monitored more closely.

For example, GP consultation rates for influenza like infections are collected each day. Exceptions relating to the colds & flu syndrome raised by the NHS24 Exceedance Reporting System are monitored closely (see Chapters 2 and 3). The proportion and location of calls to NHS24 are also monitored for other changes and clustering resulting from the progress of swine flu. Information from these and other such surveillance systems are then summarised and emailed daily within HPS. If the surveillance system implies some action should be taken, then it can be taken in a timely fashion. Each week this surveillance information is summarised alongside other medical information to produce a public situation report, used to brief the Scottish public and government about the state of swine flu within Scotland (for an example, see Health Protection Scotland (2009b) and for more detail about the surveillance systems used, see Health Protection Scotland (2009a)).

Human fatalities are among the worst outcomes for any public health threat. To reassure the public and to allow for the potential detection of swine flu becoming more lethal, HPS asked us to develop a statistical system for monitoring the daily number of deaths occurring in Scotland. We present that work in this Chapter and the next. Presently this system monitors *all* causes of death, while future work may investigate monitoring sub-categories in causes of death (for example, all of those deaths attributable to a flu related illness). However, detecting all deaths relating to influenza is not straight forward: deaths caused by influenza are often not recorded as influenza-related (Zucs et al. 2005). For this reason, a number of systems prefer to monitor deaths from all causes, as a metric for the severity of influenza deaths (this point is discussed further in Reichert et al. (2004), with examples of such systems in Zucs et al. (2005), Choi and Thacker (1981a), Serfling (1963)). Various work has been done to further elucidate the relationship between all cause mortality and the number of deaths caused by influenza (for instance, see Simonsen et al. (1997) and Choi and Thacker (1981b)). We consider this point more closely in the next Chapter; there we consider which alarm method to use with the developed models to detect potentially unusually high levels of deaths occurring and consider how to address the delay in the reporting of deaths. This Chapter focuses on the development of the models used to predict the numbers of deaths. We begin by considering a model for all deaths that is based on a Serfling model (Serfling 1963) and then use a series of Generalized Additive Models which more closely capture the Winter seasonality.

7.1 Death Data: Exploratory Analysis

Our data on deaths comes from the General Register Office for Scotland (GROS) (General Register Office for Scotland 2009a). Among a number of different functions, they are responsible for recording events such as births, deaths, marriages, civil partnerships and adoptions that occur in Scotland (General Register Office for Scotland 2009b). Through HPS, we were provided with access to the details of deaths that were registered between 1$^{\text{nd}}$ October 2006 and 14$^{\text{th}}$ May 2009. For each death reported, we have the age (at time of death), sex, date of registration and date of death, giving 145,618 records. It is important to note the difference between the date of death and the date of registration: when a death occurs there is often a delay in the death being reported. The delay can be caused by a number of factors, such as the deaths occurring at the weekend, when most registrars are closed, making it much harder to report deaths at this time. The reporting delay distribution of deaths and their cumulative frequency are shown in Figure 7.1. Most deaths (34%) are reported the day after they occur and nearly all deaths (99%) are reported within fourteen days of their occurrence. We wish to model the number of deaths that occur each day and so will focus on using the date of death for aggregations of suitable totals. However, the number of deaths recorded for the last two weeks within the data-set will be under-represented, as we expect a substantial number of the deaths occurring between April 30$^{\text{th}}$ and May 14$^{\text{th}}$ 2009 to be missing. Thus, in the modelling in this Chapter, we only fit models using records with a date of death between 1$^{\text{st}}$ October 2006 and 29$^{\text{th}}$ April 2009, during which 143,514 deaths are recorded as having occurred.

To have a timely monitoring system, it is necessary to find some way of 'correcting' for the delayed reporting that primarily occurs within the last two weeks of death data. If we do not, then we will only be able to have confidence in comparisons between the levels of reported deaths and the predicted levels for two weeks previously. This is not at all desirable when trying to inform the public about the levels of deaths occurring in a timely fashion. In Chapter 8, we propose one approach for correcting the under-reporting. In this Chapter, we focus on developing the models that will be used for prediction within this the mortality monitoring system.

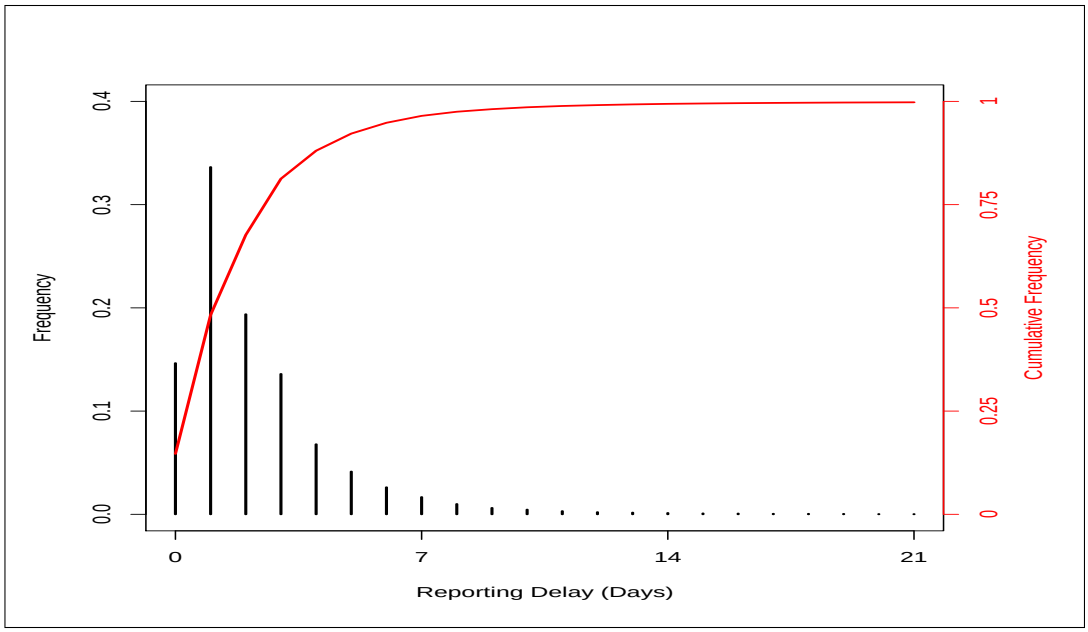


Figure 7.1: The delay distribution (left axis, black elements) associated with the reporting of deaths to GROS and their cumulative frequency (right axis, red elements). Most deaths (over 99%) are reported within fourteen days of their occurrence.

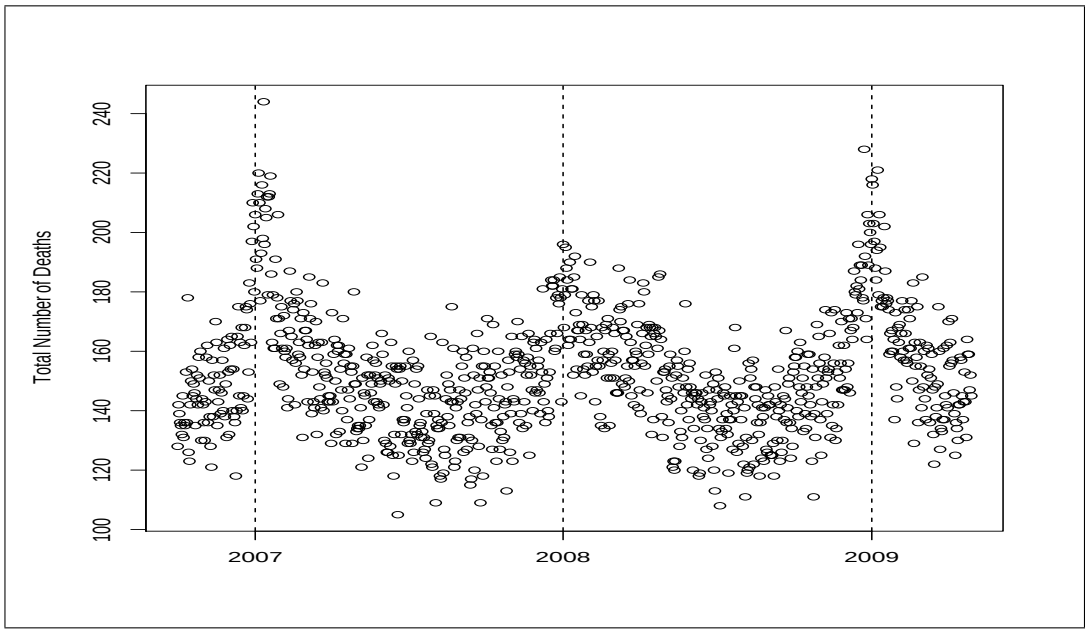


Figure 7.2: The total deaths occurring in all of Scotland between 1^{st} October 2006 and 29^{th} April 2009.

	Mon	Tues	Wednes	Thurs	Friday	Sat	Sun
Mean Deaths	152.82	151.73	152.10	153.72	153.19	153.16	149.74

Table 7.1: The mean number of deaths for each day of the week, recorded as taking place between 1st October 2006 and 29th April 2009.

The total number of deaths for each day of the period under consideration are shown in Figure 7.2. During this time, there is a mean of 152 deaths per day, with a minimum of 105 deaths and a maximum of 244 deaths. The daily total of deaths is highest during Winter, particularly at the end of year/new year period. This pattern has been noted by others as commonly occurring in developed countries with temperate climates (Reichert et al. 2004). A number of theories have been suggested for this peak in mortality during the Winter season, with among them: mortality having an inverse relationship with temperature (Donaldson and Keating 2002); hours of sunlight affecting immune strength in humans (Dowell 2001); and that sudden changes in temperatures cause increases in mortality (Bull and Morton 1978). Reichert et al. (2004) explore at length similarities in mortality levels from different conditions in the USA over the Winter period for forty years. They conclude that the peak during the Winter is most likely driven by a single source across all causes of death, and that this source is most likely influenza related. Deaths are generally lowest during the Summer months, particularly during July and August.

The mean of deaths for each day of the week are tabulated in Table 7.1. They are all of a similar level, indicating little within week seasonality. However, the mean number of deaths on a Sunday seems somewhat lower than the others. Irrespective of the reporting delay considered above, there may be a further delay linked with Sundays, as many public services will be closed on a Sunday. It can be speculated that as there may be difficulties in reporting deaths on a Sunday some are instead reported as occurring on the Monday.

In Figure 7.3, we show the proportion of total deaths within each age group for each gender. On average, men die at a younger age compared to women: the median age at death for men is 75, while it is 81 for women. The spread of age at death is greatest for men with an inter-quartile range of 19 years, while it is lower at 16 years for women. The proportion of men dying at each age group is

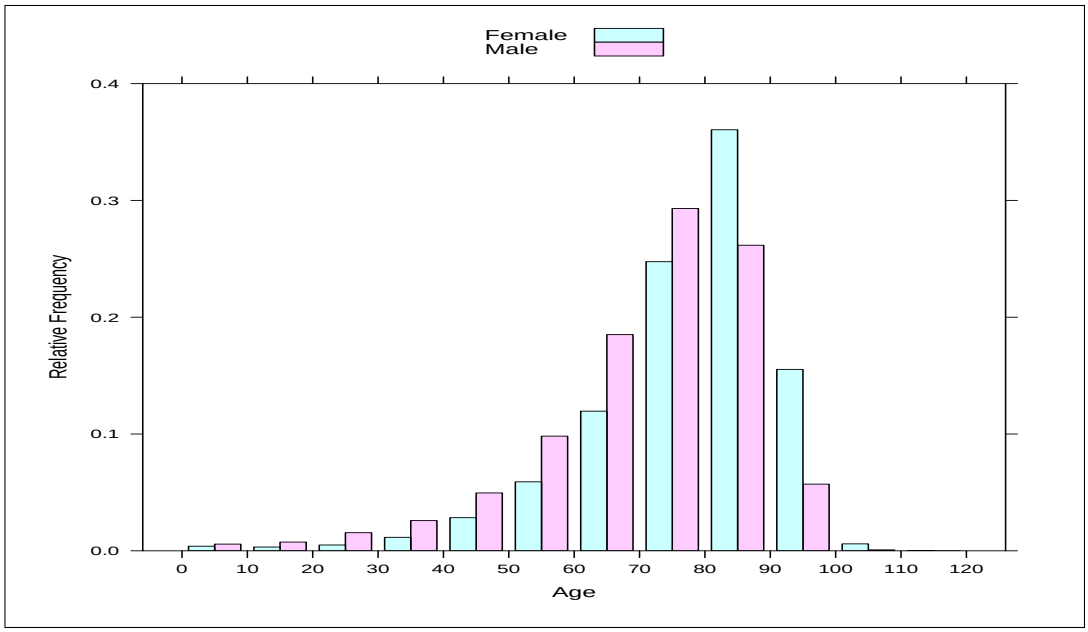


Figure 7.3: For each gender, the proportion of total deaths occurring within that age range, from individuals recorded as dying between 1st October 2006 and 29th April 2009 in Scotland.

greater than the corresponding proportion for women under the age of 80. As we might hope and expect, the proportion of deaths under thirty is very small.

HPS generally divides up age into the following ranges: 0-14, 15-44, 45-64, 65-74, 75-84 and 85+. These ranges broadly group together people expected to have similar lifestyles (0-14 children, 15-44 mobile working adults, etc.). We have seen that there is some evidence for the genders having different levels of deaths within the same age group. Thus, when fitting models to the daily numbers of deaths, we will include an interaction term between age group and gender. We will aggregate the number of deaths by age group and gender, giving twelve groups (6 age groups $\times$ 2 genders). Four records do not specify a gender, and so the number of records we now consider reduces slightly from 143,514 to 143,510. We find that the seasonality also differs between age group and sex. For instance, consider Figure 7.4, where we show the total daily deaths occurring in the oldest (85+) and youngest (0-14) age groups, split by gender. The daily counts of deaths in the 85+ group are clearly seasonal but the seasonality for females has a largest amplitude. In the youngest group, no seasonality is particularly evident, with

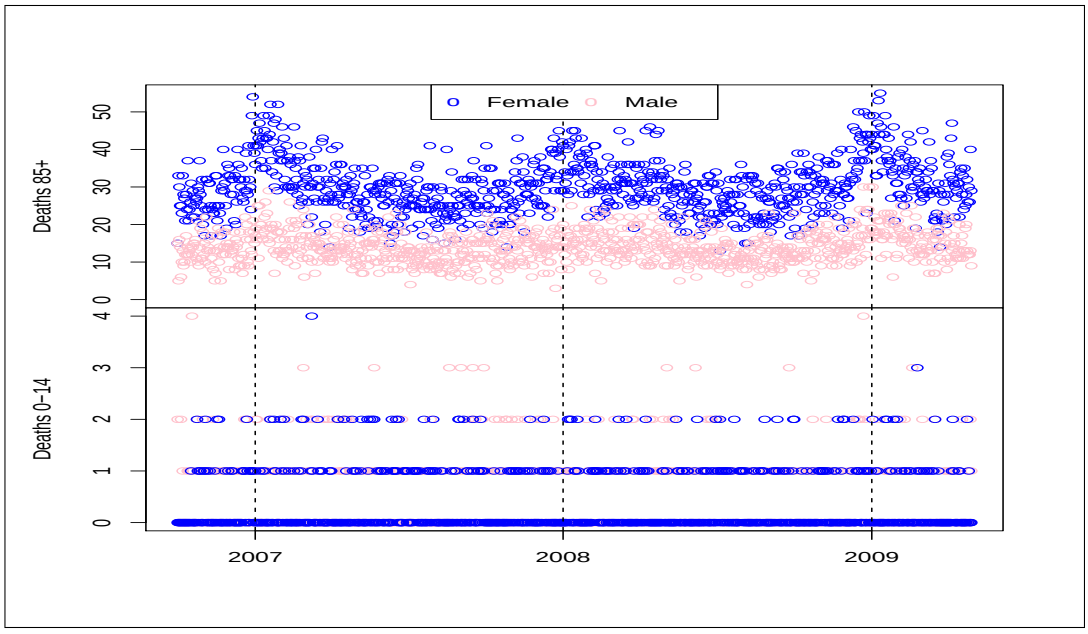


Figure 7.4: The daily totals of deaths in the oldest (85+) and the youngest (0-14) age groups, split by gender.

both genders having similar levels. More generally, the seasonality present in Figure 7.2, is much less evident in the younger age groups, but more so in the older age groups (particularly in 75-84).

We proceed by modelling the daily totals of deaths in the next Section. We are modelling counts, and so we start with a Poisson GLM. We have noted different levels of deaths between age groups and annual seasonality, so we will need to incorporate these elements within the modelling.

7.2 Modelling Death Counts with GLMS

7.2.1 Serfling Models

One of the first models that was developed for predicting of levels of deaths from all causes was presented in Serfling (1963). Serfling predicted the number of deaths $\hat{Y}_t$ for week t using the following formula:

$$\hat{Y}_t = \beta_0 + \beta_1 t + \sum_{i=1}^{n} \beta_{1+i} \sin\left(\frac{2\pi t}{13}\right) + \sum_{i=1}^{n} \beta_{1+n+i} \cos\left(\frac{2\pi t}{13}\right),$$

where the parameters β_k were estimated by least squares regression. The $\beta_1 t$ term serves to model linear trend, while the trigonometric terms are used to capture seasonality. Serfling found that only one harmonic contributed significantly to the model. The year was divided up into thirteen 4 week periods. The above model would predict the number of weekly deaths for the middle of each period. A cubic spline was then fitted between these points to allow for interpolation of values for all other weeks in the year. We adapt this basic model to deal with daily counts of deaths and include factors for age, sex, and their interaction with seasonality (similar to the model fit in Thompson, Shay, Weintraub, Brammer, Cox, Anderson, and Fukuda (2003)).

7.2.2 Initial Quasi-Poisson Model

We fit a quasi-Poisson model to the numbers of daily deaths. The daily mean, μ_{tsa}, for each age group a and gender s, is given by:

$$\begin{aligned} \log(\mu_{tas}) \;=\; & \beta_0 + \beta_1 trend_t + day_t + Sex_s + Age.Gp_a \\ & + \beta_2 \sin(2\pi p_t) + \beta_3 \cos(2\pi p_t) + Sex_s : Age.Gp_a \\ & + Sex_s : \{\sin(2\pi p_t) + \cos(2\pi p_t)\} \\ & + Age.Gp_a : \{\sin(2\pi p_t) + \cos(2\pi p_t)\}, \end{aligned} \tag{7.1}$$

where the β_k are coefficients; $trend_t$ is the number of day t as numbered from the first day ($trend_t = t$); day is a seven level factor with a different level for each day of the week; Sex_s is a factor with two levels; $Age.Gp_a$ has a level for each of the age groups specified by HPS (0-14, 15-44, 45-64, 65-74, 75-84, 85+);

p_t gives the within year time, which begins at zero on January 1st and increases by 1/365 or 1/366 increments, depending on the numbers of days in each year; $A : B$ denotes the interaction of terms A and B. The coefficients of this model are given in Table 7.2, with summary statistics for the model given in Table 7.3.

Few of the fitted coefficients are significant. The linear trend term, $trend_t$ has a very small, non-significant coefficient. If there is a trend within the daily levels of deaths, it is likely too small for us to detect with this model. The *day* term shows that there is little difference between levels of deaths on different days of the week. However, the level for Sunday is an order of magnitude smaller than the other days of the week and is significant. The Sex, $Age.Gp$ terms and their interactions are as we would expect from Section 7.1: levels of deaths increase as age increases, with more males dying in the younger ages and fewer dying as age increases. The trigonometric terms are not significant, nor are many of their interactions. The interaction between Sex and the trigonometric terms are significant. As age increases, the interaction values between $Age.Gp$ and the trigonometric terms increase too, which is as we expect: from Section 7.1 we know there is less seasonality in the younger age groups, and increasingly more as age increases.

The dispersion parameter for this quasi-Poisson model is 1.069, suggesting that there is negligible over-dispersion within the counts. Given this, we move on to fit a Poisson GLM in the next Section, adapting the terms included in the model, since so few of them are significant.

7.2.3 Poisson Models

We adapt the model defined in Equation 7.1 to the following, fitting a Poisson GLM:

$$\begin{aligned} \log(\mu_{tas}) &= \beta_0 + \beta_1 trend_t + Sunday_t + Sex_s + Age.Gp_a \\ &\quad + \beta_2 \sin(2\pi p_t) + \beta_3 \cos(2\pi p_t) + Sex_s : Age.Gp_a \\ &\quad + Sex_s : \{\sin(2\pi p_t) + \cos(2\pi p_t)\} \\ &\quad + Young_a : \{\sin(2\pi p_t) + \cos(2\pi p_t)\}, \end{aligned} \tag{7.2}$$

Variable	Coefficient	SE
β_0 (Constant)	-0.961244	0.055478
trend	0.000002	0.000010
dayTues	-0.007139	0.010199
dayWed	-0.004784	0.010193
dayThurs	0.005631	0.010185
dayFri	0.002089	0.010194
daySat	0.001858	0.010194
daySun	-0.020330	0.010233
SexM	0.278345	0.072206
Age.Gp15-44	1.737599	0.059485
Age.Gp45-64	3.190646	0.055988
Age.Gp65-74	3.481924	0.055710
Age.Gp75-84	4.162543	0.055301
Age.Gp85+	4.335582	0.055238
sin1f	0.038912	0.050213
cos1f	0.023126	0.052260
SexM:Age.Gp15-44	0.486544	0.077307
SexM:Age.Gp45-64	0.139978	0.073579
SexM:Age.Gp65-74	-0.004376	0.073291
SexM:Age.Gp75-84	-0.361611	0.072853
SexM:Age.Gp85+	-1.032579	0.073018
SexM:sin1f	-0.014547	0.007879
SexM:cos1f	-0.027195	0.008262
Age.Gp15-44:sin1f	-0.009931	0.053161
Age.Gp15-44:cos1f	0.011555	0.055330
Age.Gp45-64:sin1f	-0.000901	0.050948
Age.Gp45-64:cos1f	0.041944	0.053028
Age.Gp65-74:sin1f	-0.003664	0.050772
Age.Gp65-74:cos1f	0.075134	0.052848
Age.Gp75-84:sin1f	0.007446	0.050481
Age.Gp75-84:cos1f	0.111135	0.052544
Age.Gp85+:sin1f	0.034480	0.050558
Age.Gp85+:cos1f	0.141464	0.052629

Table 7.2: The coefficients of the quasi-Poisson model defined by Equation 7.1, which is fit to the counts of deaths recorded as occurring in Scotland between 1st October 2006 and 29th April 2009.

Model	Distribution	Null Deviance	Residual Deviance	Explained Deviance	% Deviance Explained	Null DoF	Residual DoF	Used DoF	AIC	Dispersion Parameter
Eqn. 7.1	Q	105993	12268	93725	88	11303	11271	32	–	1.069
Eqn. 7.2	P	105993	12273	93720	88	11303	11281	22	54047	–
Eqn. 7.3	P	105993	12146	93848	89	11303	11277	26	53927	–
Eqn. 7.4	P	105993	12146	93848	89	11303	11277	26	53927	–

Table 7.3: Summary of quasi-Poisson (Q) and Poisson (P) models fit in Section 7.2 to the daily total of deaths occurring in Scotland between 1st October 2006 and 29th April 2009. The models: 7.1, quasi-Poisson, one harmonic, full range of interactions with harmonic, Sex_s, $Age.Gp_a$; 7.2, Poisson, first use of $Young_a$, fits common seasonality to youngest three age groups (one harmonic); 7.3, Poisson, two harmonics; 7.4, Poisson model, reformulates 7.3, combining Sex_s and $Age.Gp_a$.

where: $Sunday$ has two levels, Sunday and all the other days; $Young$ is similar to $Age.Gp$, except that the first three age groups (0-14, 15-44 and 45-64) share the same level; and other terms are as defined previously. The coefficients resulting from this model are shown in Table 7.4, with summary statistics given in Table 7.3. In this model, all of the main terms and most of the interactions are significant. The use of $Young$ serves to improve the fit of the model by only allowing the seasonality to vary with age in the older age groups. Thus, the model is more parsimonious this way. Now the trigonometric terms and their interactions with age (via $Young$) are significant. The $Sunday$ coefficient is significant and negative, suggesting that slightly fewer deaths take place on a Sunday. The other observations from the quasi-Poisson model continue to apply here too.

The residuals of the model are reasonably well behaved. A plot of the residuals, shown in Figure 7.5, show that large positive residuals occur somewhat more frequently than we would expect in a Poisson model. A number of these large residuals will be linked with the inability of the model to adapt to the large seasonal Winter peak: as only one harmonic of trigonometric terms is used, there is

Variable	Coefficient	SE
β_0 (Constant)	-0.9650	0.0527
SundaySunday	-0.0199	0.0076
SexM	0.2788	0.0698
Age.Gp15-44	1.7387	0.0571
Age.Gp45-64	3.1964	0.0537
Age.Gp65-74	3.4864	0.0535
Age.Gp75-84	4.1670	0.0531
Age.Gp85+	4.3401	0.0530
sin1f	0.0362	0.0094
cos1f	0.0574	0.0098
SexM:Age.Gp15-44	0.4865	0.0748
SexM:Age.Gp45-64	0.1395	0.0712
SexM:Age.Gp65-74	-0.0048	0.0709
SexM:Age.Gp75-84	-0.3620	0.0705
SexM:Age.Gp85+	-1.0330	0.0706
SexM:sin1f	-0.0147	0.0076
SexM:cos1f	-0.0275	0.0080
Young65-74:sin1f	-0.0008	0.0118
Young65-74:cos1f	0.0410	0.0123
Young75-84:sin1f	0.0103	0.0105
Young75-84:cos1f	0.0770	0.0110
Young85+:sin1f	0.0373	0.0109
Young85+:cos1f	0.1073	0.0114

Table 7.4: The coefficients of the Poisson model (one trigonometric harmonic) defined by Equation 7.2, which is fit to the counts of deaths recorded as occurring in Scotland between 1st October 2006 and 29th April 2009.

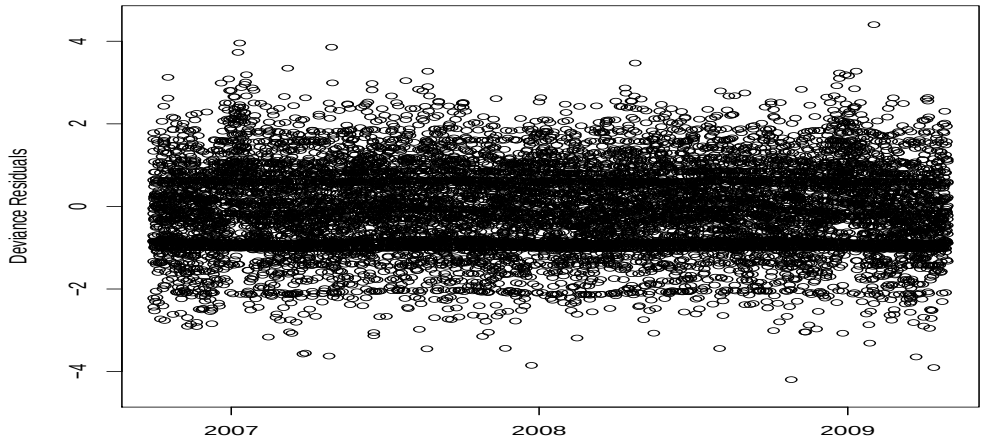


Table 7.5: The deviance residuals, grouped by date, of the model defined by Equation 7.2, which is fit to the counts of deaths recorded as occurring in Scotland between 1st October 2006 and 29th April 2009.

no way that the seasonality will able to peak as sharply as the Winter peak (see Figure 7.2). A plot of the residuals versus fitted values suggests no particular problems. The qq-plot of the residuals has minor curvature in the negative residuals and stronger curvature in the positive residuals (again, likely linked with the Winter peak). It is not possible to use a χ^2 goodness of fit on this model and data since there are very low values modelled in the youngest age groups. These violate the large sample assumptions of the χ^2 test, thus invalidating the test's applicability here (McCullagh and Nelder 1983). However, since the residuals of the model are reasonably well behaved, we consider the model a reasonable fit.

This model can be further improved by adding another trigonometric harmonic, which extends Equation 7.2 to:

$$\begin{aligned}\log(\mu_{tas}) &= \beta_0 + \beta_1 trend_t + Sunday_t + Sex_s + Age.Gp_a \\ &\quad +\beta_2 \sin(2\pi p_t) + \beta_3 \cos(2\pi p_t) + \beta_4 \sin(4\pi p_t) + \beta_5 \cos(4\pi p_t) \\ &\quad +Sex_s : Age.Gp_a + Sex_s : \{\sin(2\pi p_t) + \cos(2\pi p_t)\} \\ &\quad +Young_a : \{\sin(2\pi p_t) + \cos(2\pi p_t)\}\,.\end{aligned} \tag{7.3}$$

The coefficients for this model are given in Table 7.6. No interactions were included with the second harmonics since they did not make a significant contribution to the model. Both an ANOVA and AIC (Table 7.3) comparison between the model with one harmonic and this model with two, suggest that the larger model with an extra harmonic is preferable. The extra harmonic gives the larger model greater ability to fit more closely to the data; however, the improvement is slight (Table 7.3).

To allow for more straight forward comparison of GAM we will go on to develop, we reformulate the model defined by Equation 7.3 to the following:

$$\begin{aligned} \log(\mu_{tas}) &= \beta_0 + \beta_1 trend_t + Sunday_t + AgeSex_{as} \\ &\quad + \beta_2 \sin(2\pi p_t) + \beta_3 \cos(2\pi p_t) + \beta_4 \sin(4\pi p_t) \\ &\quad + \beta_5 \cos(4\pi p_t) + Sex_s : \{\sin(2\pi p_t) + \cos(2\pi p_t)\} \\ &\quad + Young_a : \{\sin(2\pi p_t) + \cos(2\pi p_t)\}, \end{aligned} \tag{7.4}$$

where $Age.Gp$ (five degrees of freedom), Sex (one degree of freedom) and their interactions (five degrees of freedom) have been combined into the twelve level factor $AgeSex$, which has a level for each combination of age group and gender (eleven degrees of freedom). Other terms are the same as defined previously. This model is essentially the same as that defined by Equation 7.3, as can be confirmed from Table 7.3, with coefficients given in Table 7.7. This formulation allows us to directly see the values fit to each combination of age group and sex and their corresponding standard errors; we show these values exponentiated in Figure 7.5. The Figure shows the same patterns as we have noted previously: more males die at younger ages, leading to the biggest difference in levels of deaths between females and males in the 85+ age group.

While the inclusion of the extra harmonic gives the seasonal pattern greater ability to adapt to the data, it still does not adapt well to the peak at the end of the year – see plots of the fitted values in Figures 7.9 to 7.16. Trigonometric variables are not best suited to modelling a seasonal pattern with such a sharp peak. Thus, we turn to using Generalized Additive Models, since they allow some greater flexibility in fitting a seasonal pattern.

Variable	Coefficient	SE
β_0 (Constant)	-0.9633	0.0527
SundaySunday	-0.0198	0.0076
SexM	0.2778	0.0698
Age.Gp15-44	1.7387	0.0571
Age.Gp45-64	3.1964	0.0537
Age.Gp65-74	3.4865	0.0535
Age.Gp75-84	4.1673	0.0531
Age.Gp85+	4.3404	0.0530
sin1f	0.0368	0.0096
cos1f	0.0497	0.0098
sin2f	0.0137	0.0053
cos2f	0.0513	0.0052
SexM:Age.Gp15-44	0.4865	0.0748
SexM:Age.Gp45-64	0.1395	0.0712
SexM:Age.Gp65-74	-0.0048	0.0709
SexM:Age.Gp75-84	-0.3620	0.0705
SexM:Age.Gp85+	-1.0330	0.0706
SexM:sin1f	-0.0134	0.0078
SexM:cos1f	-0.0224	0.0080
SexM:sin2f	-0.0158	0.0077
SexM:cos2f	-0.0256	0.0075
Young65-74:sin1f	-0.0009	0.0119
Young65-74:cos1f	0.0403	0.0122
Young75-84:sin1f	0.0102	0.0106
Young75-84:cos1f	0.0755	0.0109
Young85+:sin1f	0.0377	0.0110
Young85+:cos1f	0.1050	0.0113

Table 7.6: The coefficients of the Poisson model defined by Equation 7.3, which is fitted to the daily counts of deaths occurring in Scotland between 1st October 2006 and 29th April 2009. Two harmonics are used to model annual seasonality.

Variable	Coefficient	SE
β_0 (Constant)	-0.9633	0.0527
SundaySunday	-0.0198	0.0076
ageSex0-14M	0.2778	0.0698
ageSex15-44F	1.7387	0.0571
ageSex15-44M	2.5029	0.0548
ageSex45-64F	3.1964	0.0537
ageSex45-64M	3.6136	0.0533
ageSex65-74F	3.4865	0.0535
ageSex65-74M	3.7595	0.0533
ageSex75-84F	4.1673	0.0531
ageSex75-84M	4.0830	0.0531
ageSex85+F	4.3404	0.0530
ageSex85+M	3.5853	0.0534
sin1f	0.0368	0.0096
cos1f	0.0497	0.0098
sin2f	0.0137	0.0053
cos2f	0.0513	0.0052
SexM:sin1f	-0.0134	0.0078
SexM:cos1f	-0.0224	0.0080
SexM:sin2f	-0.0158	0.0077
SexM:cos2f	-0.0256	0.0075
Young65-74:sin1f	-0.0009	0.0119
Young65-74:cos1f	0.0403	0.0122
Young75-84:sin1f	0.0102	0.0106
Young75-84:cos1f	0.0755	0.0109
Young85+:sin1f	0.0377	0.0110
Young85+:cos1f	0.1050	0.0113

Table 7.7: The coefficients of the Poisson model (two trigonometric harmonics) defined by Equation 7.4, fit to the daily totals of deaths in Scotland recorded as occurring in Scotland between 1^{st} October 2006 and 29^{th} April 2009.

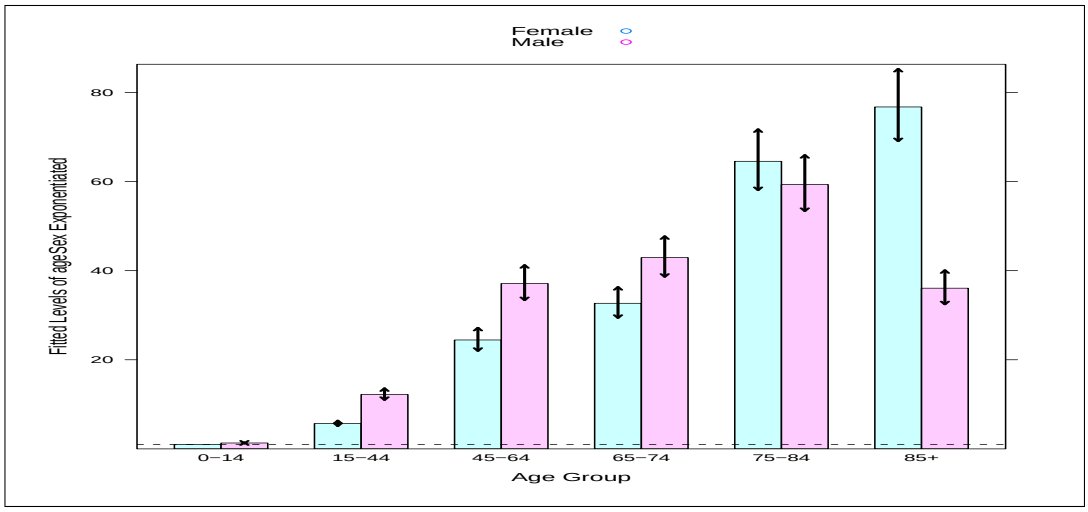


Figure 7.5: The values fitted to each combination of age group and sex on the response scale, calculated in the fitting of the Poisson model defined by Equation 7.4. Treatment contrasts were used, with the 0-14 females taken as the reference group; the dotted line gives the level of this group. The arrows give twice the standard error above and below the fitted value on the linear predictor scale, which is then exponentiated.

7.3 Modelling Death Counts with GAMs

Choosing Models Fitted Closely to Winter Peaks

The choice of using Generalized Additive Models (GAMs), to more closely model the sharp peaks in mortality during the Winter, is motivated by the context in which the mortality monitoring system was developed: we desired to monitor swine flu in case of it having greater or differently timed mortality compared to 'normal' seasonal flu. In this context, and in contrast with other mortality monitoring systems, the sharp Winter peaks are not unusual but instead represent a standard part of the seasonal pattern of mortality. Thus, it makes sense to choose a modelling paradigm, such as GAMs, that can fit more closely to the Winter peaks. There is also an argument for the GAM fit being useful beyond the swine flu epidemic for monitoring all-cause mortality for general increases, rather than increases primarily attributable to influenza. Outside of the swine flu epidemic, HPS will be interested in monitoring all-cause mortality for both influenza related increases *and* other sources of increases. Thus, the final system

developed includes both a GLM (as developed in Section 7.2) and a GAM for predictions of expected mortality.

7.3.1 GAM background

Generalized Additive Models (GAMs) are one of the many extensions of GLMs. We introduce them as in Wood (2006b). In a GLM, we have the following modelling structure:

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta},$$

where $\mu_i \equiv \mathbb{E}(Y_i)$, $g(\cdot)$ is a smooth monotonic link function, $\mathbf{X}_i$ is the i^{th} row of a model matrix $\mathbf{X}$, and $\boldsymbol{\beta}$ is a vector of unknown parameters. Typically, a GLM assumes that the Y_i have some exponential family distribution. A GAM extends a GLM by including smooth functions of the covariates:

$$g(\mu_i) = \mathbf{X}_i^*\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots, \tag{7.5}$$

where $\boldsymbol{X}_i^*$ is a row of the model matrix for any strictly parametric model components, $\boldsymbol{\theta}$ is the corresponding parameter vector, and the f_j are smooth functions of the covariates, x_k and the other terms are as above. Using functions of the covariates allows for greater flexibility in the fit of models. A number of different choices for these 'smooth functions' are possible, with a number of them being included in the `mgcv` package developed by Wood (2006b). We will use this package within `R` to fit GAMs utilising cubic splines.

When a cubic spline is fit to some data, the range of the data is typically divided up into several sections. The points that form the boundaries of these sections are known as knot points. Between each pair of knot points, a cubic polynomial is fit to the data falling within that region. To ensure that the fit over the whole range of the data is smooth, curves ending at the same knot point are required to be continuous with each other up to the second derivative. If this condition were not required, we would very likely have smooth curves in each section of the data, with discontinuities at the knot points where curves in neighbouring sections of the data meet. The weekly predictions in Serfling (1963) are made at four weekly intervals (so, thirteen predictions for the year). These predictions are joined together using a cubic spline, to allow for estimates

of the number of deaths occurring in weeks between the predictions. For further discussion on splines see Zuur et al. (2009) and Wood (2006b).

Having chosen a smooth function, in our case the cubic spline, the function can then be represented by a set of 'basis' functions. These functions define the space of functions of which the smooth function is an element (Wood 2006b). Then, the function f can be represented by:

$$f(x) = \sum_{j=1}^{q} b_j(x)\beta_j,$$

where $b_j(x)$ is the j^{th} basis function and the β_j are unknown parameters. For example, a set of basis functions for a second order polynomial could be: 1, x and x^2 This form can then be used in Equation 7.5 to turn a GAM into a linear model which can be solved to find the β_j and other parameters by using standard techniques.

It now remains to choose how much smooth the function f of the covariates should be. In terms of the cubic regression spline, we can make the spline more closely follow the data by increasing the number of knots. However, this means the spline will not be as smooth. Decreasing the number of knots makes the spline smoother, but means it does not follow the data as closely. Rather than choosing the number of knots to determine the level of smoothing, Wood proposes a different approach (Zuur et al. 2009). Representing the GAM in its linear model form, Wood suggests minimising the following quantity:

$$\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 + \lambda \int f''(x)^2 dx.$$

The first part of this quantity is the standard sum squared error which we aim to minimise in most models to improve their fit. The second term 'penalizes models that are too "wiggly"' (Wood (2006b) p. 128), since the second derivative of 'wiggly' models will be larger. The λ is called a 'smoothing parameter'. When $\lambda = 0$, there is no penalty for the the smoothers being too 'wiggly' and fitting more closely to the data. As λ becomes larger, the smoothers tend to a straight line, as a straight line's second derivative is zero. If λ is known, then it is reasonably straight forward to optimise this quantity. However, λ is usually not known.

When both $\boldsymbol{\beta}$ and λ are unknown, Wood (2006b) suggests the use of cross validation to calculate their 'best' estimates. Let the 'optimal' function for our model be denoted $f(x_i)$ and an estimate of the function $\hat{f}(x_i)$. Then, λ could be chosen so that $f(x_i)$ is as close to $\hat{f}(x_i)$ as possible, which can be measured by:

$$M = \frac{1}{n}\sum_{i=1}^{n}\Big((f(x_i) - \hat{f}(x_i)\Big)^2.$$

Given $f(x_i)$, it would be possible to find a λ such that the value M is as small as possible. However, since $f(x_i)$ is not known, M must be estimated. This estimate can then minimised as a function of λ, allowing λ to be found and then the coefficients $\boldsymbol{\beta}$ to be found. The process used to estimate M is called cross validation and we do not present it here; extensive details can be found in Wood (2006b). However, note that since we will be fitting Poisson GAMs, we assume we know the variation of the data and so can use the Unbiased Risk Estimator (UBRE) to find M. A smaller UBRE indicates a better fitting the model.

For more information on GAMs see Hastie and Tibshirani (1990), Wood (2006b).

7.3.2 GAM fitting

In the deaths data, we know that each age group and sex have a different seasonality. Thus, for each age group and sex, we will need a different seasonal pattern. To fit a GAM that can accommodate this, requires the use of 'variable coefficient models', initially proposed in Hastie and Tibshirani (1993). In these models, the smooth functions can be multiplied by known covariates. Thus, Equation 7.5 becomes:

$$g(\mu_i) = \mathbf{X}_i^*\boldsymbol{\theta} + f_1(x_{1i})x_{2i} + f_2(x_{3i}, x_{4i})x_{5i} + f_3(x_{6i})x_{ti} + \ldots. \tag{7.6}$$

The `mgcv` package allows us to fit such models through the use of the 'by' parameter in definitions of the smoothing functions (Wood 2006a).

In the model we will come to fit, we will use a cyclic cubic regression spline on the period time p to capture the annual seasonality. This cubic spline requires that the start and end of the curve must be equal (Wood 2006b).

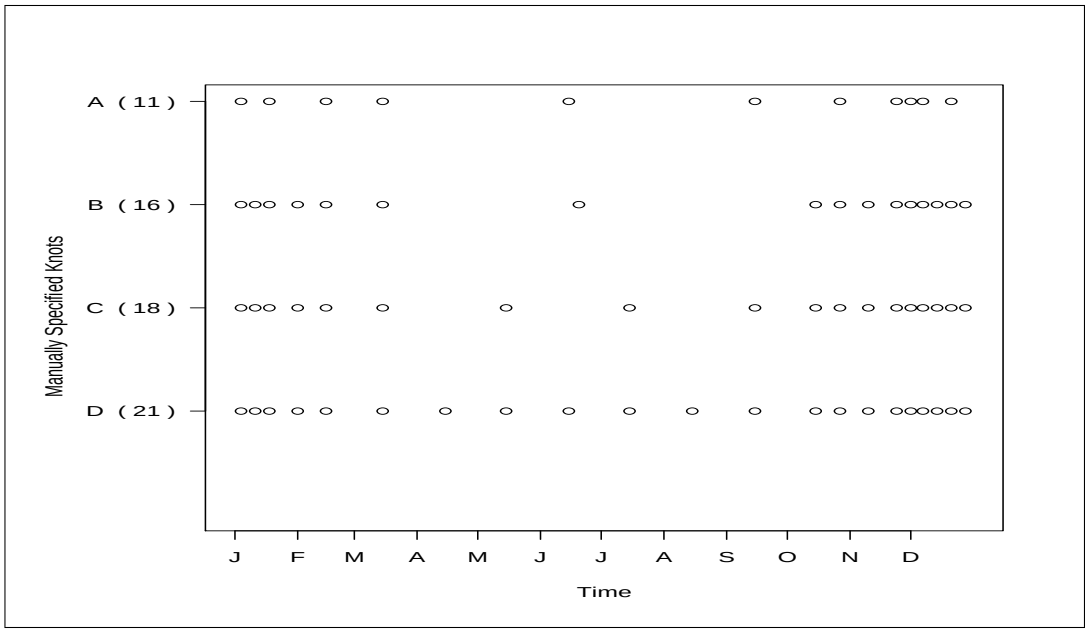


Figure 7.6: Positions of the knot points within the manually specified knot series for the GAM models fit to the totals of deaths for each age group and sex combination.

We fit the following Poisson GAM to the daily deaths recorded in Scotland using the `gam` function from the package `mgcv`:

$$log(\mu_{tas}) = \beta_0 + Sunday_t + AgeSex_{as} + AgeSex_{as} \times f(p_t), \tag{7.7}$$

where $f(p_t)$ is cyclic cubic regression spline fit on the within period time p_t and the other terms are as before. As $AgeSex$ is a factor, this results in a separate smooth for each level of the factor and thus a separate seasonal pattern for each combination of age group and sex. The `gam` function also allows us to specify the number and/or position of the knot points of the spline. We try a range of values for the knot points. If only a number of points is specified, knots are placed equally throughout the data. We try specifying 12, 26 and 52 knots to roughly place them at monthly, fortnightly and weekly intervals respectively. We also try four other definitions of knots, as shown in Figure 7.6. These knot points are placed most densely around the Winter solstice and then decrease in density through to the Summer solstice. While it is not clear the exact reason for the seasonal peak in the Winter (see Section 7.1 for a discussion of this), the amount

Knots	UBREs
A (11)	0.0670927
B (16)	0.0663166
C (18)	0.0663402
D (21)	0.0655781
Monthly (12)	0.0674307
Fortnightly (26)	0.0658283
Weekly (52)	0.0658346
Lowest UBRE	D (21)

Table 7.8: The different UBRE results from the fits of the GAM model defined by Equation 7.7 for each different set of knot points. The location of the manually specified knot points are shown in Figure 7.6.

of day light can be seen either as a proxy for the effect, or to at least to correlate in some way with the the effect. This justifies the way we have chosen our knot points.

To choose between the different sets of knot points, we consider the UBRE of the GAMs that result from fitting with each set of knot points. These are shown in Table 7.8. As mentioned briefly in Section 7.3.1, a lower UBRE indicates a better fitting model. From Table 7.8, we see that there is not much difference in the UBREs, but that the manually specified set of knot points D results in the GAM with the lowest UBRE. Thus, we use this GAM for the predictions of daily numbers of deaths within each age group and sex.

The levels of the *AgeSex* factor, with corresponding standard errors, are shown in in Figure 7.7. The differences between the groups are very similar to those we noted from the fits given by the GLMs (see Figure 7.5). The *Sunday* factor has a level of -0.0200, with standard error 0.0075, which, again, is very similar to the value fit by the Poisson GLMs (see Table 7.7). We consider the seasonal fit from this model and compare with that fit by the GLMs in the next Section. As with GLMs, we should check plots of the residuals to check the fit of the model. The `mgcv` package comes with the `gam.check` function, which produces a number of these plots automatically. We show the output of `gam.check` for our chosen GAM in Figure 7.8. These plots suggest that the model has a

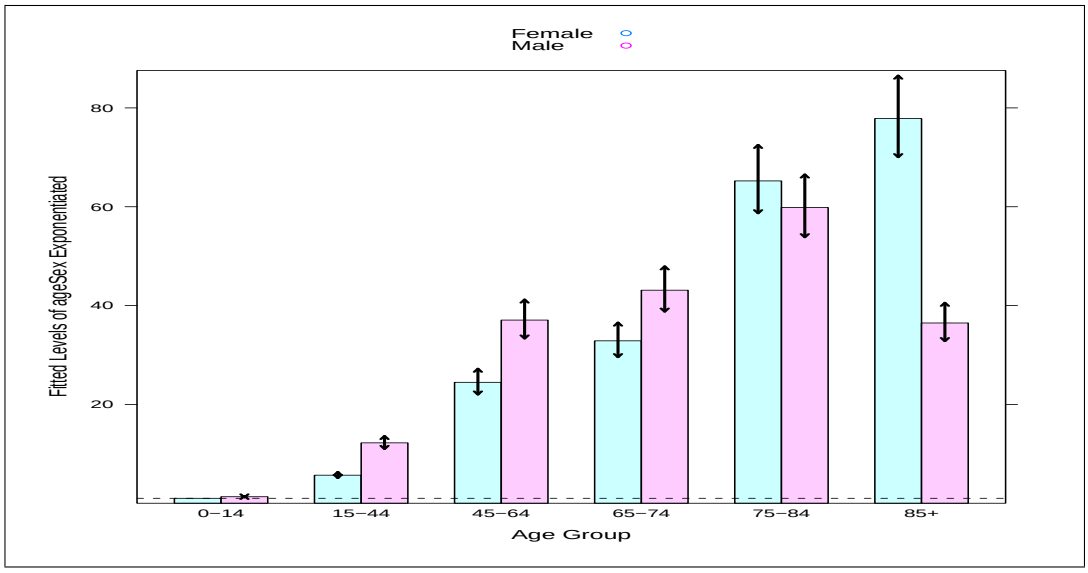


Figure 7.7: The values fitted to each combination of age group and sex on the response scale, fit in the fit in the Poisson GAM used to model the daily counts of deaths in Scotland in Section 7.3.2. These values ignore the effect of *Sunday* and seasonality. The arrows give twice the standard error above and below the fitted value on the linear predictor scale, which is then exponentiated.

reasonable fit, with variance being dealt with suitably by the model. The bottom right plot follows a reasonably straight line as we would hope.

7.4 Comparison of Models

In this Section we consider the fit of the two models and how they differ, primarily with regard to their seasonal fits. Plots of the values fitted to each age group (separated by sex and model) are shown in Figures 7.9 to 7.14 and the fits for each sex are collected in Figure 7.15 (females) and 7.16 (male). The 'bumpy' nature of the lines in the plots is caused by the Sunday factor in each model decreasing the fitted means on Sundays.

The main differences between the models can be seen in the seasonal patterns that they fit. However, comments about the seasonal fits of the models should be interpreted with caution: as with the NHS24 data, we have a great deal of data that only comes from a small number of annual cycles. Thus, comments about the seasonal fits of the models may only apply to the years under consideration.

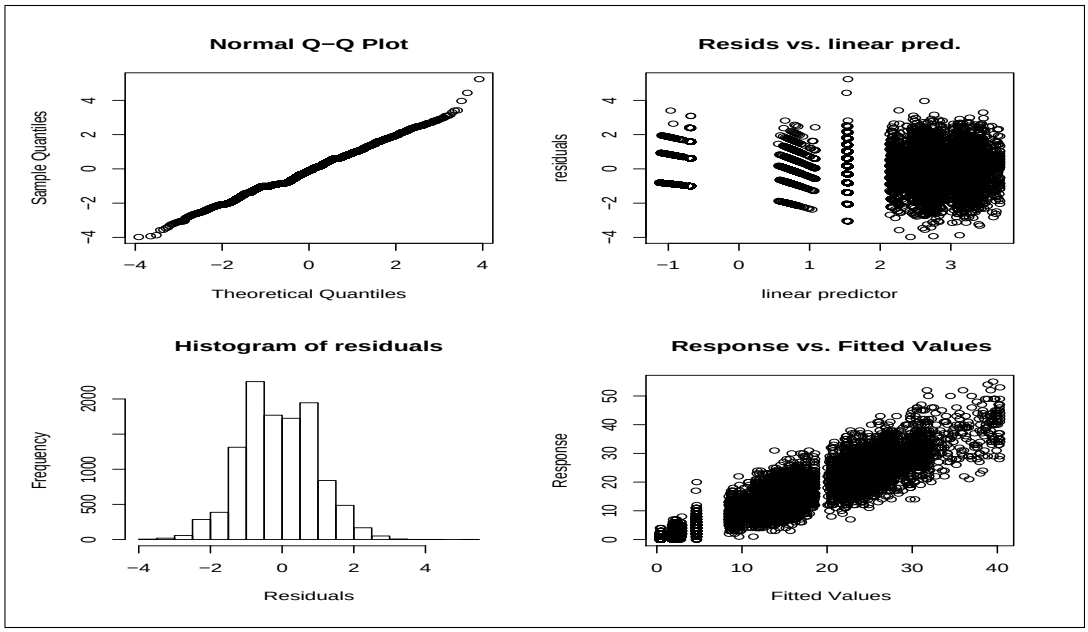


Figure 7.8: Diagnostic plots resulting from using the `gam.check` function on the GAM fit in Section 7.3.2, with knots specified by the set of positions denoted D (see Figure 7.6).

There are two main areas to comment upon: the fit of the models in the youngest two age groups and then the fit to the seasonal Winter peak.

Fitting models to the youngest age groups, 0-14 and 15-44, is difficult because of the low counts of deaths in these age groups. This is particularly true in the 0-14 groups, in which there are often less than one death per day. These groups serve to demonstrate the differences between the models. In the GLM, there is a common element of seasonality between the groups – note how the seasonality in this model is similar across the different groups of both sex and age, mostly just changing in level and amplitude. With the GAM, an essentially separate seasonal pattern is fit to each age group and sex combination. Thus, there is a marked difference between the models in the seasonal pattern fitted for 0-14 males: the GAM fits two constant levels – one for Sundays and another for the other days of the week; the GLM fits a pattern that is similar to the seasonality of the other groups. The GAM also fits an unusual pattern to the counts of females in these youngest groups. Within the 0-14 females, the largest peak occurs in the Summer. In the 15-44 females, the fit is very bumpy, with the Winter peak

Age Group and Sex	Seasonal Peaks in GLM	Seasonal Peaks in GAM
0-14F	Jan 15 (Mon)	Jul 02 (Mon)
0-14M	Jan 10 (Wed)	–
15-44F	Jan 15 (Mon)	Dec 21 (Thu)
15-44M	Jan 10 (Wed)	–
45-64F	Jan 15 (Mon)	Jan 10 (Wed)
45-64M	Jan 10 (Wed)	Jan 03 (Wed)
65-74F	Jan 13 (Sat)	Jan 03 (Wed)
65-74M	Jan 08 (Mon)	Jan 04 (Thu)
75-84F	Jan 13 (Sat)	Jan 06 (Sat)
75-84M	Jan 09 (Tue)	Jan 04 (Thu)
85+F	Jan 17 (Wed)	Jan 10 (Wed)
85+M	Jan 15 (Mon)	Jan 10 (Wed)

Table 7.9: The dates of the seasonal peaks in 2006/2007 fit by the GLM and GAM model. No peak is given for the GAM fit to the 0-14M and 15-44M groups, as there is no unique maximum fit by the GAM for these groups – see Figures 7.9 and 7.10. The fitted values for theses groups alternates between a constant mean for Sundays and another constant mean for all other days.

occurring in 2006, before the end of the year. However, differences in seasonality in these youngest groups are magnified because of the scale used in Figures 7.9 and 7.10: both models essentially fit a constant level in these age groups, as can be more clearly seen in Figures 7.15 and 7.16.

The most distinctive difference between the seasonal patterns is seen in their fit to the Winter peak, particularly among the older age groups. The seasonal pattern fit by the GAM is much 'spikier', fitting closer to the higher counts around the Winter peak (our motivating reason for using a GAM). The date of the Winter peak fit by both models for each age group and sex combination, during late 2006 and 2007, are shown in Table 7.9. The dates of the peaks for 2008 may differ very slightly, since, as this was a leap year, the seasonal cycle is fit to 366 days. In the GLM model, we find that the Winter peaks occur between January 8$^{\text{th}}$ and 17$^{\text{th}}$, with the peak for males always occurring before the peak for females. The situation with the GAM is more complex. For the 45-64 age group and above, we find that the Winter peak occurs between January 3$^{\text{rd}}$ and 10$^{\text{th}}$, generally earlier than the peaks fitted by the GLM. In the younger male age groups, 0-14 and 14-45, there is no unique seasonal peak because of two constant

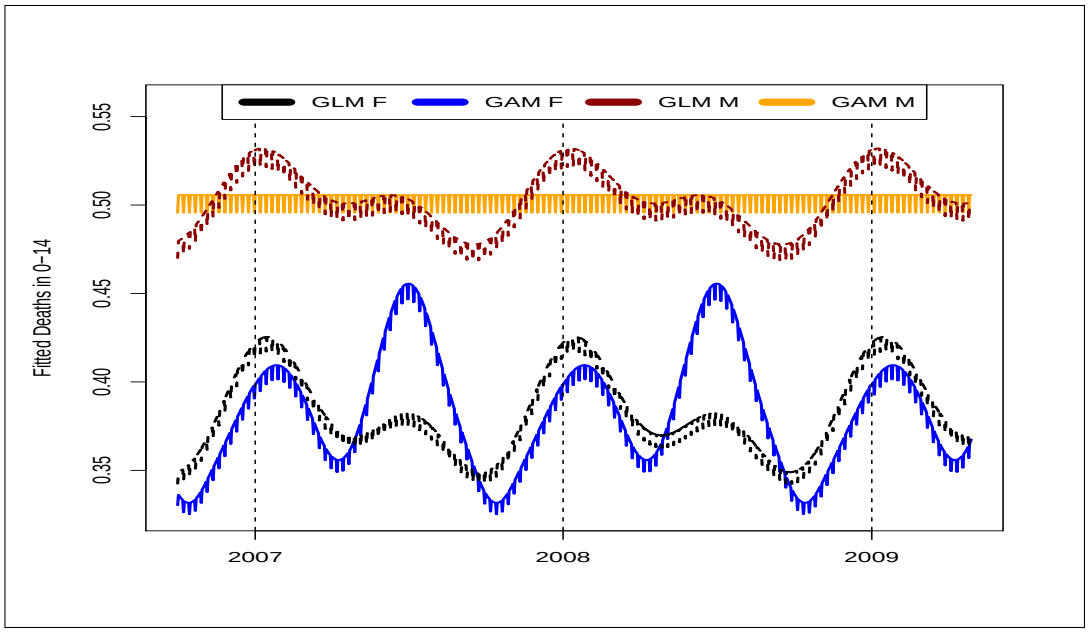


Figure 7.9: Comparative fits of the models fit to the male and female deaths in the 0-14 age group.

levels being fitted (see above). The seasonal peaks for females in these groups are also quite different.

As discussed in Section 7.3, choosing between these models depends on exactly what it is we wish to monitor. The GLM will be more consistent with previous mortality models, by considering most of the deaths around the Winter peak to be 'excesses'. The 'peakier' nature of the GAM means that it will have fewer excesses around the Winter peak. In terms of monitoring for the effect of Swine flu, we argue that it is better to use the GAM, as this fits more closely to Winter peak. In doing so, excesses from the GAM would give some indication of swine flu having a higher mortality than previous strains of influenza. The GAM may also be preferred since its residuals are less skewed than the residuals of the GLM. In the work for HPS, we developed the mortality system to include both models for prediction, allowing users to choose the model that they feel is most appropriate for their monitoring context.

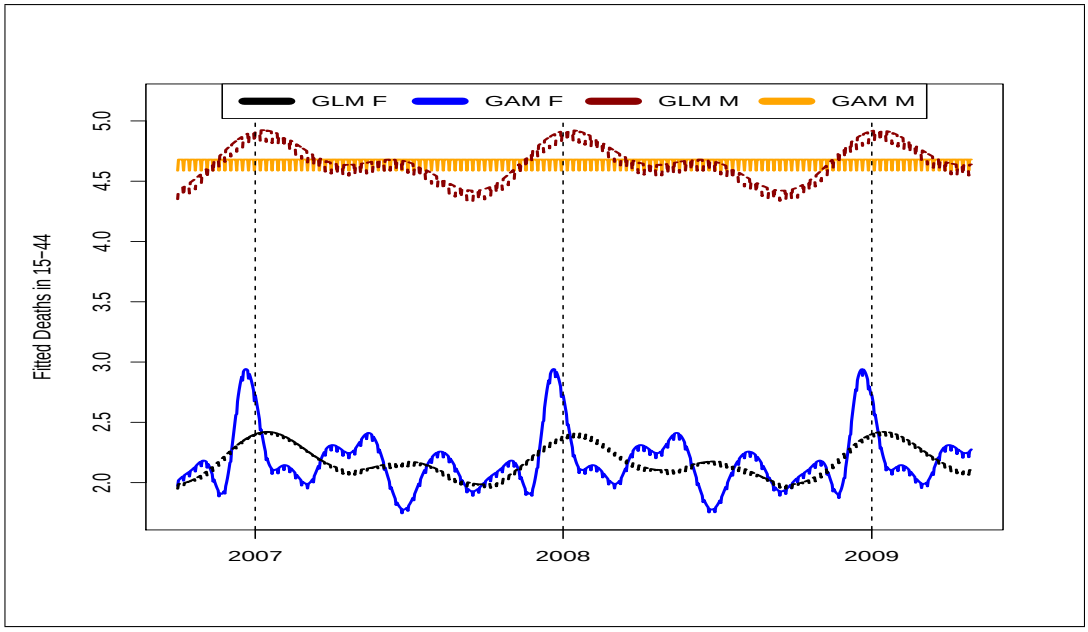


Figure 7.10: Comparative fits of the models fit to the male and female deaths in the 15-44 age group.

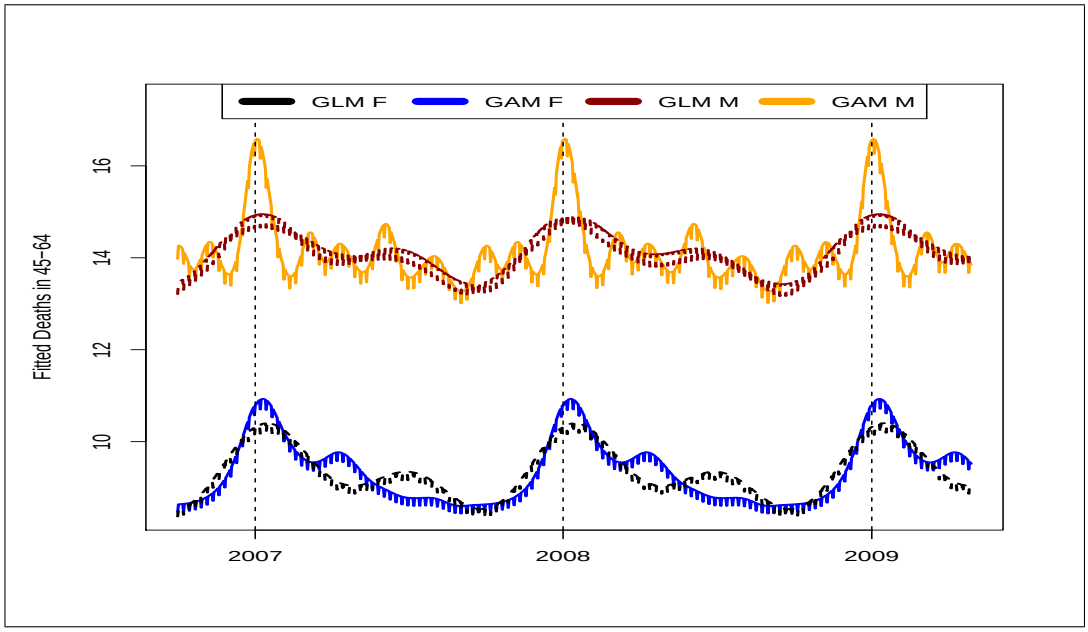


Figure 7.11: Comparative fits of the models fit to the male and female deaths in the 45-64 age group.

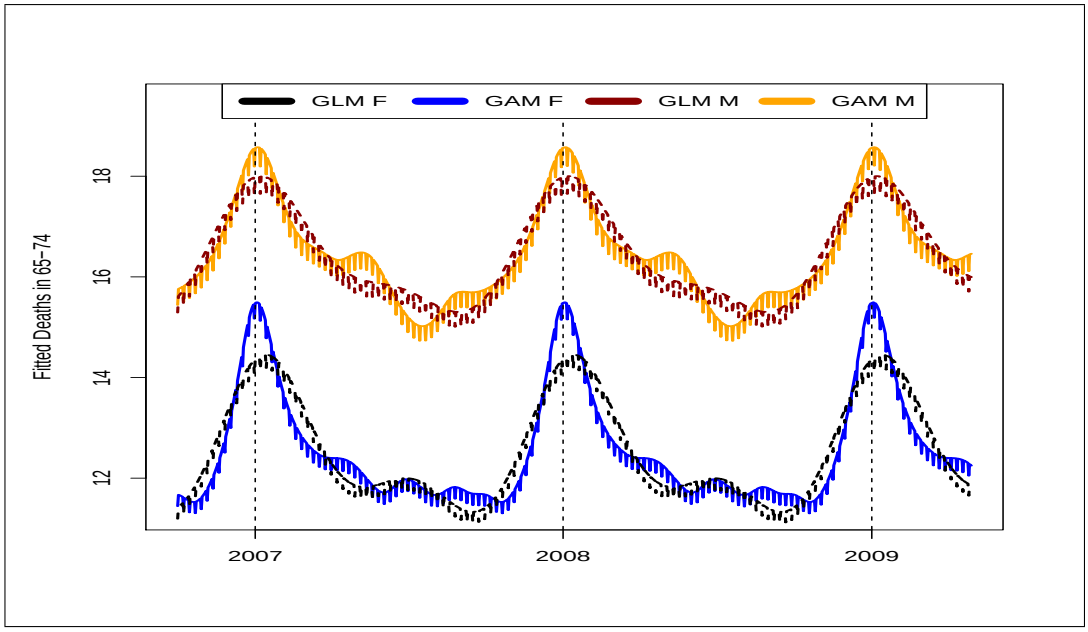


Figure 7.12: Comparative fits of the models fit to the male and female deaths in the 65-74 age group.

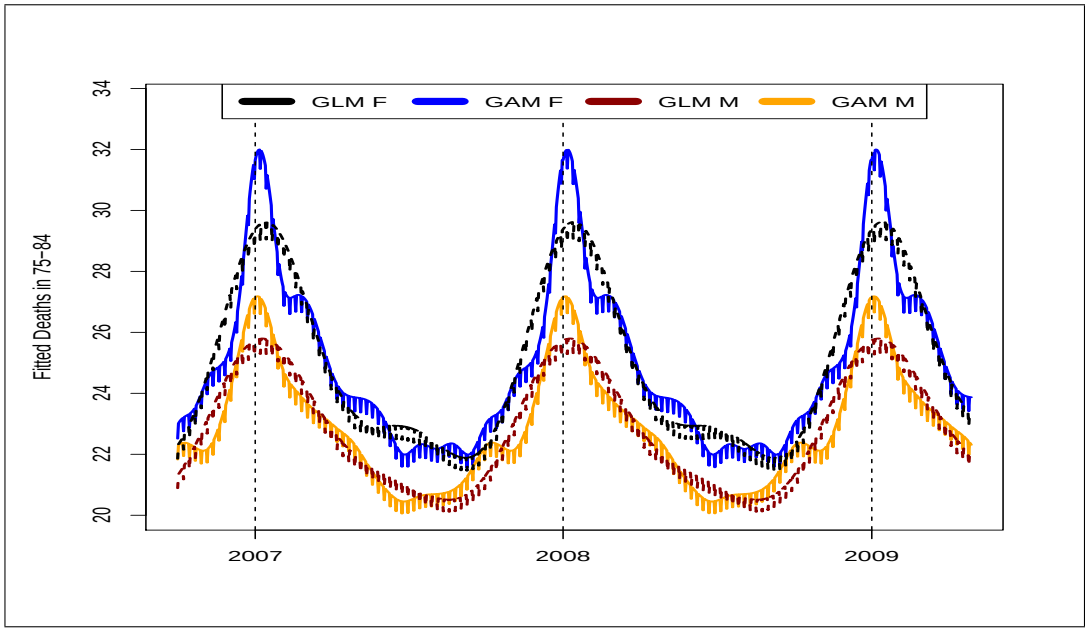


Figure 7.13: Comparative fits of the models fit to the male and female deaths in the 75-84 age group.

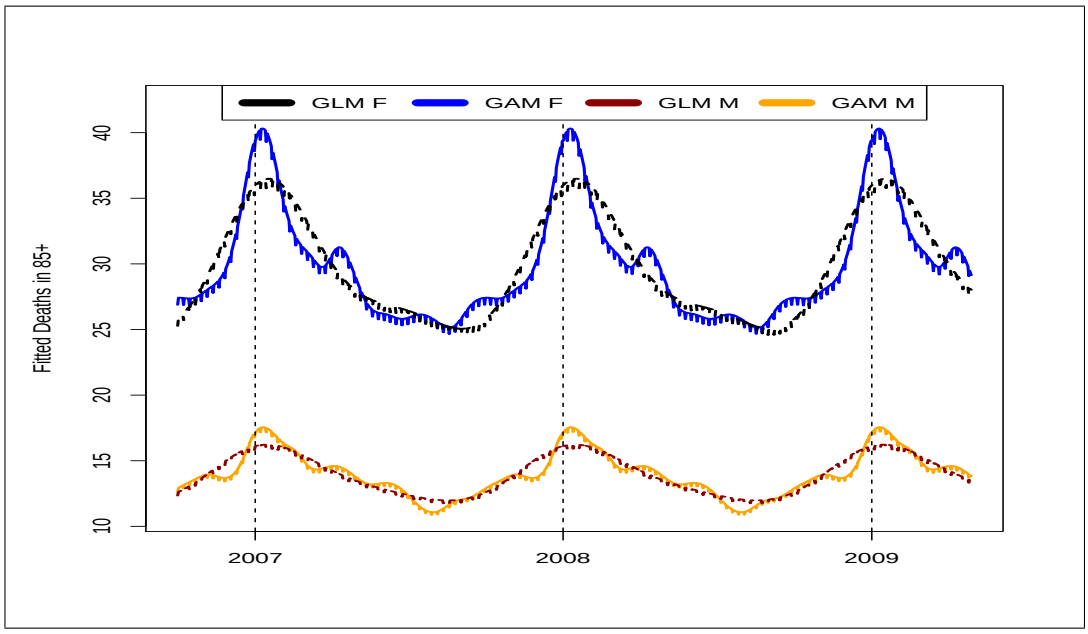


Figure 7.14: Comparative fits of the models fit to the male and female deaths in the 85+ age group.

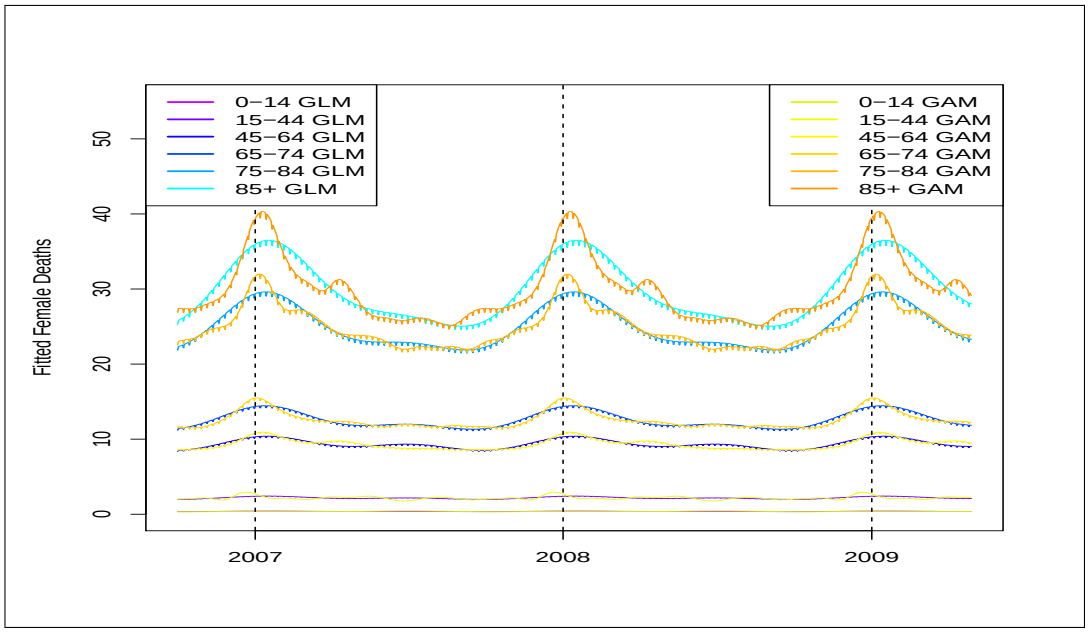


Figure 7.15: Comparative fits of the models fit to the female deaths occurring in Scotland.

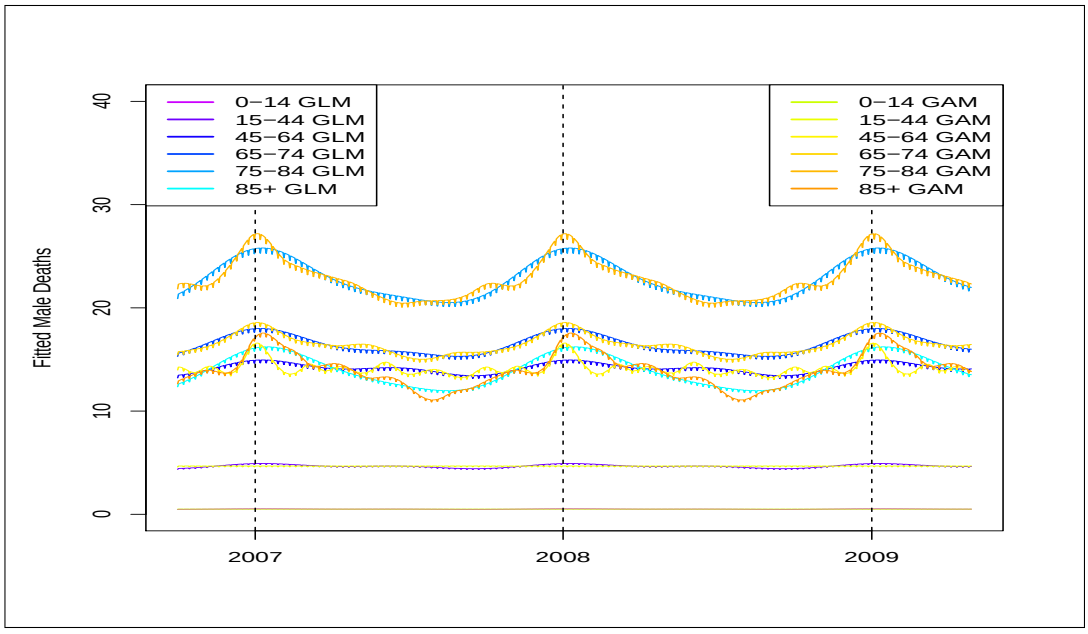


Figure 7.16: Comparative fits of the models fit to the male deaths occurring in Scotland.

7.5 Conclusions & Future Work

In this Chapter, we have fitted models to the daily totals of all those dying in Scotland. These models can then be used in a system for monitoring the numbers of deaths occurring each day. We started by using a Poisson GLM to model the daily counts of deaths, adapting a model initially used by Serfling (1963). We found that a Serfling model does not fit well to the annual peak in deaths during the Winter, since the peak is very sharp. To improve the fit of the model we fit a Poisson GAM to the counts. GAMs allow for smooth functions to be fit to data. Given these functions can be more general, the seasonal pattern in the GAM fitted more closely to the the data around the Winter peak. We argue that the closer fit to the Winter peak is both better for monitoring the effect of swine flu compared to 'normal' seasonal flu and monitoring all-cause mortality for general, non-influenza related, increases. Thus, the GAM may help extend the utility of the system beyond the swine flu epidemic. However, the final system includes both types of models to allow the user to choose the model most suitable for their monitoring context.

After the work in this Chapter had been completed, a longer period of death data was provided by GROS to HPS. Initial exploration of this data suggested that the models developed in this Chapter fitted well over the longer period. However, a small linear trend could be detected within the larger data-set, and so the models were augmented with a linear trend term. A simplified form of the GAM developed in this Chapter is fit to this larger data-set and forms the basis of the death monitoring system in use at HPS (it ignores the effect of Sundays).

An obvious direction for future work is to investigate how the Winter peak should be modelled. One thing we have not considered in the modelling here is serial correlation; addressing this may help fit models more closely to data during the Winter peak. In previous Chapters we have utilised a variable based on Holt-Winters smoothing to address serial correlation and model local trend. It may be interesting to see if that approach could work with the death data. Another approach may be to use an ARIMA model (as used in Choi and Thacker (1981a)) with covariates for each age group and sex combination. ARIMA models directly deal with serial correlation. Modelling serial correlation with either approach would allow the model to be dynamic and better adapt to the Winter peak, as the model's level would increase as the number of deaths heads towards the Winter peak.

It may be decided that the Winter peak is not something that models should be fit to. In which case, it would be considered to be an 'excess' of deaths and so an usual value. Some approaches have taken to replacing values during the Winter peak with values to be more normally expected (Choi and Thacker 1981a). Given we only had access to two full seasonal cycles, this was not an option. With the longer period of data now available to HPS, it may now be feasible to explore such replacements.

Should it become available, it may be fruitful to see if the models developed here can be augmented with cause of death. We might find that increases in certain causes of death may be a better barometer than others for predicting the total numbers of deaths. However, more accurate results may be found by modelling the different causes of death separately. Although, as we have noted previously, this has not been the experience elsewhere (Simonsen et al. 1997). Tangentially, or in addition, we could also model regional levels of deaths.

Having developed the models for monitoring deaths, we turn to addressing the reporting delay we touched on in Section 7.1. We then consider alarm methods that can be combined with the models to form a death surveillance system.

Chapter 8

Mortality Surveillance System

In Chapter 7, we developed models to be used for the prediction of deaths occurring daily within Scotland within different genders and age groups. In this Chapter, we consider the other elements necessary to complete a mortality surveillance system. We begin by considering a way of correcting for the delay associated with the reporting of deaths. This delay allows for more timely monitoring of deaths, but is not without its own difficulties. We then consider alarm methods that compare predictions from the models developed in the previous Chapter and the observed deaths, to indicate times of unusually high levels of deaths. All these elements are then brought together to give a mortality surveillance system, which we describe the use of.

8.1 Reporting Delay Across All Deaths

We briefly considered the reporting delay of deaths in Section 7.1. We found that 33% of deaths are reported the day after they occur and that nearly all deaths are reported within fourteen days ($> 99\%$) – see Figure 7.1. We would expect different days of the week to have different reporting delays associated with them; for example, deaths on Sundays will almost always have a delay of at least one day, since registrars offices are closed on Sundays. Thus. we begin by considering the reporting delays for each day of the week separately. We also separate out the delays that occur on public holidays (when again, registrars will be closed). These lead to the patterns shown in Figure 8.1. For Monday to Thursday, we find that

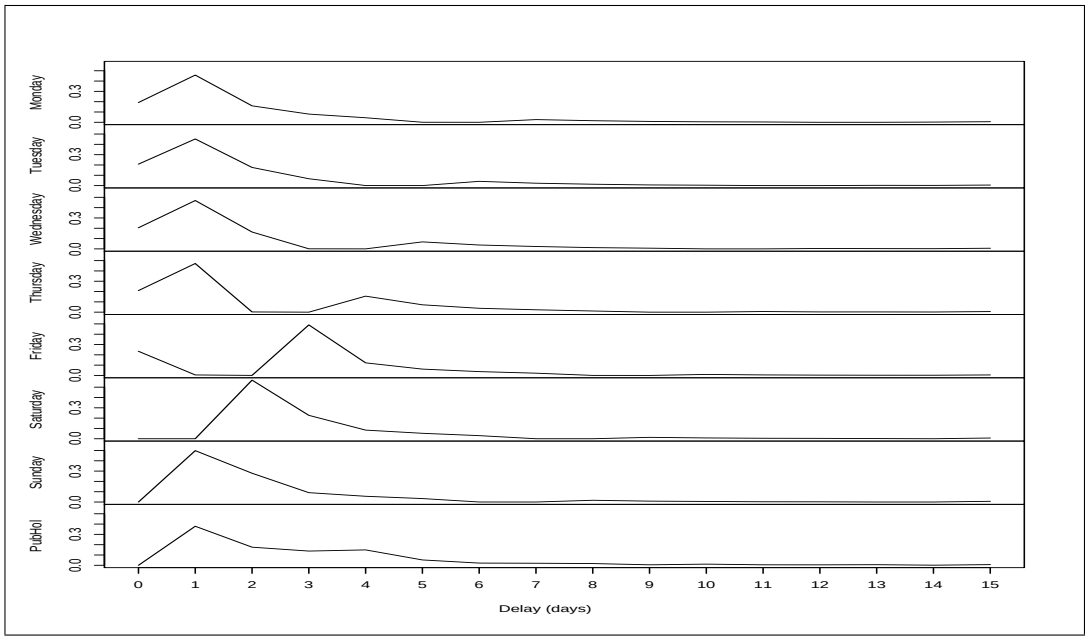


Figure 8.1: The relative proportions of reporting delay of deaths, separated by week day (including a separate category for public holidays), across all age groups and genders.

most deaths are reported the day after they occur. A similar pattern can be seen on the other days when the effects of weekends are ignored. Since the pattern in delays is broken up by weekends, we instead consider the delays measured in working days (Monday to Friday). This is perhaps more natural, since it is generally only during working days (Monday to Friday) that most registrars are open, allowing the reporting of deaths. However, it is not straight forward to calculate the working day delays for public holidays, since they occur on different days of the week. Thus, for the public holidays, we still use their raw delays, which include weekends. Public holidays are also ignored in the calculation of working day delays for the day types. This means that a small number of deaths occurring on the other days may have a longer working day delay than is actually the case.

The relative proportions of delays measured in working days for each day of the week (including public holidays) are shown in Figure 8.2. By considering the delay in measured in working days, it is much easier to see the similarities between reporting delay patterns. Indeed, when we plot the working week days

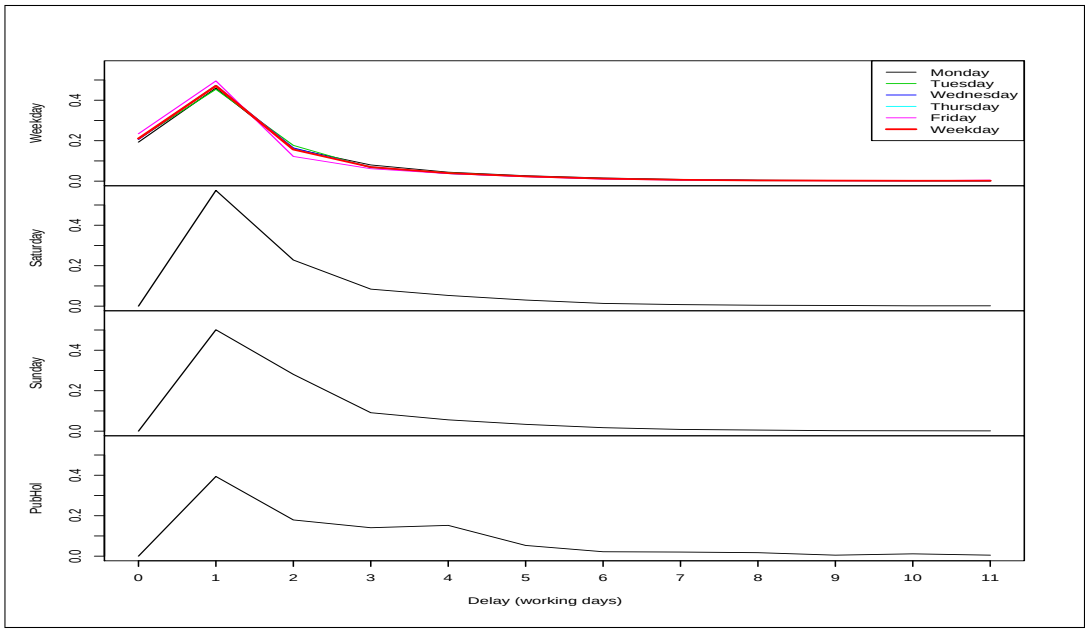


Figure 8.2: The relative proportions of reporting delay of deaths measured in working days, separated by week day (including a separate category for public holidays), across all age groups and genders.

together, as in Figure 8.2, it is hard to detect differences between the days. Friday is perhaps a little different from the other days, particularly at two days. Given the similarity in delays on the working week days, we consider all of these delays together, collapsing down the five working days delays to one set of delays. Shown by the thick red line in Figure 8.2, we see this is a reasonable fit across the working week days. There are negligible numbers of reports for Saturdays, Sundays and public holidays recorded on the day themselves, since the registrars' offices are closed on these days.

The distribution of reporting delays can be used to approximately correct the levels of reported deaths near to the present date. Consider the cumulative proportion of delays for week days, shown in Figure 8.3. Take, say, a Monday. We know that only approximately 20% of deaths that occur on the Monday will be reported that same day. The following day, HPS will only receive information on the 20% of total deaths that were reported. However, knowing about the reporting delay, and so knowing the proportion of total deaths that they expect to be reported then, they can adjust the number of recorded deaths for Monday

Date	Day	Delay	Correction	Reported	Estimated	Occurred
2008-08-14	Thu	0	0.2104	21	100	122
2008-08-13	Wed	1	0.6814	100	147	157
2008-08-12	Tue	2	0.8378	105	125	130
2008-08-11	Mon	3	0.9076	137	151	154
2008-08-10	Sun	3	0.8732	142	163	151
2008-08-09	Sat	3	0.8843	117	132	121
2008-08-08	Fri	4	0.9473	153	162	156
2008-08-07	Thu	5	0.9709	116	119	120
2008-08-06	Wed	6	0.9834	117	119	119
2008-08-05	Tue	7	0.9900	130	131	131
2008-08-04	Mon	8	0.9936	111	112	111
2008-08-03	Sun	8	0.9942	138	139	141
2008-08-02	Sat	8	0.9932	121	122	122
2008-08-01	Fri	9	0.9961	123	123	128
2008-07-31	Thu	10	0.9980	144	144	145
2008-07-30	Wed	11	1.0000	136	136	137

Table 8.1: This table shows the delay 'correction' for data that would have been received by HPS from GROS on the morning of Friday 15^{th} August, 2008. The 'Delay' column gives the number of working days from the Thursday 14^{th}. The 'Correction' is the cumulative proportion of deaths reported by the time specified in 'Delay'. We use the four day-types of working week day, Saturday, Sunday and public holiday (shown in Figure 8.2 – no public holidays occur in the period under consideration). The 'Reported' deaths are then divided by the 'Correction' to give an 'Estimated' of number of deaths that actually occurred on each day (shown in 'Occurred' column). The effect of the correction decreases the further back into the past.

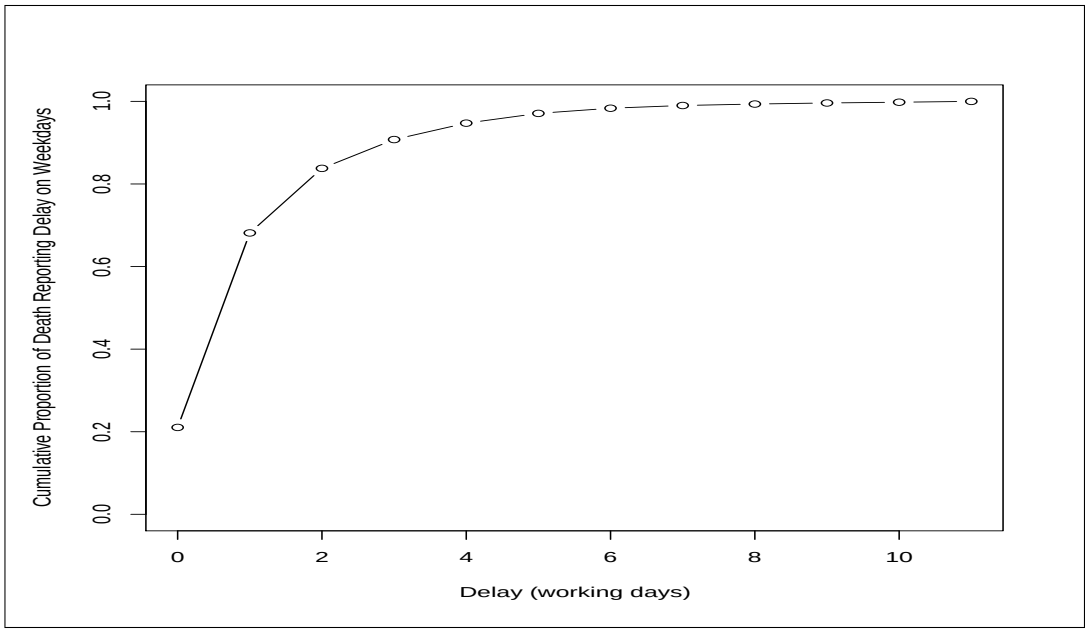


Figure 8.3: The cumulative proportion of reporting delays for week days (Monday to Friday) for all ages and genders.

to a more realistic estimate of the actual number of deaths that occurred. So, in the following day, if they know that ten deaths have been recorded, we can approximate that fifty deaths actually occurred. Such 'corrections' can be applied to the last two weeks of data, to project from the recorded number of deaths the actual number that occurred during this period. A longer period of correction could be used, but as we noted earlier, over 99% of deaths are reported within two weeks; thus, the effect of correction outside of two weeks would be very small. We give an example of the correction being applied to national counts of deaths in Table 8.1.

The models developed in the previous Chapter can be used to give predictions for the numbers of deaths occurring. The 'correction' considered here allows us to compare the prediction with the reported number of deaths in a more meaningful and timely fashion. If this, or some other correction, is not used, we would have to wait for a longer time for all deaths to be reported, before we could tell if there have been unusually high levels of death. However, the use of this correction does make any exception raised more speculative. We may find at certain times, deaths are recorded in a very timely fashion, which, once the correction has been applied,

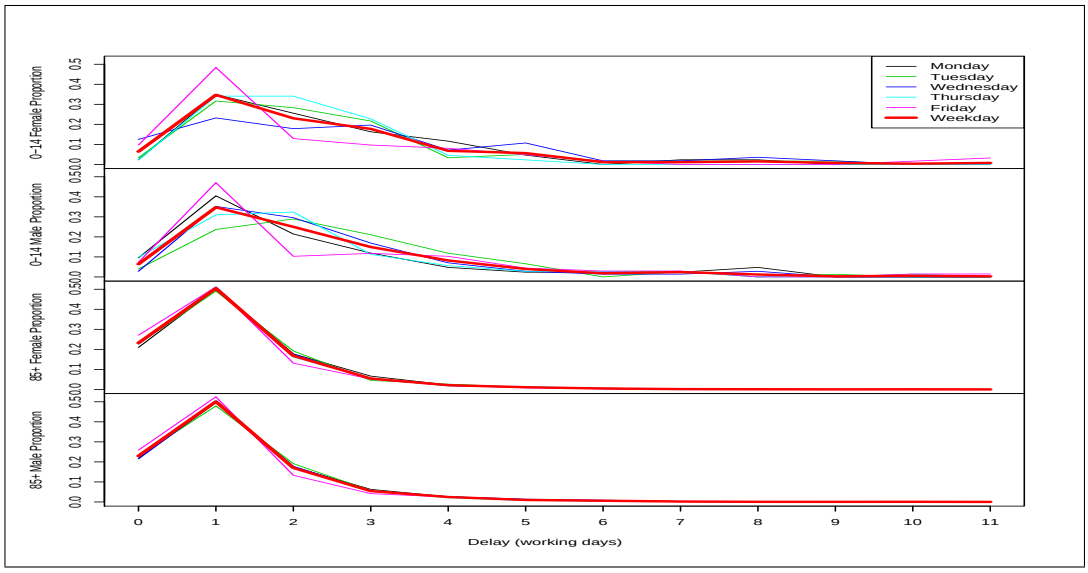


Figure 8.4: The relative proportions of reporting delay of deaths measured in working days, for Monday to Friday, with another value collecting these days together. The youngest and oldest groups split by gender are shown. Due to the much smaller number of deaths in the youngest age group, the delays in these groups are much more variable.

may give the false impression that a large number of deaths have occurred. This makes the developed mortality system somewhat harder to interpret, a point which we discuss further in Section 8.5.

We may expect a different reporting delay among the different age groups. For instance, someone who is in a working age group is more likely to be missed if they do not turn up for work and so their death may be reported in a more timely fashion. The models in the previous Chapter also give predictions for each age group and gender. Thus, we consider the different reporting delays for each age group and gender in the next Section.

8.2 Reporting Delay by Age Group and Gender

Having consider the reporting delay across all deaths, we now consider the delays within each combination of age group and gender, giving twelve groups. Within each of these 12 groups, there are potentially 8 separate days of delays (week days and public holidays), giving 96 different sets of delays. Giving this large

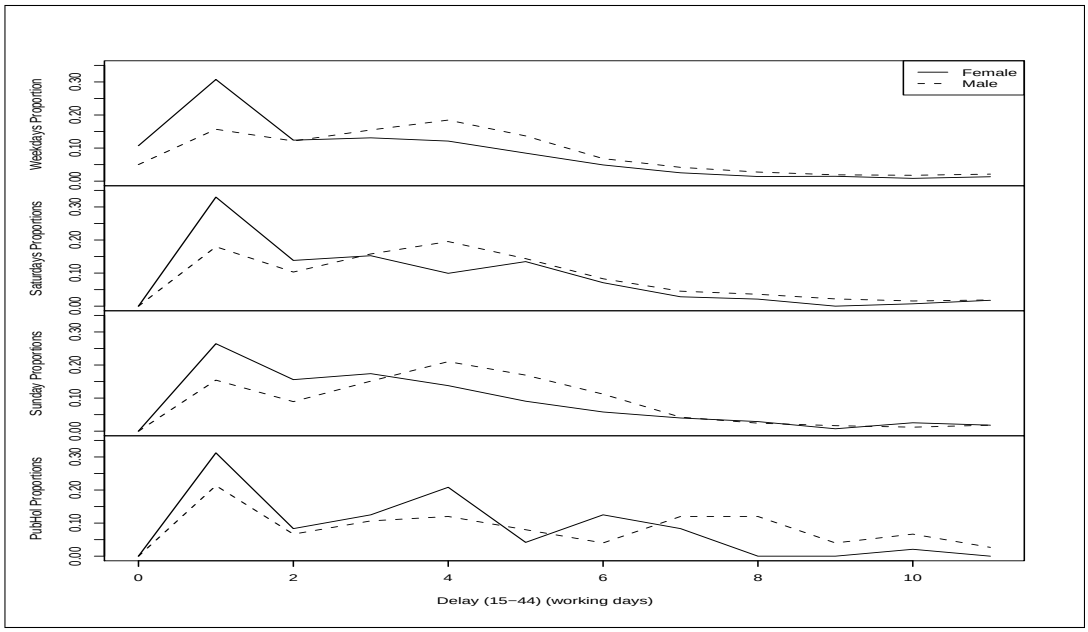


Figure 8.5: The relative proportions of reporting delay of deaths measured in working days, for the different days of the week, in the 15-44 age group, separated by gender.

number, we consider if we can group similar delays together, allowing for easier understanding of the differences between the delays.

The delays associated with the older groups are less variable than in the younger age groups, because of the larger number of deaths occurring in the older age groups. This can be seen if we consider the delays associated with working days (Monday to Friday). The working days delays for the youngest and oldest groups are shown in Figure 8.4. Generally, particularly in the older age groups, we find a pattern similar between the working days and so collect all of the working days together to give one set of delays. However, for the youngest age groups, there is a lot of variability around this collected set of delays. Given the smaller variability of the older age groups, we consider them a better indicator of the underlying pattern among the delays.

We next consider combining the delays by gender. On investigation, within each age group, the reporting delay between the genders is mostly very similar. The only age group where there are marked differences between the genders is in the 15-44 age group, shown in Figure 8.5. The delays among the males in

this group are the only ones in which there is a peak later than one working day. This is because many male deaths in this age group are caused by violence and thus require a police autopsy, increasing the delay in reporting. Given the genders among the other age groups are so similar, and for uniformity, we combine the delays for both genders within each age group. This leaves us with 24 sets of reporting delays, one for each age group (6 groups) and day type (four types: working week day, Saturday, Sunday and public holiday) combination. As described in the previous Section, these delays can be used to correct reported levels of deaths to allow for more accurate comparisons with the predicted levels of deaths given by the death count models. We show how these are included in the mortality surveillance system in Section 8.5. In the next Section, we consider if a distribution can be fitted to the delay patterns we have found.

8.3 Fitting Distributions to Reporting Delays

To produce the reporting delays considered in Section 8.2, we have just considered the raw delays in the data. This means storing the 12 different proportions for each of the 4 day types within each of the 6 different age groups – a total of 288 values. In this section, we consider if probability distributions can be used to describe the delays. If they can, then we would only need to store the parameters of these distributions. It would also allow the delays within the different age groups to be compared more easily.

To fit distributions to the delays, we use the `fitdistr` function in `R`, available from the `MASS` package. This functions fits uni-variate distributions by maximising likelihoods (Venables and Ripley 2002). We fit negative binomial distributions to the delays, since this is the most general discrete waiting time distribution. For this reason, the negative binomial is sometimes referred to as a 'binomial waiting time distribution' (Freund, Kniesner, and LoSasso 1999). The parameters, with standard errors, of the resulting negative binomial distributions for each day type (working week day, Saturdays, Sundays and public holidays) are shown in Table 8.2. Deaths that occur on Saturdays, Sundays and public holidays, cannot be reported on the day they occur – there will always be a delay associated with deaths occurring on these days. Thus, the empirical distribution of delays will always have a probability of zero for a delay of zero on these days. Such a

Age Group	μ	μ SE	θ	θ SE
Week days				
0-14	2.607	0.098	2.185	0.209
15-44	4.076	0.054	1.901	0.058
45-64	2.141	0.017	1.810	0.037
65-74	1.444	0.010	3.158	0.093
75-84	1.309	0.008	2.897	0.061
85+	1.224	0.006	1978.306	55.731
Saturdays				
0-14	1.744	0.255	0.517	0.109
15-44	3.726	0.135	0.983	0.060
45-64	1.420	0.040	0.544	0.024
65-74	0.877	0.024	0.500	0.023
75-84	0.666	0.014	0.652	0.030
85+	0.669	0.013	1.052	0.060
Sundays				
0-14	2.428	0.299	0.779	0.151
15-44	4.679	0.187	0.737	0.038
45-64	1.587	0.042	0.654	0.028
65-74	0.981	0.025	0.665	0.032
75-84	0.857	0.017	0.735	0.029
85+	0.779	0.014	1.252	0.066
Public Holidays				
0-14	2.773	0.961	0.438	0.169
15-44	5.142	0.569	0.691	0.101
45-64	2.752	0.182	0.709	0.073
65-74	1.765	0.102	0.902	0.104
75-84	1.567	0.075	0.763	0.066
85+	1.537	0.069	1.178	0.130

Table 8.2: Parameters of the negative binomial distributions fit to the reporting delays for the different age groups and day types. The parameters μ and θ are the mean and dispersion parameters, respectively, of the negative binomial distribution. Recall that the distributions for Saturdays, Sundays and public holidays are fit to the delays minus one day (See Section 8.3 for an explanation of the reason for this). SE = Standard Error.

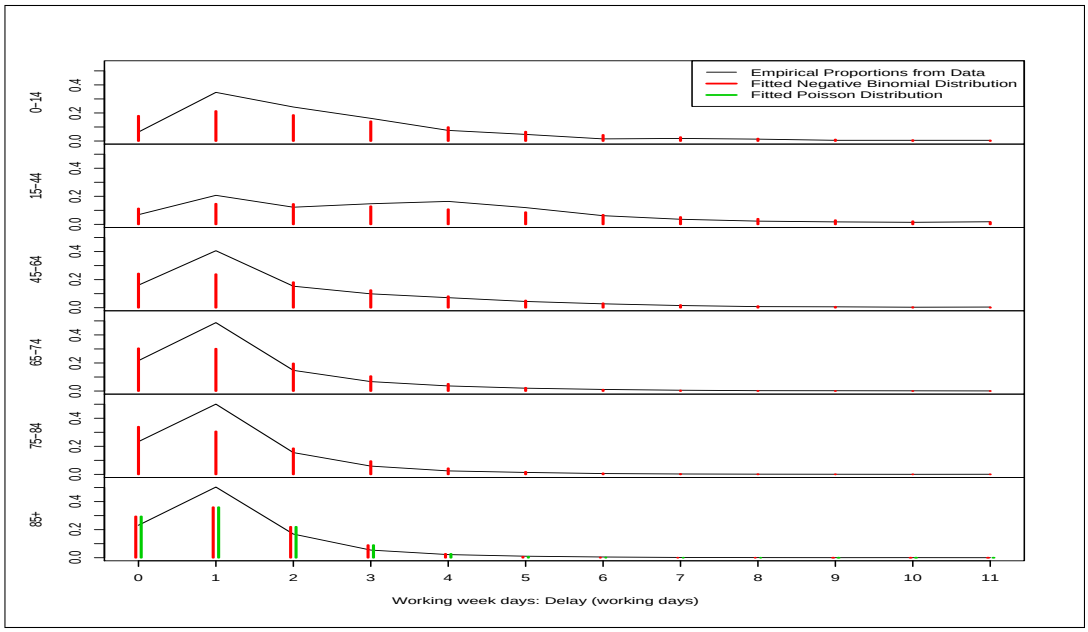


Figure 8.6: The relative proportions of reporting delays for working week days, separated by age groups, with fitted distribution values.

situation does not match with any of the probability distributions we are trying to fit to the empirical distribution of delays from these days: all of the probability distributions take the value zero with a non-zero probability. Thus, on Saturdays, Sundays and public holidays, we only consider the delays of 1 day or greater; as the probability distributions are defined to take values from zero to infinity, we use `fitdistr` to fit the probability distributions to the delays minus one day (the transformed delays then go from 0). This transform is not required on the delays associated with the other days of the week, as deaths can be, and often are, reported on the day they occur on these days. In the plots we produce, we have added one to the fitted distributions on Saturdays, Sundays and public holidays, so that the appropriate probability goes with the correct time of delay.

A plot of the fits to the delays on working week days can be seen in Figure 8.6. Generally, these distributions do not fit particularly well to the data. The main problem occurs with the fits of the distributions for 0 and 1 days of delay: at 0, the fitted distributions are always higher than the empirical proportion, while at 1, the fitted distribution is quite a bit lower than the empirical values. The dispersion parameter for the 85+ group is very large and is three orders

Age Group	Delay/No delay (B)			Delay Length (D)			
	P(No delay)	SE		μ	μ SE	θ	θ SE
Week days							
0-14	0.067	0.010		1.794	0.105	0.730	0.068
15-44	0.068	0.004		3.372	0.057	1.104	0.034
45-64	0.160	0.003		1.549	0.021	0.543	0.011
65-74	0.217	0.003		0.844	0.013	0.444	0.011
75-84	0.234	0.002		0.710	0.009	0.401	0.008
85+	0.233	0.002		0.595	0.007	0.600	0.015

Table 8.3: Parameters of the marginal distributions that form the joint distribution $Y = B(D-1)$, defined in Equation 8.1, that models the distribution of delays for working week days. The parameters μ and θ are the mean and dispersion parameters, respectively, of the negative binomial distribution. SE = Standard Error.

of magnitude larger than the dispersion parameters of the other groups. When the dispersion parameter is very large, the negative binomial tends to a Poisson distribution (Hilbe 2007). This can be seen in Figure 8.6, where we show the results of fitting a Poisson distribution to the 85+ age group: there is little difference between the fit of the two distributions.

Due to the poor fit of the negative binomial distribution to week days, we use a different approach. We consider the delays, Y, for week days, within each age group, to have the following form:

$$Y = B(D-1), \tag{8.1}$$

where: B is a Bernoulli variable, taking the value 0 with probability p and the value 1 with probability $1-p$; $D-1$ has a negative binomial distribution. This separates out modelling the delay into two parts: a report can have no delay with probability p, or the report can be delayed with probability $1-p$. If the report is delayed, then the time of its delay, minus one day, is modelled by D. This formulation gives much greater freedom to fit to the proportion of delays at 0 and 1. Conceptually, it is also reasonable: a delay of 0 days is plainly *not* a delay. The parameters for the marginal distributions B and D are shown in Table 8.3. To find B, a logistic regression model was fit to a binary form of the data, which was 0 if there was no delay and 1 otherwise. For D, the `fitdistr` function was

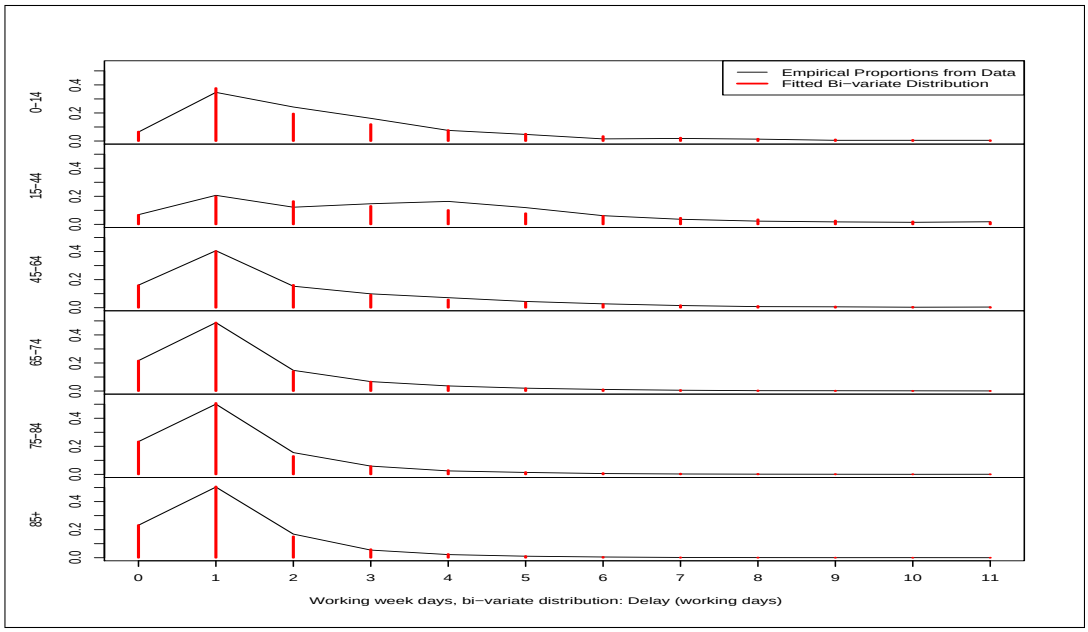


Figure 8.7: The relative proportions of reporting delays for working week days, separated by age groups, with fitted values from the distribution defined by Equation 8.1.

used as above to fit a negative binomial distribution to a subset of the data that excluded zeros.

The fit of Y for the delay on week days, for each age, can be seen in Figure 8.7. This is a much better fit than that shown in Figure 8.6. The fits for no-delay ($Y = 0$) match the data exactly but this is no surprise given the structure of B. The fits for 1 day of delay are very close to the empirical data, with the largest deviation in the 0-14 age group. Fits for longer delays are all generally good but worst in the 15-44 group. This is caused by there being a two modes in the delays of this age group, because of the difference between genders previously noted (see Figure 8.5). Comparing the values in Table 8.3 gives a number of interesting results. Generally, as age increases, more deaths are reported on the day they occur (P(No delay) increases). Part of this will be caused by younger deaths being more frequently investigated with postmortems and autopsies, delaying reporting. Levels are similar in the oldest three age groups. A similar pattern is noted in the length of delays: the shortest delays are noted in the in the oldest three age groups. The longest delay occurs in the 15-44 age group.

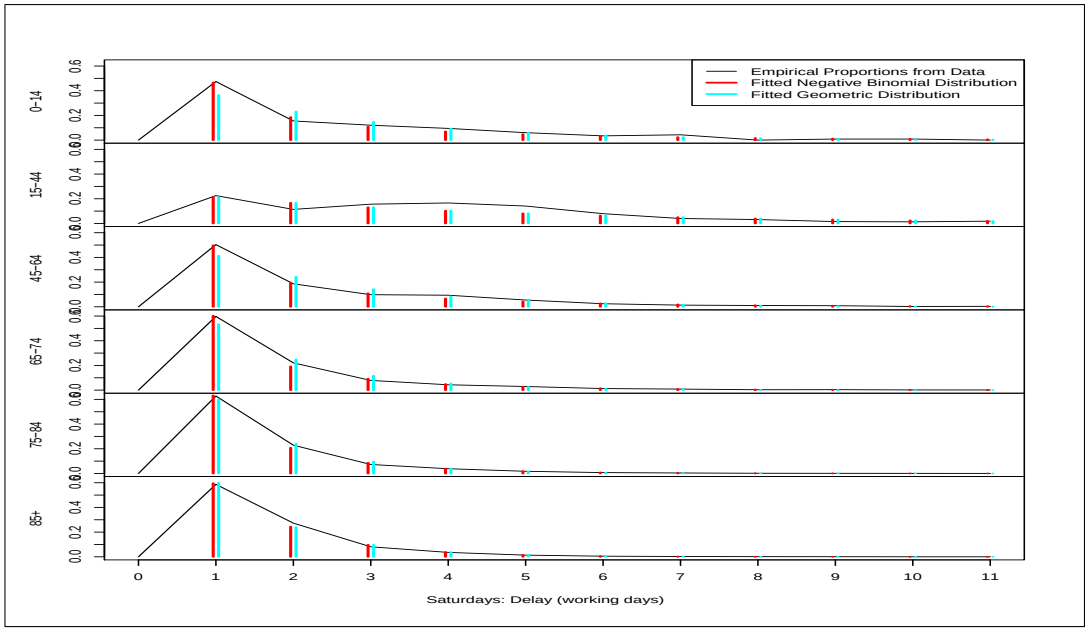


Figure 8.8: The relative proportions of reporting delays for Saturdays, separated by age groups, with fitted distribution values.

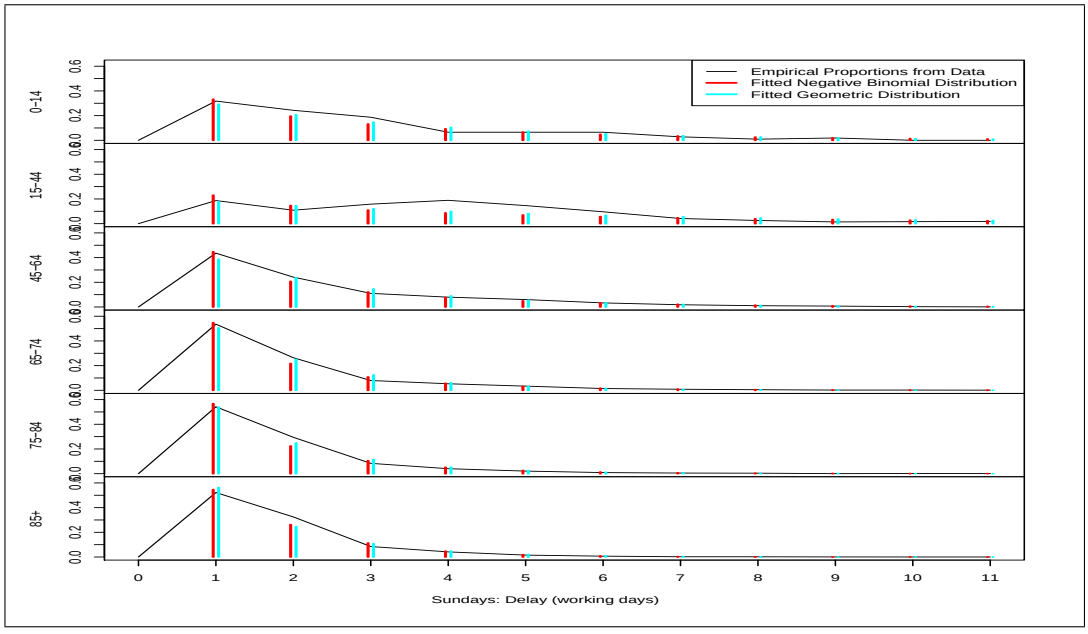


Figure 8.9: The relative proportions of reporting delays for Sundays, separated by age groups, with fitted distribution values.

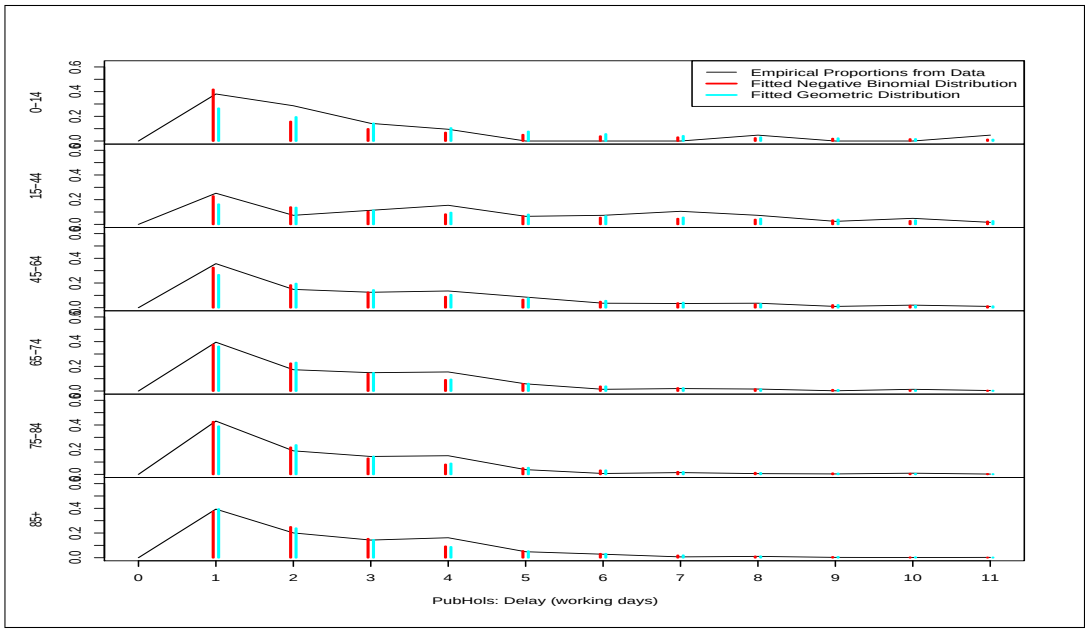


Figure 8.10: The relative proportions of reporting delays for public holidays, separated by age groups, with fitted distribution values.

The fit of the distributions to the data on the other days are also close. The fits can be seen in Figures 8.8, 8.9 and 8.10 for Saturdays, Sundays and public holidays respectively. The fits are poorest in the 15-44 age group, once more because of the underlying differences between the genders. The fits to the public holidays would be better if weekends were dealt with more precisely in the calculation of the delays measured in working week days. The same pattern in the distribution of means as noted with the working week days can be seen, with the delays smallest in the oldest three age groups, all at a similar level. The delays are greatest in the youngest age groups, with the maximum in the 15-44 age group. Some of the dispersion parameters are close to one (for instance, $\theta = 0.983$ for Saturdays in the 15-44 age group). When the dispersion parameter is one, the negative binomial becomes the geometric distribution (Freund et al. 1999). We show the corresponding fits of the geometric distribution in Figures 8.8, 8.9 and 8.10. The differences between these two distributions are small, but we prefer the negative binomial fit, since its fits are generally closer to the empirical data on day 1. This is important, since the biggest correction is applied to the data on day 1, since this is when the smallest proportion of death that have occurred are reported.

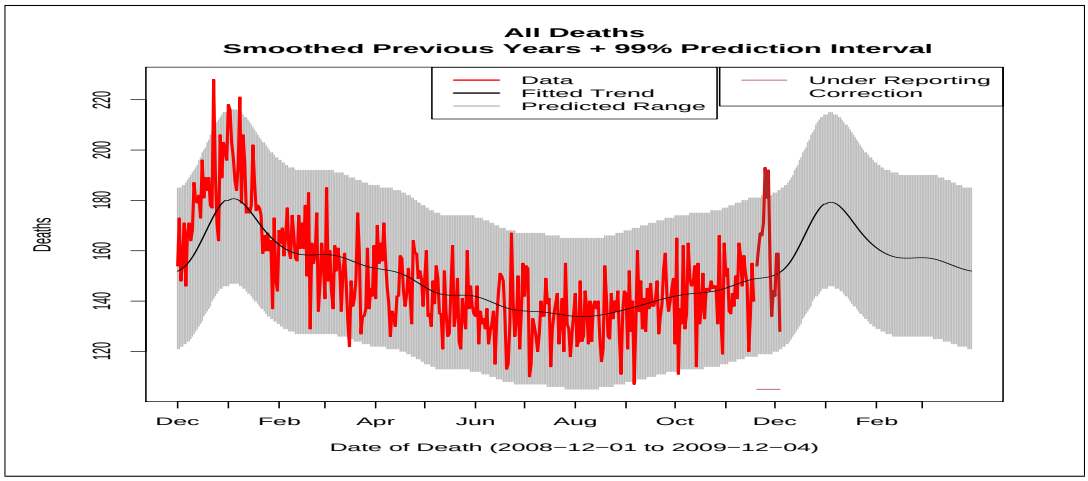


Figure 8.11: An example of the output from the mortality surveillance system that is emailed to HPS on a daily basis for the monitoring of mortality and included within the weekly report. The grey lines correspond to the 99% confidence interval. The inclusion of the line indicating the period over which the under-reporting correction is applied helps remind users that exceptions near to the current day are speculative because of the under-reporting correction.

The distributions seem to fit reasonably well to the delays and so could be used in the mortality system for producing the correction factors. However, this work was completed after the implementation of the mortality surveillance system. Thus, the system currently in use at HPS does not utilise it but rather uses the empirical values considered in Section 8.2.

8.4 Alarm Method for Detecting Unusually High Levels of Deaths

From the models developed in Chapter 7, a prediction for the number of deaths can be found. In Section 8.2, we considered an approach to correcting the number of reported deaths received in the last two weeks, to a figure closer to the actual number occurring. It remains to choose a way to compare these two figures, so that it can be decided when an unusually high number of deaths are occurring.

As with the previous systems we have developed, we use an exceedance reporting method to detect unusually high levels of deaths. This increases the

potential for detecting large results in a more timely fashion. This method has been used in other mortality surveillance systems, for example Choi and Thacker (1981a), Serfling (1963). From the distribution fitted by the GLM or GAM, we find an upper confidence limit. If the number of reported cases (after the delay correction has been applied), exceeds this limit, then an exception is reported. Typically, a 99% confidence limit is used and this can be portrayed in a graph as in Figure 8.11, for easy interpretation.

As Serfling (1963) notes, we expect a number of exceptions from natural variation. At a 99% level, we expect around 19 exceptions ($1\%/2$, $365 \times 0.05 \approx 19$). Thus, most attention is paid to consecutive days of exceptions, which should occur with a lower probability (Serfling 1963).

Another measure used in other systems is the number of 'excess' deaths. Simonsen, Clarke, Williamson, Stroup, Arden, and Schonberger (1997) defines excess mortality during an influenza season as the 'difference between the number of deaths observed and the expected baseline in the absence of influenza'. In our context, we could define excess as any positive difference between observed deaths and the predicted number of deaths. However, this is not a particularly useful measure, since it gives no idea of context: an excess of twenty deaths when the expected number is fifty is much more serious than when the expected number is one hundred. However, for easy comparison with other systems, we include an output of excess deaths in the system. We cumulatively sum excess deaths until the observed numbers are less than the expected numbers, at which point the excess count is reset to zero.

Serfling (1963) notes influenza epidemics often have 'an initial small rise in excess mortality which is not distinguishable from random fluctuations in preceding weeks' and that 'rarely does this first rise exceed the epidemic threshold' (p. 504). As we have daily data, we may stand a greater chance of detecting this 'initial small rise', but it is more likely to detected through the use of CUSUM charts (Burkom 2007). However, we leave this to be developed in future work. The alarm methods might also be augmented by considering more closely the relation between all-cause death and influenza-death (Simonsen, Clarke, Williamson, Stroup, Arden, and Schonberger 1997).

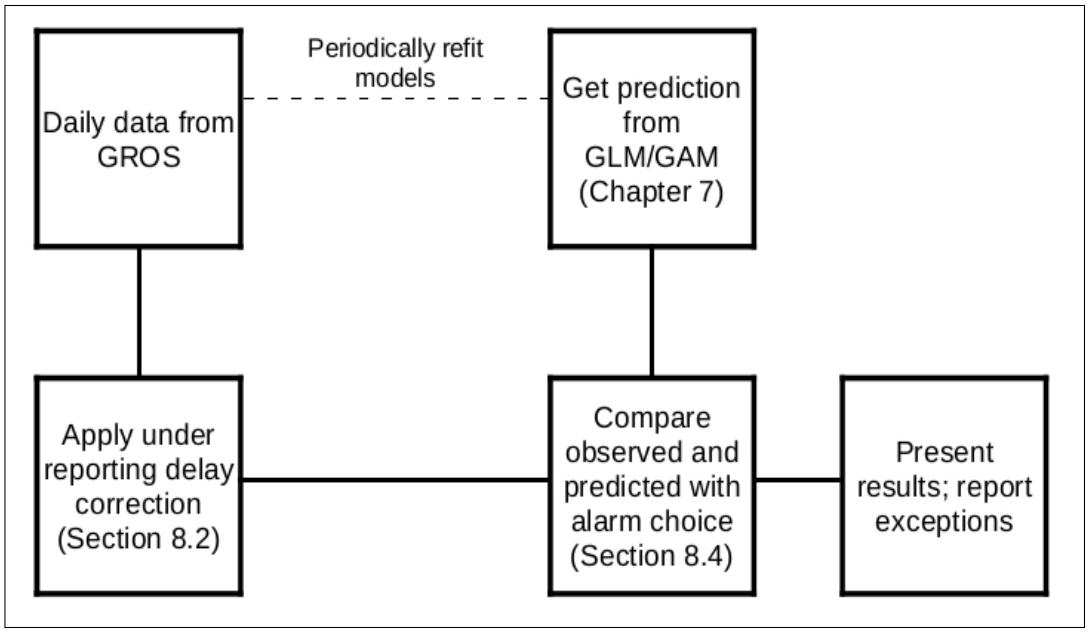


Figure 8.12: The structure of the developed mortality surveillance system.

8.5 Mortality Surveillance System

We now have the necessary elements to construct a mortality surveillance system. The structure of the system is outlined in Figure 8.12. In Chapter 7 we used a back log of historical data from GROS to fit a model to allow the daily prediction of expected deaths within six different age groups for each gender. Each day GROS provide details of the latest deaths reported to them. Due to the delay in the reporting of deaths, these reported levels will be lower than the actual numbers of death occurring. Since we know the distribution of these delays, these reported levels can be 'corrected' to a more representative level, using the work in Section 8.2 (the fitted distributions could be used, but the developed system did not – Section 8.3). The 'corrected' observed values can then be compared with the predicted values, using an alarm method such as exceedance reporting (Section 8.4). If the reported number of deaths is suitably larger than the predicted number, an exception is raised. In any case, these results are reported back to the appropriate parties.

Typically, HPS is interested in monitoring nine quantities: deaths in each age group (six); deaths in each gender (two); all deaths (one). For each age group

and gender, we have fitted Poisson distributions giving the expected number of deaths. To find the prediction for some combination of these groups, we can simply add the fitted means together, since if two Poisson variables have means μ and λ, then their sum has a Poisson distribution with parameter $(\mu + \lambda)$. Thus, for the all deaths category, we simply sum all the means; this calculation lies behind Figure 8.11. The distribution of summed means is used to calculate the exceedance limits.

Use of the system has shown that it is not entirely trivial to use, because of the correction applied for under-reporting. What can appear at first as exceptions can disappear once more data is collected and the reporting correction is smaller. To help ease the interpretation of exceptions, a model has been fit to the number of deaths reported each day. This helps separate out exceptions that are simply due to prompt reporting and those that are more likely due to unusually high levels of deaths. However, some judgement is still required, meaning that this system is best used by those with more statistical experience, particularly when the system is first introduced.

8.6 Conclusions and Future Work

In this Chapter and the preceding one, we have developed a system that can be used for the daily monitoring of all-cause mortality within Scotland. As a well developed and reasonably small nation, it is possible to collate details of all deaths occurring within the country in a reasonably timely fashion. Less developed nations, such as China, face greater challenges because of their large and isolated poor rural populations (Yang, Hu, Rao, Ma, Rao, and Lopez 2005). For nations as large as the United States of America, it is not possible to collate all this data sufficiently quickly and therefore a sample of cities must be used (Choi and Thacker 1981a). Sampling in such a way will bring with it the usual associated problems. However, there are still some smaller developed nations where such systems are only now being developed (Zucs, Buchholz, Haas, and Uphoff 2005). Thus, Scotland is one of the few places where such a system is feasible.

Developed in response to the swine flu epidemic, this system has been used to reassure the public about the severity of swine flu in Scotland. Its results

have been included in the daily reports to HPS and the official public governmental report on the state of the swine flu epidemic (Health Protection Scotland 2009b). Fortunately, at the time of writing, there have been no marked changes in mortality within Scotland during the time of the epidemic. This system, alongside others, was described in a presentation at the annual meeting of the Royal Statistical Society on the Scottish response to monitoring swine flu (Robertson, Kavanaugh, Wagner, McMenamin, Murdoch, Reynolds, Shakir, and Donaghy 2009). Our system stands out as a *daily* death monitoring system; most systems are weekly. We are also not aware of GAMs being used elsewhere for capturing the seasonal pattern in this particular context. While all-cause mortality is typically used as a barometer for seasonal influenza, there is no reason that the system cannot be used for monitoring deaths beyond the end of the swine flu epidemic.

A good area for future work would be to look more closely at the reporting delay correction. A first step would be to implement the delay corrections using the distributions considered in Section 8.3. Better results may be gained having a distribution to each of the genders in the 15-44 age group, due to the combined age group being bi-modal. Ideally, work would be done on making the delay correction more robust, so that the system becomes easier to use, requiring less statistical expertise. With sufficient amounts of data, it may be possible to subdivide the public holidays and see if there are consistent differences between them. Further, with additional data, it would also be interesting to see how, if at all, the pattern of delays have changed over time. Perhaps there may be a seasonal element to the delays: when medical staff are busier in the Winter with heavier workloads, it is conceivable that there will be increased delays associated with reporting deaths.

To help make the system more easily interpretable, we could also investigate inflating the confidence limits during the under-reporting correction period. The largest inflation in the confidence limits would happen on the limits for the day just before the system is ran and reduce to zero at the end of the correction period (currently, two weeks previously). The ‘inflation’ would help take account of the increased variation caused by the use of the correction. We would need to investigate how this inflation would be calculated and evaluated.

Different alarm methods could also be considered. We have already suggested in Section 8.4 that CUSUM charts may be useful for detecting smaller changes at an earlier points. Different alarm methods may also make it easier to deal with the Winter peak and its associated problems we have described in Chapter 7.

Chapter 9

Summary and Future Work

9.1 Summary

In this thesis we have considered work on extending and producing three surveillance systems for use in Scotland. The first of these used daily data from the initial deployment of NHS24, a twenty-four hour medical advice phone line. We considered systems for surveillance of the syndromes present in the log of calls and focused our attentions on monitoring calls related to vomiting. As we only had data from the first two years of NHS24 operations, the system developed could not be usefully tested and was somewhat speculative. The next system looked at monitoring weekly totals of positive lab test for various micro-organisms collected by Health Protection Scotland (HPS). Of the six organisms HPS asked us to consider, we chose to focus on developing a monitoring system for *Cryptosporidium*, one of the leading causes of water borne disease. This data was from a much longer period and of much better quality than the NHS24 data, allowing for the developed HPS systems to be tested and compared with the one currently used. In both the NHS24 and HPS systems, we developed regional surveillance systems for monitoring data at the health board level, which is challenging because of the small counts involved. To model these counts we used negative binomial GLMs. To deal with serial correlation and capture local trend, we considered extending the negative binomial models to utilise either past observations or a variable based on exponential smoothing of the data. Including either of these quantities allows the models to be somewhat dynamic, allowing them to adapt

to changing levels in the data. We also found that the models were improved by carefully modelling seasonality. In the HPS system, we were able to compare our regional systems against the present national system applied to regional counts. We found that the systems were of a similar quality or that our systems did better. In both the NHS24 and HPS monitoring systems, we considered if there were links between health board regions and if these could be used to improve the systems. In the NHS24 system, links were found and an approach suggested for incorporating these links in the models. A more detailed analysis was used in the HPS system, showing that the links could not be modelled easily by separate GLMs. More advanced models would be required for this.

Using data on the daily numbers of deaths from the General Registrars' Office of Scotland (GROS), we developed an all-cause mortality surveillance system. This was done in response to the swine flu pandemic during 2009, to help detect if swine flu was increasing mortality in Scotland. There were two main challenges with this system: modelling seasonality and dealing with the delay in the reporting of deaths. With limited data – only two and a half annual cycles – we needed to model a seasonal cycle that peaks sharply during the Winter. We found that the best way to do this was using a Generalized Additive Models (GAMs). To make the system timely, we also needed a way to address the delayed reporting of deaths. It was found that the delays were reasonably consistent, and so could be dealt with by considering the patterns of delays in past data. These elements were brought together to produce the mortality system that, at the time of writing, is being used daily to monitor all-cause mortality in Scotland.

9.2 Future Work

Given the opportunity, we would apply the systems developed in this thesis more widely. For NHS24, we could consider fitting monitoring systems to the other syndromes. In the HPS system, we could consider some of the other organisms suggested for study by HPS. In both systems, we would need to evaluate if the models developed here would generalise easily. This is particularly important for the NHS24 system, which was developed on a relatively small set of low quality data from the start of NHS24 operations. For the mortality surveillance system, we could investigate developing systems for monitoring deaths at the regional

health board level. It would also be fruitful to look at changing the confidence intervals during the under-reporting correction period, as discussed in Section 8.6. This should make the system more intuitive and so more widely usable.

As we have discussed previously, improvements may be gained by modelling all of the health boards simultaneously (for example, as we suggest in Section 5.6.2). This would allow us to both model links between health boards and perhaps 'share' data between them to improve the models. We might find that it is also possible to group similar syndromes (or organisms) together to further improve the models. While there are these advantages to such multi-variate modelling, there would also be some difficulties. First, multi-variate models would generally need to have common and simple structures for each health board. We would need to investigate to determine if such simple structures can capture all the elements we have dealt with in our models. Secondly, there are issues of interpretation and presentation: can the results and implications of a multi-variate system be presented easily to a non-statistical audience? It is easier, although sometimes more time consuming, to convey the results from each health board with a separate model.

In a different direction, particularly if individual board models are to be preferred, the use of zero inflated distributions could be investigated. These should give better fitting models for those organisms with very low regional counts. The residuals for these models may also be more easily interpreted (in contrast to taking steps such as we did in Section 5.3.1 to use things like modal residuals).

We could also investigate in greater detail as to what should constitute an exception. In this thesis, we have used an exceedance alarm method, where an exception is raised if the number of observed cases exceeds some upper quantile of a predicted distribution. This is quite a broad approach; better results may be gained by defining exceptions in other ways. Some of these approaches might be linked with the underlying organism that is being monitored. Also, we might have different thresholds for different organisms, depending on the relative public health threats.

A larger departure would be to look into developing systems that are more dynamic. Through the use of either past observations or the variable based on Holt-Winters smoothing, the NHS24 and HPS systems are somewhat dynamic, adapting to changes in the data. However, the parameters in the underlying

models are fixed; so, while the models can adapt to changes in the data, the models implicitly assume that the structures present (seasonality, local trend and so on) do not change. For this reason, it would be wise to periodically re-fit the models developed in this thesis, perhaps annually, to take account of structural changes. In contrast, there are modelling paradigms where the parameters of models themselves can be random and change with each new piece of data used by the model Assuming that such models work well for prediction in this area, these systems could be more automatic and require less maintenance. However, challenges may still be faced by dealing with the small counts at the regional level.

Appendix A

Perl Code Used With NHS24 Data

The code listing in this appendix gives the `Perl` program that was applied to the raw data from NHS24. It serves to check if the postcode for a particular call to NHS24 is valid or not. If it is not valid, two corrections are used to try and correct the postcode. If neither of these corrections can correct the postcode, that call is marked as having an invalid postcode.

Other rudimentary checks are carried out on the other fields for each call. If a call fails one of these checks, its details are recorded in a log file, allowing editing of the offending call to correct it.

With thanks to Ian Thurlbeck and Alan Campbell for their help on some of the more advanced regular expressions.

```
#!/usr/bin/perl
use strict;

# Stores all the valid postcode parts
my %validpcs = (
  "AB10"=>   2,
  ⋮
  # Postcodes omitted for space
  ⋮
  "PA88"=>  15,
);
```

A.1 Modules Used in the Main Code

A.1.1 Sub-module: tester

```
sub tester {
  # Checks if $testValue, from field $field satisfies
  # the regexp; if not returns an error string
  # specifying the lineNo, $field and the erroneous value

  my ($testValue, $regexp, $lineNo, $field);
  ($testValue, $regexp, $lineNo, $field)=@_;
  #print "Yay!\n";
  $_= $testValue;
  if (/$regexp/){}
  else {
    print LOG "Line $lineNo has an invalid $field, with value $testValue.\n";
    print "Line $lineNo has an invalid $field, with value $testValue.\n";
  }
}
```

A.1.2 Sub-Module: trimWhiteSpace

```
sub trimWhiteSpace($) {
    # Remove whitespace from the start and end of the string
   my $string = shift;
   $string =~ s/^\s+//;
   $string =~ s/\s+$//;
   return $string;
}
```

A.1.3 Sub-module: writeout

```
sub writeout {
  # Writes out a line to the outfile for insertion into the database
  # Hardwired values for the database
  my $database = "calls";
  my $outputString = "INSERT INTO $database VALUES (";
  my $fields = join ",", @_;
  print DATAOUT "$outputString$fields);\n";
}
```

A.1.4 Sub-module: pcfix

```
sub pcfix {
  # Checks the given postcode and if it doesn't exists, tries to fix it
  # Failing that just returns the original postcode
  my ($testpc, ) = @_;
  my ($where, $recon);
  # Set for the reg exp searching
  $_= $testpc;
  if (exists $validpcs{$testpc}) {
    # Valid postcode - return that
    $testpc;
  } elsif (!$testpc){
    # Empty postcode - return empty
    $testpc;
  } elsif (/^[A-Z]{1,2}0/) {
    # Trap errors such as XX0X-> XXX
    # Find the position of the string in $where,
    # then reconstitute the string
    $where = index($testpc, "0");
    $recon = substr($testpc, 0, $where).substr($testpc, $where+1);
    if (exists $validpcs{$recon}) {
      # Put this forward as the correct postcode
      $recon;
    } else {
      # Can't fix this postcode
      $testpc;
    }
  } elsif (/^[A-Z]{1,2}[0-9]{1,3}/) {
    # Trap errors that occur because of too many digits
    # We try to fix by dropping the last digit in the string
    $recon = substr($testpc, 0, length($testpc)-1);
    if (exists $validpcs{$recon}) {
      # Put this forward as the correct postcode
      $recon;
    } else {
      # Can't fix this postcode
      $testpc;
    }
  } else {
      # Postcode errors not caught by any of the fixes above
    $testpc;
  }
}
```

A.2 Main Code of Program

```
my (@bits, $lineNo, $calldate, $calltime, $postcode, $age, $outcome);
my ($gender, $protocol, $syndrome, $pcEmpty, $pcvalid, );

# Checks an input file for validity
# Outputs a line to insert into a database
if ($#ARGV != 2 and $#ARGV != 3) {
  print "Usage:  ", $0, " <datafile> <outputfile> <pcerrorfile> /headers/\n";
  exit 1;
}

# Set up files
open(DATAIN,          $ARGV[0])        ||die "Cannot open data file";
open(DATAOUT,    ">".$ARGV[1])         ||die "Cannot create output file";
open(PCERRORS,   ">".$ARGV[2])         ||die "Cannot create pcerror file";
open(LOG,      ">>".$ARGV[0].".log") ||die "Cannot access logfile";

# Log the repeating of the batch process in the log file
my $now = `date`;
chomp($now);
print LOG "=============================================================\n";
print LOG "Running updated $0 on the $now - now with added fixing power.\n";

# Ignore the first line of the file, if it is a header
$lineNo = 0;
if ($ARGV[3]) {
  <DATAIN>;
  $lineNo++;
}

while (<DATAIN>) {
  $lineNo++;

  @bits = split(/,/);
  #$bits[0] = $mon;
  ($calldate,$calltime,$postcode,$age,$outcome,$gender,$protocol, $syndrome)
      = @bits;
  ## NOTE: After each check, each value is put in single speech marks, if apt

  # Check the date
  &tester($calldate,"\^200[45]-(0[1-9]|[1][0-2])-(0[1-9]|[12][0-9]|3[01])\$",
    $lineNo,"Date");
  $calldate="'$calldate'";

  # Check the time
  &tester($calltime,"\^([01][0-9]|2[0-3]):[0-5][0-9]:[0-5][0-9]\$",$lineNo,"Time");
  $calltime="'$calltime'";
```

```
  # Check the postcode
  # Try to fix up the postcode first
  $postcode = &pcfix($postcode);
  if (exists $validpcs{$postcode}) {
    $postcode = "'$postcode'"; $pcvalid = "'Y'"
  } elsif (!$postcode) {
   $postcode="NULL"; $pcvalid = "'Y'";
  } else {
    # Send the invalid postcodes into a file
    print PCERRORS "Line $lineNo, has an invalid postcode $postcode.\n";
    $postcode = "'$postcode'";
    $pcvalid = "'N'";
  };

  # Check the age - assumed 119 as the maximum age
  $age=&trimWhiteSpace($age);
  if ($age) {
    &tester($age,"\^([0-9]|[1-9][0-9]|1[01][0-9])\$",$lineNo,"Age");
    $age="'$age'";
  } else {
    $age="NULL";
  }

  # Outcome is not tested as there is no structure
  if ($outcome) {
    $outcome = "'$outcome'";
  } else {
    $outcome = "NULL";
  }

  # Gender check
  if ($gender) {
    &tester($gender,"\^(Male|Female)\$",$lineNo,"Gender");
    $gender="'".substr($gender,0,1)."'";
  } else {
    $gender = "NULL";
  }

  # Protocol does not need to be tested
  if ($protocol) {
    $protocol = "'$protocol'";
  } else {
    $protocol = "NULL";
  }

  # Syndrome check
  $syndrome = &trimWhiteSpace($syndrome);
  &tester($syndrome,"\^([0-9]|1[01])\$",$lineNo,"Syndrome");
  $syndrome = "'$syndrome'";

  &writeout($calldate,$calltime,$postcode,$age,$outcome,
            $gender,$protocol,$syndrome,$pcvalid);
```

```
}

# Close the files
close(DATAIN);
close(DATAOUT);
close(PCERRORS);
# Close this log entry
print LOG "============================================================\n";
close(LOG);

# Processing number of lines is affected by if there is a header on the file
if ($ARGV[3]) {
  $lineNo--;
  print "In total, $lineNo lines were processed, with a
         header taking up another line.\n";
} else {
  print "In total, $lineNo lines were processed.\n";
}

exit;
```

Appendix B

Exponential Smoothing: Developments & Practical Considerations

B.1 Exponential Smoothing Forms & Developments

Exponential Smoothing (ES) techniques are widely used for forecasting, particularly in commerce and industry (Winklhofer, Diamantopoulos, and Witt 1996; Dalrymple 1987). Despite the simplicity and intuitiveness of the techniques, they are shown to give results comparable to '(much) more sophisticated' methods (Chatfield, Koehler, Ord, and Synder 2001; Makridakis and Hibon 2000; Chen 1997; Winklhofer, Diamantopoulos, and Witt 1996; Makridakis, Andersen, Carbone, Fildes, Hobon, Lewandowski, Newton, Parzen, and Winkler 1984; Makridakis, Andersen, Carbone, Fildes, Hobon, Lewandowski, Newton, Parzen, and Winkler 1982; Makridakis, Hibon, and Moser 1979). ES techniques can also be automated, which is useful when many time series need to be forecast (such as in inventory control), or when a subject specialist is not available (Chatfield and Yar 1988). The original method of ES is known as Simple Exponential Smoothing (SES). It has been extended to deal with trend (e.g. Holt's linear trend model) and then seasonality & trend (e.g. Holt-Winters, General Exponential Smoothing).

B.1.1 Simple Exponential Smoothing (SES)

The SES methods were developed in Brown (1959), Brown (1963), Winters (1960), Holt (2004) and others in the 1950's, but there is no single consensus on the first developer (Gardner 1985a; Chatfield, Koehler, Ord, and Synder 2001). SES is probably one of the 'best known forecasting methods', likely due to its conceptual simplicity and intuitiveness (Chatfield 2001). SES is one of the prediction methods considered in McCabe (2004), but rejected in favour of using GLMs. We now introduce SES in the manner of Chatfield et al. (2001).

We will denote an observed time series by $x_1, x_2, \ldots, x_n$. The forecast of x_{t+m} made at time t is denoted $\hat{x}_n(m)$ where the integer m is called the 'lead time' or 'forecasting horizon'. The observed one-step-ahead predictions errors, e_t, are given by $e_t = x_t - \hat{x}_{t-1}(1)$. SES updates an estimate of the local (mean) level by using a recurrence equation

$$\hat{x}_t(1) \quad = \quad \alpha x_t + (1-\alpha)\hat{x}_{t-1}(1), \tag{B.1}$$

but this is more often given in the error correction form

$$\hat{x}_t(1) \quad = \quad \hat{x}_{t-1}(1) + \alpha e_t, \tag{B.2}$$

where α is a smoothing constant or *parameter*, normally restricted to the range $(0, 1)$ (Chatfield et al. 2001). It commonly takes values between 0.1 and 0.3 (Gardner 1985b), but this is frequently not the case (Chatfield and Yar 1988). The parameter can either be manually specified, or, more usually, is chosen to minimise the squared one-step-ahead prediction errors ($\sum e_t^2$) (Chatfield et al. 2001). When this method is considered for prediction in McCabe (2004), values of α from 0.1 to 0.9 in steps of 0.1 are tested to find which value minimises $\sum e_t^2$, using data from the previous year. The optimal value of α is used to make a prediction for the week ahead. The `HoltWinters` function in `R` can be restricted to carry out SES, with α being found more precisely by using an optimisation routine to minimise $\sum e_t^2$ (Meyer 2008). For more details on fitting see Gardner (1985b).

By carrying out successive substitutions in Equation B.1, we find:

$$\begin{aligned}
\hat{x}_t(1) &= \alpha x_t + (1-\alpha)\hat{x}_{t-1}(1) \\
&= \alpha x_t + (1-\alpha)[\alpha x_{t-1} + (1-\alpha)\hat{x}_{t-2}(1)] \\
&= \alpha x_t + \alpha(1-\alpha)x_{t-1} + (1-\alpha)^2\hat{x}_{t-2}(1) \\
&= \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \alpha(1-\alpha)^3 x_{t-3} + (1-\alpha)^4\hat{x}_{t-4}(1) \\
&= \vdots \\
&= \sum_{j=0}^{\infty} \alpha(1-\alpha)^j x_{t-j}, \qquad \text{(B.3)}
\end{aligned}$$

so that a 'one step forecast ... is equivalent to computing a geometric sum of past observations' (Chatfield 2001). With the weights decaying exponentially, it is easy to see the reason for the method's name (Chatfield, Koehler, Ord, and Synder 2001). This type of decay is intuitively sensible: we expect the forecast $\hat{x}_t(1)$ to be tied most closely to those observations closest to it in time (x_t, x_{t-1}, etc.). As α increases to 1, more weight is placed on recent values and the effect of values far in the past becomes vanishingly small. When $\alpha = 1$, the forecast is the same as the last observation.

Chatfield (2001) notes that SES "is sometimes said to be applicable for series showing no seasonal variation or long-term trend but with a locally constant mean which shows some 'drift' over time". More precisely, SES has also been shown to give the optimal forecasts for certain models where this 'drift' can be described mathematically. For instance, for an ARIMA (0,1,1) model, $(1-B)X_t = Z_t - \theta Z_{t-1}$, the Minimum Mean Square Error (MMSE) forecast is given by:

$$\begin{aligned}
\hat{x}_t(1) &= x_t - \theta z_t \\
&= x_t - \theta\left[x_t - \hat{x}_{t-1}(1)\right] \qquad \text{(B.4)} \\
&= (1-\theta)x_t + \theta\hat{x}_{t-1}(1) \\
&= \alpha x_t + (1-\alpha)\hat{x}_{t-1}(1)
\end{aligned}$$

where $\alpha = 1 - \theta$, with the final form being the same as Equation B.1 (Chatfield 2001; Chatfield 1996; Gardner 1985b). SES is also optimal for forecasting random walk plus noise state-space models, where SES can be seen as a form of Kalman filter (Chatfield 2001; Chatfield, Koehler, Ord, and Synder 2001).

SES, and more generally ES methods, were originally justified by being 'intuitively sensible', as opposed to having any theoretical justification (Chatfield, Koehler, Ord, and Synder 2001). Since then, many authors have found many models for which ES techniques give MMSE predictions, and some have taken this as a justification for ES methods (Gardner 1985b; Chatfield, Koehler, Ord, and Synder 2001). There are of course advantages for having a model that describes the underlying structure of the data for which the ES methods are being applied, such as forming prediction intervals (Chatfield, Koehler, Ord, and Synder 2001). However, ES methods have been shown to work across a variety of models which have contrasting structures and assumptions. Further, there are data-sets for which a model cannot be easily specified, but that ES provides good forecasts for (Cogger 1985). For instance, McKenzie (1985) contrasts the choice between using Box-Jenkins ARIMA and ES methods for forecasting: 'The ARIMA modeller assumes that the stationarity observed in the data will be preserved into the future. ... The ES modeller, on the other hand, may note the stability of the level in the fitting period, but he fears the worst for the future. ... The ES modeller's primary aim in model building has to be robustness.' Given the prediction systems we are going to develop will be for use by non-experts, this quality of robustness is very important. Also, since data consistency is variable, sometimes subject to artificial changes (for example, see Section 4.2.1), the ability of SES to adapt to new levels in the data is very attractive. For more detailed discussions on theoretical justifications of ES, see Chatfield, Koehler, Ord, and Synder (2001), Gardner (1985b), Cogger (1985), McKenzie (1985), Gardner (1985a).

B.1.2 ES with Trend

SES takes no account of a trend being present in the data being forecast (Chatfield, Koehler, Ord, and Synder 2001). Thus, for forecasting such data SES should be extended. The first extension to deal with trend was given by Holt, who extends SES to deal with a linear trend (Holt 2004; Holt, Modigliani, Muth, and Simon 1960). In such a case, the forecast equation to predict m steps into the future becomes (using notation consistent with Meyer (2008)):

$$\hat{x}_t(m) \quad = \quad a_t + mb_t, \tag{B.5}$$

where:

$$a_t = \alpha x_t + (1-\alpha)(a_{t-1} + b_{t-1}), \tag{B.6}$$
$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}, \tag{B.7}$$

or, in their error correction form

$$a_t = a_{t-1} + b_{t-1} + \alpha e_t, \tag{B.8}$$
$$b_t = b_{t-1} + \alpha\beta e_t, \tag{B.9}$$

such that a_t is the local level of the series, b_t is the trend. The smoothing parameters α and β can be found in the same way as the SES parameter and have similar constraints on them (Gardner and McKenzie 1985). Holt's linear trend method can be carried out in `R` by use of the `HoltWinters` function to fit a non-seasonal model (Meyer 2008).

Brown gives another formulation for ES with a linear trend that utilises only one smoothing parameter:

$$\hat{x}_t(m) = a_t + \left(\frac{1-\alpha}{\alpha}\right) b_t + mb_t \tag{B.10}$$
$$a_t = \alpha x_t + (1-\alpha)b_{t-1}, \tag{B.11}$$
$$b_t = \alpha(a_t - a_{t-1}) + (1-\alpha)b_{t-1}. \tag{B.12}$$

This form is sometimes known as 'double exponential smoothing', since one smoothing parameter is used to smooth twice (Chatfield 2001). The main advantage to this method is the reduction of smoothing parameters to one. However, this is perhaps offset by the greater opacity of the method over Holt's formulation. Given contemporary computing power, the concern for this saving in efficiency is rarely a concern, leading to the method being rarely used now (Chatfield 2001). For further details see Chatfield (2001), Gardner (1985b).

Linear trend formulations have the disadvantage that for large values of m (forecasting far in to the future) the forecasted value will tend to plus or minus infinity, which is normally unrealistic. One approach to resolving this problem is to damp the trend so that for large values of m, the trend does not contribute unreasonably to the forecast. This approach is investigated in Gardner

and McKenzie (1985), where Equations B.5, B.6 and B.7 are extended by the addition of an autoregressive-damping (AD) parameter ϕ:

$$\hat{x}_t(m) = a_t + \sum_{i=1}^{m} \phi^i b_t, \tag{B.13}$$

$$a_t = \alpha x_t + (1-\alpha)(a_{t-1} + \phi b_{t-1}), \tag{B.14}$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)\phi^i b_{t-1}, . \tag{B.15}$$

We note the implications of the different values of ϕ from Gardner and McKenzie (1985). If $\phi = 0$ there is no trend and the model reduces to SES. If $0 < \phi < 1$ the trend is damped and approaches the asymptote given by the straight line $a_t + b_t\phi(1-\phi)$. When $\phi = 1$, this new formulation reduces to Holt's linear model as given by Equations B.5, B.6 and B.7. For $\phi > 1$, exponential trend will result which is usually not required. Gardner and McKenzie (1985) provide details of how to calculate the parameters: again, they are essentially chosen by minimising $\sum e_t^2$, but the authors present refinements, for efficiency, to the manner in which they are usually calculated. Damped trends have been shown on average to give very accurate long term forecasts (Gardner 1985b).

Other extensions to the basic SES method exist for dealing with trends that are non-linear (Gardner 1985b). However, these methods are rarely used; particularly not in business and commerce where higher order trends often have little relevance (Gardner 1985b). Since we will primarily be looking at predictions one time unit ahead ($m = 1$), using a linear trend is a reasonable assumption (Gardner and McKenzie 1985). Thus, we do not consider these other methods here but direct the interested reader to Gardner (1985b).

As with SES, a number of the methods considered here are optimal predictors for a number of models. Holt's linear trend model is optimal when the data has the structure of an ARIMA (0,2,2) or linear growth state-space model (Gardner 1985b; Chatfield 2001). The damped linear model is optimal for data that has an ARIMA (1,1,2) structure: the model has an autoregressive term that Holt's method's underlying model does not and is the reason for ϕ being called an 'autoregressive-damping' parameter (Gardner and McKenzie 1985). For this and other equivalences see Gardner and McKenzie (1985).

B.1.3 ES with Trend & Seasonality

Frequently, a forecaster will wish to predict series that exhibit both trend and seasonality. This is often the case in sales where the level of sales is not just affected by general trends, but also by the seasonal period. The Holt-Winters method and general exponential smoothing are two extensions of SES that deal with both trend and seasonality.

The Holt-Winters method (sometimes called Winter's seasonal method) is named in honour of C C Holt and P R Winters, but there does not seem to be any clear idea of who first developed it (Chatfield and Yar 1988). When dealing with seasonality there is a fundamental choice between treating it as additive or multiplicative in nature. We will describe the additive approach, as we deal with predicting log-counts. The counts themselves appear to have a multiplicative structure, so by dealing with the log-counts, the seasonality becomes additive. For the structure of the multiplicative form, see Gardner (1985b). Thus, the Holt-Winters additive form (as presented in Meyer (2008)) for a year of p seasons i.e. periodicity p:

$$\begin{aligned}
\hat{x}_t(m) &= a_t + mb_t + s_{t+1+(m-1) \bmod p}, && \text{(B.16)}\\
a_t &= \alpha(x_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}), && \text{(B.17)}\\
b_t &= \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}, && \text{(B.18)}\\
s_t &= \gamma(x_t - a_t) + (1-\gamma)s_{t-p}, . && \text{(B.19)}
\end{aligned}$$

with error correction forms:

$$\begin{aligned}
a_t &= a_{t-1} + b_{t-1} + \alpha e_t, && \text{(B.20)}\\
b_t &= b_{t-1} + \alpha\beta e_t, && \text{(B.21)}\\
s_t &= s_{t-p} + (1-\alpha)\gamma e_t, && \text{(B.22)}
\end{aligned}$$

where α and β are as previously, γ is the seasonal parameter for updating the seasonal factors s_t which have p factors per period.

Many people consider the Holt-Winters method to be 'robust' and 'easy-to-use' (Chatfield and Yar 1988). Holt-Winter forecasts have been found to give

comparable accuracy of forecasts to non-automatic systems which require much greater expert knowledge (such as Box-Jenkins methodology) (Chatfield 2001; Chatfield and Yar 1988). As Chatfield and Yar (1988) observes, this resonates with Schumacher (1974), who notes that sophisticated systems for short term forecasts 'rarely produce[s] significantly different results from those of a crude technique.' The additive Holt-Winters is an optimal forecaster for a very complex ARIMA model (Abraham and Ledolter 1986; McKenzie 1976).

General Exponential Smoothing (GES) was first developed by (Brown 1963). It is a good deal more complex than the Holt-Winter's method, with its main difference being that seasonality is dealt with by sinusoids (Gardner 1985b).

B.2 Practical Issues of ES

As we have seen, a number of authors commend ES over other techniques for its intuitiveness and robust results (Chatfield and Yar 1988). However, in the area of identification, other methods such as Box and Jenkins ARIMA are stronger in the sense of being more systematic (Gardner 1985a). While subjective, the most appropriate ES technique will usually be found by manually selecting the method. When manually selecting the appropriate ES method, the key decisions will be whether the data has a trend and whether it is seasonal (Gardner 1985b; McKenzie 1985). A number of very other useful guidelines are given in Chatfield and Yar (1988).

We have not considered details of starting values for ES methods. For SES, little difference has been found between using a global mean or the first observation as the starting value for the method (Clark, Cremins, and Schnaars 1987). More generally and applicable to all the methods is the approach of back-casting (Ledolter and Abraham 1984): this approach reverses the order of the data and uses the most recent observation to start a run of the ES method. The prediction given at the end of the reversed data is then used as a starting value for the ES method on the original data. When additive seasonality is required for Holt-Winters, a linear regression can be carried out on data with dummy variables to find levels for the different seasonal periods. For series with few observations, a number of other approaches exist (Gardner 1985b). For longer series, the effect of the initial values will be less noticeable. However, this is less true for the seasonal

factors, as they will only be updated once every year: thus, the weekly factors for the HPS data would only be updated twenty times (Chatfield and Yar 1988).

The `HoltWinters` function within `R` takes quite a naive approach to approach to initial values (Meyer 2008). When it is used to carry out SES, the first observation is used to initialise the level (a_t, in Equation B.17). When forecasting with level and trend, the first observation and the difference between the first two observations are used for initialisation respectively (a_t in Equation B.17 and b_t in Equation B.18). For the full Holt-Winters method with seasonality, a time series decomposition is carried out on the first three periods of data using the function `decompose`. This function separates a times series into seasonal and trend components using moving averages. The seasonal pattern found is used as the starting values for the seasonal factors s_t (Equation B.19). A straight line is fitted to the trend component using simple linear regression; the coefficients in the resulting model are used as starting values for the level and trend.

Appendix C

Mode of the Negative Binomial

In this Appendix we derive a general form for the negative binomial distribution's mode. We use the specification of the negative binomial $f_Y(y;\theta,\mu)$, with mean μ and dispersion parameter θ, as given in Venables and Ripley (1997):

$$f_Y(y;\theta,\mu) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{\theta+y}},$$

where y is a non-negative integer, $\mu > 0$ and $\theta > 0$.

We consider the difference in probability between $y = k+1$ and $y = k$:

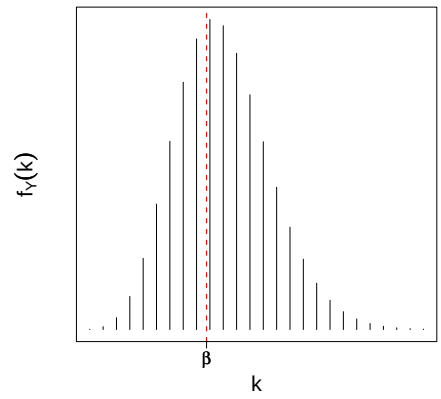
$$f_Y(k+1) - f_Y(k) = \frac{\Gamma(\theta+k+1)}{\Gamma(\theta)(k+1)!}\frac{\mu^{k+1}\theta^\theta}{(\mu+\theta)^{\theta+k+1}} - \frac{\Gamma(\theta+k)}{\Gamma(\theta)k!}\frac{\mu^k\theta^\theta}{(\mu+\theta)^{\theta+k}}$$

Then, with use of the result that $\Gamma(z+1) = z\Gamma(z)$, where $z > 0$:

$$\begin{aligned} f_Y(k+1) - f_Y(k) &= \frac{\Gamma(\theta+k)}{\Gamma(\theta)(k+1)!}\frac{\mu^k\theta^\theta}{(\mu+\theta)^{\theta+k+1}}\left(\mu(\theta+k) - (k+1)(\mu+\theta)\right) \\ &= \underbrace{\frac{\Gamma(\theta+k)}{\Gamma(\theta)(k+1)!}\frac{\mu^k\theta^{\theta+1}}{(\mu+\theta)^{\theta+k+1}}}_{>0}\left(\mu\left(1-\frac{1}{\theta}\right) - 1 - k\right) \end{aligned}$$

Let $\beta = \mu\left(1-\frac{1}{\theta}\right) - 1$. Since the braced term above is larger than 0, we get the following result:

$$f_Y(k+1) - f_Y(k) \propto (\beta - k). \tag{C.1}$$



In general, when $k < \beta$, $f_Y(k+1) - f_Y(k) > 0$ and when $\beta > k$, $f_Y(k+1) - f_Y(k) < 0$, as in the diagram. Thus from result C.1, we know that the negative binomial is uni-modal. When $\theta \geq 1$, the mode is the first integer larger than β; if $\theta < 1$, $f_Y(k+1) - f_Y(k) < 0$ for all values of k, and so the mode is at zero.

Thus, in summary:

$$mode = \begin{cases} 0 & \theta < 1, \\ \lfloor \mu \left(1 - \frac{1}{\theta}\right) \rfloor & \theta \geq 1. \end{cases}$$

References

Abraham, B. and J. Ledolter (1986). Forecast functions implied by autoregressive integrated moving average models and other related forecast procedures. *International Statistical Review 54*, 51–66.

Advances in Disease Surveillance (2007). *Utilisation of Data Derived from a Nurse-led NHS Access and Information Telephone Helpline (NHS24) in Communicable Disease Management in Scotland*, Volume 2 (126). Advances in Disease Surveillance. Available at: http://www.isdsjournal.org/article/viewPDFInterstitial/882/763 (last accessed June 28th 2009).

Agarwal, D., A. Gelfand, and S. Citron-Pousty (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics 9*, 341–355.

Alexander, L. V. and P. D. Jones (2001). Updated precipitation series for the uk and discussion of recent extremes. *Atmospheric Science Letters 1*(2), 142–150. Article and associated data available from http://hadobs.metoffice.com/hadukp/. Last accessed on July 2nd 2009.

Allardice, G. M., E. M. Wright, M. Peterson, and J. M. Miller (2001). A statistical approach to an outbreak of endophthalmitis following cataract surgery at a hospitalin the west of scotland. *Journal of Hospital Infection 49*(1), 23–29.

Association of Medical Microbiologists (1997). The facts about cryptosporidium. http://www.amm.co.uk/files/factsabout/fa_crypto.htm. Last accessed on April 4th 2008.

Awwa Research Foundation (2008). Featured topic

snapshot: Ultraviolet disinfection and treatment. http://www.awwarf.org/research/TopicsAndProjects/topicSnapshot.aspx?topic=uv. Last accessed on April 4th 2008.

Babin, S., S. Magruder, S. Hakre, J. Coberly, and J. S. Lombardo (2007). Understanding the Data: Health Indicators in Disease Surveillance. In J. S. Lombardo and D. L. Buckeridge (Eds.), *Disease Surveillance: A Public Health Informatics Approach*, Chapter 2, pp. 43–90. Wiley.

Baker, M., G. E. Smith, D. Cooper, N. Q. Verlander, F. Chinemana, S. Cotterill, V. Hollyoak, and R. Griffiths (2003). Early warning and NHS Direct: a role in community surveillance? *Journal of Public Health Medicine 25*(4), 362–368.

BBC (2009, June). WHO declares swine flu pandemic. Available at http://news.bbc.co.uk/1/hi/8094655.stm (last accessed November 16th 2009).

BBC News (2005). Health board goes under the knife. Website news story: http://news.bbc.co.uk/1/hi/scotland/4559659.stm. Last accessed June 1st 2009).

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control* (Revised ed.). Time Series Analysis and Digital Processing. Holden-Day.

Brandt, P. T., J. T. Williams, B. O. Fordham, and B. Pollins (2000). Dynamic modeling for persistent event-count time series. *American Journal of Political Science 44*(4), 823–843.

Brillman, J., T. Burr, D. Forslund, E. Joyce, R. Picard, and E. Umland (2005). Modelling emergency department visit patterns for infectious disease complaints; results and application to disease surveillance. *BMC Medical Informatics and Decision Making 5*(1), 4. Available from: http://www.biomedcentral.com/1472-6947/5/4 (last accessed on August 21st 2009).

Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*. New York: McGraw-Hill.

Brown, R. G. (1963). *Smoothing, Forecasting and Prediction*. Englewood Cliffs: Prentice Hall.

Bull, G. M. and J. Morton (1978). Environment, temperature and death rates. *Age and Ageing 7*(4), 210–224.

Burkom, H. (2007). Alerting Algorithms for Biosurveillance. In J. S. Lombardo and D. L. Buckeridge (Eds.), *Disease Surveillance: A Public Health Informatics Approach*, Chapter 4, pp. 143–192. Wiley.

Burkom, H. S., S. P. Murphy, and G. Shmueli (2007). Automated Time Series Forecasting for Biosurveillance. *Advances in Disease Surveillance 4*. Available at: http://thci.org/_documents/temp/Automated Time Series Forecasting for Biosurveillance Abstract 5-3-06.doc (last accessed June 21st 2009).

Caccò, S. M., R. C. Thompson, J. McLauchlin, and H. V. Smith (2005). Unravelling cryptosporidium and giardia epidemiology. *Trends in Parasitology 21*(9), 430–437. http://www.ncbi.nlm.nih.gov/pubmed/16046184?dopt=Abstract.

Cameron, A. C. and P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Centers for Disease Control and Prevention (2007). Division of parastitic diseases – cryptosporidiosis fact sheet. http://www.cdc.gov/NCIDOD/DPD/parasites/cryptosporidiosis/factsht_cryptosporidiosis.htm. Last accessed on April 4th 2008.

Centers for Disease Control and Prevention (2009, April). Swine Influenza A (H1N1) Infection in Two Children – Southern California, March–April 2009. pp. 400–402. Available at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5815a5.htm (last accessed November 16th 2009).

Chappell, C. L., P. C. Okhuysen, C. R. Sterling, C. Wang, W. Jakubowskui, and H. L. DuPont (1999). Infectivity of cryptosporidium parvum in healthy adults with pre-existing anti-c. parvum serum immunoglobulin. *The American Journal of Tropical Medicine and Hygiene 60*, 157–164.

Chartered Institute of Enviromental Health (2007). Previous outbreaks [of *Cryptosporidium*]. http://www.cieh.org/know.aspx?id=416. Last accessed on April 4th 2008.

Chatfield, C. (1978). The Holt-Winters Forecasting Procedure. *Applied Statistics 27*(3), 264–279.

Chatfield, C. (1996). *The Analysis of Time Series: An Introduction* (Fifth ed.). Texts in Statistical Science. Chapman & Hall.

Chatfield, C. (2001). *Time-Series Forecasting*. Chapman & Hall/CRC.

Chatfield, C., A. B. Koehler, J. K. Ord, and R. Synder (2001). A new look at models for exponential smoothing. *The Statistician 50*(2), 147–159.

Chatfield, C. and M. Yar (1988). Holt-winters forecasting: Some practical issues. *The Statistician 37*(2), 129–140.

Chen, C. (1997). Robustness properties of some forecasting methods for seasonal time series: a monte carlo study. *International Journal of Forecasting 13*, 269–280.

Choi, K. and S. B. Thacker (1981a). An Evaluation of Influenza Mortality Surveillance, 1962–1979: I. Time Series Forecasts of Expected Pneumonia and Influenza Deaths. *American Journal of Epidemiology 113*(3), 215–226. Available at http://aje.oxfordjournals.org/cgi/reprint/113/3/215 (last accessed November 23rd 2009).

Choi, K. and S. B. Thacker (1981b). An Evaluation of Influenza Mortality Surveillance, 1962–1979: II. Percentage of Pneumonia and Influenza Deaths as an Indicator of Influenza Activity. *American Journal of Epidemiology 113*(3), 227–235. Available at http://aje.oxfordjournals.org/cgi/reprint/113/3/227 (last accessed November 23rd 2009).

Clark, S. D., S. Cremins, and S. Schnaars (1987). A comparison of starting values for simple exponential smoothing. In *7th International Symposium on Forecasting*, Boston, Massachusetts.

Cogger, K. O. (1985). Comments on 'exponential smoothing: The state of the art': Introduction to the commentaries. *Journal of Forecasting 4*(1), 29–30.

Cooper, D. (2007). Case Study: Use of Tele-health Data for Syndromic Surveillance in England and Wales. In J. S. Lombardo and D. L. Buckeridge (Eds.), *Disease Surveillance: A Public Health Informatics Approach*, Chapter 4, pp. 335–365. Wiley.

Cooper, D. and F. Chinemana (2004). Nhs Direct derived data: an exciting new opportunity or an epidemiological headache? *Journal of Public Health 26*(2), 158–160.

Cooper, D. L., G. Smith, M. Baker, F. Chinemana, N. Verlander, E. Gerard, V. Hollyoak, and R. Griffiths (2004). National Symptom Surveillance Using Calls to a Telephone Health Advice Service – United Kingdom, December 2001–February 2003. *Morbidity and Mortality Weekly Report Supplement 53*, 179–183. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a33.htm (last accessed June 28th 2009).

Cooper, D. L., G. E. Smith, F. Chinemana, C. Joseph, P. Loveridge, P. Sebastionpillai, E. Gerard, and M. Zambon (2008). Linking syndromic surveillance with virological self-sampling. *Epidemiology and Infection 136*, 222–224.

Cooper, D. L., N. Q. Verlander, G. E. Smith, A. Charlett, E. Gerard, L. Willocks, and S. O'Brien (2006). Can syndromic surveillance data detect local outbreaks of communicative disease? A model using a historical cryptosporidiosis outbreak. *Epidemiology and Infection 134*, 13–20.

Dalrymple, D. J. (1987). Sales forecasting practices: Results from a united states survey. *International Journal of Forecasting 3*, 379–391.

Defra (2008). Foot and mouth disease: 2001 outbreak. http://www.defra.gov.uk/animalh/diseases/fmd/2001/index.htm. Last accessed on May 13th 2009.

Defra (2009). Foot and mouth disease: 2007 outbreak. http://www.defra.gov.uk/animalh/diseases/fmd/2007/index.htm. Last accessed on May 13th 2009.

Diggle, P. J. (1989). *Time Series: A Biostatistical Introduction*. Oxford Statistical Science Series. Oxford University Press.

Donaldson, G. C. and W. R. Keating (2002). Excess winter mortality: influenza or cold stress? *British Medical Journal 324*, 89–90.

Doroshenko, A., D. Cooper, G. Smith, E. Gerard, F. Chinemana, N. Verlander, and A. Nicoll (2005). Evaluation of Syndromic Surveillance Based on National Health Service Direct Derived Data — England and Wales. *Morbidity and Mortality Weekly Report Supplement 54*, 117–122. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/su5401a19.htm (last accessed June 28th 2009).

Dowell, S. F. (2001). Seasonal variation in host suscepitbility and cycles of certain infectious diseases. *Emerging Infectious Diseases 7*, 369–374.

Drinking Water Quality Regulator (2003). Drinking water quality regulators report following the alerts in glasgow and edinburgh in august 2002. Technical report, Drinking Water Quality Regulator. Available from: http://www.scotland.gov.uk/Resource/Doc/47049/0014630.pdf. Last accessed on April 4th 2008.

DuPont, H. L., C. L. Chappell, C. R. Sterling, P. C. Okhuysen, J. B. Rose, and W. Jakubowskui (1995). The infectivity of cryptosporidium parvum in healthy volunteers. *North England Journal of Medicine 332*, 855–859.

Effler, P., M. Ching-Lee, A. Bogard, M.-C. Ieong, T. Nekomoto, and D. Jernigan (1999). Statewide System of Electronic Notifiable Disease Reporting From Clinical Reporting From Clinical Laboratories. *Journal of the American Medical Association 282*(19), 1845.

Everitt, B. and G. Dunn (2001). *Applied Multivariate Data Analysis* (Second ed.). Arnold.

Faraway, J. J. (2006). *Extending the Linear Model with R*. Texts in Statistical Science. Chapman & Hall/CRC.

Farrington, C. P., N. J. Andrews, A. D. Beale, and M. A. Catchpole (1996). A statistical algorithm for the early detection of outbreaks of infectious diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 159*(3), 547–563.

Fayer, R., U. Morgan, and S. J. Upton (2000). Epidemiology of cryptosporidium: transmission, detection and identification. *International Journal for*

Parasitology 30, 1305–1322.

Freund, D. A., T. J. Kniesner, and A. T. LoSasso (1999). Dealing with the common econometric problems of count data with excess zeros, endogenous treatment effects and attrition bias. *Economics Letters 62*, 7–12.

Friedman, J. H. (1984). A variable span scatterplot smoother. Technical Report in the Laboratory for Computational Statistics 5, Stanford University.

Gardner, Jr., E. S. (1985a). Author's response to comments on 'exponential smoothing: The state of the art'. *Journal of Forecasting 4*(1), 37–38.

Gardner, Jr., E. S. (1985b). Exponential smoothing: The state of the art. *Journal of Forecasting 4*(1), 1–28.

Gardner, Jr., E. S. and E. McKenzie (1985). Forecasting trends in time series. *Management Science 31*(10), 1237–1246. Damping Trend.

General Register Office for Scotland (2001). 2001 census: Standard area statistics (scotland) [computer file]. Accessed through http://casweb.mimas.ac.uk/. Last viewed 22nd October 2008. Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

General Register Office for Scotland (2009a). General Register for Scotland. Available at http://www.gro-scotland.gov.uk/ (last accessed November 15th 2009).

General Register Office for Scotland (2009b). General Register for Scotland – About Us. Available at http://www.gro-scotland.gov.uk/abotgros (last accessed November 15th 2009).

Glasgow Chamber of Commerce (2009). Scottish Public Holidays. Available at http://www.glasgowchamber.org/page.asp?id=31 (last accessed August 24th 2009).

Goldstein, S. T., D. D. Juranek, O. Ravenholt, A. W. Hightower, D. G. Martin, J. L. Mesnik, S. D. Griffiths, A. J. Bryant, R. R. Reich, and B. L. Herwaldt (1996). Cryptosporidiosis: An outbreak associated with drinking water despite state-of-the-art water treatment. *Annals of Internal Medicine 124*(5), 459–468.

Gross, D. and R. J. Craig (1974). A comparison of maximum likelihood, exponential smoothing and bayes forecasting procedures in inventory modelling. *International Journal of Production Research 12*(5), 607–622. Forecasting Means.

Guerrant, R. L. (1997). Cryptosporidiosis: an emerging, highly infectious threat. *Emerging Infectious Diseases 3*(1), 51–57. Available from: http://www.cdc.gov/ncidod/EID/vol3no1/guerrant.htm.

Gupta, P. L., R. C. Gupta, and R. C. Tripathi (1996). Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis 23*, 207–218.

Hardin, J. W. and J. M. Hilbe (2007). *Generalized Linear Models and Extensions*. Texas: Stata Press.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B 55*(4), 757–796.

Hauenstein, L., R. Wojcik, W. Loschen, R. Ashar, C. Sniegoski, and N. Tabernero (2007). Putting It Together: The Biosurveillance Information System. In J. S. Lombardo and D. L. Buckeridge (Eds.), *Disease Surveillance: A Public Health Informatics Approach*, Chapter 5, pp. 193–261. Wiley.

Health Protection Agency (2008). Cryptosporidium. http://www.hpa.org.uk/infections/topics_az/crypto/menu.htm. Last accessed on April 4th 2008.

Health Protection Agency: Primary Care Surveillance Team (2005). Comparison of NHS24 call data for Scotland against NHS Direct call data for England and Wales. Draft version. Duncan Cooper, duncan.cooper@hpa.org.uk listed as contact.

Health Protection Scotland (1994). Cryptosporidios (Perth, leisure facilities). HPS weekly report 1994/47, Health Protection Scotland (HPS). ISSN: 1357-4493. Details of report found from this search: http://www.hps.scot.nhs.uk/Search/default.aspx?search=cryptosporidiosis%20perth Last accessed on July 29th 2008.

Health Protection Scotland (2000a). Cryptosporidiosis in travelleres returning from callas de majorca. HPS weekly report 2000/30, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2000/0030.pdf. Last accessed on July 29th 2008.

Health Protection Scotland (2000b). Cryptosporidium in gghb area. HPS weekly report 2000/21, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2000/0021.pdf. Last accessed on July 29th 2008.

Health Protection Scotland (2002a). Cryptosporidiosis in west lothian (livingstone, swimming pool). HPS weekly report 2002/46, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2002/0246.pdf. Last accessed on April 4th 2008.

Health Protection Scotland (2002b). Cryptosporidiosis outbreak, tayside. HPS weekly report 2002/32, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2002/0232.pdf. Last accessed on July 29th 2008.

Health Protection Scotland (2002c). Cryptosporidium in glasgow's water supplies. HPS weekly report 2002/31, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2002/0231.pdf. Last accessed on July 29th 2008.

Health Protection Scotland (2002d). *Cryptosporidium* in water supplies. HPS weekly report 2002/43, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2002/0243.pdf. Last accessed on April 4th 2008.

Health Protection Scotland (2003a). Cryptosporidium outbreak in east renfrewshire. HPS weekly report 2003/37, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from:

http://www.documents.hps.scot.nhs.uk/ewr/pdf2003/0337.pdf. Last accessed on July 29th 2008.

Health Protection Scotland (2003b). *Cryptosporidium* in majorca - update. HPS weekly report 2003/31, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2003/0331.pdf. Last accessed on April 4th 2008.

Health Protection Scotland (2004a). About health protection scotland. http://www.hps.scot.nhs.uk/about/index.aspx. Last accessed on April 4th 2008.

Health Protection Scotland (2004b). Declining level of salmonella contamination in eggs. HPS weekly report 2004/12, Health Protection Scotland (HPS). ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2004/0412.pdf. Last accessed on April 4th 2008.

Health Protection Scotland (2004c). Health protection scotland brochure. http://www.documents.hps.scot.nhs.uk/about-hps/hps-brochure.pdf. Last accessed on April 4th 2008.

Health Protection Scotland (2005). *Cryptosporidium* outbreak in tayside. HPS weekly report 2005/16, Health Protection Scotland (HPS). ISSN: 1746-6695. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf2005/0516.pdf. Last accessed on April 4th 2008.

Health Protection Scotland (2008a). About hps weekly report. Technical report, Health Protection Scotland (HPS). Available from: http://www.hps.scot.nhs.uk/ewr/about.aspx. Last accessed on August 5th 2008.

Health Protection Scotland (2008b). Basic search - hps weekly report. Technical report, Health Protection Scotland (HPS). Available from: http://www.hps.scot.nhs.uk/ewr/search.aspx. Last accessed on August 5th 2008.

Health Protection Scotland (2009a, October). Technical Description of Anal-

ysis Used in the Weekly Influenza Situation Report (Including H1N1v). Available at http://www.documents.hps.scot.nhs.uk/respiratory/swine-influenza/technical-annex-2009-10-29.pdf (last accessed November 16th 2009).

Health Protection Scotland (2009b, November). Weekly Influenza Situation Report (Including H1N1v). Available at http://www.documents.hps.scot.nhs.uk/respiratory/swine-influenza/situation-reports/weekly-influenza-sitrep-2009-11-12.pdf (last accessed November 16th 2009).

Henning, K. J. (2004). Overview of Syndromic Surveillance: What is Syndromic surveillance? *Morbidity and Mortality Weekly Report Supplement 53*, 5–11. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm (last accessed June 28th 2009).

Hilbe, J. M. (2007). *Negative Binomial Regression: Modeling Overdispersed Count Data*. Cambridge University Press.

Holt, C. C. (2004, January-March). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting 20*(1), 5–10. Reprint of the 1957 article of the same name. Available at http://www.sciencedirect.com/science/article/B6V92-4BJVV07-3/2/75e73118cfeba36d23df13dd9c445f3e (last accessed on May 26th 2009).

Holt, C. C., F. Modigliani, J. F. Muth, and H. A. Simon (1960). *Planning Production, Inventories, and Work Force*. Englewood Cliffs, New Jersey: Prentice-Hall.

Hunter, P. R. and R. C. Thompson (2005). The zoonotic transmission of giardia and cryptosporidium. *Internation journal for parasitology 35*(11–12), 1181–1190. http://www.ncbi.nlm.nih.gov/pubmed/16159658?dopt=Abstract. Last accessed on April 4th 2008.

Jackman, S. (2008). *`pscl`: Classes and Methods for `R` Developed in the Political Science Computation Laboratory, Standford University*. Department of Political Science, Stanford University, Standford. Available at http://cran.r-project.org/web/packages/pscl/ (last accessed October 24th 2009).

Kavanagh, K., C. Robertson, and J. McMenamim (2007). Exception Reporting Systems for 'Flu Like' Syndromes in Scotland. *Advances in Disease Surveillance 4*. Available at: http://www.isdsjournal.org/article/viewFile/1996/1550 (last accessed June 27th 2009).

Korich, D. G., J. R. Mead, M. S. Madore, N. Sinclair, and C. R. Sterling (1990). Effects of ozone, chlorine dioxide, chlorine, and monchloramine on cryptosporidium parvum oocyst viability. *Applied Environmental Microbiology 56*, 1423–1428.

Laberge, I. and M. W. Giffiths (1996). Prevalence, detection and control of cryptosporidium parvum in food. *International Journal of Food and Microbiology 31*, 1–26.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics 15*(3), 209–225.

Ledolter, J. and B. Abraham (1984). Some comments on the initialisation of exponential smoothing. *Journal of Forecasting 3*, 79–94.

Leung, C. S., M. S. Patel, and C. A. McGilchrist (1999). A distribution-free regional cumulative sum for identifying hyperendemic periods of disease incidence. *Journal of the Royal Statistical Society Series D - The Statistician 48*(2), 215–225.

Lombardo, J. S. (2007). Disease Surveillance, a Public Health Priority. In J. S. Lombardo and D. L. Buckeridge (Eds.), *Disease Surveillance: A Public Health Informatics Approach*, Chapter 1, pp. 1–40. Wiley.

Magruder, S. F. (2003). Evaluation of over-the-counter pharmaceutical sales as possible early warning indicator of human disease. *Johns Hopkins APL Technical Digest 24*(4), 349–354.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hobon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler (1982). The accuracy of extrapolation (time series) methods: results of a forecasting compeition. *Journal of Forecasting 1*, 111–153.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hobon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler (1984). *The Forecasting Accuracy of Major Time Series Methods*. New York: Wiley.

Makridakis, S. and M. Hibon (2000). The m3-competition: results, conclusions and implications. *International Journal of Forecasting 16*, 451–476.

Makridakis, S., M. Hibon, and C. Moser (1979). Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society, Series A (General) 142*(2), 97–145.

Mandel, B. H. (1969). The regression control chart. *Journal of Quality Technology 1*(1), 1–9.

McCabe, G., D. Greenhalgh, G. Gettinby, E. Holmes, and J. Cowden (2003). Prediction of infectious diseases: An exception reporting system. *Journal of Medical Informatics and Technologies 5*, 67–74.

McCabe, G. J. (2004). *A National Exception Reporting System for Infectious Diseases in Scotland*. Ph. D. thesis, University of Strathclyde.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall.

McKenzie, E. (1976). A comparison of some standard seasonal forecasting systems. *The Statistician 25*(1), 3–14.

McKenzie, E. (1985). Commentaries on 'exponential smoothing: The state of the art'. *Journal of Forecasting 4*(1), 32–36.

McKenzie, E. (2003). Discrete variate time series. *Handbook of Statistics 21*, 573–605.

McLauchlin, J., C. Amar, S. Pedraza-Diaz, and G. L. Nichols (2000). Molecular epidemiological analysis of cryptosporidium spp. in the united kingdom: results of genotyping cryptosporidium spp. in 1,075 fecal samples from humans and 105 fecal samples from livestock animals. *Journal of Clinical Microbiology 38*, 3984–3990.

Meinhardt, P. L., D. P. Casemore, and K. B. Miller (1996). Epidemiological aspects of human cryptosporidiosis and the role of waterborne transmission. *Epidemiological Review 18*, 118–136.

Messner, M. J., C. L. Chappell, and P. C. Okhuysen (2001). Risk assessment for cryptosporidium: a hierarchial bayesian analysis of human dose response data. *Water Researcg 35*(16), 3934–3940.

MET Office (2008a). 1971-2001 averages. http://www.metoffice.gov.uk/climate/uk/averages/19712000/index.html. Last accessed on April 23th 2008.

MET Office (2008b). Met office: Weather and climate change. http://www.metoffice.gov.uk/. Last accessed on April 18th 2008.

MET Office (2008c). Met office: Weather data for school and university students. http://www.metoffice.gov.uk/education/data/catalogue.html. Last accessed on April 18th 2008.

MET Office (2008d). Scotland 1971-2001 averages. http://www.metoffice.gov.uk/climate/uk/averages/19712000/areal/scotland.html. Last accessed on April 23th 2008.

Meyer, D. (2008). *R Documentation: Holt-Winters Filtering*. Berkeley. Available online at http://sekhon.berkeley.edu/stats/html/HoltWinters.html, last accessed 08/07/2008, or from within R by typing `?HoltWinters`.

Miron, D. and J. D. K. Kenes (1991). Calves as a source of an outbreak among young children in an agricultural closed community. *Pediatrics Infectious Diseases Journal 10*, 438–441.

Mukerjee, A. (2002). Cryptosporidium Outbreak: Grampian– January – March 2002. Available at http://www.hps.scot.nhs.uk/haiic/ic/presentations.aspx?id=120 (last accessed October 11th 2009).

NHS Direct (2009). What is NHS Direct - Information about NHS Direct's services. http://www.nhsdirect.nhs.uk/article.aspx?name=WhatIsNHSDirect. Last accessed on April 4th 2008.

NHS24 (2006). Working for a Healthier Scotland: Our Strategy, 2006-2009. http://www.nhs24.com/content/mediaassets/doc/Working for a Healthier Scotland 2006 - 2009.pdf. Last accessed June 27th 2009.

NHS24 (2008). About Us - NHS24.

http://www.nhs24.com/content/default.asp?page=home_About%20Us. Last accessed on April 4th 2008.

NHS24 (2009). NHS24 2009/10 – 2011/2012: Draft Strategic Framework – "Delivering and Moving Forward": Discussion Document. http://www.nhs24.com/content/default.asp?page=s21_47. Last accessed June 27th 2009.

Nime, F. A., J. D. Burek, D. L. Page, M. A. Holsher, and J. H. Yardley (1976). Acute enterocolitis in a human being infectes with protozoan cryptosporidium. *Gastroenterology 70*, 592–598.

O'Cathain, A., J. Nicholl, F. Simpson, S. Walters, A. McDonnell, and J. Munro (2004). Do different types of nurses give different triage decisions in NHS Direct? A mixed methods study. *Journal of Health Services Research & Policy 9*(4), 226–233.

Okhuysen, P. C., C. L. Chappell, J. H. Crabb, C. R. Sterling, and D. HL (1999). Virulence of three distinct cryptosporidium parvum isolates for healthy adults. *Journal of Infectious Diseases 180*, 1275–1281.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika 41*, 10–115.

Parliament (1988). Public health (notification of infectious diseases) (scotland) regulations 1988. Published by the Queen's Printer of Acts of Parliament and is available from The Stationery Office Limited, with ISBN 0110875508. Available online from http://www.opsi.gov.uk/si/si1988/Uksi_19881550_en_1.htm. Last accessed on April 4th 2008.

Peng, M. M., L. Xiao, A. R. Freeman, M. J. Arrowood, A. A. Escalante, A. C. Weltman, C. S. L. Ong, W. R. MacKenzie, A. A. Lal, and C. B. Beard (1997). Genetic polymorphism among cryptosporidium parvum isolates: evidence of two distinct human transmission cycles. *Emerging Infectious Diseases 3*, 567–573.

Perron, P. (1988). Trends and random walks in macroeconomic time series : Further evidence from a new approach. *Journal of Economic Dynamics and Control 12*(2-3), 297–332. Available at

http://ideas.repec.org/a/eee/dyncon/v12y1988i2-3p297-332.html (last accessed on May 26th 2009).

Pollock, K. G. J. (2009). Re: Dates for collection. Email from Kevin Pollock on May 18th 9:39am including electronic version of requested figure.

Pollock, K. G. J., H. E. Ternent, D. J. Mellor, H. V. Smith, C. N. Ramsay, and G. T. Innocent (2009). Spatial and temporal epidemiology of sporadic human cryptosporidiosis in scotland. *Zoonoses and Public Health*, Epublish ahead of print.

Pollock, K. G. J., D. Young, H. V. Smith, and C. N. Ramsay (2008). Cryptosporidiosis and filtration of water from loch lomond, scotland. *Journal of Emerging Infectious Diseases*. Available from: http://www.cdc.gov/EID/content/14/1/115.htm. Last accessed on April 4th 2008.

Raubertas, R. F. (1989). An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Statistics in Medicine 8*(3), 267–271.

Reichert, T. A., L. Simonsen, A. Sharma, S. A. Pardo, D. S. Fedson, and M. A. Miller (2004). Influenza and the Winter Increase in Mortality in the United states, 1959–1999. *American Journal of Epedimiology 160*(5), 492–502.

Reis, B. Y. and K. D. Mandl (2003). Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making 3*(2).

Ridout, M., C. G. B. Demétrio, and J. Hinde (1998). Models for count data with many zeros. In *International Biometric Conference*.

Robertson, B., M. I. Sinclair, A. B. Forbes, M. Veitch, M. Kirk, D. Cunliffe, J. Willis, and C. K. Fairley (2002). Case-control studies of sporadic cryptosporidiosis in melbourne and adelaide, australia. *Epidemiology and Infection 128*, 419–431.

Robertson, C. (2006). Protecting the leaders – syndromic surveillance for the g8 summit in Scotland. *Significance 3*(2), 69–72.

Robertson, C., K. Kavanagh, E. McKenzie, and A. Wagner (2008). Syndromic Surveillance at Health Protection Scotland. Presentation at a conference on 'Statistics for Public Health Surveillance', organised by 'The

Open University Statistics Group' on May 21st 2008. Conference details available at http://statistics.open.ac.uk/PHsurv/. Presentation available at: http://statistics.open.ac.uk/PHsurv/Robertson.pdf. Both last accessed June 27th 2009.

Robertson, C., K. Kavanaugh, A. Wagner, J. McMenamin, H. Murdoch, A. Reynolds, E. Shakir, and M. Donaghy (2009). H1n1 in scotland: Epidemiology and surveillance. Presentation at the Royal Statistical Society Annual Conference, 2009.

Robertson, L. J. (2009). Giardia and Cryptosporidium infections in sheep and goats: a review of the potential for transmission to humans via environmental contamination. *Epidemiology and Infection 137*, 913–921.

Rochelle, P. A., D. Fallar, M. M. Marshall, B. A. Montelone, S. J. Upton, and K. Woods (2004). Irreversible uv inactivation of cryptosporidium spp. despite the presence of uv repair genes. *The Journal of Eukaryotic Microbiology 51*(5), 553–562. http://www.ncbi.nlm.nih.gov/pubmed/15537090?dopt=Citation. Last accessed on April 4th 2008.

Ryan, T. P. (1989). *Statistical Methods of Quality Improvement*. Wiley.

Santín, M., J. M. Trout, and R. Fayer (2007). Prevalence and molecular characterization of cryptosporidium and giardia species and genotypes in sheep in maryland. *Veterinary Parasitology 146*(1-2), 17–24.

Sayers, G. M., M. C. Dillion, and E. Connolly (1996). Cryptosporidiosis in children who visited an open farm. *Commuicable Disease Report Review 6*, 140–144.

Schumacher, E. F. (1974). *Small is Beautiful*. London: Sphere Books.

Scottish Centre for Infection and Enviromental Health (1998). Seasonal increase in *Cryptosporidium* cases. SCIEH weekly report 98/41, Scottish Centre for Infection and Enviromental Health. ISSN: 1357-4493. Available from: http://www.documents.hps.scot.nhs.uk/ewr/pdf1998/9841.pdf. Last accessed on April 4th 2008.

Scottish Neighbourhood Statistics (2009). Scottish Neighbourhood Statistics, Information about Scotland's Areas. Available at http://www.sns.gov.uk

(last accessed October 21st 2009).

Scottish Parliament (2006). Executive Note: The National Health Service (Constitution of Health Boards (Scotland) Amendment Order 2006 SSI/2006/32. The National Health Service (Variation of the Areas of Greater Glasgow and Highland Health Boards) (Scotland) Order 2006 SSI/2006/23. Published by the Office of Public Sector Information. Available from http://www.opsi.gov.uk/legislation/scotland/sen2006/ssien_20060032_en.pdf (last accessed June 1st 2009).

Scottish Parliament Information Centre (SPICe) (2002). Contamination of glasgow's water with cryptosporidium. Briefing Note 02/100, Scottish Parliament Information Centre (SPICe). Available from: http://www.scottish.parliament.uk/business/research/pdf_res_brief/sb02-100.pdf. Last accessed on April 4th 2008.

Serfling, R. E. (1963). Methods for Current Statistical Analysis of Excess Pneumonia-Influenza Deaths. *Public Health Reports 78*(6), 494–506. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1915276/ (last accessed November 23rd 2009).

Shaddick, G. and J. Wakefield (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of Applied Statistics 51*(3), 351–372.

Simon, L. J. (1962). Fitting Negative Binomial Distributions by the Method of Maximum Likelihood. In *Proceedings of the Casualty Actuarial Society*, Volume 49, pp. 1–8. Casualty Actuarial Society. Available at http://www.casact.org/pubs/proceed/proceed62/62001.pdf (last accessed August 27th 2009).

Simonsen, L., M. J. Clarke, G. D. Williamson, D. F. Stroup, N. H. Arden, and L. B. Schonberger (1997). The Impact of Influenza Epidemics on Mortality Index. *American Journal of Public Health 87*(12), 1944–1950.

Smerdon, W. J., T. Nichols, R. M. Chalmers, H. Heine, and M. Reacher (2003). Foot and mouth disease in livestock and reduced cryptosporidiosis in humans, england and wales. *Emerging infectious Diseases 9*, 22–28.

Smith, G. E., D. L. Cooper, P. Loveridge, F. Chinemana, E. Gerard, and N. Verlander (2006). A national syndromic surveillance system for England

and Wales using calls to a telephone helpline. *Eurosurveillance 11*, 220–224.

Smith, G. J. D., D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, , S. K. M. ad Chung Lam Cheung, J. Raghwani, S. Bhatt, J. S. M. Peiris, Y. Guan, and A. Rambaut (2009, June). Origins and evolutionary genomics of the 2009 swine origin H1N1 influenza A epidemic. *Nature 459*, 1122–1125. Available at http://www.nature.com/nature/journal/v459/n7250/full/nature08182.html#B1 (last accessed November 16th 2009).

Smith, H. V., R. A. B. Nichols, M. Mallon, A. MacLeod, A. Tait, L. M. Reilly, W. J. amd Browning, D. Gray, S. W. J. Reid, and J. M. Wastling (2005). Natural Cryptosporidium hominis infections in Scottish cattle. *The Veterinary Record 156*, 1239–1255.

Smith, H. V., L. J. Robertson, and J. E. Ongerth (1995). Cryptosporidiosis and giardiasis: the effect of waterborne transmission. *Journal of Water Supply: Research and Technology-Aqua 44*, 258–274.

Smith-Palmer, A. and J. Cowden (2003). Gastro-intestinal infections (rotavirus; norovirus; adenovirus; astrovirus; calicivirus; hepatitis a; cryptosporidium; giardia). Health Protection Scotland Surveillance Report 37/05, Health Protection Scotland Surveillance Report. Available in: http://www.documents.hps.scot.nhs.uk/ewr/pdf2003/0305.pdf. Last accessed on April 6th 2008.

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn (2005). *WinBugs User Manual Version 2.10*. Cambridge: MRC Biostatistics.

Strachan, N. J. C., I. D. Ogden, A. Smith-Palmer, and K. Jones (2003). Foot and mouth epidemic reduces cases of human cryptospodiosis in scotland. *The Journal of Infectious Diseases 188*, 783–786.

Thompson, W. W., D. K. Shay, E. Weintraub, L. Brammer, N. Cox, L. J. Anderson, and K. Fukuda (2003). Mortality Associated With Influenza and Respiratory Syncytical Virus in the United States. *Journal of the America Medical Association 289*(2), 179–186.

UK Drinking Water Inspectorate (2001). Consumer information: *Cryptosporidium* and cryptosporidiosis.

http://www.dwi.gov.uk/consumer/consumer/crypto.htm. Last accessed on April 4th 2008.

Ungar, B. L. P. (1990). Cryptosporidiosis in humans (homo sapiens). In J. P. Dubey, C. A. Speer, and R. Fayer (Eds.), *Cryptosporidiosis of man and animals*. Boca Raton, FL: CRC Press.

Venables, W. N. and B. D. Ripley (1997). *Modern Applied Statistics with S-PLUS* (Second ed.). Statistics and Computing. Springer.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). Statistics and Computing. Springer.

Wagner, A., E. McKenzie, and C. Robertson (2009). Regional Disease Detection in Scotland. In *The Royal Statistical Society's Young Statisticians' Meeting*.

Wald, A. (1945). Sequential tests of statistical hypothesis. *The Annals of Mathematical Statistics 16*(2), 117–186. Available at http://projecteuclid.org/euclid.aoms/1177731118 (last accessed September 28th 2009).

Washington Post (2006). September 11, 2001. Accessed from http://www.washingtonpost.com/wp-dyn/content/linkset/2006/03/30/LI2006033000769.html. Last accessed June 28th 2009).

Welsh, A. H., R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling 88*, 297–308.

Williamson, S. (2006). NHS24: Review of keywords with callreason field for Health Protection Scotland daily report. Internal NHS24 document, compiled by 'Decision Support' at NHS24, under the direction of Stella Williamson.

Wilson, L. (2006). Evaluation of NHS24 data for influenza surveillance. Report provided by Professor Robertson.

Winklhofer, H., A. Diamantopoulos, and S. F. Witt (1996). Forecasting practice: a review of the empirical literature and an agenda for future research. *Internation Journal of Forecasting 12*, 193–221.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science 6*(3), 324–342.

Wojcik, R., L. Hauenstein, C. Sniegoski, and R. Holtry (2007). Obtaining the Data. In J. S. Lombardo and D. L. Buckeridge (Eds.), *Disease Surveillance: A Public Health Informatics Approach*, Chapter 3, pp. 91–142. Wiley.

Wood, S. N. (2006a). Defining smooths in gam formulae (R documentation). Available at http://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/s.html (last accessed November 30th 2009).

Wood, S. N. (2006b). *Generalized Additive Models: An Introdution with R*. Texts in Statistical Science. Chapman & Hall/CRC.

World Health Organization (2009a, May). Assessing the severity of an influenza pandemic. Available at http://www.who.int/csr/disease/swineflu/assess/disease_swineflu_assess_20090511/en/index.html (last accessed November 16th 2009).

World Health Organization (2009b, July). Transcript of virtual press conference with Dr. Keiji Fukuda, Assistant Director-General as Interim for Health Security and Environment, World Health Organization. Available at http://www.who.int/mediacentre/pandemic_h1n1_presstranscript_2009_07_13.pdf (last accessed November 16th 2009).

World Health Organization (2009c). Vaccines for pandemic influenza A (H1N1). Available at http://www.who.int/csr/disease/swineflu/frequently_asked_questions/vaccine_preparedness/en/index.html (last accessed November 16th 2009).

Xiao, L. and R. Fayer (2008). Molecular characterisation of species and genotypes of *cryptosporidium* and *giardia* and assessment of zoonotic transmission. *International Journal of Parasitology 38*(11), 1239.

Xie, M., B. He, and T. N. Goh (2001). Zero-inflated poisson model in statistical process control. *Computational Statistics & Data Analysis 38*, 191–201.

Yang, G., J. Hu, K. Q. Rao, J. Ma, C. Rao, and A. D. Lopez (2005). Mortality registration and surveillance in China: History, current situation and challenges. *Population Health Metrics 3*(3). Available at

http://www.pophealthmetrics.com/content/3/1/3 (last accessed November 23rd 2009).

Yee, T. W. (2007). *VGAM: Vector Generalized Linear and Additive Models*. Available at http://www.stat.auckland.ac.nz/ yee/VGAM/ (last accessed October 24th 2009).

Zeileis, A., C. Kleiber, and S. Jackman (2008). Regression Models for Count Data in R. *Journal of Statistical Software 27*(8).

Zucs, P., U. Buchholz, W. Haas, and H. Uphoff (2005). Influenza associated excess mortality in Germany, 1985-2001. *Emerging Themes in Epidemiology 2*(6). Available at http://www.ete-online.com/content/pdf/1742-7622-2-6.pdf (last accessed November 23rd 2009).

Zuur, A., E. Ieno, N. Walker, A. Saveliev, and G. Smith (2009). *Mixed Effects Models and Extensions in Ecology with R*. Statistics for Biology and Health. Springer.