Statistical Analysis of the Incidence and Mortality of African Horse Sickness in South Africa

A thesis presented to

The University of KwaZulu Natal

in fulfilment of the requirement for the degree

of

Master of Science in Statistics

by

Rebecca Burne

School of Statistics and Actuarial Science

UNIVERSITY OF KWAZULU-NATAL

May 2011

Abstract

African Horse Sickness (AHS) is a viral disease of equids. It is transmitted between animals by insect vectors, predominantly by the midge *Culicoides imicola* Kieffer (Diptera: Ceratopogodinae) although other potential vectors have been identified. It is a World Organization for Animal Health (OIE) listed disease, and as such cases are reportable to the State Veterinarian. The disease is endemic to southern Africa and each year causes large numbers of mortalities - especially in the summer months when conditions are favourable for the propagation of *C. imicola*. Outbreaks of the disease are affected by many factors, including rainfall, temperature and vaccination coverage of the national herd. The outbreaks have many direct and indirect effects on the equine and human population. Subsistence farmers and those in previously disadvantaged communities depend on their horses for work and transport. In these communities there is little education on AHS and vaccination, and these deaths are rarely reported. The competition and horse-racing fraternities are also hugely affected, as there are strict export regulations in place to avoid potential spread to AHS-free countries, and these industries create huge revenue based on the movement and performance of animals.

Outbreaks of AHS have, however, been known to occur in non-endemic regions such as Europe and the Middle East. Due to the severity of the disease, particularly in serologically naïve populations, it is important to form and understand models of the disease so that future severe outbreaks can be predicted and controlled. It will also be useful to understand the factors affecting disease on individual animal and population levels, so that disease and mortality can potentially be reduced.

Data from the African Horse Sickness Trust and South African Weather Service were used to develop a Generalized Linear Model (GLM) with a Poisson distribution relating incidence of the disease in South Africa to rainfall and temperature variables. A GLM utilizing the Binomial distribution was used to model the individual probability of mortality for individuals in KZN, given various explanatory variables. Further, Generalized Estimating Equations (GEEs) and Generalized Linear Mixed Models (GLLMs) were used to control for heterogeneity by place in the model for mortality. These models are a useful introduction to epidemiological models for early warning systems for African horse sickness in this country. This investigation platforms further work on interactions between factors in the models, and necessitates improvements in data quality and integrity from the equine owners to improve predictive capacity of the models.

Declarations

I, Rebecca Burne, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   (a) Their words have been re-written but the general information attributed to them has been referenced.

   (b) Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifcally acknowledged, and the source being detailed in the thesis and in the References sections.

Signed:

____________________ ________________

Rebecca Burne Date
Student number 206508038

____________________ ________________

Prof. Henry Mwambi Date

____________________ ________________

Ms Marion Young Date

Acknowledgements

I would like to thank, firstly, my supervisors Prof. Henry Mwambi and Ms Marion Young. Without your enthusiasm and guidance this project would never have been possible. Particularly your willingness to pick up something unusual to work on, and agreeing to a different sort of collaboration, has made this project enjoyable and hopefully successful. Working with you both was a true pleasure, and knowing I had your support and help motivated me when I was feeling overwhelmed by the quantity of work. I cannot thank you both enough for your tireless support and enthusiasm.

Without the financial and academic support of SACEMA I certainly would not have been able to complete this project. Their words of support and guidance on the research days made me aspire to make this research the best it could possibly be. It should also be mentioned that without their input by sending me to holiday courses during my undergraduate years I may not have ended up in Biostatistics at all! I truly am indebted to everyone at SACEMA who has played a role in this.

Without data, this project would have been over before it began. For their generous help, advice and access to their data I am indebted to the African Horse Sickness Trust, and in particular Douglas Welsh. I hope that this piece of research will be able to assist their work in some small way. The South African Weather Service also deserve many thanks for allowing me access to their data.

I certainly have to thank my family for their love and support during this project. Thank you for putting up with me during this time!

Lastly, I wish to thank all the horses which have shaped my life, and to dedicate this to all horses who have lost their battle to African Horse Sickness. In particular my very first horse, Ilizwe Fantastique, who succumbed to this dreadful disease. If any of the work I have done can help further the understanding of African Horse Sickness it has truly been worthwhile.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 African Horse Sickness

African horse sickness (AHS) is an infectious, non-contagious virus which affects members of the Equidae family. It affects horses primarily, with donkeys and mules having a lower mortality rate, and zebra are thought to be a reservoir host although not being affected themselves. It is enzootic to sub-Saharan Africa, but has had recorded outbreaks in various countries in North Africa, Europe and the Middle East (Mellor, 2004). It is classified as an Office International des Epizooties (OIE: the World Organization for Animal Health) listed disease, and as such is notifiable to the OIE.

African Horse Sickness is caused by the African Horse Sickness virus (AHSV). It is from the genus Orbivirus and family Reoviridae. There are nine known serotypes of the disease. Serotypes 1 to 8 are known to be highly pathogenic in horses, causing mortality of around 95%, while serotype 9 is slightly less fatal with mortality at approximately 70% (Coetzer and Erasmus, 1994). There is cross-relatedness between some AHSV serotypes of the virus, namely 1 and 2; 3 and 7; 5 and 8; and 6 and 9 (Mellor and Hamblin, 2004). There exist different presentations of the disease. The cardiac, or sub-acute, form is characterized with oedema of the head and neck. It has a mortality rate of around 50% (Mellor and Hamblin, 2004). The pulmonary, or peracute, form has a higher mortality which can exceed 95%. It presents with a high fever (39-41°C) and depression, followed by severe respiratory distress (Mellor and Hamblin, 2004). Often the animal will die before showing any clinical signs of the disease. The third form of the disease is the mixed form, so called because it presents as a mixture between the cardiac and pulmonary forms of the disease. Its mortality rate is approximately 70%. The most mild form of the disease is the AHS fever, which can occur in horses partially immune to the serotype with which they are challenged, or in donkeys and zebra. It presents with a temperature of between 39-40°C which lasts for up to six days (Coetzer and Erasmus, 1994). Horses will almost always recover from this presentation of the disease, and it is very often not diagnosed. This form of the disease is usually the only form which will affect donkeys and zebras, which have a higher resistance to the clinical disease (Coetzer and Erasmus, 1994). It is unknown at this stage whether there is a relationship between serotype and presentation of the disease (Young, 2011 *Personal Communication*).

1.2 History

AHS was first recorded in an outbreak in Yemen in 1327, although it is believed to have originated in Africa (Mellor and Hamblin, 2004). The first observation of the disease in Africa was made by a monk by the name of Father Monclaro in 1569. It was recognized in South Africa only after the import of horses to the region in 1657, and the first outbreak of major proportions was seen in 1719 (Mellor and Hamblin, 2004). Since then, many major and minor outbreaks have occurred. However, as a result of the declining horse and zebra populations over the past 100 years, and the invention of vaccinations for the disease, the severity of the outbreaks has declined.

1.3 *Culicoides* midges and African Horse Sickness

Culicoides are a genus of biting midges. They belong to the diptera Ceratopogodinae. They are small in size, ranging from 1-3mm in length. They are crepuscular - meaning that their activity is mainly around sunset and sunrise (Mellor, 2000). While both females and males of the genus drink nectar, the females require blood-meals in order for their eggs to develop, making them vectors for several diseases including Oropouche virus, African horse sickness virus, bluetongue virus, equine encephalosis virus (EEV), Akabane virus (AKAV), and epizootic hemorrhagic disease virus among many others. African horse sickness and Bluetongue viruses are OIE Listed Diseases (previously List A). (Mellor *et al.*, 2000).

Species of *Culicoides* have been discovered in almost all parts of the world, excluding only "the extreme polar regions, New Zealand, Patagonia, and the Hawaiian islands" (Mellor *et al.*, 2000). Several *Culicoides* species have been found to be competent vectors for AHSV. Wild-caught *Culicoides* species were first found to be infected with the virus by Du Toit in 1934. Wetzel (1970) then showed that *Culicoides* species were capable of transmitting the virus between infected and susceptible horses. Since then research by many contributors has shown *Culicoides imicola* to be the major vector of AHSV (Mellor and Hamblin, 2004), although evidence has been found of other species including *C. bolitinos* (Meiswinkel and Paweska, 2002; Venter *et al.*, 2000) being competent vectors of the disease. In particular in the study performed by Meiswinkel and Paweska in 2002, in regions in the eastern Free State where there were fatalities from AHSV, the virus was isolated only from *C. bolitinos* and not from any of the other *Culicoides* species captured. *C. bolitinos* was also the most abundant of the species in these regions.

The North American *C. variipennis sonorensis* has also been shown to be efficient at transmitting the virus in laboratory tests (Boorman *et al.*, 1975 ; Wellby *et al.*, 1996). The widespread worldwide distribution of *Culicoides* spp., along with the uncertainty of which subspecies may be successful vectors for the disease, indicates why AHS should be treated as a global concern.

1.4 Disease Dynamics - as illustrated by deterministic models

The disease dynamics of AHS can be illustrated simply by the deterministic models of Lord *et al.,* (1996a, 1996b, 1997). The basic model from these authors is set up as shown in Figure 1.1 and explained below.
The hosts are equids, and the disease follows a simple SIR model. The animal moves, once bitten and infected by an infective midge, from the Susceptible (x) to the Infected (y) class. From there, the animal will either die or move to the Immune (z) class. However this is a very simplistic model, which would work for a single serotype of AHS but not for the full nine strains, as a horse recovering from a certain strain gains immunity to that serotype only, but is susceptible to all other strains. Thus a more elaborate model would be one that allows possible re-infection by a different strain. If the model is run for only a short time it is not necessary to investigate natural birth and death rates of the host as its lifetime is sufficiently long. Mortality is only considered for the Infected class.
The vector's lifespan is sufficiently short that natural birth and death rates must be included into the model. Transovarian infection is presumed not to occur in *Culicoides* species (Jones and Foster, 1971 as cited in Lord *et al.*, 1996), therefore the recruitment rate enters the Susceptible (S) vector class. The vector, after biting and acquiring the virus from an Infected horse, moves then onto the Latent (L) class, where it stays for a brief period after they have bitten an infected horse but the virus has not yet replicated up to a sufficient titre for infecting further horses. The latent period was predicted by Du Toit (1994) to be 10 days, and by E.M. Nevill (as cited from personal correspondence in Braverman, 1985) to be between 7 and 11 days. After this they move into the Infective (V) class. The vectors are presumed, once infected, to stay infective for the length of their lifetime (in other words they do not move back to the susceptible class). The disease does not affect the midge, therefore natural mortality, μ, does not change between the 3 classes. There is no justification to add excess mortality due to infection.

The differential equations are then set up logically. Explanations of the parameters used are given in Table 1.1.
The rate at which hosts move from susceptible class to infective is given by the biting rate of vectors on the hosts, a, the proportion of hosts which become infected after being

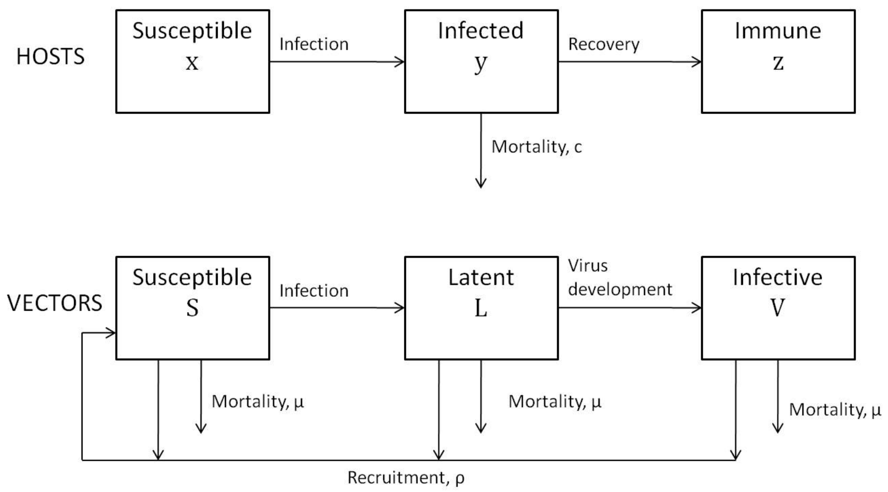

Figure 1.1: Compartmental model as given in papers by Lord *et al.*, ([27] [28] [29]). Dynamics for the vectors (*Culicoides* midges) and hosts (equids) are given. This is a simplistic model that considers only one serotype of the disease.

bitten by an infective vector, b, the ratio of vectors to hosts, (N/H), the proportion of susceptible hosts x and the proportion of infective vectors v. Thus the equation that governs the proportion of susceptible hosts is

$$dx/dt = -ab\left(\frac{N}{H}\right)xv. \tag{1.1}$$

The rate at which the infected class changes is given by the rate at which susceptibles enter, minus the mortality rate of infectives, cy, minus the recovery rate, ry. Thus

$$dy/dt = ab\left(\frac{N}{H}\right)xv - ry - cy. \tag{1.2}$$

The rate at which hosts enter the immune class is given by the recovery rate of infectives, ry. Thus

$$dz/dt = ry. \tag{1.3}$$

Overall, since there is no natural birth or death considered in the model, the total number of horses decreases by the mortality rate of infecteds. Thus, overall

$$dH/dt = -cY. \tag{1.4}$$

The susceptible class of vectors is increased by recruitment from all three classes, as transovarian infection does not occur. The recruitment rate is the rate of new female midges entering the adult population, as males do not play a role in transmission of the diseaseas they do not feed on blood (Mellor, Boorman and Baylis, 2000). This is because bloodmeals are required for the females to produce eggs. The decrease is given by the rate at which susceptible vectors move to the latent class, and the natural mortality

Table 1.1: Description of symbols used in deterministic model equations 1.1 to 1.7 as used in the compartmental models of Lord *et al.* 1996a, 1996b, 1997.

Symbol	Description
H	Total number of AHS hosts
x	Proportion of hosts susceptible
y	Proportion of hosts infected
z	Proportion of hosts immune due to recovery
Y	Number of hosts infected $= y \times H$
S	Number of susceptible vectors
L	Number of latent vectors
V	Number of infective vectors (proportion$= v = V/N$)
N	Total number of vectors $(S + L + V)$
ρ	Daily rate of female midges entering the adult population (not considering larval stages)
α	Interval between bloodmeals on an AHS host
a	Biting rate of vectors on AHS hosts $(1/\alpha)$
β	Proportion of vectors which become infected after biting an infective host
b	Proportion of hosts which become infected after a bite by an infective vector
μ	Daily mortality rate of vectors
γ	Virus development rate
r	Recovery rate for hosts
c	Mortality rate for infected hosts

rate. The rate at which they are infected is given by the interval between bloodmeals on an AHS host, α, the proportion of vectors which become infected after a bloodmeal on an infected host, β, multiplied by the number of susceptible vectors times the proportion of infected hosts. Thus

$$dS/dt = \rho(S + L + V) - \alpha\beta Sy - \mu S. \tag{1.5}$$

The rate at which the Latent vector class moves onto the infective class is then governed by the virus development rate, γ. Mortality also depletes this class. Thus

$$dL/dt = a\beta Sy - \gamma L - \mu L. \tag{1.6}$$

Once a vector is infected, it is presumed to stay infected for the remainder of its life. The only depletion of this class is therefore by natural mortality.

$$dV/dt = \gamma L - \mu V \tag{1.7}$$

It is important to note that ρ, α, μ and γ are all temperature dependant and also dependant on the subtype of *Culicoides*.

An important parameter of such a deterministic disease model is called R_0, the basic reproduction number, which is defined as the average number of secondary cases which will be caused by a single primary case of the disease in a wholly susceptible population. In their 1996 paper, Lord *et al.* found R_0 for the above AHS disease model. This was derived by a heuristic method as follows. The duration of a host's infectivity is expected to be $1/(r+c)$. During this time, they will get an average of $a(N/H)$ bites by susceptible midges per day. A proportion β of these bites will result in an infected midge. Thus there will be $a(N/H)\beta/(r+c)$ midges moving into the Latent class per infected host. A proportion $\gamma/(\gamma+\mu)$ of these latently infected midges are expected to move on to the Infectious class. These will survive for on average $1/\mu$ days, and bite at a rate a. Finally, b of these bites will result in an infected host. This leads to the following equation for R_0.

$$R_0 = \frac{a^2 b\beta}{(r+c)\mu}\left(\frac{N}{H}\right)\left(\frac{\gamma}{\gamma+\mu}\right) \tag{1.8}$$

Control measures can target some key parameters in this equation which R_0 is sensitive to. These are:

a: The biting rate of vectors on AHS hosts,

(N/H): The ratio of vectors to AHS hosts.

Reducing either or both of these parameters will directly reduce R_0. Some of the ways in which these can be controlled are shown in Section 1.5.

1.4.1 Vector Seasonality

The daily female midge emergence rate exhibits a strong seasonal dependence or forcing as is the case for many disease vectors. A method used to describe the seasonal fluctuation in ρ was to equate it to a sinusoidal function. In their work, Lord *et al.* (1997) used the equation $\rho = \mu(1+\delta\cos(\theta t))$ to capture seasonal dependence where δ described the amplitude of the function, and θ the scaling factor for seasonal length. The seasonal length is considered to be 1 year, and therefore $\theta = 2\pi/365 = 0.0172$. The vector population varying with time is found to be given by $N(t) = N_0 \exp^{\mu\delta\cos(\theta t)}$. The graph for $N(t)$ is shown in Figure 1.2. Under these conditions, R_0 cannot be directly estimated. In the current work seasonal dependence will also be integrated into the statistical disease incidence model by means of a term carefully defined to capture this effect.

1.4.2 Vaccination

The effect of vaccination in the Lord *et al.* 1997 model was studied by moving a varying fraction of susceptible hosts into the Immune class. This strategy was used to investigate whether protecting horses only, or both horses and donkeys would have an effect on the number of outbreaks which occurred. The same approach was used to find out whether vaccination after virus introduction would be able to prevent epidemics. Because the

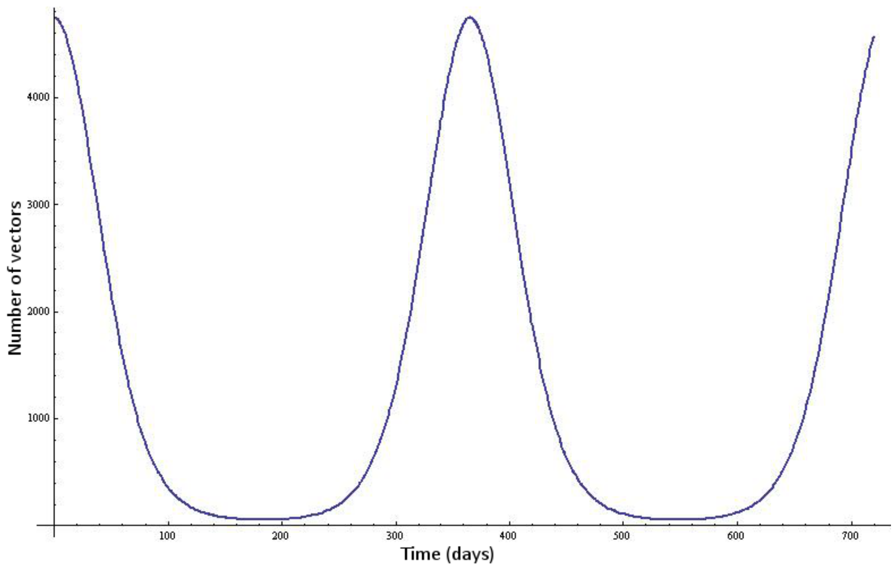

Figure 1.2: Vector population over time as given by the equation $N(t) = N_0 \exp^{\mu\delta\cos(\theta t)}$ in Lord *et al.* (1997) where $\mu = 0.25$, $\theta = 0.0172$, $\delta = 9$ and $N_0 = 500$.

model was for the outbreak in Spain, a serologically naïve population was considered (meaning that no previous immunity or vaccination exists in the population) and also no reservoir hosts such as zebra. It was found that vaccinating donkeys as well as horses was most effective, and that vaccination after the introduction of the virus into the population was not.

1.5 Methods for Prevention and Control

There are various methods employed to prevent AHS in individual animals, and also to control the disease in the country. As discussed previously, the two disease parameters which can be controlled are the biting rate of vectors on AHS hosts, and the proportion of vectors to hosts. Another more practical strategy if available is to protect susceptible hosts from acquiring the infection. This can be achieved through vaccination to immunize the host from the disease.

1.5.1 Vaccination as a Control Strategy

In South Africa, the only immunization is in the form of two polyvalent, attenuated vaccines, the first of which contains serotypes 1, 3, and 4 and the other serotypes 2, 6, 7 and 8. Vaccination is required by law, although by Onderstepoort Biological Products estimates only around 50% of the national herd is currently vaccinated annually (W. Botha, *Personal Communication*, 2010). There are two serotypes not included in the vaccines; serotype 5 and serotype 9. AHSV5 was withdrawn from the vaccines in 1993 due to reports of it causing fatal side-effects in some animals (Mellor and Hamblin, 2004). AHSV 9 is not included since there is considered to be adequate cross-immunisation between strains 6 and 9, and because strain 9 is not often found in South Africa (Coetzer and Erasmus, 1994; Mellor and Hamblin, 2004). Attenuating a virus is done in order to decrease the virulence of the virus. The most common method is by passing the virus through a host. In the past the AHS virus has been attenuated by passing it through suckling mouse brain (Mellor and Hamblin, 2004).

In South Africa, immunization is performed annually in Spring, during August - September (Onderstepoort Biological Products, 2010), which is some time before peak AHS season in order to allow for the animals' immunity to build up to a maximum during the high risk months of February-March and before vector populations increase in November and December. The two injections of the vaccines are administered in three week intervals, and the immunity of the animal will start to develop only three to four weeks thereafter (Onderstepoort Biological Products, 2010). However, vaccinating with several serologically different strains simultaneously can lead to insufficient immune response to all strains. Thus it is considered that it will take at least 2 to 3 vaccinations before an animal is immune to all strains of the disease, and a horse may never be entirely immune despite annual vaccination (Onderstepoort Biological Products, [41]; Mellor and Hamblin, 2004).

In countries outside of endemic regions where outbreaks have occurred, usually the outbreaks are attributable to one strain only. In the 1966 epizootic in Spain, it was found that the cause was AHSV serotype 9, while the 1987 outbreak in Spain which spread to Portugal was attributed to AHSV serotype 4 (Mellor *et al.*, 1990; Portas *et al.*, 1999). In these cases, monovalent vaccines were employed to guard against the specific serotype of the outbreak. A rigorous campaign of eradication in Portugal included banning of import of horses from Spain, vaccination, and slaughter of infected horses (Portas *et al.*, 1999). In West Africa, where serotype 9 is the only strain known to be active, monovalent attenuated vaccines are used (Mellor and Hamblin, 2004).

1.5.2 Stabling

Since *Culicoides* species are mostly crepuscular, stabling between sunset and sunrise has been believed to offer protection against AHS. In a study by Meiswinkel *et al.* (2000), it was found that *C. imicola* were far more abundant outside than inside stables, regardless of the interventions used inside the stable (doors open/closed, fans on/off, windows open/gauzed). Approximately 82% of the catch of *C. imicola* was outside, with only 18% inside the stables. By contrast, *C. bolitinos*, also implicated as a vector of AHS, was found to be more abundant inside of the 'open' stables (ungauzed windows) than outside. Closing the doors and gauzing the windows of a stable, however, led to a 14-fold reduction in the abundance of *C. bolitinos*. It was concluded that the stabling of horses would protect them from encounters with *Culicoides* species, but only if the stables were closed and even more so if the remaining openings were gauzed. However it was found to be virtually impossible to exclude the midges from an area entirely. Jenkins (2008), in a study on the abundance of *Culicoides* midges around stables in KwaZulu Natal, found that 37% of *Culicoides* midges of all species were caught inside of stables or under the eves. By species, however, 50.3% of the catch of *C. bolitinos* was from within the stables, verifying the work of Meiswinkel *et al.* (2000), and *C. imicola* had 39.5% of the catch indoors. *C. imicola* was the most prevalent species, comprising 32.6% of the catch, with *C. bolitinos* making up 20.7%. It cannot be concluded that stabling on its own is an effective protective measure against AHS.

1.5.3 Vector Control

There are various methods of controlling the vector population. These methods aim to reduce the abundance of the vector, or reducing the bite-load and possible infection with AHS that the animals are exposed to. The underlying aim in all methods of vector control is to minimize the host-vector contact rate.

Pesticides and Repellents

Insecticides have been used either to kill the adult *Culicoides* vectors of the disease or, through application of the chemicals to breeding sites, to kill the midges in their larval stages. Ivermectin is a broad-spectrum antiparasitic medication commonly used as a dewormer in horses which may kill biting *Culicoides* midges. Additionally, when excreted in the faeces, it is known to act as a larvicide (Mellor and Hamblin, 2004).
Insect repellents may also be used topically, although there are few repellents proven to be effective on *Culicoides* midges. The most effective of these repellents is a substance called pyrethroid-T which is able to repel midges throughout the night (Braverman and Chizov-Ginzburg, 1997). Simpkin (2009) found that cypermethrin, another pyrethroid containing substance, was the best performing repellent specifically for *Culicoides imicola*. It acts as a neurotoxin on insects. Some mosquito repellents, for example citronella based repellents, have been found to attract rather than repel *Culicoides imicola* (Braverman *et al.*, 1999), although conflicting results have been found claiming that although not efficient repellents, they do not attract midges (Simpkin, 2009). These repellents are widely used to repel flies from horses.

Habitat Control

Another method of reducing *Culicoides* populations is by habitat control. This is done by removing possible breeding sites for the midges. *Culicoides imicola* breed in moist areas, and prefer clay soils. Water and animal dung provide perfect breeding habitats for these midges. Ensuring that animal dung is removed, and wet areas drained, will interrupt the breeding cycle of the midge.

Wind Speed

It is known that *Culicoides* midges are very weak fliers, and midge activity is decreased for wind speeds greater than $3ms^{-1}$ (Mellor *et al.*, 2000). Therefore the use of fans inside stables has been tested to reduce the numbers of *Culicoides* inside the stables (Simpkin, 2008). It is also theorized that fans will assist in dispersing the odour of the horses, which may attract the vectors (AHS Trust, 2005). This coupled with stabling between dusk and dawn was proposed to be effective in reducing exposure to midges by Braverman in 1989. However, Meiswinkel *et al.*(2000) found that fans had no effect on the number of *Culicoides* midges inside stables.

Alternate Hosts

A method of decreasing the bite load on horses is by introducing alternate sources of blood-meal which the midges may bite. The *Culicoides* midges are known to transmit a wide range of viruses including Bluetongue virus (BTV) which affects all species of ruminant. Specifically, *C. imicola*, the major vector of AHSV, is known to be a vector of BTV and to feed on species of ruminant including sheep, cattle and antelope (Mellor *et al.*, 2004). Introducing these species to the area surrounding horses may prove to be beneficial in reducing the bite-load on the horses. Simpkin (2009) found that both *C. imicola* and *C. bolitinos* showed no host preference between horses, sheep and cattle. Thus alternate hosts, in the form of either cattle or sheep, may reduce the bite load on horses and help to prevent AHS. More female midges were caught near the cattle than the sheep, and therefore cattle may be the superior alternate host. However, since *Culicoides* are known to breed in dung, and in particular *C. bolitinos* breeds in cattle dung (Meiswinkel and Paweska, 2003), it is necessary to be vigilant in removing potential breeding sites if this method is used. This method of control may also have implications for the epidemiology of Bluetongue Virus, which affects all species of ruminant and shares the vector *C. imicola* (Mellor, 2000).

However such a strategy may be counterproductive by helping maintain the midge population, unless the alternate host cannot harbour the virus. Other equids, for example, should not be used as an alternate host, as donkeys can act as a reservoir host by remaining subclinical for the disease. This has been speculated to act as an overwintering mechanism of the virus.

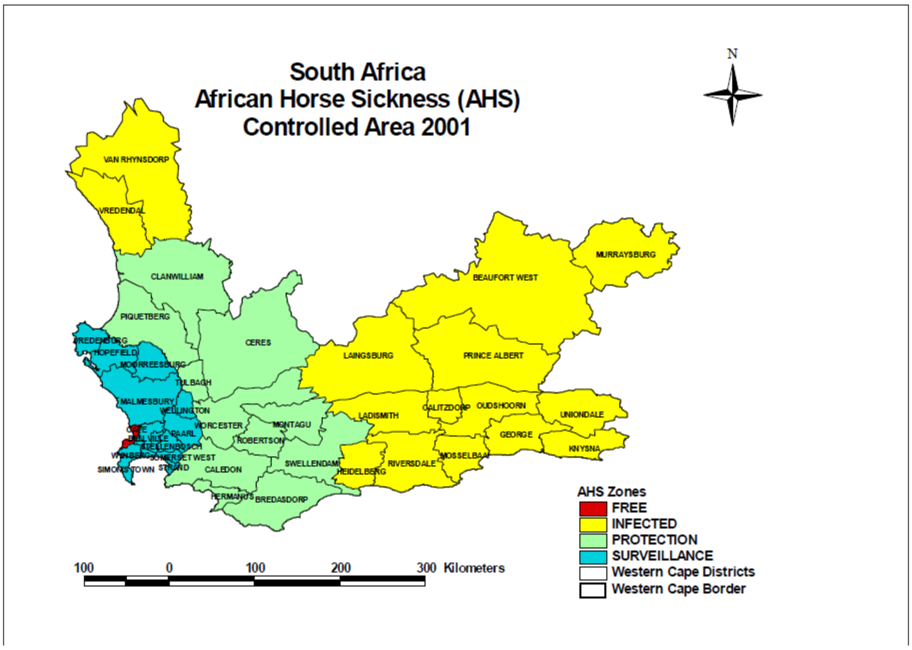

Figure 1.3: Map showing the surveillance, protection, and free zones for African Horse Sickness. (African Horse Sickness Trust Pamphlet)

Movement Control Policy and the AHS Free Zone

There are strict movement policies regarding horses in sub-saharan Africa into non-endemic regions. However, areas of the Western Cape have been shown to be AHS free, and vigilant monitoring of these regions has made it possible for export to occur. The AHS-free zone is surrounded by the AHS-surveillance zone and AHS-protection zone, shown in Figure 1.5.3. In March 2011, however, an outbreak occurred in Mamre in the AHS surveillance zone, approximately forty kilometers from the AHS free zone. In this outbreak, as of 28th March 2011, 46 horses were confirmed as infected, and 26 more suspected (http://www.africanhorsesickness.co.za/Documents/doc_45.pdf, accessed 14/05/2011). The total deaths stood at 52 as of the above date.

Horses are required to be fully vaccinated, and be certified healthy by a veterinarian before they may enter any of these zones from the rest of South Africa. Before a horse may be exported from the country, it must reside in the AHS-free zone for 20 days and be quarantined for a further 40 before departure at the Kenilworth quarantine station. The station is protected from vectors, and was designed so that the export of the racehorse London News was possible (Racing South Africa, http://www.racingsouthafrica.co.za/view_page.aspx?ID=134, accessed 18/10/2010). Subsequently, many horses enter and exit the country from its quarantine station.

1.6 Temperature and African Horse Sickness

Both the life cycle of the *Culicoides* vector and the development of the AHS virus are highly temperature dependent. At lower temperatures, the midge experiences a longer lifetime, but virogenesis is slower. At higher temperatures, although the lifetime of the vector decreases, virogenesis is faster and therefore transmission occurs more rapidly (Mellor and Hamblin, 2004).

The duration of the life cycle of the *Culicoides* midge depends on both the species of midge and climate variables from as short as 7 days to as long as 7 months (Wittman, 2000). Their adult life is usually in the region of 10 - 20 days, but may last as long as 3 months (Mellor *et al.*, 2000). Veronesi *et al.*(2009) showed that both the lifespan and reproductive capabilities of *C. imicola* Kieffer were affected by temperature in lab-reared colonies. The time from blood-feeding to adult offspring was longest at lower temperatures. At 20°C this cycle took 34-56 days, at 25°C 15-21 days, and at 28°C 11-16 days. Coupled with this was the temperature dependence of the number of eggs laid per female (fecundity), and the survival rate of the offspring to adult form. The number of surviving adult offspring per female was found to be highly variable at 28°C (between 0.1 and 3 adult offspring per female), but more concordant for 25°C (0.7-1.0) and 20°C (0.7-1.7). Since the females feed on blood which is required for egg development, and one bloodmeal is usually required for each batch of eggs (Wittman and Baylis, 2000), it follows that the more rapidly the life cycle progresses, the higher the bite-load on hosts, and therefore the higher the possible transmission rate of AHS where the virus is present. Braverman *et al.*(1985) showed this seasonality in the period between bloodmeals for *C. imicola* in Israel, ranging between 3 days and almost 5 days. They also showed the seasonality in the abundance of the midges, which began in summer (July) peaked mid-summer (August - September), and declined into winter (November).

Wellby *et al.*(1996) found that in *C. variipennis sonorensis* and *C. nubeculosus* (Meigen) that an infected vector could not transmit the virus at temperatures below 15°C, and that virogenesis could not be detected at 10°C. However, midges that survived and were then reintroduced to warmer temperatures were able to transmit the disease once more. Mellor *et al.*(1998) similarly found that *Culicoides* kept at 10°C were almost free of infection at 13 days post infection (dpi). However when these midges were kept at 10°C for 35 days and brought back to a warmer temperature for 3 days they were once again infective. This indicates that the virus is able to survive at low titres in midges at lower temperatures, and that although the midge will not be able to transmit the virus, when temperatures increase, it will then be able to replicate within the midge to a level where it can be transmitted. This ability for the virus to survive within the midge at colder temperatures, as well as the ability for the midge to live far longer at these colder temperatures, indicates a possible overwintering mechanism for the virus (Mellor and Hamblin, 2004).

1.7 AHS and Rainfall

The breeding sites of *Culicoides* species require moisture for the larval stages to develop (Mellor *et al.*, 2000). Sites which they may choose to breed include moist soil, swamps and bogs, animal dung and rotting vegetation. Rainfall is therefore advantageous to the midge, as it increases the potential breeding sites. Jenkins (2008) found that, in KwaZulu Natal, the duration of wetness of the ground was highly correlated with increased *Culicoides* midge catches, for example with overflow from a reservoir moistening the surrounding ground. He also found that modifying the surroundings so that this overflow did not occur was very effective in reducing the midge numbers.

1.8 Epidemiological Modeling

Much work has been done, predominantly by Lord and Woolhouse, on the mathematical modeling of AHS as discussed in Section 1.4. However, much of this modeling was specific to the outbreak in Spain between 1987 and 1990. This outbreak proved to be ideal for modeling purposes, as it began in a naïve population with no vaccination or immunity to the disease. Papers have been published to model AHS using a deterministic compartmental model (Lord *et al.*, 1996b), to simulate vaccination strategies (Lord *et al.*, 1997), and to calculate the basic reproduction number (Lord *et al.*, 1996a). Although in these outbreaks the protection of horses was considered paramount, the authors considered donkeys' ability to transmit the disease an important factor in the invasion and persistence of the virus.

Modeling has also been done to better understand the distribution of *Culicoides imicola*. In 2001, Wittman, Mellor and Baylis published a paper that used logistic modeling to predict the presence or absence of *C. imicola* in Europe based on the climate of an area. The climate variables considered were temperature, saturation deficit, rainfall and altitude. The model used was a logistic regression model, aiming to predict the probability of presence in an area in Iberia, and these results were extrapolated to the whole of Europe. Their model found useful climatic variables to be: minimum of the monthly minimum temperatures, maximum of the monthly maximum temperatures, and number of months per year with a mean temperature $\geq 12.5°$C (Wittman *et al.*, 2001). Their model had a high degree of accuracy in correctly predicting the presence of *C. imicola* in Iberia.

Baylis, Meiswinkel and Venter (1999) used a similar technique to relate climate data and satellite imagery to the distribution of *C. imicola* in southern Africa. Here 34 sites in South Africa were sampled for *Culicoides*. The satellite imagery variables investigated were the normalised difference vegetation index (NDVI), the land surface temperature (LST), cold cloud duration (CCD). The NDVI is a measure of photosynthetic activity of vegetation, and is correlated with functions of moisture. LST is correlated with temperature, and CCD with rainfall. Climate variables used were annual mean daily maximum and minimum temperature, annual minimum temperature, number of days with temperatures below $0°C$, October-March rainfall, April-September rainfall, and total annual rainfall. The sampled estimates of *C. imicola* abundance were normalised using the transformation $\log(n + 1)$, and then regressed on the climate and satellite imagery variables. The best model used the variables minimum LST and minimum NDVI.

1.9 Implications for AHS in South Africa

It is the opinion of the author that insufficient work has been done to understand the disease in South Africa. The modelling work by Lord and others, regarding the outbreak in Spain, Portugal and other epizootic areas, has helped to understand the disease process, but has not increased understanding of the disease in this country specifically. The outbreak in Spain affected horses with no previous exposure to the disease or vaccination, and thus no acquired immunity. In South Africa there is a certain amount of immunity within the national herd, which varies for vaccinated / unvaccinated horses. Because it is such a severe disease, having broad-reaching implications for the community, there is need to better understand the factors and processes which affect outbreak severity within South Africa, as well as those measures which might be taken on an individual horse level to decrease the probability of mortality.
The African Horse Sickness Trust (AHST) was formed in 2005 in order to bring together major stakeholders of the horse community in order to reduce the threat of the disease to South African equines. Their aims are to:

- Improve reporting of the disease,
- Increase vaccination coverage of the national herd,
- Improve vaccines available to South Africa, and to
- Increase the body of scientific research into the disease.

They also aim to introduce an "early warning system" which will indicate which areas are likely to come under threat from the disease. The AHST also serves as a repository of information and data collected from 2005 to the present, from which it is possible to improve our epidemiological understanding of AHS and its transmission.

1.10 Objectives of the Study

Specific objectives of this study are

- To explore and develop the techniques for modelling incidence and mortality of African Horse Sickness,
- To gather and use data with which these modelling techniques can be utilized,
- To investigate and discern factors and processes affecting both the incidence and mortality of the disease from the results of these models,
- To further understand what processes may affect the outbreaks of the disease, in the case of models which are not perfect.

Statistical modelling of the African Horse Sickness Trust data will help to further the knowledge on AHS in South Africa, and to assist in providing early warning systems. To this end, data is statistically modelled to better understand these factors correlated with the disease, and to give direction for future study.

Chapter 2

Exploratory Data Analysis

2.1 African Horse Sickness Trust

There are two sources of data used in this project. The first, kindly provided by the African Horse Sickness Trust (AHST), records cases of African Horse Sickness in South Africa from the end of 2005 until May 2010. Each case in the database was reported by the attending veterinarian or the owner of the animal, and for each case certain variables were recorded. However, after the primary case, if additional cases occurred in the same area they were simply listed as the number of additional cases which survived (AdditionalAlive), and the number of additional mortalities (AdditionalDead). Therefore the information on the case was given for the **primary case only**. A Case Identity number is assigned to each primary case for ease of reference. For each case several variables were recorded. The variables considered in this research are outlined and briefly described below.

Horse Status

It was recorded in a variable named “HorseStatus” whether the horse was still alive or had died due to the disease. A few cases also listed “Euthanised” as an option - but since interest was in modeling this as a binomial variable, these few cases were changed to being listed as having died. This is a realistic measure to take, since firstly there were very few (28) “Euthanised” observations, and secondly that only cases which were severe (and therefore likely to result in death) would be euthanised. There were 461 primary cases listed as “Alive”, and 486 listed as “Dead”.

Province

The province of the occurrence of the case was recorded. The list of provinces and their abbreviations are given in Table 2.1. The third column gives the total number of cases recorded from each province.

Place

The place within the province in which the case occurred was also recorded. There were

Table 2.1: List of provinces and their abbreviations used in the data, and the total number of cases reported for each province.

Province	Abbreviation	Cases
Eastern Cape	ECP	101
Free State	FS	21
Gauteng	GAU	370
Kwa-Zulu Natal	KZN	191
Limpopo	LIM	48
Mpumalanga	MPU	78
Northern Cape	NCP	23
North West Province	NWP	62
Western Cape	WCP	53

239 places recorded.

Case Classification

A case classification variable was included. It recorded whether the case was Suspected (SUS), Confirmed by a veterinarian (VETC) or Confirmed by a sample sent to the lab (LABC) or if it was suspected but the lab testing proved negative (SUSN).

Other Cases

It was recorded whether there were other cases in the surrounding area in the same outbreak. It was recorded as a binary variable with 1 for "other cases occurred" and 0 for "there were no other cases".

Vaccination

A further two variables recorded whether the horse had been vaccinated or not (Vaccinated = 1, Not Vaccinated = 0), and whether or not the vaccination had been performed late (Vaccinated Late = 1, Not Vaccinated Late = 0). Routine vaccination for AHS is performed between August and October, as it takes time to build up efficacy in terms of immunity and it is optimal that the animal's immunity is at its peak over the AHS outbreak months (February to April). If a horse was vaccinated late (after October) this may have an effect on its susceptibility to the disease.

These two variables were for the purpose of the current analysis combined into one variable with three levels; 0 = Not Vaccinated, 1 = Vaccinated Late, 2 = Vaccinated Timeously.

Further Preventative Measures

Further preventative measures were also recorded. The person entering the data had the opportunity to state whether the horse was stabled, whether pesticides were applied, and any further preventative measures used. The data does not indicate whether the pesticides were applied to the horse or the surroundings.

Binary variables were created from this information as Stabled (1 = stabled, 0 = not stabled) and Pesticides (1 = pesticides used, 0 = no pesticides used).

Treatment

Four types of intervention strategies were recorded. The possible types of treatment are Conventional (CONV), Homeopathic (HOM), Alternative (ALT) or None (NONE). Conventional treatment is that usually administered by a veterinarian, including anti-inflammatory drugs, antibiotics, and vitamins. Homeopathic treatments are those suggested by a homeopath. Alternative covers those treatments not considered conventional or homeopathic, but often include use of marijuana or "dagga", which was specifically stated for some of the cases under additional information.

Presentation

The variable Presentation records which presentation of AHS occurred in the animal. It is either Cardiac (CARD), Pulmonary (PULM), Mixed (MIX), or Mild (MILD). There is also an option"Don't know" (DK) for those entering data who were unsure of how the disease presented.

Isolation

A binary variable recorded whether a horse was isolated / quarantined once symptoms were noticed.

The complete list of variables with associated categories is given in Table 2.2.

2.1.1 Chi-Square tests of Association

To test which variables were associated, Chi-Square tests were employed using SAS FREQ procedure. Firstly we compared HorseStatus with all other categorical variables. The Chi-Square probabilities are shown in Table 2.3.

The variables which do not have significant association with HorseStatus are Stabled and Pesticides. This makes biological sense, as these are prevention strategies which may have an effect on whether the horse contracts AHS or not, but would be unlikely to have an effect on the outcome of the disease once the horse has contracted it. Both Presentation and Treatment had very strong relationships with HorseStatus ($p < 0.0001$). Province and OtherCases have marginal associations ($p \approx 0.05$). It is also to be noted that the test for association between HorseStatus and Classification variables was based on expected values of less than 5, and therefore may not be reliable. Contingency tables showing the frequencies and percentage mortalities for each of the significant interactions with HorseStatus are shown in Tables 2.4, 2.5, 2.7, 2.8 and 2.9. 95% confidence intervals are given, and negative values are censored to zero. Bar graphs showing the relationships are shown in Figures 2.1, 2.2, 2.3, 2.4 and 2.5.

In Figure 2.1, we can see clearly that Eastern Cape, Gauteng, and Mpumalanga, have mortality rates slightly below 50%. The rest of the provinces have rates above 50% (more deaths than survivals). However from Table 2.4 we see that the only provinces whose 95% confidence intervals for mortality are entirely above 50% are Northern Cape

Table 2.2: List of categorical variables, their levels and abbreviations used for the AHST data

Variable	Number of levels	Levels	
Province	9	ECP	Eastern Cape
		FS	Free State
		GAU	Gauteng
		KZN	KwaZulu Natal
		LIM	Limpopo
		MPU	Mpumalanga
		NCP	Northern Cape
		NWP	North West Province
		WCP	Western Cape
Place	239	Not listed due to large numbers of levels	
Classification	4	SUS	Suspected
		VETC	Confirmed by Veterinarian
		LABC	Confirmed by Lab Sample
		SUSN	Suspected but Lab Negative
OtherCases	2	1	Other Cases
		0	No Other Cases
Vaccination	3	2	Vaccinated Timeously
		1	Vaccinated Late
		0	Not Vaccinated
Stabled	2	1	Stabled
		0	Not Stabled
Pesticides	2	1	Pesticides Used
		0	No Pesticides Used
Treatment	4	CONV	Conventional
		HOM	Homeopathic
		ALT	Alternative
		NONE	None
Presentation	5	CARD	Cardiac
		PULM	Pulmonary
		MIX	Mixed
		MILD	Mild
		DK	Don't Know
Isolation	2	1	Isolated
		0	Not Isolated

Table 2.3: Chi-Square Probabilities for Interactions between HorseStatus and other variables

Variable	df	χ^2 value	p-value
Province	8	16.9409	0.0307
Classification	3	14.5175	0.0023
OtherCases	1	3.9813	0.0460
Vaccinated	2	8.5732	0.0138
Presentation	4	86.9568	<.0001
Treatment	3	65.1324	<.0001
Stabled	1	1.1569	0.2821
Pesticides	1	1.7390	0.1873
Isolation	1	6.6722	0.0098

and North West Province. The other confidence intervals all include 50. In Figure 2.2, we see that more unvaccinated horses died than survived. Horses vaccinated late or timeously both had a greater proportion surviving, however the "Vaccinated Late" group was based on a very small number of observations (11). The 95% confidence interval for mortality for "Vaccinated Late" was entirely below 50, while the interval for "Vaccinated Timeously" included 50. Table 2.6 shows the contingency table if vaccination status is pooled into classes "Vaccinated" and "Unvaccinated", in other words timing not taken into account. If a simple chi squared test of independence is done on this it is found that the probability is 0.0325 - in other words mortality is not independent of vaccination status even if the timing is not taken into account. Figure 2.3 indicates that a larger proportion are expected to survive with Mild and Cardiac presentations, while over 50% died in Mixed and Pulmonary presentations. In Figure 2.4, it shows all treatments do reduce mortality. No treatment gave 85% mortality, while Homeopathic, Alternative and Conventional treatments had 40, 24 and 49% mortality respectively. Figure 2.5 shows us that a horse that was isolated seems more likely to die than one which was not isolated - which may be due to the stress of being separated from other horses.

Although these comparisons give us crude results on the effects that these treatments and preventative measures have on mortality, the strength and nature of these associations can be further investigated in models such as GLMs, where the effect of multiple variables can simultaneously be assessed within the same model.

2.1.2 Condensed Data

In order to have the ability to model the number of cases, the AHST data was condensed into a form that would be able to show the number of cases in a given month. For this

Table 2.4: Contingency Table for Province by HorseStatus

	ECP	FS	GAU	KZN	LIM	MPU	NCP	NWP	WCP	Total
Alive	59	7	187	94	24	40	7	20	23	461
Dead	42	14	183	97	24	38	16	42	30	486
Total	101	21	370	191	48	78	23	62	53	947
% mortality	41.58	66.67	49.46	50.79	50.00	48.72	69.57	67.74	56.60	51.32
Lower Limit 95% CI	31.97	46.50	44.36	43.70	35.85	37.63	50.76	56.11	43.26	48.14
Upper Limit 95% CI	51.20	86.83	54.55	57.88	64.15	59.81	88.37	79.38	69.95	54.50

Table 2.5: Contingency Table for Vaccination Status by HorseStatus

	Unvaccinated	Vaccinated Late	Vaccinated Timeously	Total
Alive	208	9	235	452
Dead	254	2	223	479
Total	462	11	458	931
% mortality	54.98	18.18	48.69	51.45
Lower Limit 95% CI	50.44	0.00	44.11	48.24
Upper Limit 95% CI	59.52	40.97	53.27	54.66

Table 2.6: Contingency Table for Pooled Vaccination Status by HorseStatus

	Unvaccinated	Vaccinated	Total
Alive	208	244	452
Dead	254	225	479
Total	462	469	931
% mortality	54.98	47.97	51.45
Lower Limit 95% CI	59.52	52.50	54.66
Upper Limit 95% CI	50.44	43.45	48.24

Table 2.7: Contingency Table for Presentation by HorseStatus

	Mild	Cardiac	Mixed	Pulmonary	Unknown	Total
Alive	35	174	25	22	196	452
Dead	3	135	65	86	190	479
Total	38	309	90	108	386	931
% mortality	7.89	43.69	72.22	79.63	49.22	51.45
Lower Limit 95% CI	0.00	30.22	72.03	62.97	44.24	48.24
Upper Limit 95% CI	2.06	39.73	87.23	81.48	54.21	54.66

Table 2.8: Contingency Table for Treatment by HorseStatus

	None	Homeopathic	Alternative	Conventional	Total
Alive	15	50	29	358	452
Dead	88	33	9	349	479
Total	103	83	38	707	931
% mortality	85.44	39.76	23.68	49.36	51.45
Lower Limit 95% CI	10.17	45.68	29.23	78.62	48.24
Upper Limit 95% CI	37.20	53.05	50.29	92.25	54.66

Table 2.9: Contingency Table for Isolation by HorseStatus

Not Isolated	Isolated	Total	
Alive	321	131	452
Dead	302	177	479
Total	623	308	931
% mortality	48.48	57.47	51.45
Lower Limit 95% CI	44.55	51.95	48.24
Upper Limit 95% CI	52.40	62.99	54.66

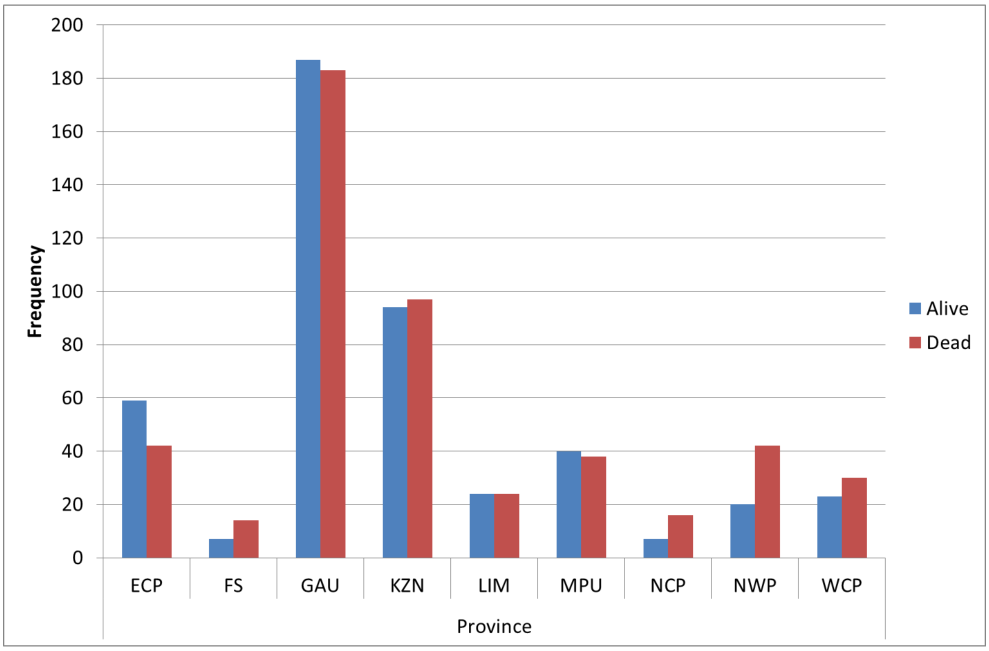

Figure 2.1: Bargraph showing interaction between Province and HorseStatus. Mortality rates are greater than 50% for FS, NCP, NWP and WCP

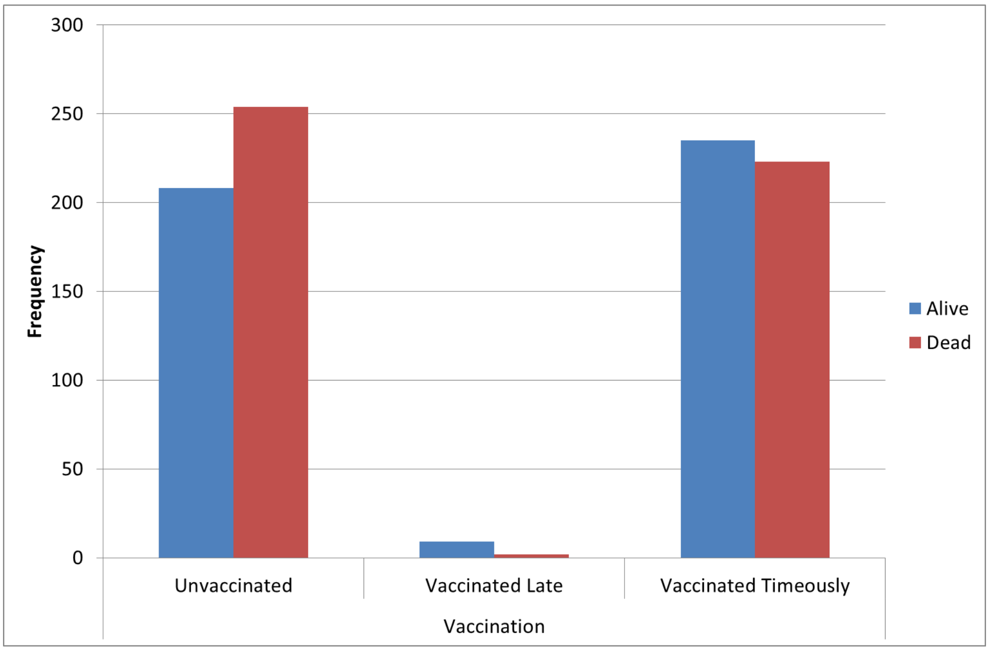

Figure 2.2: Bargraph showing interaction between Vaccination and HorseStatus. There are very few observations for Vaccinated Late. Mortality is greater than 50% only for Unvaccinated class.

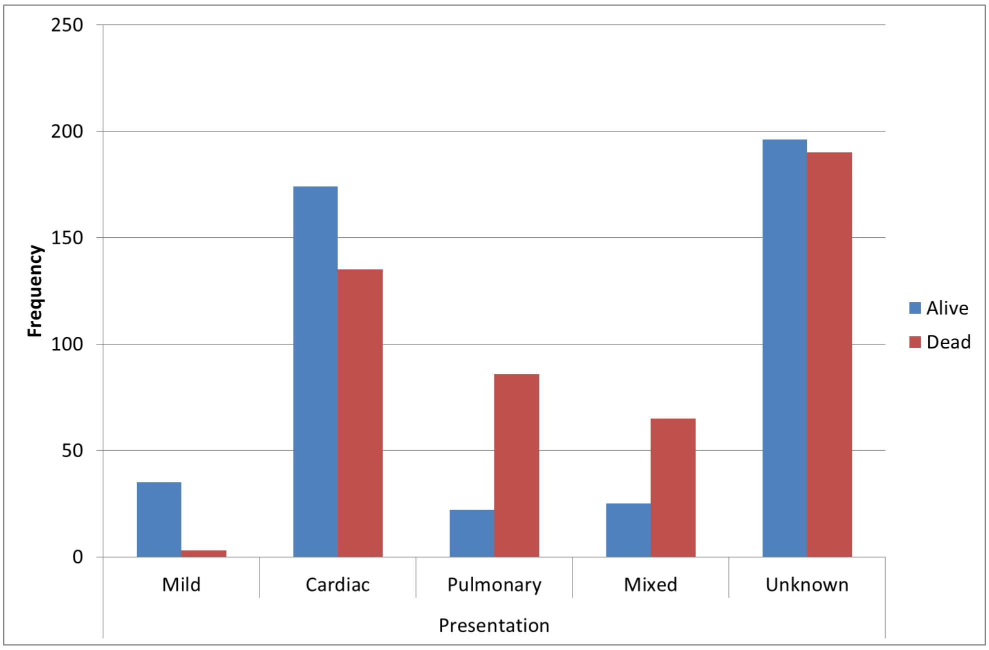

Figure 2.3: Bargraph showing interaction between Presentation and HorseStatus. Mortality is greater than 50% for Mixed and Pulmonary forms. Unknown presentation makes up a large proportion of observations.

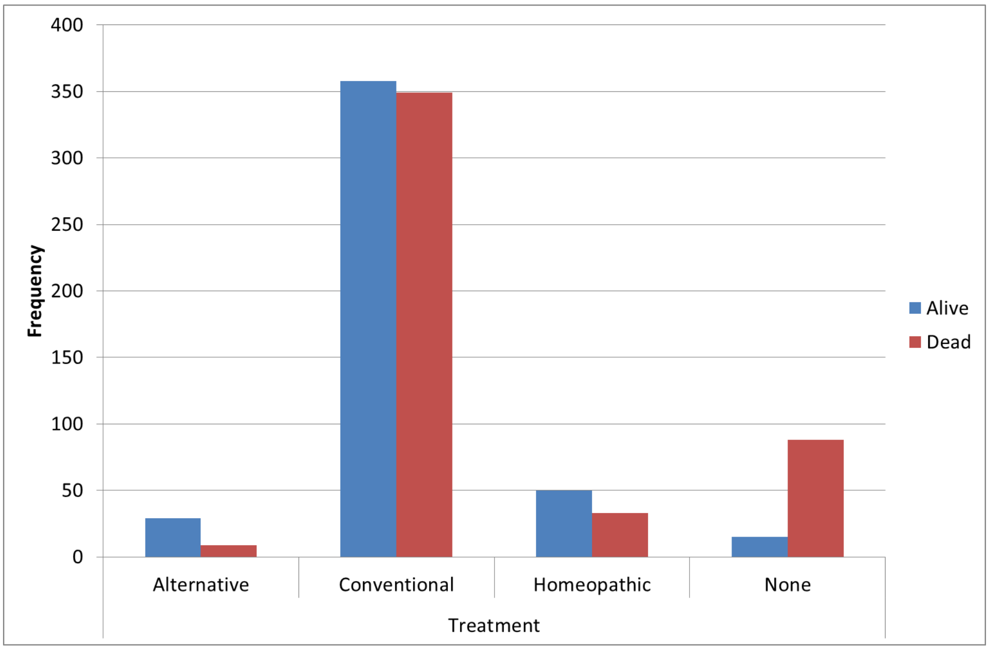

Figure 2.4: Bargraph showing interaction between Treatment and HorseStatus. Conventional treatment makes up the majority of the observations, and Conventional, Homeopathic and Alternative treatments are all found to be protective. Mortality is far greater than 50% where no treatment was administered.

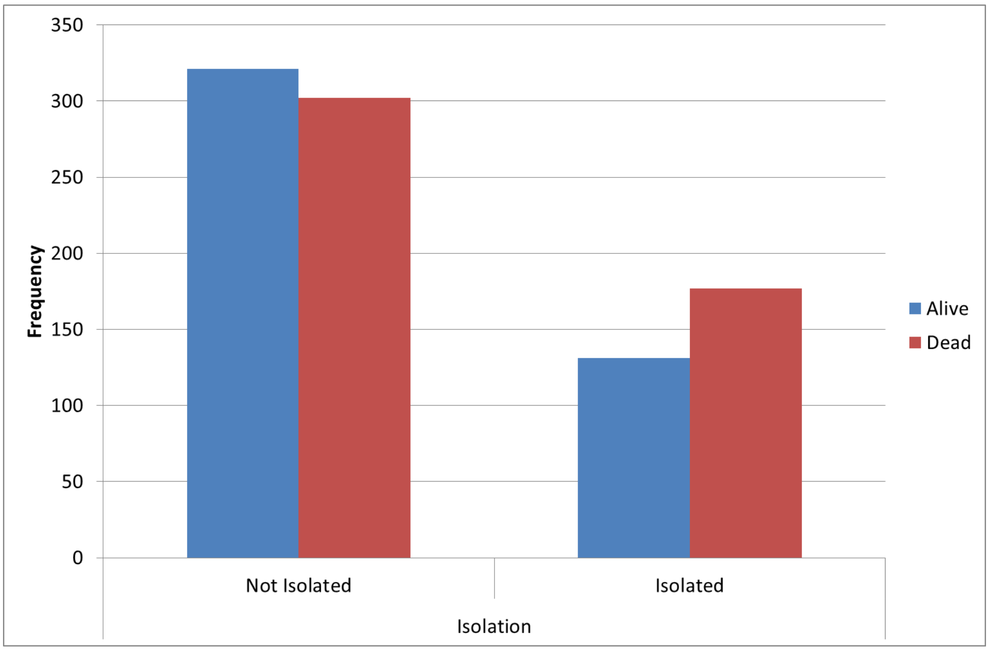

Figure 2.5: Bargraph showing interaction between Isolation and HorseStatus. Most animals were not isolated, and mortality was lower amongst these animals. Isolated horses had a mortality greater than 50%.

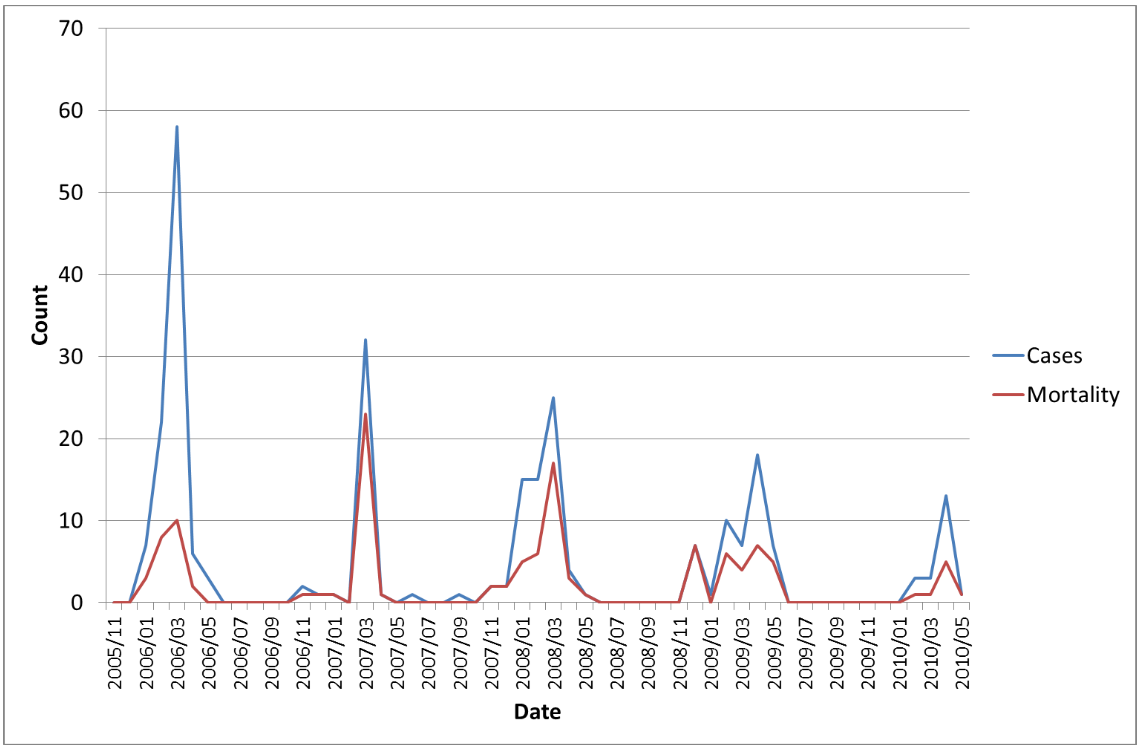

Figure 2.6: Cases and Mortality for KwaZulu Natal.

modelling purpose, we chose to focus on Kwa-Zulu Natal, and therefore only the cases whose Province code was listed as KZN were included in this data set. Note that the primary and additional cases were aggregated to form this dataset.

For each month, the number of cases, number of confirmed cases (where 'confirmed' was taken to have " Classification" listed either as Confirmed by Vet or Confirmed by Lab), number of mortalities and number of confirmed mortalities were included in the data. Several time variables were included - including Year (2005 to 2010), Month (1 to 12), Date (2005/11 to 2010/05). Time in fractions of a year was also calculated (1/12, 2/12,...,12/12,13/12,...) in order to have a continuous time variable for use in modeling.

Plots of Cases and Mortality are included for both KwaZulu Natal and South Africa in Figure 2.6 and Figure 2.8 respectively. Figure 2.7 is a simple bar graph showing average numbers of cases per month over the entire dataset.

In Figure 2.6 it is clear that there exists a highly seasonal pattern in the occurrence of cases. Most of the cases occur between October and May each year, reaching a peak intensity of outbreak between December and March. The peak occurrence differs slightly in timing between the years with peaks occurring in March 2006, March 2007, March

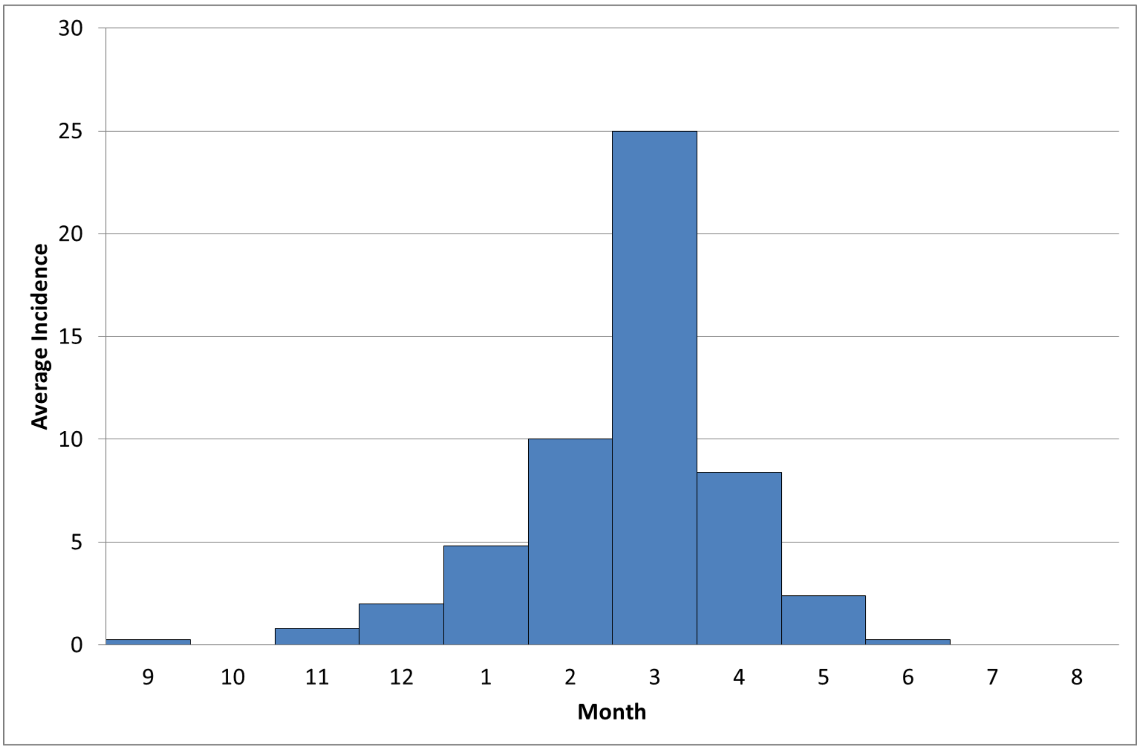

Figure 2.7: Simple bar graph showing the average number of cases per month for the five outbreaks in the AHS Trust data for KwaZulu Natal. This exhibits the seasonal pattern, with cases starting in October and increasing in number until March, and then decreasing until the off season between June and September.

Table 2.10: Table displaying summarized values of outbreaks of AHS for KwaZulu Natal between 2005 and 2010

Outbreak	Cases	Mortalities	Percentage Mortalities	[95% CI]	Duration (months)
2005-2006	96	23	23.96	[15.42, 32.50]	5
2006-2007	38	27	71.05	[56.63, 85.47]	8
2007-2008	65	36	55.38	[43.30, 67.47]	9
2008-2009	50	29	58.00	[44.32, 71.68]	6
2009-2010	20	8	40.00	[18.53, 61.47]	4

2008, April 2009 and April 2010. The number of cases in each peak differs greatly too - with a maximum total in 2006 of 58, 2007 of 32, 2008 of 25, 2009 of 18 and 2010 of 13. The mortality follows the same trend but with only a proportion of the animals dying.

Summarized values from the outbreaks are displayed in Table 2.10. The most severe outbreak for KwaZulu Natal occurred in the 2005-2006 season with a total case count of 96, and the minimum occurred in the 2009 - 2010 season with 20. The observed probability of mortality per outbreak ranges from 0.2396 to 0.7105. The duration of the outbreaks was between 4 and 9 months.

The line graphs in Figure 2.8 have much the same pattern as that in Figure 2.6, with highly seasonal trends and all cases occurring within a narrow band of months. However whereas in the KwaZulu Natal instance all cases occurred between October and May, in the South African 2007-2008 outbreak a single case occurred as early as September. This was also the outbreak with the longest duration of 10 months (as outlined in table 2.11). The latest cases were observed in June. The most severe outbreak appears to have occurred in the 2005-2006 season.

In these plots it is even more clear how the intensity of the outbreaks differs in different years. Possible causes of this disparity are differences in temperature, rainfall, midge population and many more.

Summarized values of the South African outbreaks are shown in Table 2.11. The maximum cases observed over the course of an outbreak is 849 (in 2005-2006), and the minimum is 83 (in 2009-2010). Once again it is of interest to note that the outbreaks differ so greatly, with the largest outbreak following the smallest. The mortality ranges from 17.31% to 58.72%. The length of outbreak ranges between 7 and 10 months.

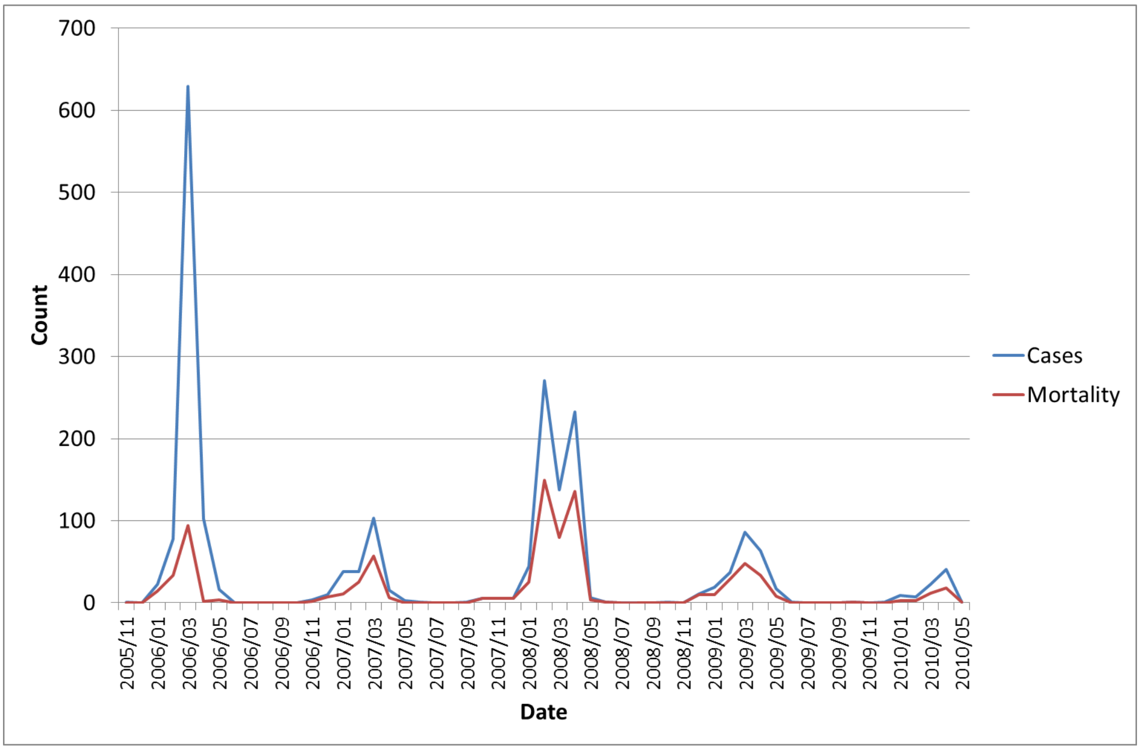

Figure 2.8: Cases and Mortality for South Africa

Table 2.11: Table displaying summarized values of outbreaks of AHS for South Africa between 2005 and 2010

Outbreak	Cases	Mortalities	Percentage Mortalities	[95% CI]	Duration (months)
2005-2006	849	147	17.31	[14.77, 19.86]	7
2006-2007	212	108	50.94	[44.21, 57.67]	8
2007-2008	709	410	57.83	[54.19, 61.46]	10
2008-2009	235	138	58.72	[52.43, 65.02]	9
2009-2010	83	38	45.78	[35.06, 56.50]	8

Data Limitations

There are some limitations to the data from this source. Firstly, although the disease is notifiable by law, there is poor reporting of the disease and it is difficult to know what percentage of the cases are actually recorded. Specifically in the rural areas, where vaccination is not routinely practised and education about the disease is low, disease and deaths are likely to go unreported. Not having access to a veterinarian is another factor which may reduce the likelihood of reporting the disease. It is also uncertain what percentage of the cases recorded are truly AHS, as there can be some confusion with Equine Encephalitis (particularly in the Pulmonary form), and only a small percentage of the cases are confirmed by laboratory testing. There are also certain horse owners who are more likely to report cases. Those who are competing have an interest in reporting and control of the disease being improved as this facilitates the ability to move horses around the country.

Secondly, some of this data were poorly reported, as shown in Figure 2.9. Here the percentage of cases where Vaccination Status, Age, Presentation, Confirmation of Case, and GPS coordinates have been reported are given. By Presentation filled in, we mean where Presentation was not filled in as "Don't Know", and by Confirmation of Case we mean where Classification was not "Suspected". The reporting of these fields can be seen to be relatively poor. In Figure 2.10 the same information is shown grouped into the different outbreaks. None of these fields were reported for the 2005-2006 outbreak, as this was the first year that the African Horse Sickness Trust had begun reporting cases. From the 2006-2007 outbreak reporting was much improved in most fields, however the low reporting of Age in the first two outbreaks makes us sceptical of its use for modelling purposes. The reporting of GPS coordinates is too scarce in use in spatial models. Improved reporting is an important consideration for future studies, and the data's limitations are an important consideration in any conclusions drawn from models developed.

This data set does, however, represent the only one of its kind and is therefore extremely important to platform further research and improved reporting.

2.2 South African Weather Service Data

Meteorological data were obtained from the South African Weather Service (SAWS). Data were only available for five locations, and so the locations were selected based on how widespread they were (whether they could be said to be representative of the entire province) and how many cases were reported for each area. The locations chosen were Ixopo, Ladysmith, Pietermaritzburg, Newcastle and Vryheid. Their locations are shown in Figure 2.11.

The variables available from each of these locations were Average Daily Maximum

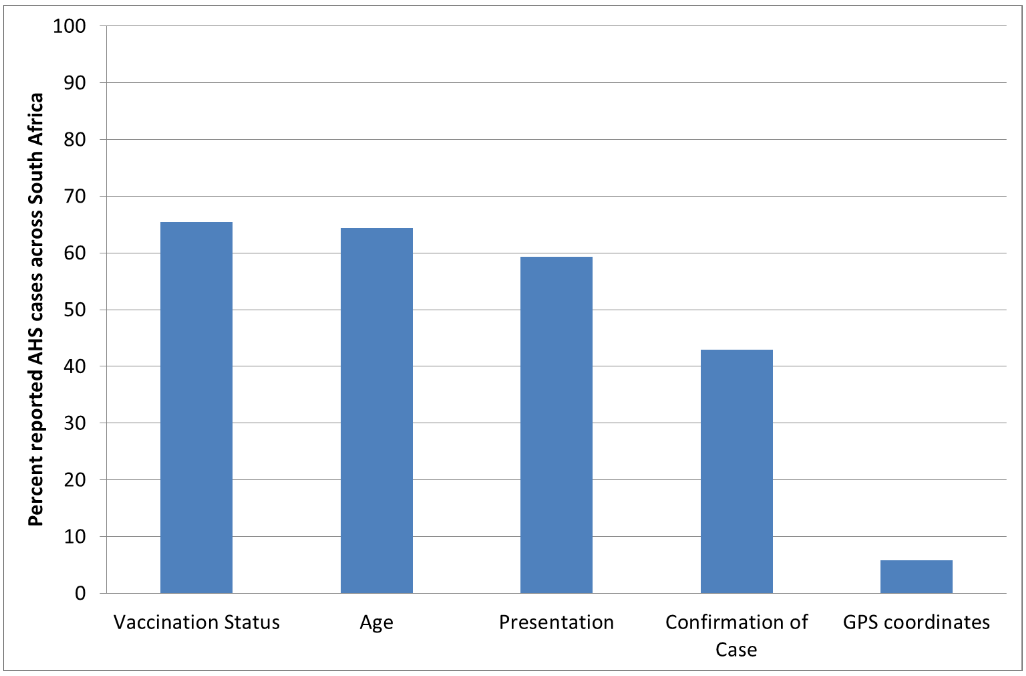

Figure 2.9: Bargraph displaying percentage of cases where the fields Vaccination Status, Age, Presentation, Confirmation of Case, and GPS co-ordinates were recorded. By Confirmation of case, we mean where the Classification was not “Suspected”.

Temperature, Average Daily Minimum Temperature, and Monthly Rainfall. The Average Daily Maximum Temperature and Average Daily Minimum Temperatures were the monthly averages of the maximum and minimum daily temperatures respectively. Monthly Rainfall was the cumulative rainfall in millimeters over the month.

In order to get average measures across the province, the three variables were averaged over the five locations. Since the locations are fairly well spread out over the province this was taken as an indicator of what the weather variables were over the entire province. This was used in Chapter 3 for the analysis. The plots of Average Daily Maximum Temperature, Average Daily Minimum Temperature and Average Monthly Rainfall are shown in Figures 2.12 and 2.13. Plots for individual location temperatures and rainfalls are given in Appendix A.

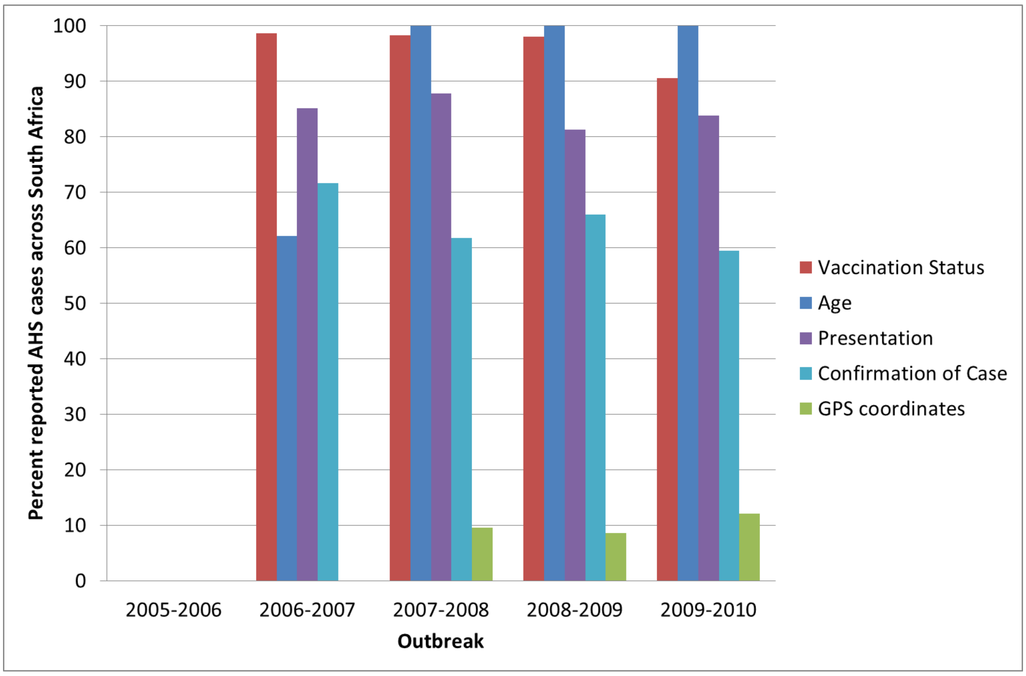

Figure 2.10: Bargraph showing percentage of cases by outbreak where the fields Vaccination Status, Age, Presentation,and GPS co-ordinates were recorded, and where Classification was not "Suspected" (Confirmation of Case). None were completed in the first outbreak shown. Vaccination status is thereafter quite well completed at over 90%. Age has been fully completed for the last three outbreaks. Presentation has fluctuated between 80 and 90 %, and Confirmation of case has not been well completed at between 60 and 70 %. GPS coordinates are consistently low.

Table 2.12: Basic descriptive statistics for Average Maximum Daily Temperatures for the five locations

	IXO	*LDS*	*PMB*	*NWC*	*VRY*
Number of observations	112	120	126	126	126
Missing	5	10	0	0	0
Minimum Observation	18.2	17.8	20.8	18.6	16.6
Maximum Observation	28.4	31.60	31.2	31.6	28.1
Range	10.2	13.8	10.4	13	11.5
Mean	23.5268	25.2733	26.0468	25.3222	23.5632
Median	23.75	25.65	26.25	25.60	23.90
Mode	24.80	21.90	26.60	24.40	23.5
Variance	5.6113	10.5763	5.2118	9.3443	7.2088
Standard Deviation	2.3688	3.2521	2.2829	3.0568	2.6849

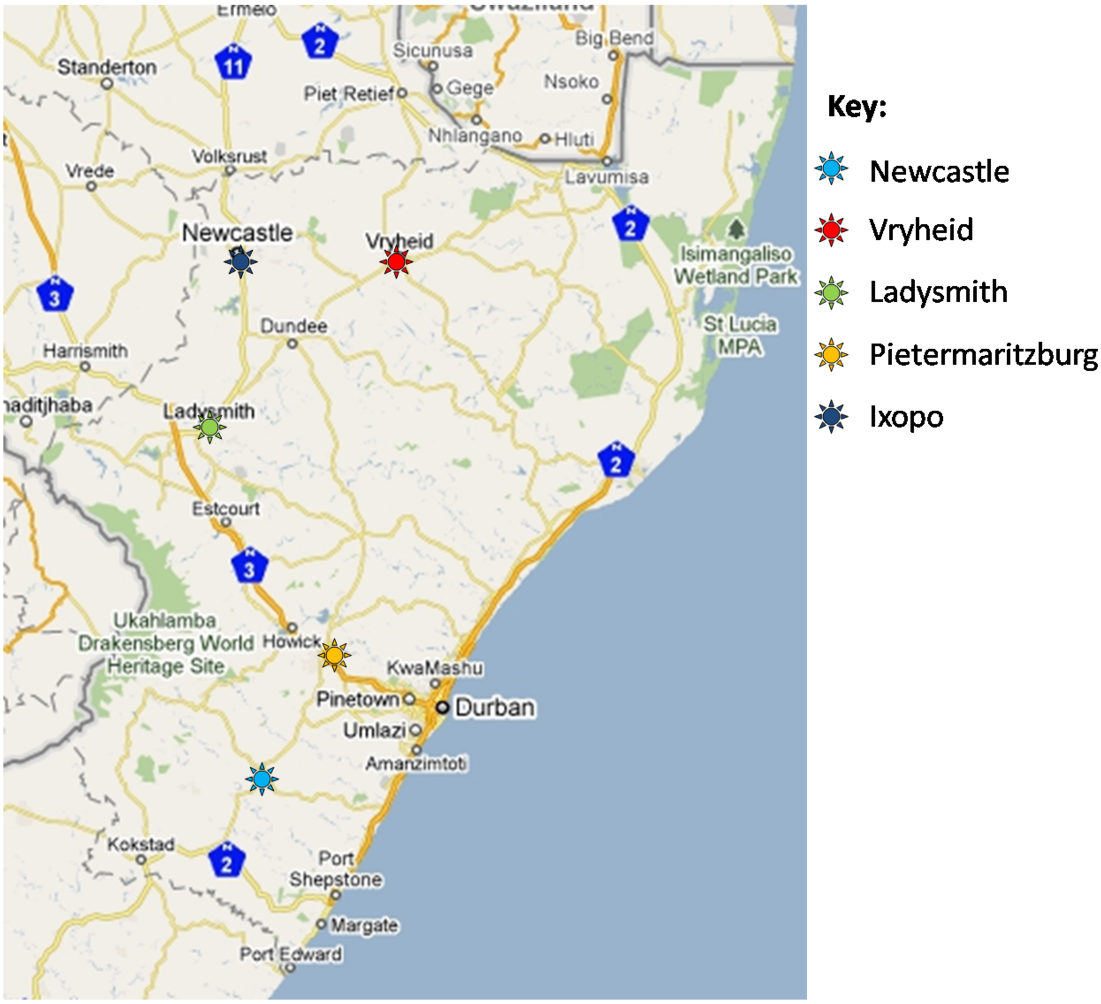

Figure 2.11: Map showing approximate locations from which weather data was available in KwaZulu Natal (©2010 Google - Map Data ©2010 AfriGIS (Pty) Ltd, Tele Atlas, Tracks4Africa). The weather stations' locations (Ixopo, Ladysmith, Newcastle, Pietermaritzburg, Vryheid) are marked with a sun symbol as shown in the Key.

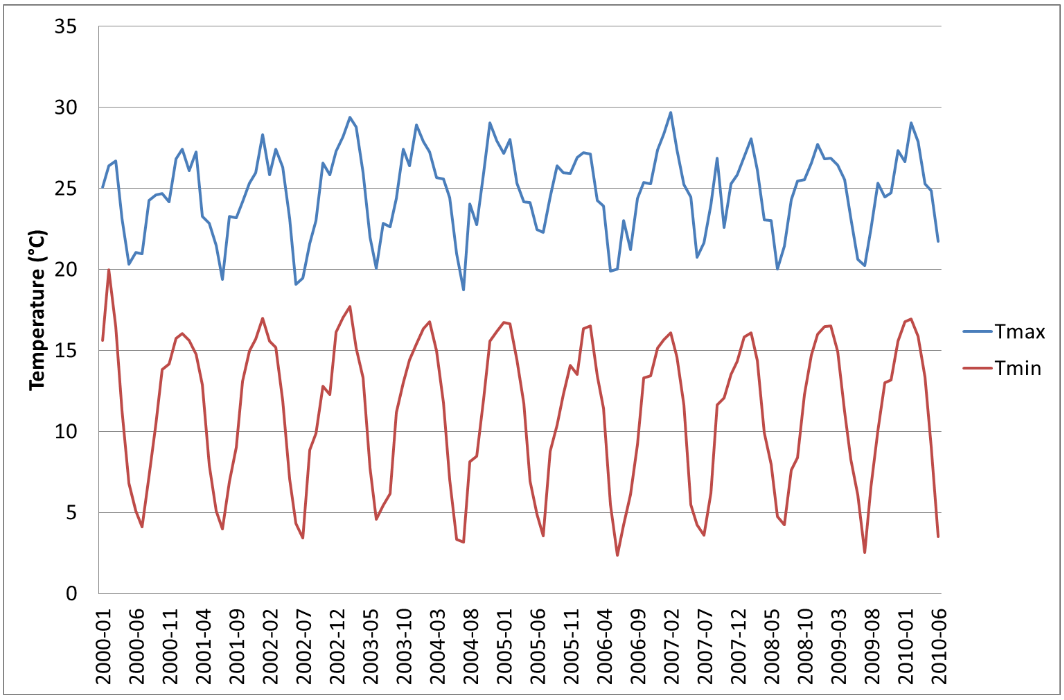

Figure 2.12: Average of the temperature variables of Ixopo, Ladysmith, Newcastle, Pietermaritzburg and Vryheid.

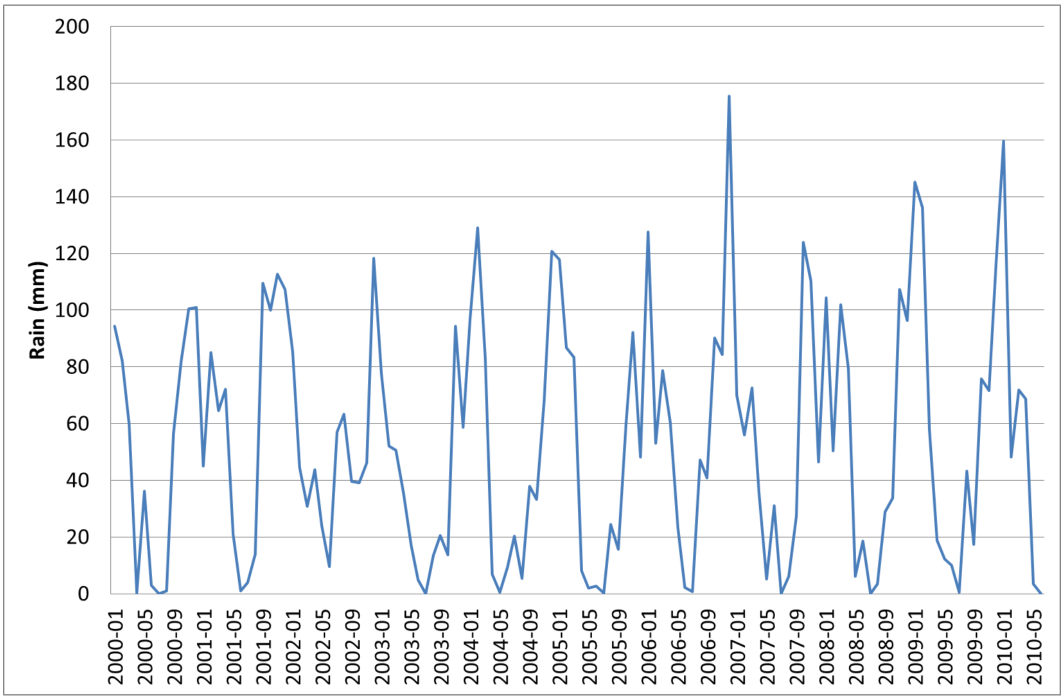

Figure 2.13: Average of the rainfall in millimeters of Ixopo, Ladysmith, Newcastle, Pietermaritzburg and Vryheid.

Table 2.13: Basic descriptive statistics for Average Minimum Daily Temperatures for the five locations

	IXO	***LDS***	***PMB***	***NWC***	***VRY***
Number of Observations	112	126	126	126	126
Missing	5	6	0	0	0
Minimum Observation	2.0	0.8	4.8	1.7	2.0
Maximum Observation	17.3	27.2	20.0	17.4	17.3
Range	15.3	26.4	15.2	15.7	15.3
Mean	10.3861	10.5817	13.3754	11.0206	10.5810
Median	11.20	12.05	14.40	12.30	11.35
Mode	11.10	14.20	18.50	15.50	14.70
Variance	22.2523	26.6602	19.5495	4.5668	16.9044
Standard Deviation	4.7172	5.1633	4.4215	20.8561	4.1115

Table 2.14: Basic descriptive statistics for Monthly Rainfall for the five locations

	IXO	***LDS***	***PMB***	***NWC***	***VRY***
Number of Observations	112	126	126	126	126
Missing	18	6	0	0	0
Minimum Observation	0.0	0.0	0.0	0.0	0.0
Maximum Observation	254.0	293.6	192.8	221.8	199.8
Range	254.0	293.6	192.8	221.8	199.8
Mean	60.8500	54.7159	67.5603	43.5760	34.9952
Median	53.50	35.60	58.90	26.60	16.80
Mode	0.00	0.00	0.00	0.00	0.00
Variance	2458.0413	3318.1321	3240.3359	2315.2991	1895.4486
Standard Deviation	49.5786	57.6032	56.9240	48.1176	43.5368

Chapter 3

Generalized Linear Models

3.1 Introduction

Consider the classical linear model for a continuous Gaussian or Normal response Y. Let $y_1, y_2, ..., y_n$ be a random sample from Y with corresponding predictor variables $x_1, x_2, ..., x_p$.
A linear model for such data is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \epsilon_i = \boldsymbol{x}_i' \boldsymbol{\beta}, \tag{3.1}$$

where $\boldsymbol{x}_i' = (1, x_{i1}, ..., x_{ip})$ and $\beta_0, \beta_1, ..., \beta_p$ are the regression coefficients. The general linear model allows the inclusion of both continuous and categorical predictor variables.

In matrix form the model for all the data is compactly written as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is the $n \times p$ design matrix and $\boldsymbol{\epsilon}$ is a vector of measurement errors. Let $\boldsymbol{I}$ be the $n \times n$ identity matrix. It is commonly assumed that $\boldsymbol{\epsilon} \sim MVN(0, \sigma^2 \boldsymbol{I})$ which implies that the y_i's are mutually independent. If the independence assumption is not necessarily true then we let $\boldsymbol{\epsilon} \sim MVN(0, \boldsymbol{V})$ where $\boldsymbol{V}$ is the variance-covariance matrix. An alternative form of the model above is

$$E(y_i) = \boldsymbol{x}_i' \boldsymbol{\beta}. \tag{3.3}$$

Here the mean is directly related to a linear predictor $\boldsymbol{x}_i'\boldsymbol{\beta}$. This version of the model specification easily extends to the case of non-Normal data, leading to Generalized Linear Models. (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989).

As an example, consider the case where data represents disease incidence and we wish to model this as a response. In this case, the response can take on discrete values above zero. Therefore the use of Normal distribution to describe it would be inaccurate, as the Normal distribution is continuous while the distribution for counts is discrete. In the case where the response is binary it too could not be related to a Normal variable. Therefore Generalized Linear Models (GLM's) due to Nelder and Wedderburn (1972) are used, where other assumptions can be made about the data's distribution. A more extensive account of GLMs is given by McCullagh and Nelder (1989). GLMs have three components associated with them. The random component specifies the distribution of the response variable. The systematic component encompasses the explanatory variables in the predictor function. The link function specifies the function which equates the mean of the response to the systematic component.

Suppose we wish to model a set of response variables Y_i with means μ_i $i = 1, 2, ..., N$ which are known to follow a certain distribution (not necessarily Normal) as a function of several variables $X_{1i}, X_{2i}, ..., X_{pi}$. GLMs are models of the form

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{3.4}$$

$g(\mu_i)$ is a monotonic, differentiable function known as the link function, which is chosen dependent on the distribution of Y. β_j, $j = 0, 1, 2, ..., p$ are the set of $p + 1$ regression parameters for the model, and the sum $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ is the systematic component of the model.

In matrix form, this can be written as $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ where

$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ . \\ . \\ . \\ \eta_N \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}$, and $\boldsymbol{X}$ is the $N \times (p+1)$ matrix of explanatory variables.

The i^{th} component of the vector η relates $g(\mu_i)$ to the linear predictor with covariates from the i^{th} unit or observation.

3.2 Exponential Family of Distributions

For the classical GLM (Nelder and Wedderburn, 1972) it is necessary that the response variable be from a distribution that belongs to the exponential family. In other words, the probability mass function can be written in the form:

$$f(y_i, \theta) = \exp[(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)], \tag{3.5}$$

where θ is referred to as the natural parameter, and for the canonical link function $g(\mu_i) = \theta_i$ and therefore $\theta_i = \beta_0 + \sum_j \beta_j x_{ij}$. It is usually sufficient to assume $a(\phi) = \phi$. ϕ is called the dispersion parameter, and affects the variance of the response Y

(see section 3.4). If $\phi \neq 1$, the model is said to be either overdispersed ($\phi > 1$) or underdispersed ($\phi < 1$). However, for standard distributions such as the Binomial and Poisson it is usually assumed that $\phi = 1$. Many distributions including Normal, Exponential, Gamma, Chi-Square, Poisson, and Binomial belong to the Exponential family of distributions.

In the case of counts data for Poisson or Binomial models we normally assume $\phi = 1$. If $\phi \neq 1$ it means we cannot fully specify the probability models and hence the likelihood of the data except perhaps to make assumptions about the first two moments. This is precisely the reason that led Wedderburn (1974) to develop the idea of quasi-likelihood. Jorgensen (1986) showed that there is no GLM family on the positive integers that satisfies the mean-variance relationship $V(\mu) = \phi\mu$ with $\phi > 1$.

3.3 Log-Likelihood Equation and Deviance

In what follows, we assume the likelihood of interest exists. The log-likelihood equation is denoted as $\ell = \sum_i l_i$, where $\ell_i = \log f(y_i; \theta_i, \phi_i)$. If the distribution is from the exponential family, this can be simplified using Equation (3.5) to

$$\ell_i = \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi) \tag{3.6}$$

We denote the log likelihood for means $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_N)'$ by $l(\boldsymbol{\mu}, \boldsymbol{y})$.

The Deviance is then defined as shown in Equation (3.7), where $\ell(\hat{\boldsymbol{\mu}}, \boldsymbol{y})$ is the maximum possible log-likelihood for the model in question, and $\ell(\boldsymbol{y}; \boldsymbol{y})$ is the log-likelihood for the saturated model. Because the saturated model has a parameter for every point it will have perfect fit, and hence the estimated mean for each observation will be the point itself, that is $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$.

Since the deviance shows the difference between the log-likelihoods of the saturated model and the model to be tested, it is said to be a measure of 'lack of fit' of the model. It is therefore important in checking the fit of the model as a smaller deviance indicates a model with better fit. Deviance is given by

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2\log\{\frac{\text{maximum.likelihood.for.the.model}}{\text{maximum.likelihood.for.saturated.model}}\} = -2[\ell(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - \ell(\boldsymbol{y}; \boldsymbol{y})]. \tag{3.7}$$

From Equation (3.6) we can see that if the maximum likelihood estimates of θ and μ from the model of interest are denoted $\hat{\theta}$ and $\hat{\mu}_i$, and the estimate of θ from the saturated model as $\tilde{\theta}$, then from Equation (3.7) we can write the deviance in the form:

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= -2\{\ell(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - \ell(\boldsymbol{y}; \boldsymbol{y})\}, \qquad (3.8) \\ &= 2\sum_i \left\{ \frac{[y_i\tilde{\theta}_i - b(\tilde{\theta}_i)]}{a(\phi)} \right\} - 2\sum_i \left\{ \frac{[y_i\hat{\theta}_i - b(\hat{\theta}_i)]}{a(\phi)} \right\}, \end{aligned}$$

$$= \frac{2}{a(\phi)} \sum_i \{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}.$$

If we assume that $a(\phi) = \phi/\omega_i$, which is usually a reasonable assumption, then from Equation (3.8) we can write:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_i \frac{\omega_i}{\phi}[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]. \tag{3.9}$$

The statistic $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/\phi$ is called the scaled deviance. This is a more general approach and holds when the dispersion parameter is not equal to 1. As with the deviance in Equation (3.7), a smaller scaled deviance will indicate a model with a better fit. To compare nested models therefore we can use change in scaled deviance.

3.4 Mean and Variance of the GLM

From the definition of the log-likelihood given in Equation (3.6), we can get the first and second order partial derivatives with respect to θ_i:

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{[y_i - b'(\theta_i)]}{\phi} \quad \text{and} \quad \frac{\partial^2 \ell_i}{\partial^2 \theta_i} = \frac{-b''(\theta_i)}{\phi}, \tag{3.10}$$

assuming $a(\phi) = \phi$.

The general likelihood results state that

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0, \quad \text{and} \quad -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = E\left(\frac{\partial \ell}{\partial \theta}\right)^2. \tag{3.11}$$

For observation i having log-likelihood equation $\ell_i = \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi)$, we have $E[(Y_i - b'(\theta_i))/a(\phi)] = 0$ and so:

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{3.12}$$

and from the second equation:
$b''(\theta_i)/\phi = E[(Y_i - b'(\theta_i))/\phi]^2 = \text{var}(Y_i)/[\phi]^2$ and therefore:

$$\text{var}(Y_i) = b''(\theta_i)\phi \tag{3.13}$$

If $\phi > 1$, in other words overdispersion exists, then $var(Y_i) > b''(\theta_i)$. This is how overdispersion affects the variance function. For example under a Poisson model $b(\theta_i) = e^{\theta_i}$ which means that

$$\mu_i = b'(\theta_i) = e^{\theta_i}. \tag{3.14}$$

It follows therefore that

$$Var(Y_i) = \phi b''(\theta_i) = \phi e^{\theta_i} = \phi \mu_i. \tag{3.15}$$

Therefore when $\phi = 1$, $Var(Y_i) = v(\mu_i) = \mu_i$ and the variance is equal to the mean. But under an overdispersed model $var(Y) > \mu$ if $\phi > 1$. Thus standard errors assuming $\phi = 1$ when in reality $\phi > 1$ will be underestimated.

3.5 Asymptotic Covariance Matrix for $\boldsymbol{\beta}$

To derive the asymptotic covariance matrix for $\boldsymbol{\beta}$, first we must define the score equations for the GLM formulation. The score equations are of the form

$$\partial \ell(\boldsymbol{\beta})/\partial \beta_j = \sum_i \partial \ell_i / \partial \beta_j = 0. \tag{3.16}$$

Using the chain rule of differentiation, and Equation (3.6), we find the likelihood equations to be

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{\mathrm{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad j = 1, 2, ..., p \tag{3.17}$$

A useful result from Cox and Hinkley (1974) states that for distributions from the exponential family:

$$E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) = -E\left(\frac{\partial \ell_i}{\partial \beta_j}\right)\left(\frac{\partial \ell_i}{\partial \beta_k}\right). \tag{3.18}$$

Using this, and the result in Equation (3.17), we find that

$$E\left(-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{3.19}$$

In a compact matrix form, the information matrix $\boldsymbol{J}$ has the form

$$\boldsymbol{J} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} \tag{3.20}$$

where $\boldsymbol{W}$ is a diagonal matrix with elements $w_i = (\partial \mu_i / \partial \eta_i)^2 / var(Y_i)$. Thus it can be seen that the elements of $\boldsymbol{W}$ depend on the link function. The asymptotic covariance matrix for $\boldsymbol{\beta}$ is the inverse of this matrix, estimated by

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{J}}^{-1} = (\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}. \tag{3.21}$$

3.6 Fitting the GLM

For Normal data, Ordinary Least Squares (OLS) method estimates are equivalent to finding Maximum Likelihood Estimates (MLE's) for the parameters β_i. This method serves to minimize the sum of squared deviances of the fitted line. The MLE for $\boldsymbol{\beta}$ is found to be

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y. \tag{3.22}$$

However, when Normality of the response data does not hold, the score equations are usually non-linear in $\boldsymbol{\beta}$ and therefore iterative methods are necessary. This requires the use of Iterative Weighted Least Squares methods (IWLS).
The model can be fitted using the Newton-Raphson method, amongst other possible methods. This is an iterative method shown in Equation (3.23).

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{u}^{(t)}, \tag{3.23}$$

where $\boldsymbol{\beta}^{(t)}$ is the estimate for $\boldsymbol{\beta}$ at the t^{th} iteration, $\boldsymbol{H}^{t}$ is called the Hessian matrix (which is assumed to be non-singular), having components

$$h_{ab} = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\beta_a \partial\beta_b}, \tag{3.24}$$

and

$$\boldsymbol{u}^{(t)} = \begin{pmatrix} \partial\ell(\boldsymbol{\beta})/\partial\beta_1 \\ \partial\ell(\boldsymbol{\beta})/\partial\beta_2 \\ . \\ . \\ . \\ \partial\ell(\boldsymbol{\beta})/\partial\beta_p \end{pmatrix}. \tag{3.25}$$

The Newton-Raphson method comes about by taking the Taylor expansion of $L(\boldsymbol{\beta})$ and taking the first order derivative, and then equating to zero. That is,

$$\ell(\boldsymbol{\beta}) \approx \ell(\boldsymbol{\beta}^{(t)}) + \boldsymbol{u}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + (\frac{1}{2}((\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})'\boldsymbol{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) \tag{3.26}$$

$$\partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \approx \boldsymbol{u}^{(t)} + \boldsymbol{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = \mathbf{0} \tag{3.27}$$

which, when solved for $\boldsymbol{\beta}$, gives us the equation shown in (3.23), provided $\boldsymbol{H}^{(t)}$ is invertible.

An alternative method is the Fisher Scoring equation which is given in Equation (3.28). Let $\boldsymbol{J}$ denote the information matrix with elements $-E(\partial^2\ell(\boldsymbol{\beta})/\partial\beta_i\partial\beta_j)$, in other words $-\boldsymbol{J} = E(\boldsymbol{H})$. The algorithm in Equation (3.23) can now be written as:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\boldsymbol{J}^{(t)})^{-1}\boldsymbol{u}^{(t)}, \tag{3.28}$$

where the information matrix $\boldsymbol{J}$ is defined as in Equation (3.20) and $\boldsymbol{W}$ is a diagonal matrix with elements $w_i = (\partial\mu_i/\partial\eta_i)^2/var(Y_i)$. Then

$$\boldsymbol{J}^{(t)} = \boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X}, \tag{3.29}$$

where $\boldsymbol{W}^{(t)}$ is evaluated at $\boldsymbol{\beta}^{(t)}$.

An initial estimate for $\boldsymbol{\beta}$ must be substituted into the equation for $\boldsymbol{\beta}^{(0)}$. The iterative cycle is computed until changes in $\ell(\boldsymbol{\beta}^{(t)})$ are small. The local maximum of $\ell(\boldsymbol{\beta})$ will then have been reached, and the derivative will be approximately zero. At this stage, since $\boldsymbol{u}^{(t)} \approx 0$, $\boldsymbol{\beta}^{(t+1)} \approx \boldsymbol{\beta}^{(t)}$, the iterative process may halt. The number of iterations required to calculate the maximum likelihood estimate of $\boldsymbol{\beta}$ depends on the accuracy of the initial estimate.

SAS proc GENMOD utilizes the Fisher scoring method up to a certain iteration (by default 1) which can be specified by the SCORING option in the MODEL statement, after which it uses the Newton-Raphson method until convergence.

When the canonical link is used, however, it can be shown that the two methods are identical since $\boldsymbol{H} = -\boldsymbol{J}$. To show this, note that the log-likelihood contribution from observation i is given by

$$\ell_i = \frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c, \tag{3.30}$$

and

$$\theta_i = \sum \beta_j x_{ij}, \tag{3.31}$$

then

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i x_{ij} - b'(\theta_i)x_{ij}}{a(\phi)} = \frac{(y_i - \mu_i)x_{ij}}{a(\phi)} \tag{3.32}$$

and

$$\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k} = -\left(\frac{\partial \mu_i}{\partial \beta_k}\right)\frac{x_{ij}}{a(\phi)}. \tag{3.33}$$

It can be seen that since this does not depend on the observation y_i, the value of $\partial\ell(\boldsymbol{\beta})/\partial\beta_j\partial\beta_k$ will be equal to its expectation, and $\boldsymbol{H} = -\boldsymbol{J}$, therefore making the Fisher scoring and Newton-Raphson methods equivalent.

3.7 Testing Goodness of Fit

For the special case where $\phi = 1$ (ie. there is no overdispersion), nested models can be tested against each other. Consider two models, where model 2 is nested in model 1. In other words, Model 1 has p explanatory variables $X_1, X_2, ..., X_p$ and Model 2 has q explanatory variables $X_1, X_2, ..., X_q$ where $q < p$ and the $X_1, X_2, ..., X_q$ are a subset of $X_1, X_2, ..., X_p$. Model 1 has fitted values $\hat{\boldsymbol{\mu}}_{\mathbf{1}}$, deviance $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_{\mathbf{1}})$, and log-likelihood $\ell(\hat{\boldsymbol{\mu}}_{\mathbf{1}})$. Model 2 has fitted values $\hat{\boldsymbol{\mu}}_{\mathbf{2}}$, deviance $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_{\mathbf{2}})$, and log-likelihood $\ell(\hat{\boldsymbol{\mu}}_{\mathbf{2}})$. Since Model 2 has fewer explanatory variables, it cannot have a greater log-likelihood. Thus $\ell(\hat{\boldsymbol{\mu}}_{\mathbf{2}}) \leq \ell(\hat{\boldsymbol{\mu}}_{\mathbf{1}})$. It will then follow that $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_{\mathbf{2}}) \geq D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_{\mathbf{1}})$.

To test the models against each other we use:

$$\begin{aligned} -2[\ell(\hat{\boldsymbol{\mu}}_2; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y})] &= -2[\ell(\hat{\boldsymbol{\mu}}_2; \boldsymbol{y}) - \ell(\boldsymbol{y}; \boldsymbol{y})] - 2[\ell(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y}) + \ell(\boldsymbol{y}; \boldsymbol{y})] \quad (3.34) \\ &= D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_2) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1). \end{aligned}$$

When this difference is large, Model 2 fits poorly when compared to Model 1. If one finds this difference close to zero, it implies that Model 2 fits very nearly as well as Model 1. This test statistic is approximately χ^2 distributed with degrees of freedom equal to $q-p$ (Agresti, 2002). Thus if $\chi^2_{\text{calculated}} > \chi^2_{\alpha;p-q}$, we will conclude that Model 2 is significantly worse in fit than Model 1.

3.8 Binomial GLM

Suppose we are interested in n identical trials which have a binary outcome - denoted as either success or failure. Define a response variable y_i where $y_i = 1$ if success, and $y_i = 0$ if failure for $i = 1, 2, ..., n$. The probability of success $P(Y_i = 1) = \pi$ and the probability of failure $P(Y_i = 0) = 1 - \pi$.
The Binomial distribution has random variable which is the total number of successes in n trials, that is $Y = \sum_{i=1}^{n} y_i$, and has a probability mass function given as

$$p(y; \pi, n) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y = 1, 2, ..., n, \tag{3.35}$$

where y is the number of successes, $\binom{n}{y} = \frac{n!}{y!(n-y!)}$, and n is the number of trials.
Equation (3.35) can be written in the form of (3.5) as:

$$p(y; \pi, n) = \exp\{y \log(\frac{\pi}{1-\pi}) + n \log(1-\pi) + \log \binom{n}{y}\}, \tag{3.36}$$

with $\theta_i = \log(\frac{\pi}{1-\pi})$, $b(\theta_i) = -n\log(1-\pi)$. Hence we find that $\mu_i = b'(\theta_i) = n\pi$, and $var(Y_i) = b''(\theta_i)a(\phi) = n\pi(1-\pi)$ as expected for a Binomial distribution, with $a(\phi) = 1$.

Relating the probability directly to a normal regression model would incur major problems. Take, for example, a simple linear regression model. If the equation was structured as $\pi(x) = \beta_0 + \beta_1 x$, then we cannot rule out values of x for which $\pi(x) < 0$ or $\pi(x) > 1$. This has no statistical meaning, as the range of probabilities is expected to lie between 0 and 1. The GLM formulation of this problem, with the logit link, has the distinct advantage that the probability is restricted to its natural range.
The GLM can be written as

$$\text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{3.37}$$

or in matrix form as:

$$\text{logit}(\boldsymbol{\pi}) = \boldsymbol{X\beta}, \tag{3.38}$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}, \tag{3.39}$$

are regression coefficients, and $\boldsymbol{X}$ is the $n \times (p+1)$ matrix of explanatory variables, with the first column consisting of ones corresponding to the intercept coefficient β_0.

The probability $\hat{\pi}(\boldsymbol{x}_i)$ can be found using the following equation:

$$\hat{\pi}(\boldsymbol{x}_i) = \frac{\exp(\sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\sum \beta_j x_{ij})}, \tag{3.40}$$

where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$. Note that the regression coefficients β_j $j = 0, 1, 2, ..., p$ are interpreted in terms of the logit scale.

3.9 Poisson GLM for Counts Data

Very often the distribution used for counts data is the Poisson distribution, characterized by the probability mass function given in Equation (3.41) where the mean is given by μ. This is because counts may take on non-negative integer values, and therefore the Poisson distribution is the more realistic distribution given by

$$p(y_i; \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \qquad y_i = 1, 2, \ldots \tag{3.41}$$

Equation (3.41) can be re-written in the form of Equation (3.42), which has natural exponential form with $\theta_i = \log(\mu)$, and $b(\theta_i) = \mu$. That is,

$$f(y_i, \mu_i) = \exp\{y_i \log(\mu_i) - \mu_i - \log(y_i!)\}. \tag{3.42}$$

The mean can therefore be found to be μ_i, and the variance μ_i which is the expected outcome for the Poisson distribution. The canonical link function θ is the log link. Poisson distributed data can therefore be modeled using a log-link GLM. Therefore, the GLM equation for Poisson data is given by

$$\log \mu_i = \sum_{j=1}^{p} \beta_j x_{ij}, \qquad i = 1, 2, \ldots, N. \tag{3.43}$$

or in matrix form given by

$$\log(\boldsymbol{\mu}) = \boldsymbol{X\beta}, \tag{3.44}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$, are regression coefficients, and $\boldsymbol{X}$ is the $n \times (p+1)$ matrix of explanatory variables. One advantage of the GLM formulation is that the predictor variables can include both continuous and categorical variables.

3.10 Poisson Generalized Linear Model for African Horse Sickness disease incidence over time

3.10.1 African Horse Sickness Trust Data

We intend to model the number of cases using a Generalized Linear Model for counts data. Assuming a Poisson distribution for the counts data - we define the response as the number of cases in a particular month - we model the response as a function of the year and month of occurrence as categorical variables.

First we attempt to model the total cases for KZN as a function of year and month. Since month runs from 1 to 12 for each year, and we expect the monthly effects to be relatively equal for each year, we treat it as a nested variable in year. However, in modeling this scenario, it is found that the model is saturated, meaning there are as many estimated parameters as data points. The fact that there are zero degrees of freedom for the deviance indicates the fact that the model is saturated.

Since this model is found to be saturated, we try to model the cases as a function of additive effects of year and month. Both year and month are treated as categorical variables, but unlike the above model we do not treat Month as though it is nested in Year. The results are shown in Table 3.1.

This model is unsaturated, however we do have the problem that the negative of the Hessian is not positive definite. Therefore the fit of this model will be questionable.

In the previous two models both month and year were treated as categorical variables. However this does not take into account that predicting a categorical effect for a year in the future is impossible, and therefore these models have no value as predictive models. It is therefore important to treat year as a continuous variable, as prediction is an important aspect of this model.

We try to model instead treating month as categorical and year as continuous. However it is clear that year cannot be taken as a linear term, as the number of cases fluctuates greatly within each year. We therefore wish to introduce a trigonometric term for year which will take into account these seasonal fluctuations.

In a similar manner to that used in Lord (1996), we use the term $sin(2\pi t)$ in the GLM. In the paper by Lord, the abundance of the vector was modeled using the function $N(t) = N(0)e^{\mu\delta\sin(\theta t)}$ or $\log(N(t)) = \log(N(0)) + \mu\delta\sin(\theta t)$ where μ and δ were constants, $N(0)$ the initial population of the vector, and θ the scaling parameter (to ensure the period of one year). In this case t is a continuous variable in years (1/12, 2/12, 3/12...), and $\theta = 2\pi = 6.28319$. Since we use the log link, we can fit the term $sin(2\pi t)$ to the same effect. The results of fitting this as a single term are shown in Table 3.2. The term

Table 3.1: SAS results for the model expressing cases of AHS in KZN as a function of year and month.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	38	152.8661	4.0228
Scaled Deviance	38	152.8661	4.0228
Pearson Chi-Square	38	161.5248	4.2507
Scaled Pearson X2	38	161.5248	4.2507
Log Likelihood		423.9170	

WARNING: Negative of Hessian not positive definite.

$sin(2\pi t)$ when fitted alone is found to be significant (p<0.0001), with a coefficient of 3.0250. A plot of the observed and predicted cases over time is shown in Figure 3.10.1.

Table 3.2: SAS output for GLM expressing AHS cases in KZN as a funtion of Sin(2πt)

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	53	263.7222	4.9759
Scaled Deviance	53	263.7222	4.9759
Pearson Chi-Square	53	318.6727	6.0127
Scaled Pearson X2	53	318.6727	6.0127
Log Likelihood		368.4890	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Error	Wald 95% CI		χ^2	Pr > χ^2
Intercept	1	-0.1004	0.2011	-0.4946	0.2937	0.25	0.6175
sinyr	1	3.0250	0.2347	2.5650	3.4849	166.16	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

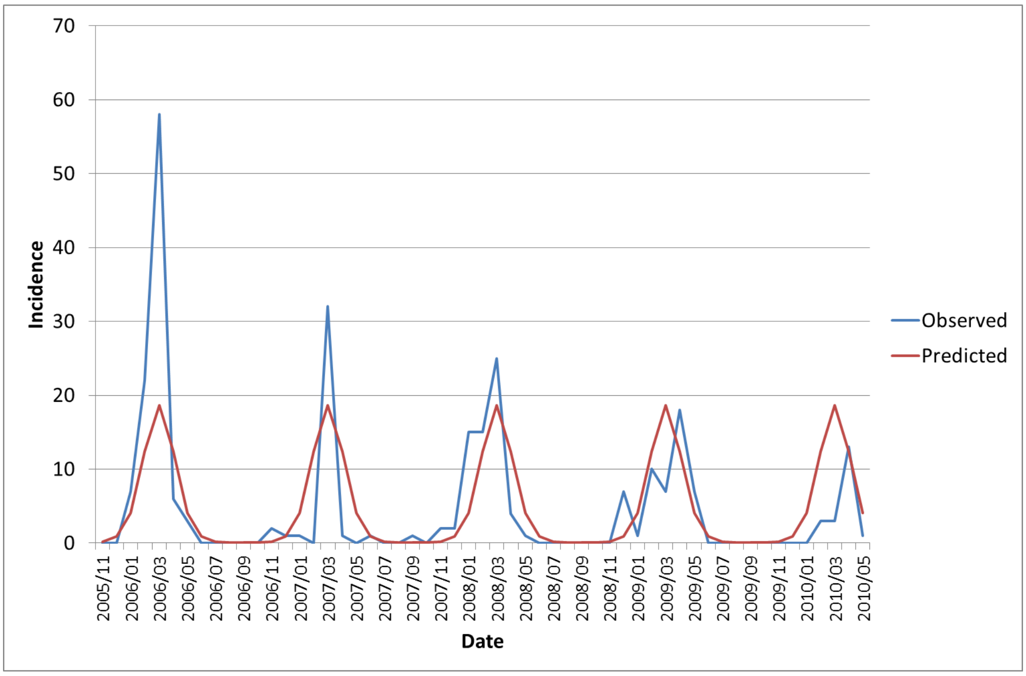

Figure 3.1: Plot of predicted and observed cases against time for the model $\log\{E(Cases)\} = \beta_0 + \beta_1 sin(2\pi t)$

Next we fit the categorical month term along with $sin(2\pi t)$. The results are shown in Table 3.3. Again we are warned that the negative of the Hessian is not positive definite, and we should therefore not use this model.
A model with categorical month as the only explanatory variable is then fitted. The results are shown in Table 3.4. Month 12 is held as the reference category, and the months that are found to be significantly different from the reference category are months 1, 2, 3, 4 and 9. This correlates well with what we see in the plot of cases against time - as these are the months which have the highest number of cases.

Although the model with month as the explanatory variable appears to fit better than that with $sin(2\pi t)$, as it has a lower deviance (222.1145 against 263.7222) and a higher log-likelihood (389.2928 as opposed to 368.4890), the fit is questionable since the negative of the Hessian is not positive definite. We also find that the model with month as explanatory variable does not fit significantly better than that with $sin(2\pi t)$ only, as the difference in log likelihood is $182.5461 - 171.9037 = 10.6424 < \chi^2_{(0.05,11df)} = 19.6752$.

For these reasons we prefer to use the term $sin(2\pi t)$, as it takes account of the seasonal trend. Since it is a continuous variable it uses fewer degrees of freedom and so is also a more parsimonious model than the model with categorical variable month (which has 11 degrees of freedom). It also acts as a smoothing variable. It is expected that once the weather variables are added to the model, it will explain more of the variation.

Table 3.3: SAS output for GLM expressing cases of AHS in KZN as a function of Month + Sin(2πt)

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	43	222.1145	5.1655
Scaled Deviance	43	222.1145	5.1655
Pearson Chi-Square	43	208.1408	4.8405
Scaled Pearson X2	43	208.1408	4.8405
Log Likelihood		389.2928	

WARNING: Negative of Hessian not positive definite.

Table 3.4: SAS output for GLM expressing cases of AHS in KZN as a function of month

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	43	222.1138	5.1654
Scaled Deviance	43	222.1138	5.1654
Pearson Chi-Square	43	208.1400	4.8405
Scaled Pearson X2	43	208.1400	4.8405
Log Likelihood		389.2932	

WARNING: Negative of Hessian not positive definite.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Error	Wald 95% CI		χ^2	Pr > χ^2
Intercept		1	0.6931	0.3162	0.0734	1.3129	4.80	0.0284
Month	1	1	0.8755	0.3764	0.1378	1.6132	5.41	0.0200
Month	2	1	1.6094	0.3464	0.9305	2.2884	21.59	<.0001
Month	3	1	2.5257	0.3286	1.8816	3.1698	59.07	<.0001
Month	4	1	1.4351	0.3519	0.7454	2.1247	16.63	<.0001
Month	5	1	0.1823	0.4282	-0.6569	1.0215	0.18	0.6702
Month	6	1	-2.0794	1.0488	-4.1351	-0.0238	3.93	0.0474
Month	7	1	-24.3863	69802.59	-136835	136786.2	0.00	0.9997
Month	8	1	-24.3863	69802.59	-136835	136786.2	0.00	0.9997
Month	9	1	-2.0794	1.0488	-4.1351	-0.0238	3.93	0.0474
Month	10	1	-24.3863	69802.59	-136835	136786.2	0.00	0.9997
Month	11	1	-0.9163	0.5916	-2.0758	0.2432	2.40	0.1214
Month	12	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

3.10.2 South African Weather Service Data

The South African Weather Service data from Newcastle, Ixopo, Pietermaritzburg, Ladysmith and Vryheid weather stations formed the basis for the weather variables used. We were limited in the number of cities or towns from which data could be gained. These locations were chosen due to their relative abundance of cases of AHS, as well as their spread over Kwa-Zulu Natal. This ensured that the averages used would be as close to representative of the entire province as was achievable.

For each of the locations, monthly average maxima and average minima of temperatures were supplied, as well as the total monthly rainfall in millimeters. The monthly average temperature was then calculated by taking the mean of the maximum and minimum temperatures for that month. The temperature variables were averaged in order to get the variables TMax and TMin. The rainfall variable was also averaged to get the variable Rain.

As explained in the previous section, the variable $sin(2\pi t)$ was chosen as the best variable to account for the seasonal variation, and therefore was fitted along with the defined weather variables in a Poisson GLM. Results are shown in Table 3.5. It can be seen that all variables are significant at a confidence level of $\alpha = 0.05$.

The overall model can be expressed as:

$$\log(\mu) = 9.3447 + 2.5690\sin(2\pi t) - 0.5674.TMax + 0.4399.TMin - 0.0060.Rain \quad (3.45)$$

A plot of the observed as well as predicted values according to this model are shown in Figure 3.2. It can be seen that the variation in the number of cases is explained fairly well, with some small differences in the predicted peaks. These differences are difficult to accurately account for and capture in the model.

There may also, however, be dependencies that cannot be accurately measured. For example, it is unknown if the herd immunity was for some reason increased in years which had small outbreaks despite the climatic variables being favourable for propagation of the disease. There are also, as described in Baylis, Mellor and Meiswinkel (1999), correlations between large outbreaks and the warm phase of the El Niño Southern Oscillation (ENSO).

3.10.3 Model Checking

McCullagh and Nelder (1983) discuss ways of checking that the model assumptions have not been violated. Figures 3.3 and 3.4 show the model checking plots for the above GLM in Equation (3.45). Figure 3.3 is the Q-Q Plot for the standardized residual deviance for the model. This shows slight deviation from a straight line in the center, but elsewhere it seems to adhere well to a line with slope approximately equal to 1. This fit validates our choice of distribution, as well as showing that the GLM assumptions were not seriously violated.

Table 3.5: SAS output for Poisson GLM for cases of AHS in KZN as a function of time and weather variables

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	50	207.8348	4.1567
Scaled Deviance	50	207.8348	4.1567
Pearson Chi-Square	50	252.8398	5.0568
Scaled Pearson X2	50	252.8398	5.0568
Log Likelihood		396.4327	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Error	Wald 95% CI		χ^2	Pr > χ^2
Intercept	1	9.3447	1.4826	6.4388	12.2505	39.73	<.0001
sinyr	1	2.5690	0.2380	2.1025	3.0355	116.52	<.0001
Tmax	1	-0.5674	0.0846	-0.7331	-0.4017	45.03	<.0001
Tmin	1	0.4399	0.0694	0.3038	0.5759	40.15	<.0001
Rain	1	-0.0060	0.0025	-0.0110	-0.0010	5.58	0.0182
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Figure 3.4 is a plot of the Standardized Residual Deviance for the model. We require the scatter of the points to be completely random and centered around zero. Although there does appear to be some clustering towards the bottom, a systematic component was not identified, showing that the variance function used was adequate.

3.10.4 Interaction and Quadratic Effects

Although the model given in Equation 3.45 appears to behave reasonably well, we have not investigated interaction effects or quadratic terms. We have not considered the possibility that the effect of the variables may not be strictly linear on the incidence.

We therefore investigate certain other terms for their significance. The terms included in the initial model are $sinyr, Tmax, Tmin, Rain$, quadratic terms $Tmax^2, Tmin^2, Rain^2$, and interaction terms $Tmax \times Rain, Tmin \times Rain$. We proceed in the same manner as before, iteratively dropping the least significant term. The iterative fit statistics are shown in Table 3.6. The final model results are shown in Table 3.7.

Table 3.6: Table showing model information for 'Stepwise' process removing insignificant terms from Poisson GLM for disease incidence with constant Scale parameter

	Model Information			Model Checking		Variable to be dropped		
	Log-Likelihood	Deviance	DF	Change in Deviance	$P > \chi^2_{(df)}$	Variable	Type III p-value	df
1	412.8146	175.071	45			Tmax*Rain	0.9347	1
2	412.8112	175.0777	46	0.0067	0.9348	Tmax	0.4482	1
3	412.5266	175.6469	47	0.5692	0.4506	.	.	

Table 3.7: SAS proc GENMOD results for the Poisson model for disease incidence with constant Scale parameter

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Error	Wald 95% CI		χ^2	Pr > χ^2
Intercept	1	-0.8226	1.4862	-3.7355	2.0904	0.31	0.5799
$sinyr$	1	2.1965	0.2142	1.7767	2.6163	105.15	<.0001
$Tmin$	1	1.0603	0.2500	0.5703	1.5503	17.99	<.0001
$Rain$	1	-0.0434	0.0196	-0.0819	-0.0049	4.88	0.0272
$Tmax^2$	1	-0.0099	0.0017	-0.0133	-0.0066	33.66	<.0001
$Tmin^2$	1	-0.0403	0.0112	-0.0622	-0.0185	13.07	0.0003
$Rain^2$	1	-0.0004	0.0001	-0.0006	-0.0002	11.75	0.0006
$Tmin \times Rain$	1	0.0065	0.0021	0.0024	0.0106	9.69	0.0018
$Scale$	0	1.0000	0.0000	1.0000	1.0000		

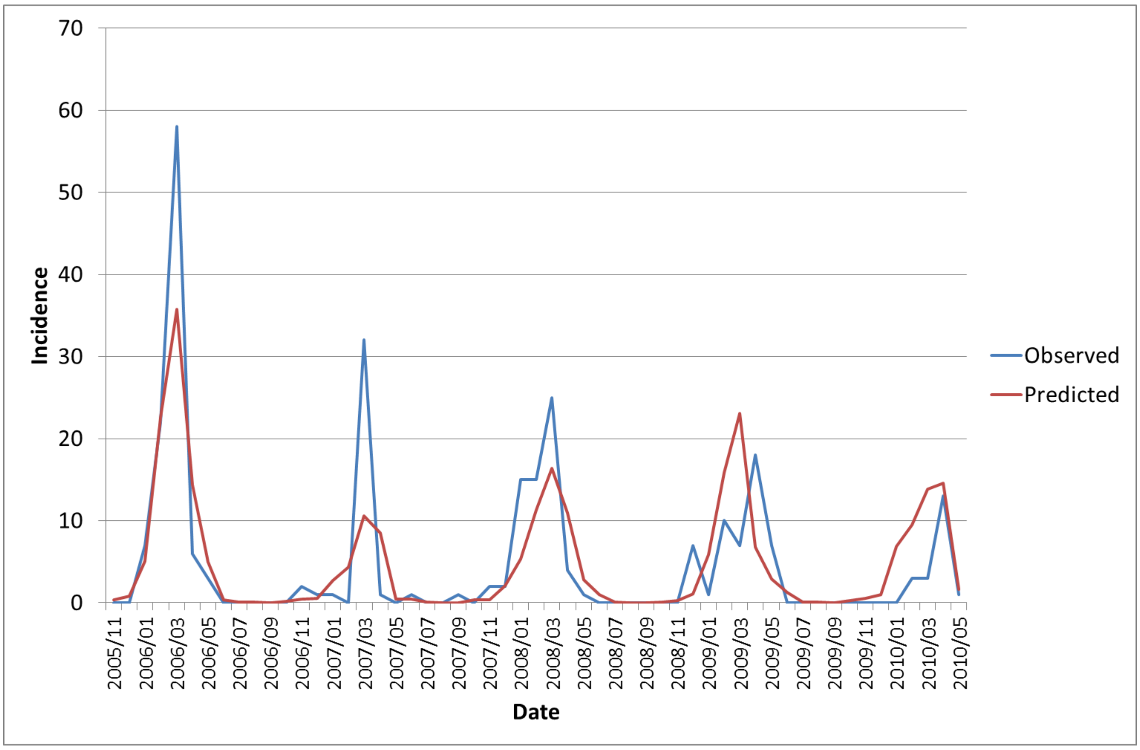

Figure 3.2: Plot of predicted and observed cases against time for the model $\log(\mu) = 9.3447 + 2.5690\sin(2\pi t) - 0.5674.TMax + 0.4399.TMin - 0.0060.Rain$

The final model equation is therefore

$$\begin{aligned} \log(\mu) &= -0.8226 + 2.1965.sinyr + 1.0603Tmin - 0.0434.Rain \\ &\quad -0.0099Tmax^2 - 0.0403.Tmin^2 - 0.0004.Rain^2 + 0.0065.Tmin \times Rain \end{aligned} \tag{3.46}$$

A plot of the predicted incidence against that which was observed is shown in Figure 3.5.

Model Checking

The same model checking procedure is followed as before. Similar results are obtained. Figures 3.6 and 3.7 show the QQ-plot and Residual plot respectively for the model in Equation (3.46). The QQ-plot in Figure 3.6 shows slight deviation from a straight line in the center, but the rest appears to adhere well to a straight line with slope 1. Figure 3.7, the plot of standardized deviance residuals, appears to show good random scatter. There is slight clustering in one section of the graph but it is not so severe that we believe the model assumptions are severely violated.

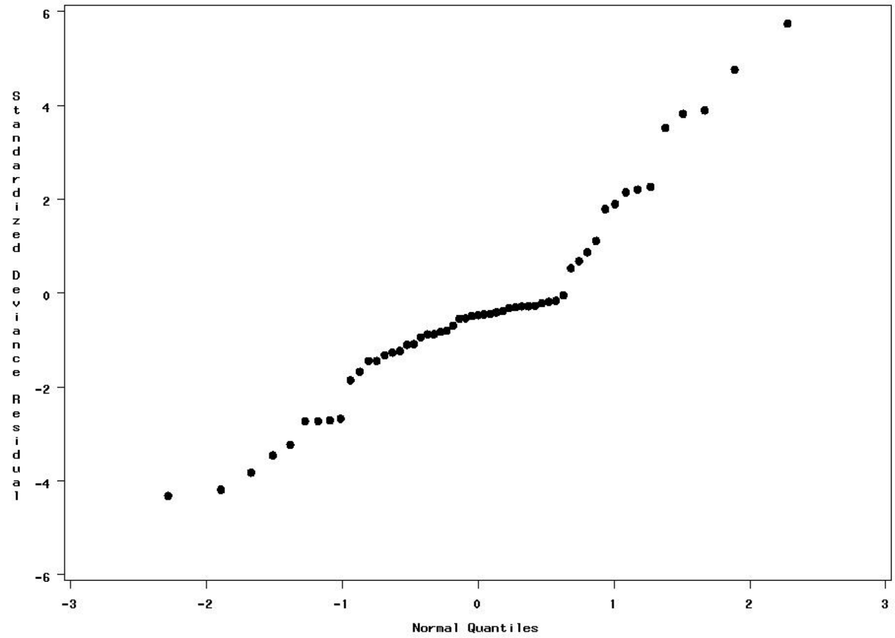

Figure 3.3: Q-Q Plot for Poisson Generalized Linear Model

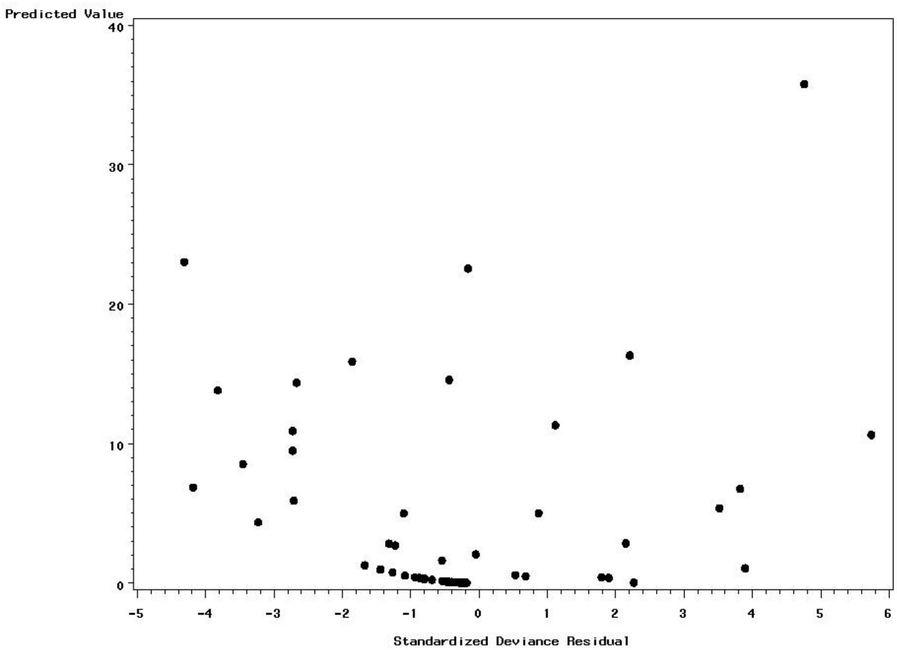

Figure 3.4: Residual Plot for Poisson Generalized Linear Model

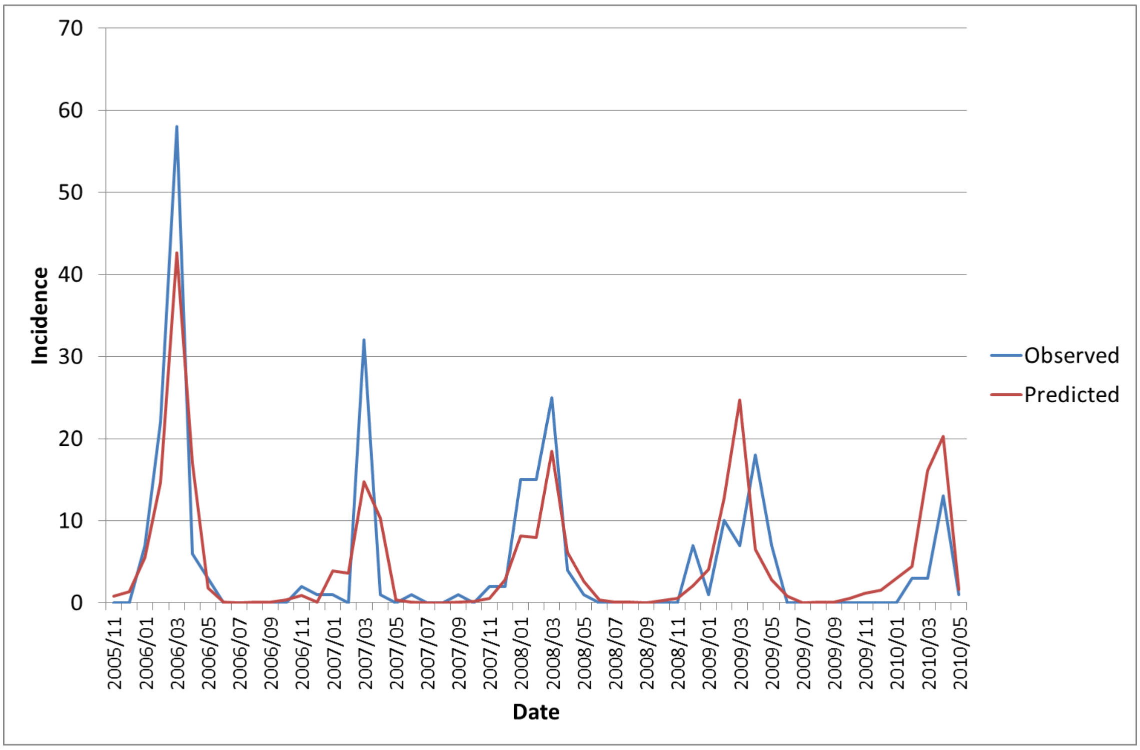

Figure 3.5: Plot of observed and predicted incidence for the model $log(\mu) = -0.8226 + 2.1965.sinyr + 1.0603Tmin - 0.0434.Rain - 0.0099Tmax^2 - 0.0403.Tmin^2 - 0.0004.Rain^2 + 0.0065.Tmin \times Rain$

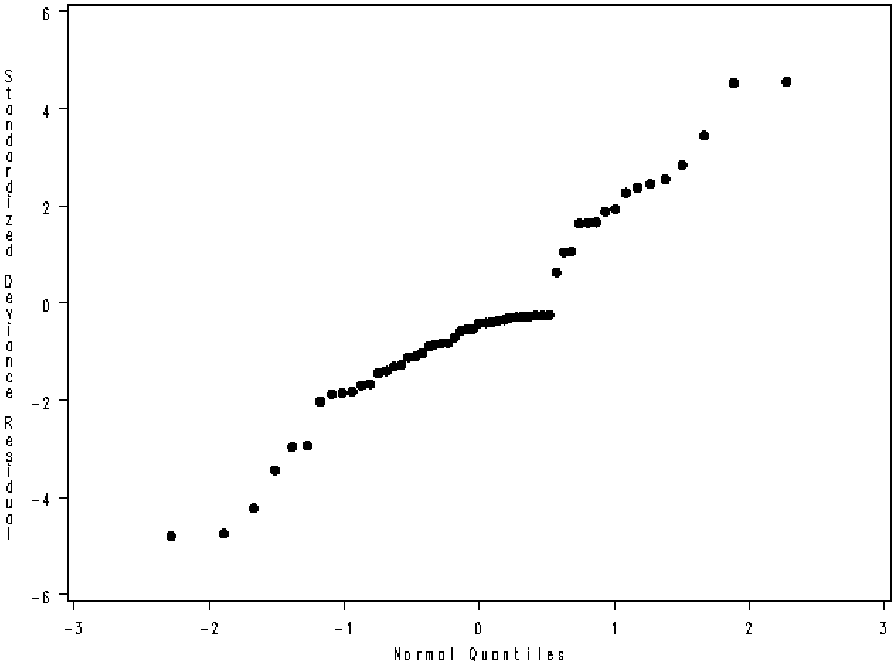

Figure 3.6: Q-Q Plot for Poisson Generalized Linear Model with interaction effects

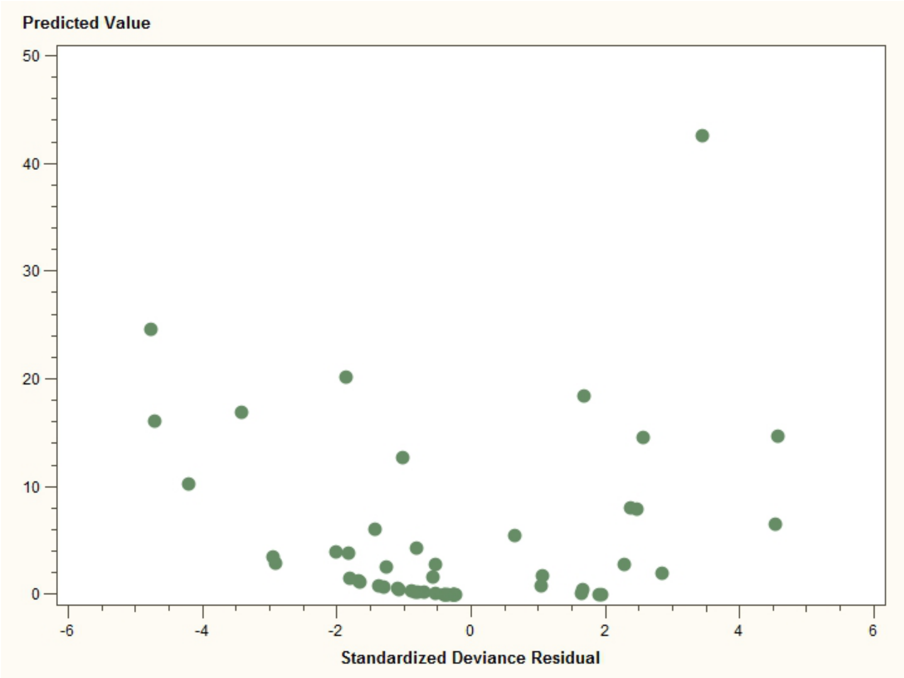

Figure 3.7: Residual Plot for Poisson Generalized Linear Model with interaction effects

Dependence of Incidence on the Explanatory Variables

Figure 3.8 shows the relation of incidence to $Tmin$ and $Rain$, holding $sinyr$ and $Tmax$ constant. A range of 0 to 20°C was chosen for $Tmin$, and 0 to 200mm for $Rain$. These ranges were chosen from their apparent natural ranges from the data. It can be seen that for extreme low minimum temperatures, incidence is minimal regardless of the rainfall. Incidence is also minimal for extremely high rainfall. As $Tmin$ increases one begins to see the dependence of incidence on both $Tmin$ and $Rain$. Favourable conditions for increased incidence appear to be with increased minimum temperatures and moderate rainfall. Incidence is maximised for maximum $Tmin$ within our range, and for $Rain$ of just over 100mm. Using Mathematica to find the exact maximum, it is found to be at $Tmin = 20°C$ and $Rain = 108.25mm$. However, if we allow $Tmin$ to vary in a wider range, the maximum is found at $Tmin = 25.4674$ and $Rain = 152.6730$, although we do not expect $Tmin$ to exceed 20°C based on our data. The maximum observed $Tmin$ was 16.96°C.

The dependence on $Tmax$ is more straightforward. Since the only dependence on this variable is $-0.0099Tmax^2$, we know that above zero this is a monotonically decreasing function. Thus minimised maximum temperatures will maximise the incidence and vice versa. Realistically, however, in our data $Tmax$ only inhabited the range from 19.88 to 29.68°C.

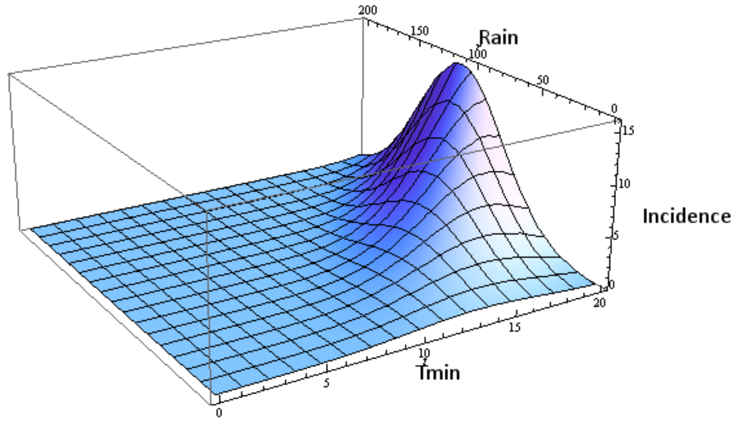

Figure 3.8: Incidence μ plotted against Tmin and Rain, with sinyr = 0 and Tmax = 25 (constant) according to the model $log(\mu) = -0.8226 + 2.1965.sinyr + 1.0603Tmin - 0.0434.Rain - 0.0099Tmax^2 - 0.0403.Tmin^2 - 0.0004.Rain^2 + 0.0065.Tmin \times Rain$. It can be seen that incidence is maximised at relatively high minimum temperatures, and with moderate rainfall.

3.10.5 Estimating the Scale Parameter

In the previous model, however, we did not account for the fact that overdispersion may have occurred. We therefore re-fit the model with this in mind, starting with the same set of explanatory variables. We choose to allow the Pearson scale parameter to be estimated. The iterative process is shown in Table 3.8.

Table 3.8: Table showing model information for 'Stepwise' process removing insignificant terms from Poisson GLM for disease incidence, with Pearson Scale parameter

	Model Information			Model Checking			Variable to be dropped		
	Log-Likelihood	Deviance	Scale	DF	Change in Deviance	$P > \chi^2_{(df)}$	Variable	Type III p-value	df
1	106.1093	175.0710	1.9724	45			$Tmax \times Rain$	0.9669	1
2	108.4622	175.0777	1.9509	46	0.0067	0.9348	$Tmax$	0.6975	1
3	110.3848	175.6469	1.9332	47	0.5692	0.4506	$Rain$	0.2533	1
4	108.6669	181.0309	1.9420	48	5.3840	0.0203	$Tmin \times Rain$	0.2010	1
5	106.1967	187.5882	1.9566	49	6.5573	0.0104	$Rain^2$	0.1833	1
6	103.5693	194.5838	1.9727	50	6.9956	0.0082	$Tmin^2$	0.0822	1
7	96.3699	209.3860	2.0262	51	14.8022	0.0001			

If a significance level of 0.05 were used, we would halt iterations at the 4th model. After *Rain* is dropped, we see the change in deviance is 5.3840 with a probability of 0.0203, which is significant. Hence we would halt iterations and use model 3, where the value of the scale parameter is 1.9332. This model minimises the scale parameter (and hence the variance), and maximises the log-likelihood - indicating that it is indeed the best model. The results for model 3 are shown in Table 3.9. Model 3 is given as

$$\begin{aligned}\log(\mu) &= -0.8226 + 2.1965.sinyr + 1.0603Tmin - 0.0434.Rain \\ &\quad -0.0099Tmax^2 - 0.0403.Tmin^2 - 0.0004.Rain^2 + 0.0065.Tmin \times Rain\end{aligned} \quad (3.47)$$

which is the same as the model without estimating the scale parameter, and the plot is the same as in Figure 3.5.

Table 3.9: SAS proc GENMOD results for the Poisson model for disease incidence with Pearson Scale parameter being estimated

Analysis Of Parameter Estimates

Parameter	Estimate	Std Error	Wald 95% CI		χ^2	p-value
Intercept	-0.8226	2.8731	-6.4538	4.8086	0.08	0.7746
sinyr	2.1965	0.4141	1.3849	3.0081	28.14	<.0001
$Tmin$	1.0603	0.4833	0.1130	2.0076	4.81	0.0283
Rain	-0.0434	0.0380	-0.1178	0.0310	1.30	0.2533
$Tmax^2$	-0.0099	0.0033	-0.0164	-0.0034	9.01	0.0027
$Tmin^2$	-0.0403	0.0216	-0.0826	0.0019	3.50	0.0614
$Rain^2$	-0.0004	0.0002	-0.0007	0.0000	3.14	0.0763
$Tmin \times Rain$	0.0065	0.0040	-0.0014	0.0144	2.59	0.1073
Scale	1.9332	0.0000	1.9332	1.9332		

3.10.6 Summary

In this section the incidence of AHS in KwaZulu-Natal has been modelled using a GLM with Poisson distribution. After examining different combinations of explanatory variables, it is ascertained that the best model for our data is:

$$\begin{aligned}\log(\mu) &= -0.8226 + 2.1965.sinyr + 1.0603Tmin - 0.0434.Rain \\ &\quad -0.0099Tmax^2 - 0.0403.Tmin^2 - 0.0004.Rain^2 + 0.0065.Tmin \times Rain.\end{aligned} \quad (3.48)$$

Although there is some evidence of overdispersion, estimation of the scale parameter does not change the optimal model.

3.11 Binomial Generalized Linear Model for African Horse Sickness Mortality

In this section we model the probability of death as a binary variable or outcome for each case. Therefore the Binomial distribution is used, with a logit link. We use all possible explanatory variables at first, and drop them in a stepwise fashion according to their Type III p-values in SAS. In Table 3.10, the general information about model variables and variable levels is given. Table 3.11 shows the stepwise procedure for dropping model terms. For each iteration, the log-likelihood and deviance are shown, along with the variable with the highest Type III p-value. Then for each step where a variable has been dropped, the difference in the deviances is shown and compared to the relevant χ^2 value. The variables "Pesticides" and then "Stabled" are dropped successfully, but when "OtherCases" is dropped the model is significantly worse in fit ($p = 0.0499$). Hence we use model 3, with significant variables Province, Classification, OtherCases, Vaccination, Presentation, Treatment and Isolation. Results for parameter estimates are shown in Table 3.12.

Table 3.10: Model Information for Binomial Generalized Linear model

Class Level Information

Class	Levels	Values	Reference Category
Province	9	ECP; FS; GAU; KZN; LIM; MPU; NCP; NWP; WCP	WCP
Classification	4	LABC; SUS; SUSN; VETC	VETC
OtherCases	2	0; 1	1
Vaccinated	3	0; 1; 2	0
Presentation	5	CARD; DK; PULM; MILD; MIX	MILD
Treatment	4	ALT; CONV; HOM; NONE	NONE
Stabled	2	0; 1	1
Pesticides	2	0; 1	1
Isolation	2	0; 1	1

Response Profile

Value	Horse Status	Total
1	Dead	474
2	Alive	449

3.11.1 Results

In Table 3.13 the estimates β_i along with the adjusted odds ratio (OR) e^{β_i} and the exponential of the Wald Confidence Interval are shown for the each of the levels. Significant

Table 3.11: Stepwise Regression for Binomial Generalized Linear Model for Probability of Mortality

	Model Information			Model Checking		Variable to be dropped		
	Log-Likelihood	Deviance	DF	Change in Deviance	p-value	Variable	Type III p-value	df
1	-534.4895	1068.9791	898			Pesticides	0.4424	1
2	-534.7844	1069.5689	899	0.5898	0.4425	Stabled	0.2025	1
3	-535.5987	1071.1974	900	1.6285	0.2019	OtherCases	0.0504	1
4	-537.5207	1075.0414	901	3.844	0.0499			

levels are indicated with an asterisk (*). e^{β_i} will be the odds ratio of mortality between level i and the base category for each of the categorical variables. An explanation of these figures along with potential reasons for the differences is given below. In each case the 95% confidence interval is quoted in square brackets following the odds ratio.

A horse in Eastern Cape, Gauteng and KwaZulu Natal had odds of mortality of 0.3252 [0.1521; 0.6952], 0.5239 [0.2769; 0.9912] and 0.4902 [0.2496, 0.9627] respectively times that of one in the Western Cape. It is possible that different serotypes are prevalent in these provinces which have a lower probability of mortality than in other provinces.

A horse where the case was confirmed by a sample sent to the laboratory (LABC) had odds around twice the odds (CI [1.2789; 5.0249]) of one where the case was confirmed by a veterinarian (VETC). This is unsurprising, since an owner is far more likely to send a sample to the laboratory if their horse has died. Both Suspected and Suspected but lab negative were not significantly different in likelihood of mortality from VETC.

If there were no other cases recorded in the surrounding area (OtherCases = 0), then the horse had odds of mortality 1.3682 [0.9994, 7.8729] times that if there were other cases recorded. However the odds of mortality was not significantly different between these two levels, as the confidence interval includes one.

Vaccinating a horse timeously halved the odds of the case ending up in mortality compared with one which was not vaccinated (Vaccinated = 0) (CI [0.3154, 0.7698]. This indicates that vaccination protects the horse from mortality even in the case of it contracting the disease. This is a biologically sound finding, as the horse would have circulating antibodies which may help the immune system to fight off the disease.
A horse that was vaccinated late (Vaccinated = 1) had odds of 0.0685 [0.0103, 0.4559] times that of an unvaccinated horse. This means that the odds ratio for a horse that was vaccinated late was 0.0685/0.4927 = 0.1390 times that for one which was vaccinated on time. This is a very interesting finding, since one would assume that vaccinating timeously would give the horse greater protection from the disease and therefore also from mortality. Although we cannot test (with this data) whether the horse will have greater protection from contracting the disease when vaccinated timeously, one might

Table 3.12: Model Information for Binomial GLM

Analysis Of Parameter Estimates

Parameter		Estimate	Std Error	Wald 95% CI		χ^2	p-value
Intercept		3.6240	0.5435	2.5587	4.6893	44.46	<.0001
Province	WCP	Ref	-	-	-	-	-
	ECP	-1.1232	0.3876	-1.8829	-0.3635	8.40	0.0038
	FS	-0.0210	0.5716	-1.1414	1.0994	0.00	0.9707
	GAU	-0.6464	0.3253	-1.2840	-0.0088	3.95	0.0469
	KZN	-0.7129	0.3444	-1.3879	-0.0380	4.29	0.0384
	LIM	-0.6134	0.4522	-1.4997	0.2728	1.84	0.1749
	MPU	-0.5368	0.3991	-1.3190	0.2454	1.81	0.1786
	NCP	0.0345	0.6052	-1.1516	1.2207	0.00	0.9545
	NWP	0.1887	0.4256	-0.6454	1.0228	0.20	0.6575
Classification	VETC	Ref	-	-	-	-	-
	LABC	0.9302	0.3491	0.2460	1.6144	7.10	0.0077
	SUS	-0.1623	0.2155	-0.5846	0.2601	0.57	0.4515
	SUSN	-0.8467	1.0812	-2.9658	1.2724	0.61	0.4336
OtherCases	1	Ref	-	-	-	-	-
	0	0.3135	0.1602	-0.0006	0.6275	3.83	0.0504
Vaccinated	0	Ref	-	-	-	-	-
	1	-2.6810	0.9671	-4.5766	-0.7855	7.68	0.0056
	2	-0.7078	0.2277	-1.1540	-0.2616	9.67	0.0019
Presentation	MILD	Ref	-	-	-	-	-
	CARD	2.1453	0.6475	0.8762	3.4143	10.98	0.0009
	DK	2.1441	0.6601	0.8504	3.4378	10.55	0.0012
	PULM	3.5882	0.6909	2.2341	4.9424	26.97	<.0001
	MIX	3.6606	0.6847	2.3187	5.0025	28.59	<.0001
Treatment	NONE	Ref	-	-	-	-	-
	ALT	-2.8911	0.5282	-3.9264	-1.8559	29.96	<.0001
	CONV	-1.9780	0.3282	-2.6212	-1.3348	36.33	<.0001
	HOM	-2.1601	0.4122	-2.9680	-1.3521	27.46	<.0001
Isolation	1	Ref	-	-	-	-	-
	0	-0.4443	0.2056	-0.8472	-0.0415	4.67	0.0306
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

assume from these results that when vaccinated late a horse has greater chance of fighting off the disease if it is contracted. However, since vaccinated late had such a small sample size, and vaccinated late and timeously have overlapping confidence intervals, we can presume that this finding is not significant.

However another possible explanation for this lies in the sport-horse industry. Those owners that compete their horses often compete over the summer months, and will only vaccinate once the competition season is over. This is because the vaccination requires that the horse is only minimally worked for six weeks during the vaccinations. Often this is done only in December, over the Christmas period. These owners are also very likely to vaccinate late every year. This is because competing horses can be very expensive, and therefore the owner is likely to take every possible precaution against disease. Since these owners do vaccinate every year their horses will build up immunity to AHS. Thus they will be less likely to die than an animal that was vaccinated timeously, but is not vaccinated routinely every year. These owners will also be the ones very likely to report a case or death due to AHS, which biases the data slightly.

An animal that had the Cardiac presentation of the disease had odds of 8.5446 [2.4018, 30.3957] times that of one that had the Mild or AHS fever presentation. A horse with the Pulmonary form of the disease had an odds ratio of 36.1689 [9.3381, 140.1061]. A horse with a Mixed (Cardio-Pulmonary) presentation of the disease had odds 38.8847 [10.1625, 148.7847] times higher than one which presented with the Mild presentation. A horse with an unknown presentation (Presentation = DK), had 8.5344 [2.3406, 31.1184] times the odds of death when compared to an animal with the Mild form. Coetzer and Erasmus (1994) considered that the Cardiac form had mortality of around 50%, the Mixed around 70%, and the Pulmonary around 95%. However, our findings are that a horse presenting the Mixed form had higher odds of death than the Pulmonary form: 38.8847/36.1689 = 1.0751. This is in contradiction with Coetzer and Erasmus's estimates. The odds ratio for their estimates comparing Mixed to Pulmonary would be

$$\frac{(0.7/0.3)}{(0.95/0.05)} = 0.1228. \tag{3.49}$$

From our data a Pulmonary case had odds 36.1689/8.5446 = 4.2330 times higher than a Cardiac case. The odds ratio according to Coetzer and Erasmus would have been

$$\frac{(0.95/0.05)}{(0.5/0.5)} = 19. \tag{3.50}$$

However, Coetzer and Erasmus (1994) do not mention whether these estimates come from studies in naïve populations or populations where vaccination strategies are in place. It is known (and verified by this data) that vaccination will protect the horse from mortality and will therefore affect these estimates. The number of unknown presentations in our data may also have affected our estimates.

All of the treatment possibilities had lower probabilities of mortality than no treatment at all (Treatment = NONE). An animal treated with Alternative remedies, Conven-

tional treatment and Homeopathic treatments were respectively 0.0555 [0.0197, 0.1563], 0.1383 [0.0727, 0.2632] and 0.1153 [0.0514, 0.2587] times as likely to die when compared with one which received no treatment at all. The surprising fact is that Conventional treatments, ie. those prescribed by a trained veterinarian, did not perform better than Alternative and Homeopathic remedies. A horse on Alternative treatment had odds $\frac{0.0556}{0.1383} = 0.4020$ times that of one on Conventional treatments, and one on Homeopathic treatment had odds $\frac{0.1153}{0.1383} = 0.8337$ times Conventional treatments. However, since Homeopathic, Alternative and Conventional treatments have overlapping confidence intervals, we cannot find that their is a significant difference between them.
An interesting fact is that one known "Alternative" treatment is to treat the horse with marijuana. The possible benefit of such an alternative treatment is that it may reduce stress in the animal to allow it to recuperate. It is unknown, though, whether marijuana was used in all cases where treatment was given as Alternative, although in a few cases it was given in extra information. It is also unknown whether marijuana was given in conjunction with other alternative treatments.

A horse which was not isolated had lower odds of mortality than those isolated (OR = 0.6413 [0.4286, 0.9593]). This may be due to the fact that a horse with more severe case of the disease would be more likely to be isolated than one without, and therefore would be more likely to die due to the severity of the disease. Isolating a horse may also increase stress in the animal, which is naturally herd-bound, and thereby increase the heart-rate of the animal. Since AHS often affects the heart, this may have an adverse affect on the horse.

3.11.2 Summary

This analysis has provided a good base from which to continue the analysis of this data. However, we have not addressed the possible problem of overdispersion. When we use proc GENMOD to estimate the deviance dispersion parameter, it gives the estimate as $\phi = 1.0910$. This is only very slight overdispersion, however it could be accounted for by taking into account the differences by location. There are various methods which can be utilized to take into account this heterogeneity, some of which will be explained in the next chapter.

Table 3.13: Parameters for binomial model for probability of mortality

Parameter	Level		Estimate	Adjusted OR	95% CI (OR)	
Province	WCP	Ref				
	ECP	*	-1.1232	0.3252	0.1521	0.6952
	FS		-0.0210	0.9792	0.3194	3.0024
	GAU	*	-0.6464	0.5239	0.2769	0.9912
	KZN	*	-0.7129	0.4902	0.2496	0.9627
	LIM		-0.6134	0.5415	0.2232	1.3136
	MPU		-0.5368	0.5846	0.2674	1.2781
	NCP		0.0345	1.0351	0.3161	3.3896
	NWP		0.1887	1.2077	0.5245	2.7810
Classification	VETC	Ref				
	LABC	*	0.9302	2.5350	1.2789	5.0249
	SUS		-0.1623	0.8502	0.5573	1.2971
	SUSN		-0.8467	0.4288	0.0515	3.5694
OtherCases	1	Ref				
	0	*	0.3135	1.3682	0.9994	1.8729
Vaccinated	0	Ref				
	1	*	-2.6810	0.0685	0.0103	0.4559
	2	*	-0.7078	0.4927	0.3154	0.7698
Presentation	MILD	Ref				
	CARD	*	2.1453	8.5446	2.4018	30.3957
	DK	*	2.1441	8.5344	2.3406	31.1184
	PULM	*	3.5882	36.1689	9.3381	140.1061
	MIX	*	3.6606	38.8847	10.1625	148.7847
Treatment	NONE	Ref				
	ALT	*	-2.8911	0.0555	0.0197	0.1563
	CONV	*	-1.9780	0.1383	0.0727	0.2632
	HOM	*	-2.1601	0.1153	0.0514	0.2587
Isolation	1	Ref				
	0	*	-0.4443	0.6413	0.4286	0.9593

Chapter 4

Accounting for Cluster to Cluster Heterogeneity and Within Cluster Correlation

4.1 Introduction

Clearly our problem involves modeling data where observations occur in clusters. The premise of modeling clustered data is that observations within clusters will be alike; in our case the probability of mortality may be similar within each place (as they are probably affected with the same serotype of the disease and environmental conditions). We start by accounting for within cluster correlation. Then we address the question of cluster to cluster heterogeneity by means of models allowing for subject specific effects.

4.2 Generalized Estimating Equations

Under Generalized Linear Models theory, it is assumed that the observations are all independent. Generalized Estimating Equations, developed by Liang and Zeger (1986), extend the theory of GLM to be able to deal with data where a correlation structure exists. This method is most useful in cases such as the current problem, where observations from the same place are likely to be more similar than those occurring in different places; known as clustered data. These GEE models can also be referred to as marginal models, since the population average fixed effects are the effects of interest. In other words, GEE models ensure that the correlation between observations from the same cluster is accounted for in the estimation of parameters $\boldsymbol{\beta}$. In Liang and Zeger's GEEs, assumptions are made about the correlation structure - called 'working' assumptions - in order to account for within cluster correlation. A distinct advantage of the GEE formulation is that, even if the correlation structure has not been correctly specified, the estimates are still consistent. The advantage comes in the estimation of standard errors.

If missingness occurs, it is assumed that the data is missing completely at random (MCAR) (Agresti, 2002 and Hedeker *et al.*, 2006). This is necessary as GEE is not a likelihood-based procedure, and therefore the missingness can only be ignored if it is MCAR.

Let $y_{i1}, y_{i2}, ..., y_{in_i}$ be the observations from cluster i for $i = 1, 2, ..., m$. Assuming that the response y_{ij} follows an exponential family distribution given in Equation (3.5), the equation for $\mu_{ij} = E(Y_{ij})$ is given, as in the GLM, as:

$$g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}. \tag{4.1}$$

It is also assumed as before that $\text{var}(y_{ij}) = \phi\psi''(\theta_{ij}) = \phi v(\mu_{ij})$.

We define a working covariance matrix for the cluster of observations $y_i = (y_{i1}, y_{i2}, ..., y_{in_i})'$ as

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}, \tag{4.2}$$

where $\boldsymbol{R}(\boldsymbol{\alpha})$ is the "working" correlation matrix, and $\boldsymbol{A}_i = \text{diag}(\text{var}(Y_{ij})) = \text{diag}(\phi\psi''(\theta_{ij}))$. In other words, we assume that the within cluster correlation is dependent on some additional parameters $\boldsymbol{\alpha}$. The working correlation matrix is chosen based on the assumed realistic correlation structure of the data. $\boldsymbol{V}_i$ is called the "working" covariance matrix, as it is understood that it is an approximation and in all likelihood not equal to the true covariance. The generalized estimating equations are then:

$$\sum_{i=1}^{m} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0, \tag{4.3}$$

where $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ is an $n_i \times p$ matrix. For non-identity link, this requires iterative algorithms to solve as there exists no closed form solution. The "sandwich" method is most often used, iteratively solving Equations (4.2) and (4.3) with estimates for $\boldsymbol{\alpha}$ and ϕ given by Liang and Zeger to be:

$$\begin{aligned} \hat{\phi} &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} e_{ij}^2}{N\sum_{i=1}^{m} n_i}, \\ \hat{\alpha}_{jk} &= \frac{1}{\hat{\phi}m}\sum_{i=1}^{m} e_{ij}e_{ik}, \end{aligned} \tag{4.4}$$

where residuals are defined as $e_{ij} = \frac{y_{ij}-\mu_{ij}}{\sqrt{v(\mu_{ij})}}$.

The estimator from this method $\hat{\boldsymbol{\beta}}$ is consistent, regardless of whether the covariance structure of the cluster was correctly modeled. This is an advantage of the GEE method, and the reason that the "working" covariance structure is adequate for estimation.

SAS proc GENMOD can be utilized to model GEE's. In this case, the REPEATED statement is used to specify the cluster variable together with the assumed correlation structure.

4.2.1 Specification of Working Correlation Structure

There are several choices for correlation structure, which can usually be chosen based on some realistic assumption about the structure. In our instance, where "Place" is the cluster variable, we expect that the observations within each place will be alike, and have the same correlation with every other observation within the same place. In other words the correlations between all observations within a place are considered homogeneous. Thus the compound symmetry structure will be used, which has the following structure for a single cluster with four observations ($n_i = 4$)

$$\boldsymbol{R(\alpha)} = \sigma^2 \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \rho \\ \alpha & \alpha & \alpha & 1 \end{pmatrix} = \sigma^2 R, \tag{4.5}$$

where R gives the correlation structure.

Therefore the covariance between any two observations within a place would be $\sigma^2\alpha$, where σ^2 is the variance for each observation.
A GEE with this correlation structure will use the estimated Pearson residuals $\hat{r}_{ij}$ to estimate α as follows:

$$\hat{r}_{ij} = (y_{ij} - \hat{\mu}_{ij})/\sqrt{V(\hat{\mu}_{ij})}, \tag{4.6}$$

$$\hat{\alpha} = \sum_{i=1}^{m} \left\{ \frac{\sum_{u=1}^{n_i} \sum_{v=1}^{n_i} \hat{r}_{iu}\hat{r}_{iv} - \sum_{u=1}^{n_i} \hat{r}_{iu}^2}{n_i(n_i - 1)} \right\}. \tag{4.7}$$

$$\tag{4.8}$$

Another consideration when choosing the correlation structure is that, in the case of GEEs, incorrect choice will not affect the consistency of the estimates it gives. Thus it is advisable to choose the structure with the smallest number of parameters to avoid having too many parameters to estimate. Therefore our choice of compound symmetry, with only two parameters, seems the best choice for our case.

4.2.2 Model Selection

Since the GEE is not likelihood based in its formulation methods of model selection based on likelihoods, such as likelihood ratio tests, cannot be utilized. Pan (2001) describes an information criterion based on the Quasi-likelihood that can be used to compare GEE models.

The log quasi-likelihood is defined by McCullagh and Nelder (1989) to be:

$$Q(\mu, \phi; y) = \int_{y}^{\mu} \frac{y - t}{\phi V(t)} dt, \tag{4.9}$$

where $var(y) = \phi V(\mu)$ is the relationship between the variance and mean for the distribution of y. The quasi-likelihood was first suggested by Wedderburn (1974).

The QIC is then defined as

$$QIC(R) = -2Q(\hat{\boldsymbol{\beta}}(\boldsymbol{R}), \phi) + 2trace(\hat{\Omega}\hat{V}_R), \tag{4.10}$$

where $\boldsymbol{R}$ is the working correlation matrix, $\hat{\Omega}$ is the inverse of the model-based covariance estimate under assumption of independent working correlation ($\boldsymbol{R} = I$), and $\hat{V}_R$ is the robust covariance estimate.

An approximation to $QIC(R)$ which can be used in variable selection can be defined as follows:

$$QIC_u(R) = -2Q(\hat{\boldsymbol{\beta}}(R), \phi) + 2p. \tag{4.11}$$

The term $2p$ serves as a penalty for increasing the number of parameters. A small value of QIC_u indicates a model which has an adequate fit while not having needless parameters, and thus the model with smallest QIC_u is chosen as the optimal model.

4.2.3 Applications of Generalized Estimating Equations in Modeling AHS Mortality

We model the Binomial data using Generalized Estimating Equations, with "Place" serving as a clustering variable. The full model is fitted initially, with all variables shown in Table 3.10, and including "Place" with 239 levels in the REPEATED statement to model it as a cluster effect. The variables are then dropped in a stepwise fashion according to their Type III p-values as given in the SAS output. The steps are shown in Table 4.1. As can be seen in Table 4.1, whether we select the best model based on QIC or QICu makes no difference, as both are minimized for model 4. Final GEE model estimates are shown Table 4.2. Empirical standard errors were used.

In this Table, we see that the working correlation is estimated to be 0.1109. This shows us that there exists a slight positive correlation between observations which occur in the same location, as was expected. This could be due to different serotypes being prevalent in different locations, which may have different mortality rates.

Table 4.3 shows the odds ratios as well as the 95% confidence intervals for the odds ratios for the GEE model. Significant levels are indicated with an asterisk. The odds of death for a horse whose disease status is confirmed by a laboratory (Classification = LABC) was more than twice that of one which was confirmed by a veterinarian. A horse vaccinated timeously had odds of mortality of about 0.5 that of an unvaccinated horse, while the odds for a horse which was vaccinated late was approximately 0.07

Table 4.1: Stepwise Procedure for Binomial GEE for Probability of Mortality

	QIC	QICu	Correlation	Variable to Drop	p-value	df
1	1152.0096	1125.1855	0.0914	Province	0.4932	8
2	1141.8379	1126.9904	0.1150	Stabled	0.4555	1
3	1139.9052	1125.2851	0.1133	Pesticides	0.2265	1
4	1138.9316	1124.6464	0.1109	OtherCases	0.0978	1
5	1141.7889	1127.7383	0.1159	Isolation	0.1098	1
6	1144.0096	1130.5156	0.1215	.	.	.

times that of one which was unvaccinated. As explained before, this is probably due to the vaccination habits of competitive horse owners vaccinating late when out of the competition season, but vaccinating every year.

A horse with Cardiac symptoms had odds of mortality of 9 times that of one with a Mild case. Mixed and Pulmonary cases had odds respectively 39.12 and 36.98 times that of a Mild case.

Once again, Alternative treatments were observed to perform the best, reducing odds of mortality by 0.045 times that of an untreated horse. The next best treatment was Homeopathic, reducing odds of mortality by 0.09 times, and then Conventional, with odds of mortality 0.11 times that of an untreated horse. However their overlapping confidence intervals cause us to conclude that their is no significant difference between the treatments.

4.3 Generalized Linear Mixed Models

Another way in which the heterogeneity between observations in different places can be accounted for is by specifying "Place" as a random effect. The usual categorical variables in GLMs are termed fixed effects, and apply to all the levels of interest. By contrast, a random effect applies to a sample of all of the categories of interest. Since we know that not all of the cases of AHS have been reported, the places listed are almost certainly only a random sample of all of the places in which AHS cases occurred. We therefore wish to model the data using a random effect for each cluster/place. Because the data we are interested in is Binomial in nature, we choose to use extensions of Generalized Linear Models that can account for random effects, and thus use Generalized Linear Mixed Models (GLMMs).

Let $y_{i1}, y_{i2}, ..., y_{in_i}$ be the observations from cluster i for $i = 1, 2, ..., m$. Then the Generalized Linear Mixed Model introduces a $q \times 1$ vector of random effects $\boldsymbol{b}_i$ for each cluster i to the usual GLM equation for y_{ij} as follows:

$$g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i, \tag{4.12}$$

Table 4.2: Analysis of GEE parameter estimates for mortality data

Parameter		Estimate	SE	95% CI		Z	p-value
Intercept		0.3004	1.18	-2.0125	2.6132	0.25	0.7991
Classification	VETC	Ref	-	-	-	-	-
	LABC	0.8236	0.3605	0.1170	1.5302	2.28	0.0223
	SUS	-0.1855	0.2045	-0.5863	0.2153	-0.91	0.3643
	SUSN	-0.7135	0.7688	-2.2204	0.7933	-0.93	0.3534
OtherCases	1	Ref	-	-	-	-	-
	0	0.2656	0.1668	-0.0613	0.5926	1.59	0.1112
Vaccinated	0	Ref	-	-	-	-	-
	2	-0.6523	0.1972	-1.0388	-0.2659	-3.31	0.0009
	1	-2.7183	0.8841	-4.4511	-0.9854	-3.07	0.0021
Presentation	MILD	Ref	-	-	-	-	-
	CARD	2.2045	0.8476	0.5432	3.8658	2.6	0.0093
	DK	2.1206	0.8709	0.4137	3.8275	2.43	0.0149
	MIX	3.6666	0.8719	1.9577	5.3755	4.21	<.0001
	PULM	3.6103	0.9120	1.8227	5.3978	3.96	<.0001
Treatment	NONE	Ref	-	-	-	-	-
	ALT	-3.0997	0.9900	-5.0401	-1.1593	-3.13	0.0017
	CONV	-2.2075	0.8031	-3.7816	-0.6334	-2.75	0.0060
	HOM	-2.4233	0.8846	-4.1570	-0.6895	-2.74	0.0062
Isolation	1	Ref	-	-	-	-	-
	0	-0.3968	0.2197	-0.8274	0.0338	-1.81	0.0709

Exchangable working correlation = 0.11087

Table 4.3: Odds Ratios and confidence intervals for final GEE model. Significant levels are marked with an asterisk (*).

Parameter	Level		Adjusted OR	95% CI (OR)	
Classification	VETC		Ref	-	-
	LABC	*	2.2787	1.1241	4.6191
	SUS		0.8307	0.5564	1.2402
	SUSN		0.4899	0.1086	2.2107
OtherCases	1		Ref	-	-
	0		1.3042	0.9405	1.8087
Vaccinated	0		Ref	-	-
	1	*	0.0660	0.0117	0.3733
	2	*	0.5208	0.3539	0.7665
Presentation	MILD		Ref	-	-
	CARD	*	9.0657	1.7215	47.7415
	DK	*	8.3361	1.5124	45.9475
	MIX	*	39.1187	7.0830	216.0479
	PULM	*	36.9771	6.1885	220.9199
Treatment	NONE		Ref	-	-
	ALT	*	0.0451	0.0065	0.3137
	CONV	*	0.1100	0.0228	0.5308
	HOM	*	0.0886	0.0157	0.5018
Isolation	1		Ref	-	-
	0		0.6725	0.4372	1.0344

where $\mu_{ij} = E[y_{ij}|\boldsymbol{b}_i]$ is now a conditional mean specific to cluster i, $\boldsymbol{x}_{ij}$ is a $(p+1)\times 1$ vector of fixed covariates, $\boldsymbol{\beta}$ is the $(p+1)\times 1$ vector of fixed effects, $\boldsymbol{z}_{ij}$ is a $q\times 1$ vector of covariates for random effects, and $\boldsymbol{b}_i$ is a $q\times 1$ vector of random effects. In our case, since each cluster has only one random effect, $q = 1$, $\boldsymbol{z}_{ij} = 1$ and the vector $\boldsymbol{b}_i$ becomes the scalar b_i, and the model can be re-written as $g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + b_i$, $i = 1, 2, ..., m$.

The matrix form of Equation (4.12) (for each cluster i) is given as:

$$g(E(\mathbf{Y}_i|b_i)) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i b_i, \tag{4.13}$$

where $\boldsymbol{X}_i$ is the $n_i \times p$ matrix for the regressors with the j^{th} row equal to $\boldsymbol{x}'_{ij}$. $\boldsymbol{Z}_i$ is an $n_i \times q$ design matrix for the i^{th} cluster, where the j^{th} row is $\boldsymbol{z}'_{ij}$. In our case, however, there is only one random effect and therefore $q = 1$ and $\boldsymbol{Z}_i = \underline{\mathbf{1}}_{(n_i\times 1)}$ is a vector of 1's of length n_i. Therefore the model can be re-written as

$$g(E(\mathbf{Y}_i|b_i)) = \boldsymbol{X}'_i\boldsymbol{\beta} + \boldsymbol{b_i}, \tag{4.14}$$

where $\boldsymbol{b_i}$ is a $n_i \times 1$ vector where each element is b_i: $\boldsymbol{b_i} = (b_i, b_i, ..., b_i)'$

We assume that the conditional distribution of Y_{ij} given b_i has a pdf following the exponential family of distributions, with probability density function (pdf) given by

$$f(y_{ij}|b_i) = \exp\{(y_{ij}\theta_{ij} - \psi(\theta_{ij}))/\phi + c(y_{ij})\}. \tag{4.15}$$

Note that the term previously referred to as $b(\theta_{ij})$ in the exponential family in Chapter 3 is now re-named $\psi(\theta_{ij})$ in Equation (4.15) to avoid confusion with the random effect b_i.

We also assume that the b_i are Normally distributed with constant variance. That is, $b_i \sim N(0, \sigma_s^2)$. Other types of distribution for b_i can be assumed, but that is not the focus of the current analysis.

4.3.1 Estimation in GLMM's

There are various methods for estimation of GLMM's, which will be briefly outlined below. Some methods are computationally very intense and others are more achievable. However they mostly require the use of specialized statistical software, especially where the matrices are large.

4.3.2 Conditional Likelihood Method

Since we know that the conditional pdf follows an exponential family distribution (see Equation (4.15)), we have that the conditional likelihood is given by

$$L = \prod_{i=1}^{m}\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, b_i) \propto \prod_{i=1}^{m}\prod_{j=1}^{n_i} \exp\{\theta_{ij}y_{ij} - \psi(\theta_{ij})\}. \tag{4.16}$$

Assuming the canonical link, $g(\mu_{ij}) = \theta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + b_i$, and substituting θ_{ij} in the above equation gives

$$L \propto \exp\{\boldsymbol{\beta}' \sum_{ij} \boldsymbol{x}_{ij} y_{ij} + \sum_{i} b_i \sum_{j} y_{ij} - \sum_{ij} \psi(\theta_{ij})\}. \tag{4.17}$$

Thus we have that sufficient statistics for $\boldsymbol{\beta}$ and b_i are $\sum_{ij} \boldsymbol{x}_{ij} y_{ij}$ and $\sum_j y_{ij}$ respectively. The expression for L above assumes $\phi = 1$.

Let us call the target parameters $\boldsymbol{\delta} = (\sigma_s^2, \boldsymbol{\beta}) = (\alpha, \boldsymbol{\beta})$ for simplicity. In order to find the marginal likelihood estimate $\boldsymbol{\delta}$ we have to integrate b_i out from the conditional likelihood function. To do this we should solve

$$L(\boldsymbol{\delta}, \boldsymbol{y}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij}|b_i; \boldsymbol{\beta}) f(b_i) db_i. \tag{4.18}$$

This requires numerical integration methods, as there is no closed form solution for non-Normal response. The SAS NLMIXED procedure uses a dual Quasi-Newton Algorithm to solve it iteratively.

4.3.3 Maximum Likelihood Estimation

We wish to find the score equation $S(\boldsymbol{\delta})$ for $\boldsymbol{\beta}$ given b_i, that is $\frac{\mathrm{d}l}{\mathrm{d}\boldsymbol{\delta}} = 0 = S(\boldsymbol{\delta})$. Differentiating (4.18) with respect to $\boldsymbol{\beta}$ we get:

$$S_{\boldsymbol{\beta}}(\boldsymbol{\delta}|\boldsymbol{y}, \boldsymbol{b}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \{y_{ij} - \mu_{ij}(b_i)\} = 0, \tag{4.19}$$

where $\mu_{ij}(b_i) = E(y_{ij}|b_i) = g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + b_i)$. Differentiating with respect to $\sigma_s^2 = \alpha$ gives:

$$S_{\alpha}(\boldsymbol{\delta}|\boldsymbol{y}) = \frac{1}{2}\{\sum_{i=1}^{m} E[b_i b_i'|\boldsymbol{y}_i]\}\alpha^{-1} - \frac{m}{2}\alpha^{-1} = 0. \tag{4.20}$$

Solving Equations (4.19) and (4.20) for the Maximum Likelihood Estimates requires the use of expectation maximization (EM) algorithm of Dempster *et al.* (1977). However this method is not the focus of this thesis.

4.3.4 Penalized Quasi-Likelihood

Approximating the score equations can avoid having to integrate the conditional likelihood. The conditional model is used rather than conditional means to yield an approximation of the conditional distribution of b given y that resembles a Normal distribution, which is preferable to work with. The method was introduced by Breslow and Clayton (1993).

This method is given as described in Diggle *et al.* (1994) as follows:
Let $v_{ij} = \mathrm{var}(y_{ij}|b_i)$, and $\boldsymbol{Q}_i = \mathrm{diag}[v_{ij} g'(\mu_{ij})^2]$. Then we define a "working" response

w_{ij} as $w_{ij} = g(\mu_{ij}) + (y_{ij} - \mu_{ij})g'(\mu_{ij})$. This is a linearization process. We then define the vector $\boldsymbol{w}_i = (w_{i1}, w_{i2}, ..., w_{in_i})'$ for $i = 1, 2, .., m$.

The $n_i \times n_i$ variance-covariance matrix for cluster i is given by $\boldsymbol{V}_i = \boldsymbol{Q}_i + \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i'$ for fixed $\boldsymbol{G} = \alpha$, in our case becomes $\boldsymbol{V}_i = \boldsymbol{Q}_i + \alpha\boldsymbol{J}$ with $\boldsymbol{J}$ being a matrix of 1's of dimension $n_i \times n_i$. Given that the matrix form of the GLMM equation is $g(E(\boldsymbol{Y}_i|b_i)) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{b}_i$, the updated values of $\boldsymbol{\beta}$ and b_i are given iteratively as given in the following equations.

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1} \sum_{i=1}^{m} (\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{w}_i) && (4.21)\\
\hat{b_i} &= \hat{\boldsymbol{G}}\boldsymbol{Z}_i\boldsymbol{V}_i^{-1}(\boldsymbol{w}_i - \boldsymbol{X}_i\boldsymbol{\beta}) && (4.22)\\
&= \hat{\boldsymbol{\alpha}}\boldsymbol{V}_i^{-1}(\boldsymbol{w}_i - \boldsymbol{X}_i\boldsymbol{\beta}) && (4.23)\\
\hat{\alpha} &= m^{-1} \sum_{i=1}^{m} E(\hat{b_i}\hat{b_i'}|\boldsymbol{y}_i) && (4.24)\\
g(\mu_{ij}) &= \boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b_i}} && (4.25)\\
& && (4.26)
\end{aligned}$$

There are two main methods of this iterative process. Marginal Quasi-Likelihood (MQL) assumes that since $b_i \sim N(0, G)$ that we can update $g(\mu_{ij}) = x_{ij}'\hat{\boldsymbol{\beta}}$. Penalized Quasi-Likelihood (PQL) does not assume mean 0 for b_i, and thus updates using $g(\mu_{ij}) = x_{ij}'\hat{\boldsymbol{\beta}} + Z_{ij}'\hat{b_i}$. PQL is slower to converge and less accurate.

The SAS GLIMMIX procedure uses this method of estimation as default. The MQL or PQL methods can be specified. However, it should be noted that PQL can yield biased estimates for variance components (Breslow, 2003), and so should be used with caution.

4.3.5 Adaptive Gauss-Hermite Quadrature

A method which is preferable and which can be specified in the GLIMMIX procedure is the method of adaptive Gauss-Hermite quadrature.

Ordinary Gauss-Hermite quadrature is a method used to approximate any integral of the form:

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x)dx \approx \sum_{r=1}^{R} w_r f(t_r), \tag{4.27}$$

where q denotes the order of the approximation, w_r are the weights defined by a R-order Hermite polynomial, and t_r are the quadrature points. The weights and quadrature points can be found in tables such as Abramowitz and Stegun (1972).

If it is assumed that the random effects are normally distributed with mean zero, the marginalized likelihood can be written as:

$$\begin{aligned} L_i &= \int \prod_{j=1}^{n_i} f(y_{ij}|b_i, \boldsymbol{\beta})\phi(b_i; \mu_b, \sigma_b^2) db_i, \qquad (4.28) \\ &= \int L_i^c(b) \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(b_i - \mu_b)^2}{2\sigma^2}\} db_i, \end{aligned}$$

where $\phi(b_i; \mu_b, \sigma_b)$ is the normal probability function, $L_i^c(b) = \prod_{j=1}^{n_i} f(y_{ij}|b_i, \boldsymbol{\beta})$ is the conditional likelihood of the response, and σ_b^2 is the variance of b. Using the parameterization $\delta_i = (b_i - \mu_b)/\sqrt{2\sigma_b^2}$, the above equation can be re-written:

$$L_i = \frac{1}{\sqrt{\pi}} \int L_i^c(b).exp(\delta_i^2) d\delta_i, \qquad (4.29)$$

which is in the required form in 4.27, and therefore can be estimated under ordinary Gauss-Hermite quadrature as:

$$L_i \approx \frac{1}{\sqrt{\pi}} \sum_{r=1}^{R} w_r L_i^c(\mu_b + \sqrt{2\sigma_b^2} t_r). \qquad (4.30)$$

However, using this ordinary Gauss-Hermite quadrature, the function which is to be integrated is sampled at fixed points, irrespective of the range of the actual function. Therefore Adaptive Gauss-Hermite quadrature was proposed by Liu and Pierce (1994) to rectify this.

Suppose that we wish to find the integral of the form:

$$\int g(b) db = \int L_i^c(b) \phi(b; \mu, \sigma) db. \qquad (4.31)$$

The adaptive Gauss-Hermite quadrature method approximates μ and σ as follows:

$$\begin{aligned} \hat{\mu} &= \text{mode}(g(b)), \qquad (4.32) \\ \hat{\sigma} &= \frac{1}{\sqrt{\hat{j}}}, \qquad (4.33) \\ \hat{j} &= -\frac{\partial^2}{\partial b^2} \log(g(\hat{\mu})). \qquad (4.34) \\ & \qquad (4.35) \end{aligned}$$

Then defining

$$h(b) = \frac{g(b)}{\phi(b; \hat{\mu}, \hat{\sigma})}, \qquad (4.36)$$

the integral can be approximated as follows:

$$\begin{aligned}\int g(b)db &= \int h(b)\phi(b;\hat{\mu},\hat{\sigma})db, \qquad (4.37)\\ &\approx \frac{1}{\sqrt{\pi}}\sum_{r=1}^{R} w_r h(\hat{\mu}+\sqrt{2\hat{\sigma}_b^2}t_r),\\ &= \sum_{r=1}^{R}\frac{w_r}{\sqrt{\pi}}\frac{g(\hat{\mu}+\sqrt{2\hat{\sigma}^2}t_r)}{\phi(b;\hat{\mu},\hat{\sigma})},\\ &= \sqrt{2}\hat{\sigma}\sum_{r=1}^{R} w_r \exp(t_r^2) g(\hat{\mu}+\sqrt{2\hat{\sigma}^2}t_r).\end{aligned}$$

This yields better estimates than ordinary Gauss-Hermite quadrature.

4.3.6 Applications of Generalized Linear Mixed Models to Modeling of Probability of Mortality

We wish to model the binomial data, with the variable "Place" added as a random cluster effect. We employ SAS proc GLIMMIX in order to fit GLMMs to the data, using the Gauss-Hermite method of fitting. Table 4.4 shows the categorical variables, the levels and reference categories for all the binomial models. Beginning by including all possible variables, variables were dropped in a stepwise fashion according to their Type III p-values, until all effects were found to be significant. The best model was chosen as the one with minimal AIC, which is shown in Table 4.5 to be the model in step 4. The final model solutions are shown in Table 4.6, where it can be seen that the significant variables are Classification, Vaccinated, Presentation, Treatment and Isolation.

Table 4.7 shows the estimates of β_i as well as the odds ratio e^{β_i} for each level of the categorical variables. Significant levels are indicated with an asterisk.

In a given place, a horse that was vaccinated timeously had odds of mortality 0.4750 [0.2887, 0.7815] times that of one that was un-vaccinated, and a horse that was vaccinated late had odds ratio 0.0531 [0.0071, 0.3954]. A horse with Cardiac presentation had odds of mortality 11.5479 [3.0374, 43.9038] times higher than one with a Mild or fever presentation. An unknown presentation caused a horse to have 10.2379 [2.6253, 39.9289] times the odds of mortality compared to Mild. Horses with mixed presentation and Pulmonary presentation had respectively 59.9014 [14.4010, 249.1861] and 55.1855 [13.0241, 233.8313] times the odds of mortality of one with Mild presentation.

A horse treated with Alternative remedies had odds of 0.0306 [0.0092, 0.1023] times that of an untreated horse. Conventional and Homeopathic treatments respectively caused horses to have odds 0.0818 [0.0363, 0.1841] and 0.0644 [0.0241, 0.1720] times that of untreated.

Table 4.4: Class Level Information for Binomial Models

Class	Levels	Values	Reference Category
Province	9	ECP; FS; GAU; KZN; LIM; MPU; NCP; NWP; WCP	WCP
Classification	4	LABC; SUS; SUSN; VETC	VETC
OtherCases	2	0; 1	1
Vaccinated	3	0; 1; 2	0
Presentation	5	CARD; DK; PULM; MILD; MIX	MILD
Treatment	4	ALT; CONV; HOM; NONE	NONE
Stabled	2	0; 1	1
Pesticides	2	0; 1	1
Isolation	2	0; 1	1
Place	235	not printed	-

Table 4.5: Stepwise Regression Steps for Binomial GLMM for Probability of Mortality

	AIC	Covariance Estimate	Standard Error	Variable to drop	p-value
1	1093.58	0.7421	0.2596	Stabled	0.4346
2	1092.19	0.7525	0.2622	Province	0.3624
3	1085.01	0.8503	0.2737	Pesticides	0.2356
4	1084.42	0.8521	0.2735	OtherCases	0.0764
5	1085.57	0.8720	0.2768	Isolation	0.0745
6	1086.78	0.8898	0.2796		

A horse which was not isolated reduced odds of mortality 0.6375 [0.4078, 0.9967] times that for one which was, although this was only marginally significant with a p-value of 0.0484.

However, although the variable “Place” is accounting for some of the heterogeneity between cases, we speculate that a further random variable could be the outbreak. Note that the variance component for Place is estimated as 0.8524 with standard error of 0.2735 under model 4 with minimum AIC. It could be the case that a specific place may have different serotypes of the disease in different outbreaks, and as a consequence mortality rates may be different. A categorical variable was created which assigns a number to each outbreak (1 = 2005/2006, 2 = 2006/2007, 3=2007/2008, 4=2008/2009, 5 = 2009/2010). Outbreak nested in place is added as a further random factor to the model, and the stepwise procedure repeated using proc GLIMMIX to fit the models. The stepwise procedure is shown in Table 4.8. Choosing the model with lowest AIC, we use the model indicated by step 4 in the table. The final results for this model are shown in Table 4.9, and the odds ratios in Table 4.10.

If we compare the results between the two GLMMs with only Place as a random

Table 4.6: Solutions for Fixed Effects for Binomial GLMM for the Probability of Mortality

Effect	Level	Estimate	CI		Std Error	DF	t Value	p-value
Intercept		0.4196	0.8171	-1.1903	2.0294	234	0.51	0.6081
Classification	VETC	Ref	-	-	-	-	-	-
	LABC	0.8871	0.3864	0.1284	1.6458	674	2.30	0.0220
	SUS	-0.2166	0.2406	-0.6890	0.2559	674	-0.90	0.3684
	SUSN	-0.7763	1.1815	-3.0961	1.5435	674	-0.66	0.5113
OtherCases	1	Ref	-	-	-	-	-	-
	0	0.3141	0.1770	-0.0335	0.6616	674	1.77	0.0764
Vaccinated	0	Ref	-	-	-	-	-	-
	1	-2.9349	1.0222	-4.9420	-0.9279	674	-2.87	0.0042
	2	-0.7445	0.2536	-1.2424	-0.2466	674	-2.94	0.0034
Presentation	MILD	Ref	-	-	-	-	-	-
	CARD	2.4465	0.6802	1.1110	3.7820	674	3.60	0.0003
	DK	2.3261	0.6931	0.9652	3.6871	674	3.36	0.0008
	MIX	4.0927	0.7260	2.6673	5.5182	674	5.64	<.0001
	PULM	4.0107	0.7354	2.5668	5.4546	674	5.45	<.0001
Treatment	NONE	Ref	-	-	-	-	-	-
	ALT	-3.4862	0.6144	-4.6926	-2.2799	674	-5.67	<.0001
	CONV	-2.5036	0.4133	-3.3151	-1.6921	674	-6.06	<.0001
	HOM	-2.7423	0.5003	-3.7246	-1.7601	674	-5.48	<.0001
Isolation	1	Ref	-	-	-	-	-	-
	0	-0.4502	0.2276	-0.8970	-0.0033	674	-1.98	0.0484

Table 4.7: Odds ratios and confidence intervals from binomial GLMM. Significant levels are indicated with an asterisk.

Parameter	Level		Estimate	Adjusted OR	95% CI (OR)	
Classification	VETC		Ref	-	-	-
	LABC	*	0.8871	2.4281	1.1370	5.1852
	SUS		-0.2166	0.8053	0.5021	1.2916
	SUSN		-0.7763	0.4601	0.0452	4.6809
OtherCases	1		Ref	-	-	-
	0		0.3141	1.3690	0.9671	1.9379
Vaccinated	0		Ref	-	-	-
	1	*	-2.9349	0.0531	0.0071	0.3954
	2	*	-0.7445	0.4750	0.2887	0.7815
Presentation	MILD		Ref	-	-	-
	CARD	*	2.4465	11.5479	3.0374	43.9038
	DK	*	2.3261	10.2379	2.6253	39.9289
	MIX	*	4.0927	59.9014	14.4010	249.1861
	PULM	*	4.0107	55.1855	13.0241	233.8313
Treatment	NONE		Ref	-	-	-
	ALT	*	-3.4862	0.0306	0.0092	0.1023
	CONV	*	-2.5036	0.0818	0.0363	0.1841
	HOM	*	-2.7423	0.0644	0.0241	0.1720
Isolation	1		Ref	-	-	-
	0	*	-0.4502	0.6375	0.4078	0.9967

Table 4.8: Stepwise Regression Steps for Binomial GLMM for Probability of Mortality, with "Place" and "Outbreak" nested in place as random effects.

		Place		Outbreak(Place)			
	AIC	Covariance Estimate	Standard Error	Covariance Estimate	Standard Error	Variable to drop	p-value
1	1088.84	0.0343	0.3486	1.0824	0.5242	Stabled	0.4747
2	1087.35	0.0301	0.3529	1.1043	0.5308	Province	0.1969
3	1082.37	0.3333	0.3132	0.8724	0.4597	Pesticides	0.1648
4	1082.32	0.3679	0.3075	0.8151	0.4432	OtherCases	0.0571
5	1083.99	0.3923	0.3092	0.7992	0.4433	Isolation	0.0576
6	1085.67	0.4170	0.3113	0.7644	0.4358		

effect, and with Place and Outbreak(Place), we see that the AIC for the latter model is marginally smaller (1082.32 compared to 1084.42). We also see that Outbreak(Place) has a covariance estimate which is not close to zero (0.8151). Accounting for outbreak as a random effect seems to be useful in the model. The standard errors in the latter model are marginally larger than for the first, which is to be expected since it is accounting for an extra source of variation.

4.4 Comparison of Techniques

Table 4.11 shows the odds ratios (e^{β_i}), standard errors (for original estimate) and 95% confidence intervals for the odds ratios for each of the three models formulated; GLM, GEE and GLMM. Two GLMMs are shown: GLMM1 refers to the model with only "Place" as random effect, and GLMM2 refers to the model with "Place" and "Outbreak(Place)" as random effects. Only the levels which were found to be significant are shown. The three estimates all agree to within a relatively small range of values. Neither the GEE nor the GLMMs found "Province" to be significant. This could be due to the fact that they accounted for "Place", as either a random or a clustering effect, and this made the "Province" variable unnecessary.

Although the estimates themselves may differ slightly between the three models, they all predict the same relationship between the levels in each case. 95% confidence intervals are shown in square brackets for clarity. For example all three models predict that Alternative treatment performs best (odds ratios 0.0555 [0.0197, 0.1563], 0.0451 [0.0065, 0.3137], 0.0306 [0.0092, 0.1023] and 0.0248 [0.0068, 0.0902] for the GLM, GEE and GLMM1 and 2 respectively), then Homeopathic (0.1153 [0.0514,0.2587], 0.0886 [0.0157, 0.5018], 0.0644 [0.0241, 0.1720] and 0.0531 [0.0188, 0.1505]), and then Conventional (0.1383 [0.0727, 0.2632], 0.1100 [0.0228, 0.5308], 0.0818 [0.0363, 0.1841] and 0.0721 [0.0309, 0.1682]), with all three treatment strategies performing significantly bet-

Table 4.9: Solutions for Fixed Effects for Binomial GLMM for the Probability of Mortality with Place and Outbreak(Place) as random effects.

Effect	Level	Estimate	CI		Std Error	DF	t Value	p-value
Intercept		0.2946	-1.4023	1.9915	0.8597	171	0.34	0.7323
Classification	VETC	Ref	-	-	-	-	-	-
	LABC	1.0822	0.2541	1.9103	0.4215	503	2.57	0.0105
	SUS	-0.1946	-0.7039	0.3148	0.2592	503	-0.75	0.4533
	SUSN	-0.6431	-3.2067	1.9205	1.3048	503	-0.49	0.6223
OtherCases	1	Ref	-	-	-	-	-	-
	0	0.3635	-0.01099	0.7380	0.1906	503	1.91	0.0571
Vaccinated	0	Ref	-	-	-	-	-	-
	1	-3.2313	-5.2890	-1.1736	1.0473	503	-3.09	0.0021
	2	-0.7953	-1.3278	-0.2628	0.2711	503	-2.93	0.0035
Presentation	MILD	Ref	-	-	-	-	-	-
	CARD	2.6025	1.1475	4.0575	0.7406	503	3.51	0.0005
	DK	2.5725	1.0745	4.0706	0.7625	503	3.37	0.0008
	MIX	4.4794	2.8845	6.0743	0.8118	503	5.52	<.0001
	PULM	4.3025	2.7100	5.8950	0.8106	503	5.31	<.0001
Treatment	NONE	Ref	-	-	-	-	-	-
	ALT	-3.6985	-4.9913	-2.4056	0.6581	503	-5.62	<.0001
	CONV	-2.6296	-3.4767	-1.7825	0.4312	503	-6.10	<.0001
	HOM	-2.9350	-3.9762	-1.8937	0.5300	503	-5.54	<.0001
Isolation	1	Ref	-	-	-	-	-	-
	0	-0.5167	-0.9951	-0.0384	0.2435	503	-2.12	0.0343

Table 4.10: Odds ratios and confidence intervals from binomial GLMM with Place and Outbreak(Place) as random effects. Significant levels are indicated with an asterisk.

Parameter	Level		Estimate	Adj. OR	95% CI (OR)	
Classification	VETC		Ref	-	-	-
	LABC	*	1.0822	2.9512	1.2893	6.7551
	SUS		-0.1946	0.8232	0.4947	1.3700
	SUSN		-0.6431	0.5257	0.0405	6.8244
OtherCases	1		Ref	-	-	-
	0		0.3635	1.4384	0.9891	2.0917
Vaccinated	0		Ref	-	-	-
	2	*	-0.7953	0.4514	0.2651	0.7689
	1	*	-3.2313	0.0395	0.0050	0.3093
Presentation	MILD		Ref	-	-	-
	CARD	*	2.6025	13.4974	3.1503	57.8296
	DK	*	2.5725	13.0985	2.9285	58.5921
	MIX	*	4.4794	88.1817	17.8946	434.5452
	PULM	*	4.3025	73.8843	15.0293	363.2168
Treatment	NONE		Ref	-	-	-
	ALT	*	-3.6985	0.0248	0.0068	0.0902
	CONV	*	-2.6296	0.0721	0.0309	0.1682
	HOM	*	-2.9350	0.0531	0.0188	0.1505
Isolation	1		Ref	-	-	-
	0	*	-0.5167	0.5965	0.3697	0.9624

ter than no treatment at all. All models also predicted that vaccinating late had odds of mortality between $\pm$ 0.04 and 0.07 times the odds of not vaccinating at all, while vaccinating timeously only halved the odds of mortality. Each model also predicted Mixed presentation as being the most severe (odds ratios 38.8847 [10.1625, 148.7847], 39.1197 [7.0830, 216.0479], 59.9014 [14.4010, 249.1861] and 88.1817 [17.8946, 434.5452] for the GLM, GEE and GLMMs respectively), with next being Pulmonary (36.1689 [9.3381, 140.1061], 36.9771 [6.1885, 220.9199], 55.1855 [13.0241, 233.8313] and 73.8843 [15.0293, 363.2168]), then Cardiac (8.5446 [2.4018, 30.3957], 9.0657 [1.7215, 47.7415], 11.5479 [3.0374, 43.9038] and 13.4974 [3.1503, 57.8296]) compared to Mild as the presentation reference.

In terms of the standard errors, the GLM standard errors were consistently smaller than the GLMM1, but only very slightly with the greatest difference being 0.0862. The reason for this is that the GLMM model accounts for an extra source of variability compared to the GLM. The GEE possessed standard errors both smaller and larger than those from the GLMMs and GLM. However, these standard errors were in all cases not largely different, and probably only due to the different methods. We therefore have confidence in the accuracy of all models, but probably most in the GLMMs and GEE since they are accounting for more of the variability by location. The choice of whether to use either GLMM or the GEE lies only in the particular interest of the model - the GEE should be used if one is interested in population averaged or marginal effects. The GLMM should be used if one wishes to use a random intercept model, and may be interested in the particular effect of a certain place. That is, if the focus is on cluster specific effects.

Table 4.11: Table showing the Adjusted Odds Ratio, Standard error (of the original estimate), and 95 % Confidence Interval for the Adjusted Odds Ratio for the final GLM, GEE and both GLMM models for the probability of mortality. GLMM1 refers to the GLMM with only “Place” as random effect, GLMM2 refers to the model with “Place” and “Outbreak(Place)” as random effects. Only estimates for significant levels are shown.

		GLM			GEE			GLMM1			GLMM2		
Parameter		OR	SE	CI for OR	OR	SE	CI for OR	OR	SE	CI for OR	OR	SE	CI for OR
Province	ECP	0.3252	0.3876	[0.1521, 0.6952]									
	GAU	0.5239	0.3253	[0.2769, 0.9912]									
	KZN	0.4902	0.3444	[0.2496, 0.9627]									
Classification	LABC	2.5350	0.3491	[1.2789, 5.0249]	2.2787	0.3605	[1.1241, 4.6191]	2.4281	0.3864	[1.1370, 5.1852]	2.9512	0.4215	[1.2893, 6.7551]
Vaccinated	2	0.4927	0.2277	[0.3154, 0.7698]	0.5208	0.1972	[0.3539, 0.7665]	0.4750	0.2536	[0.2887, 0.7815]	0.4514	0.2711	[0.2651, 0.7689]
	1	0.0685	0.9671	[0.0103, 0.4559]	0.0660	0.8841	[0.0117, 0.3733]	0.0531	1.0222	[0.0071, 0.3954]	0.0395	1.0473	[0.0050, 0.3093]
Presentation	CARD	8.5446	0.6475	[2.4018, 30.3957]	9.0657	0.8476	[1.7215, 47.7415]	11.5479	0.6802	[3.0374, 43.9038]	13.4974	0.7406	[3.1503, 57.8296]
	DK	8.5344	0.6601	[2.3406, 31.1184]	8.3361	0.8709	[1.5124, 45.9475]	10.2379	0.6931	[2.6253, 39.9289]	13.0985	0.7625	[2.9285, 58.5921]
	MIX	38.8847	0.6847	[10.1625, 148.7847]	39.1187	0.8719	[7.0830, 216.0479]	59.9014	0.7260	[14.4010, 249.1861]	88.1817	0.8118	[17.8946, 434.5452]
	PULM	36.1689	0.6909	[9.3381, 140.1061]	36.9771	0.9120	[6.1885, 220.9199]	55.1855	0.7354	[13.0241, 233.8313]	73.8843	0.8106	[15.0293, 363.2168]
Treatment	ALT	0.0555	0.5282	[0.0197, 0.1563]	0.0451	0.9900	[0.0065, 0.3137]	0.0306	0.6144	[0.0092, 0.1023]	0.0248	0.6581	[0.0068, 0.0902]
	CONV	0.1383	0.3282	[0.0727, 0.2632]	0.1100	0.8031	[0.0228, 0.5308]	0.0818	0.4133	[0.0363, 0.1841]	0.0721	0.4312	[0.0309, 0.1682]
	HOM	0.1153	0.4122	[0.0514, 0.2587]	0.0886	0.8846	[0.0157, 0.5018]	0.0644	0.5003	[0.0241, 0.1720]	0.0531	0.5300	[0.0188, 0.1505]
Isolation	0	0.6413	0.2056	[0.4286, 0.9593]	0.6725	0.2197	[0.4372, 1.0344]	0.6375	0.2276	[0.4078, 0.9967]	0.5965	0.2435	[0.3697, 0.9624]

Chapter 5

Climate Change and Predictions

5.1 Climate Change

The Intergovernmental Panel on Climate Change (IPCC) is the leading organization driving research into climate change. They are developing models on how climatic variables are likely to be affected by certain future emissions scenarios; namely A1B, A2, B1, B2, A1FI and A1T. The scenario we will be interested in is the A1 scenario, which is described as follows:

"The A1 storyline and scenario family describes a future world of very rapid economic growth, global population that peaks in mid-century and declines thereafter, and the rapid introduction of new and more efficient technologies. Major underlying themes are convergence among regions, capacity building and increased cultural and social interactions, with a substantial reduction in regional differences in per capita income. The A1 scenario family develops into three groups that describe alternative directions of technological change in the energy system. The three A1 groups are distinguished by their technological emphasis: fossil intensive (A1FI), non-fossil energy sources (A1T), or a balance across all sources (A1B) (where balanced is defined as not relying too heavily on one particular energy source, on the assumption that similar improvement rates apply to all energy supply and end-use technologies)." (IPCC, 2007)

Climate change predictions are available from the IPCC Regional Climate Projections utilizing MMD (multi-model dataset)-A1B. The MMD consists of 21 climate change prediction models. The predictions are given in the form of increase in degrees Celsius for Temperature, and percentage increase for Precipitation between 1980 to 1999 period and those predictions for 2080-2099. Minimum and Maximum predictions are given, along with 25th, 50th and 75th percentiles for the 21 models, each for the seasons DJF (December, January ,February), MAM (March, April, May), JJA (June, July,

Table 5.1: Minimum and Maximum, 25th, 50th, and 75th percentiles for IPCC MMD-A1B predictions, grouped into predictions from December - February (DJF), March - May (MAM), June - August (JJA), and September - November (SON).

	Temperature (increase $^\circ C$)					Precipitation (increase %)				
	Min	25	50	75	Max	Min	25	50	75	Max
DJF	1.8	2.7	3.1	3.4	4.7	-6	-3	0	5	10
MAM	1.7	2.9	3.1	3.8	4.7	-25	-8	0	4	12
JJA	1.9	3.0	3.4	3.6	4.8	-43	-27	-23	-7	-3
SON	2.1	3.0	3.7	4.0	5.0	-43	-20	-13	-8	-3

August) and SON (September, October, November). The predictions for South Africa (co-ordinates 35S,10E - 12S,52E) are given in Table 5.1.

The two models given in equations 5.1 and 5.1 were tested with the 25 different combinations of scenarios. The letters a, b, c, d, and e were assigned to the Minimum, 25, 50, 75th percentiles and Maximum respectively, so that for example pred1ad represents the prediction from model 1, with the Minimum of temperature, and the 75th percentile for precipitation. This was done by taking the average monthly climatic variables from our data, and altering them according to the different scenarios and using our models in Equations 3.45 and 3.46 to predict the incidence.

$$\log(\mu) \quad = \quad 9.3447 + 2.5690\sin(2\pi t) - 0.5674.TMax + 0.4399.TMin - 0.0060.Rain$$

$$\begin{aligned}\log(\mu) \quad = \quad & -0.8226 + 2.1965.sinyr + 1.0603Tmin - 0.0434.Rain \\ & -0.0099Tmax^2 - 0.0403.Tmin^2 - 0.0004.Rain^2 + 0.0065.Tmin \times Rain\end{aligned}$$

The model which predicted the worst outbreak was model 1 with minimum temperature and precipitation predictions - although the predictions were no worse than the observed average from the data. The plots of all model 1 and model 2 predictions are shown in Figures 5.1 and 5.2. The highest 5 predictions for each are then shown in Figures 5.3 and 5.4. For ease of reference, the axes are maintained constant with maximum of 18. For comparison the plot of the observed average incidence is shown in Figure 5.5. The model 2 predictions all had a strange pattern, with a decrease in the number of cases in February, followed by a sharp increase in March, and the usual declining pattern thereafter. One notable difference is that the observed average for

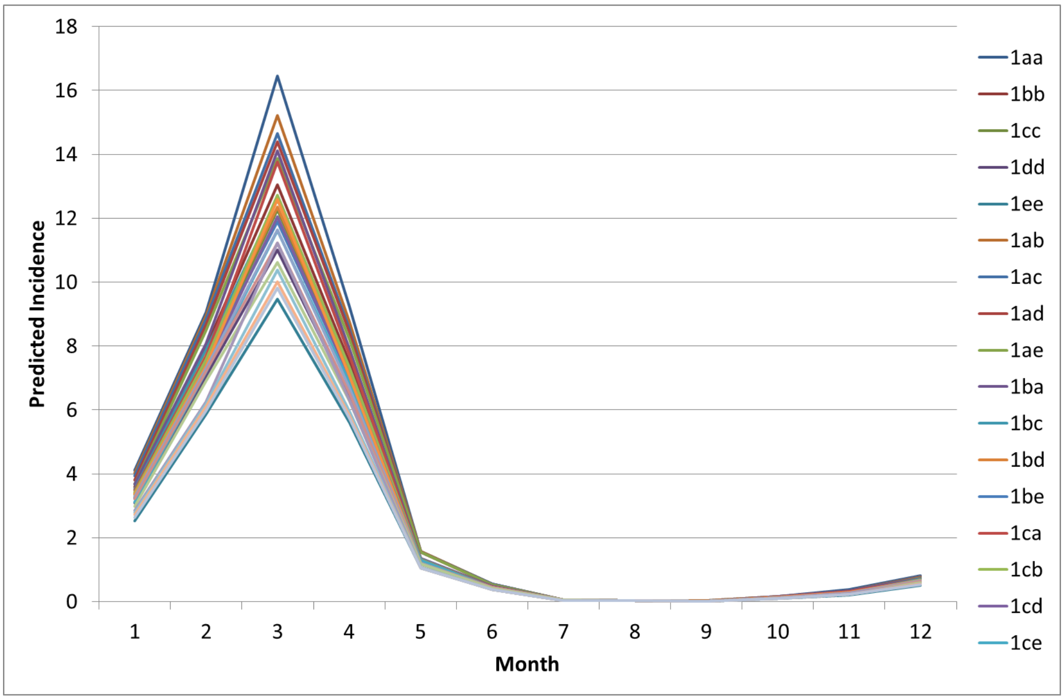

Figure 5.1: Predictions for incidence from model 1 with climate change following MMD-A1B predictions. The letters a, b, c, d, e refer to the minimum, 25th percentile, 50th percentile, 75th percentile and maximum respectively from the MMD-A1B predictions for climate change.

both models in month 6 is above zero, when there had been only one observed case for this month in the data. This shows that with climate change, the disease may have a longer season. In fact, the predictions show only 3 'zero' months, in July to September.

However, it is important to take these predictions with caution. Our models predict the incidence based on the climatic variables. However, the relationship between the two is indirect, as the climate variables drive the vector population and virogenesis, which directly drive the disease incidence. This relationship is shown schematically in Figure 5.6. While these models may predict best conditions for the current major vector, *C. imicola*, it is possible as discussed by Wittman and Baylis (2000), that changing climatic conditions may make it possible for other potential vector species to flourish. Thus, unless all potential vector species are taken into account, these models can serve as a guideline only. In particular, Mellor (2000) discusses the possible vector capacity of *C.obsoletus* and *C.pulicaris* which are two of the most abundant *Culicoides* species around the world.

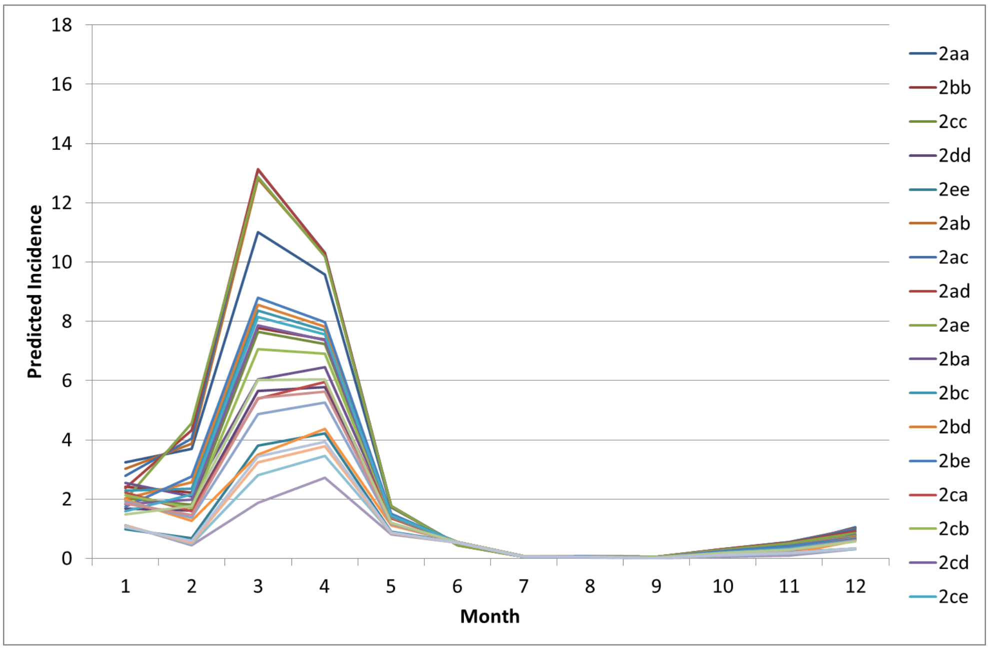

Figure 5.2: Predictions for incidence from model 2 with climate change following MMD-A1B predictions. The letters a, b, c, d, e refer to the minimum, 25th percentile, 50th percentile, 75th percentile and maximum respectively from the MMD-A1B predictions for climate change.

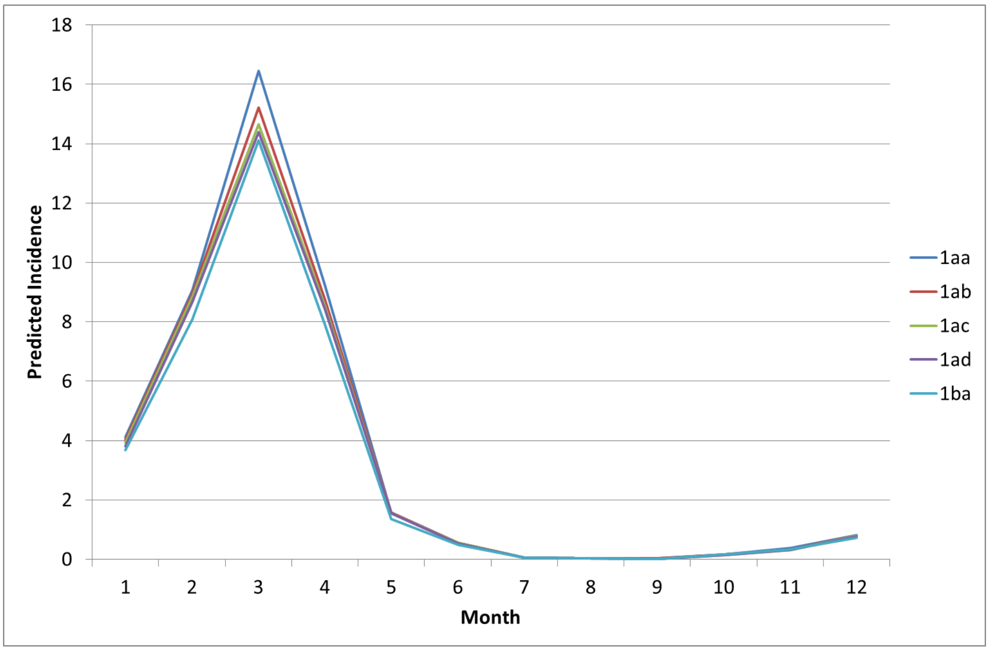

Figure 5.3: The highest five predictions for incidence from model 1 with climate change following MMD-A1B predictions.

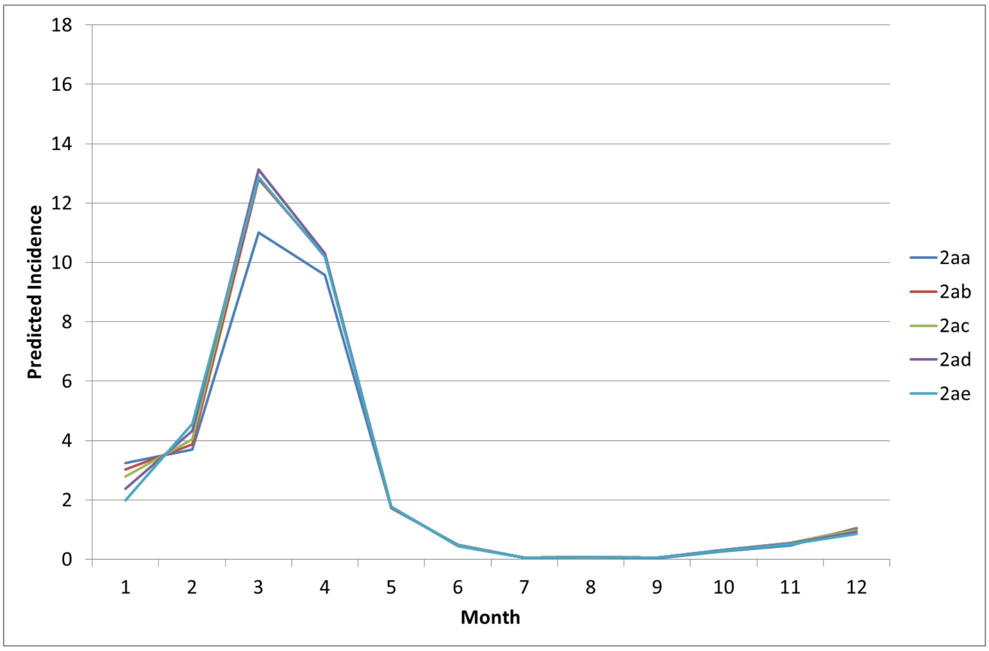

Figure 5.4: The highest five predictions for incidence from model 2 with climate change following MMD-A1B predictions

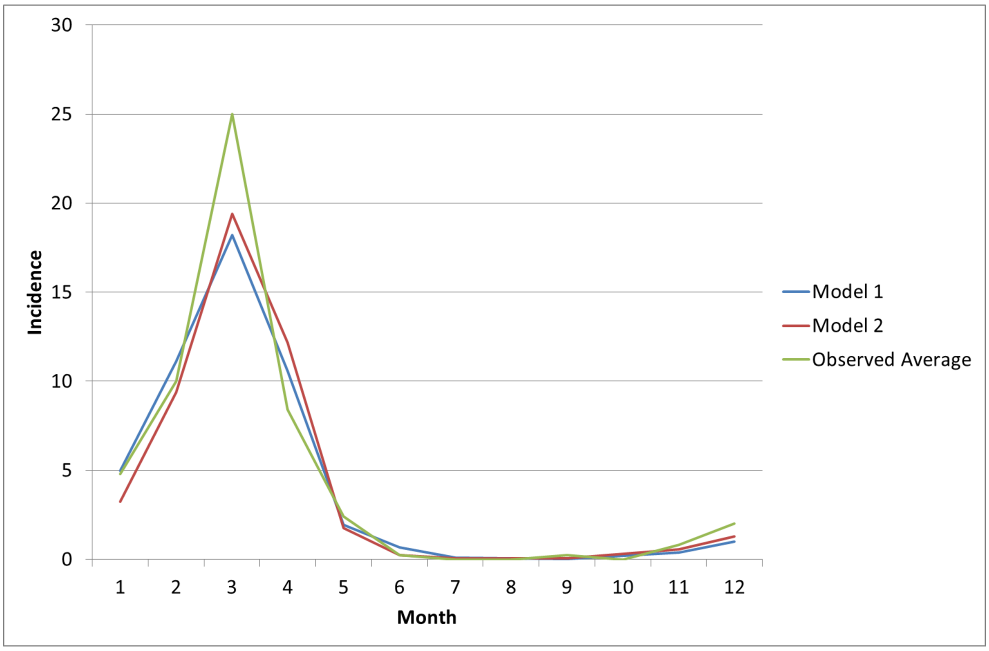

Figure 5.5: Observed average incidence shown over a year, exhibiting the seasonal pattern, along with the predicted averages from Model 1 and Model 2 for the observed weather data.

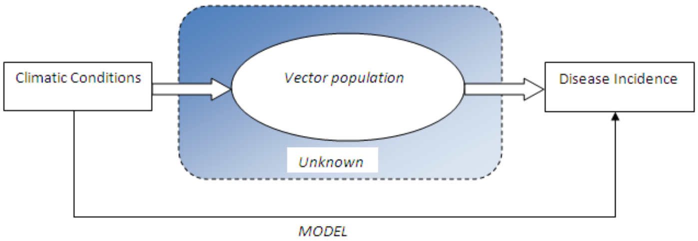

Figure 5.6: Schematic diagram representing the disease dynamics. Our model describes the disease incidence based on climatic variables. However, in reality, these climatic variables drive other factors (which are un-quantified) on which the incidence depends. This illustrates why the predictions for climate change from these models should be taken with caution.

5.2 Refractory Periods

Koelle et al. (2005) discuss a "refractory period" in cholera epidemiology, which is described as the period "when population susceptibility levels are low as the result of immunity and the size of cholera outbreaks only weakly reflects climate forcing". They discovered that there existed this refractory period within the cholera epidemics, where an outbreak of cholera following a particularly severe outbreak may be much smaller, despite the climatic conditions being favourable for the disease. There is a possibility that this refractoriness may explain some of the difference between our existing model for disease incidence and the true observed values (Young, 2010, *Personal Communication*), and we wish to investigate this further.

John and Samuel (2000) suggested the definition of 'herd immunity' to be "the proportion of subjects with immunity in a given population". They propose that the effect of the immune segment of the population protecting the susceptible portion should be described as "herd effect". The herd effect is therefore dependent on the herd immunity and the force of infection of the disease. Since we cannot quantify these two effects in our case we presume both to have an effect on the refractory period if it is found to exist.

We therefore wish to attempt to model the incidence of AHS in Johannesburg introducing variables which may explain this refractory period. Johannesburg is chosen for two reasons. Firstly, it is a much more localized area and therefore the equine population of the Johannesburg area could be said to constitute a "herd", where we better understand the effects of herd immunity and herd effect. The locations where cases had occurred are shown in Figure 5.7, which clearly shows the highly localized pattern of the cases. Secondly, we have a more accurate measure of the population size. This is due to the fact that most equines in the area are registered with the Gauteng Horse Society (GHS). The GHS registration number is around 2600 (www.thsinfo.co.za/THSOverview.html).

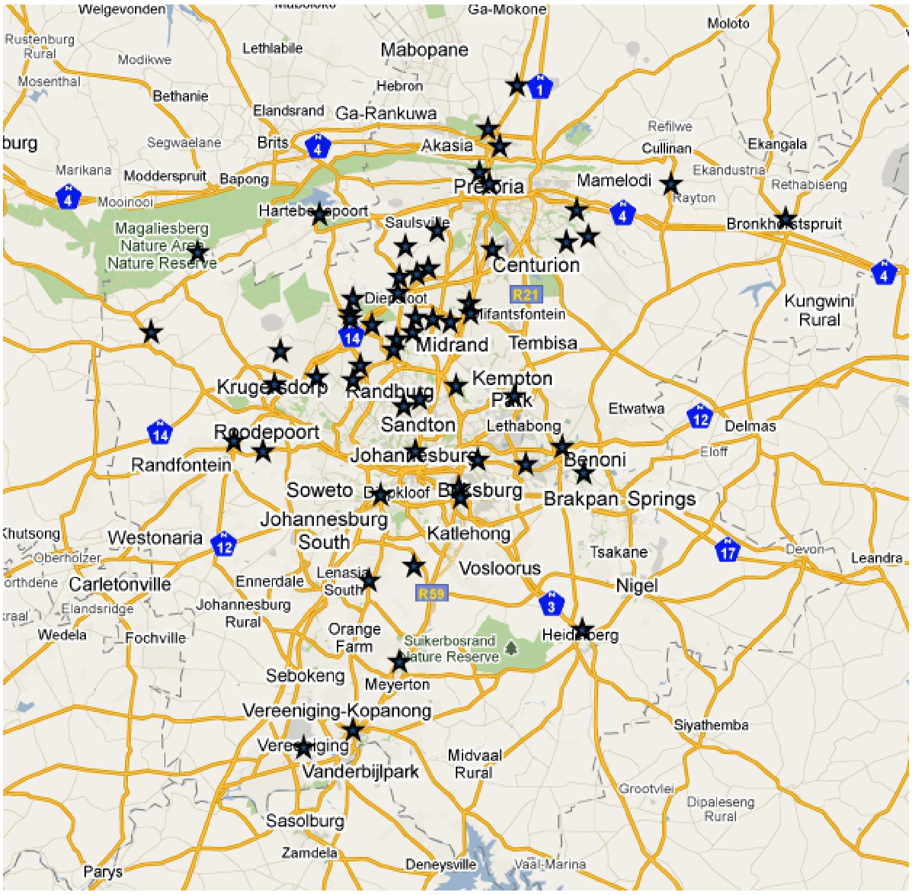

Figure 5.7: Locations of cases of AHS in the Johannesburg and surrounds area for the AHST data. Each star shows an incidence of cases, not the number observed or date. Map approximately -25.4 to -26.9 latitude 27.3 to 28.9 longitude. (©2010 Google - Map Data ©2010 AfriGIS (Pty) Ltd, Tele Atlas, Tracks4Africa)

Since the cost of keeping a horse in Johannesburg is very high at often over 3000 South African Rand per month, most horse owners in the area keep horses in order to compete. To compete they must be registered with the Gauteng Horse Society - and thus the population size is likely to be close to the number of registrations with GHS. Climatic information was acquired once again from the South African Weather Service, and the locations from which the data came are shown in Figure 5.8.

We wish to add to the models a term which reflects the severity of the previous outbreaks. To this end we created a ‘severity index’. This was calculated by taking the total number of cases in an outbreak (usually between October and May of the following year) and dividing by the estimate of the population size. Even if the estimate of population size is not accurate, it still serves to give a fraction indicating the severity of an outbreak, and is kept constant over all outbreaks. This severity index was assumed

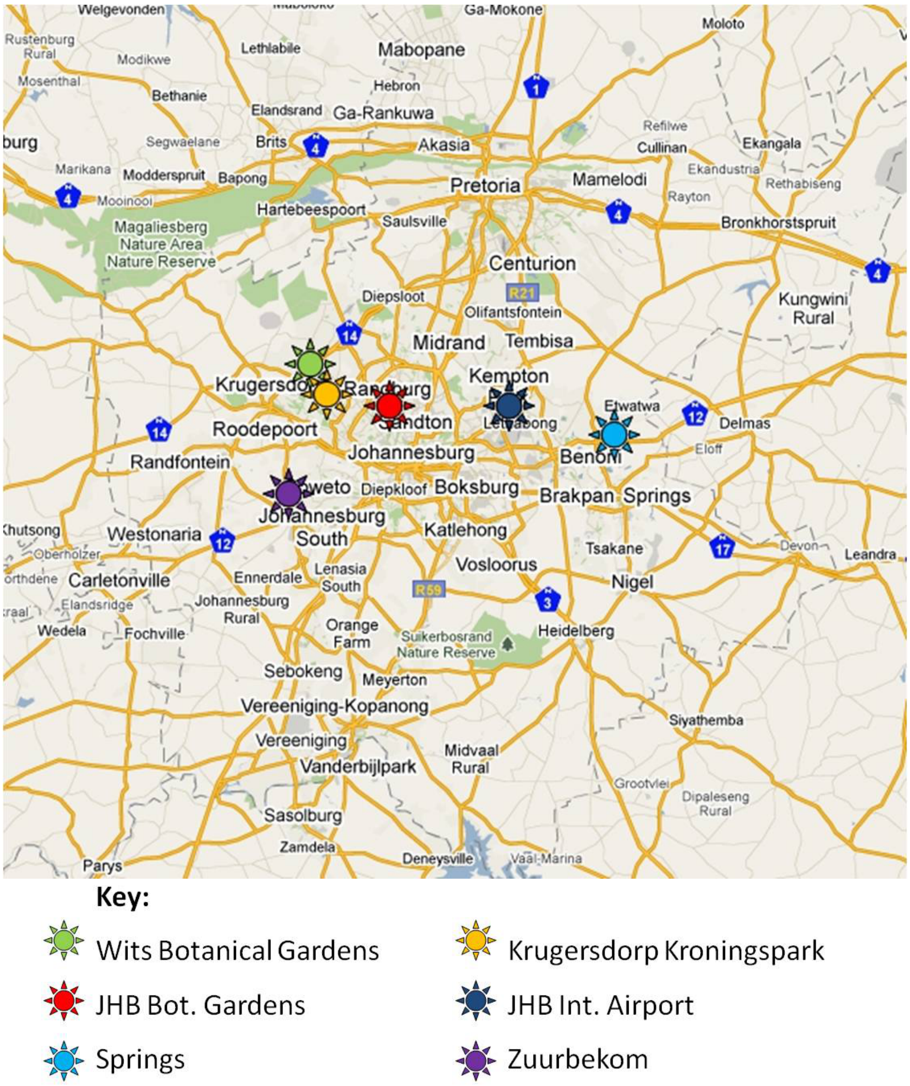

Figure 5.8: Locations of the weather stations for which data was acquired from the South African Weather Service. Map approximately -25.4 to -26.9 latitude 27.3 to 28.9 longitude. (©2010 Google - Map Data ©2010 AfriGIS (Pty) Ltd, Tele Atlas, Tracks4Africa)

to apply to the year between September of one year and August of the next. It was then added at lags of one year and two years to the model - SI1 was the severity index from the previous year's outbreak, and SI2 the severity index from the outbreak two years prior. Estimates for the total number of cases for 2003/2004 and 2004/2005 seasons outbreaks in JHB were given as 112 and 49 respectively from the Department of Agriculture website (http://www.daff.gov.za/), since they were not available from the AHST data. The impact of seasons prior to this are not possible to explore with the current data.
Four models were then fitted. The first was fitted with only time and climate variables. In the second model the variable SI1 was added, in the third SI2, and in the fourth both SI1 and SI2 were used. In each case the climatic and time variables used were $sinyr, Tmax, Tmin, Rain, Tmax^2, Tmin^2, Rain^2, Tmax \times Rain, Tmin \times Rain$. In each case the stepwise procedure followed in the previous chapters was used. The tables of the stepwise procedures and final model fit are shown in Table 5.2 and Table 5.3, and the plots in Figures 5.9, 5.10 and 5.11.

Figure 5.9 shows the best model for Johannesburg incidence with only climatic variables included (Model 1). Although it certainly explains some of the variation, it has similar lack of fit features as those found in the model for KZN. Figure 5.10 shows the plot of the model when SI1 is included (Model 2). The plot shows a better fit than the previous model. There are still some differences in the observed and predicted peaks, but they are substantially decreased. When SI2 is fitted it shows an increased accuracy (Model 3) (See Figure 5.11). However, when both SI1 and SI2 are fitted along with the other variables, SI1 becomes insignificant and is dropped from the model. Thus Model 4 is the same as Model 3, and SI2 is a superior explanatory variable to SI1. This shows that the outbreak severity from two years prior explains much of the variability in the current outbreak, and thus that the refractory period has a lag of at least two years.

It has been shown by these models that some sort of refractory period exists, within which the herd immunity protects the population from severe outbreak despite climatic conditions being favourable for the disease. We cannot reliably explore the length of this refractory period with only five years of data, but we postulate that it may be further than the two years discovered here since Model 3 still did not have perfect fit. The length of the immunity should be further investigated for predictive purposes. For the purposes of early warning systems knowledge of the refractory periods is vital. For example, the African Horse Sickness Trust may be advised, after a severe outbreak in a particular area, that blanket vaccination efforts may not be entirely necessary. On the contrary, when an area has not experienced severe outbreaks for some time they may know that blanket vaccinations in that area will be vital. If the length of the refractory period is determined to be, for example, three years, they may also know that blanket vaccinations should be practiced every three years or less in order to prevent large numbers of mortalities.

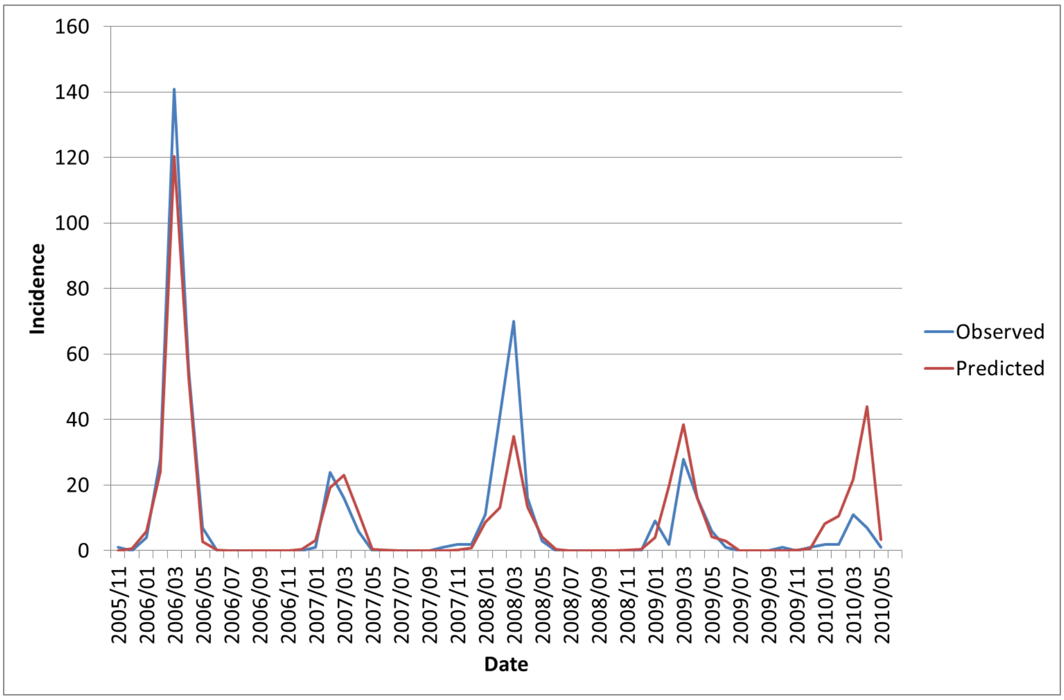

Figure 5.9: Model 1 : $\log(\mu_i) = 31.7656 + 3.4912 sinyr - 2.9563.Tmax + 1.4629Tmin - 0.0199.Rain + 0.0511Tmax^2 - 0.0468Tmin^2 + 0.0001Rain^2$

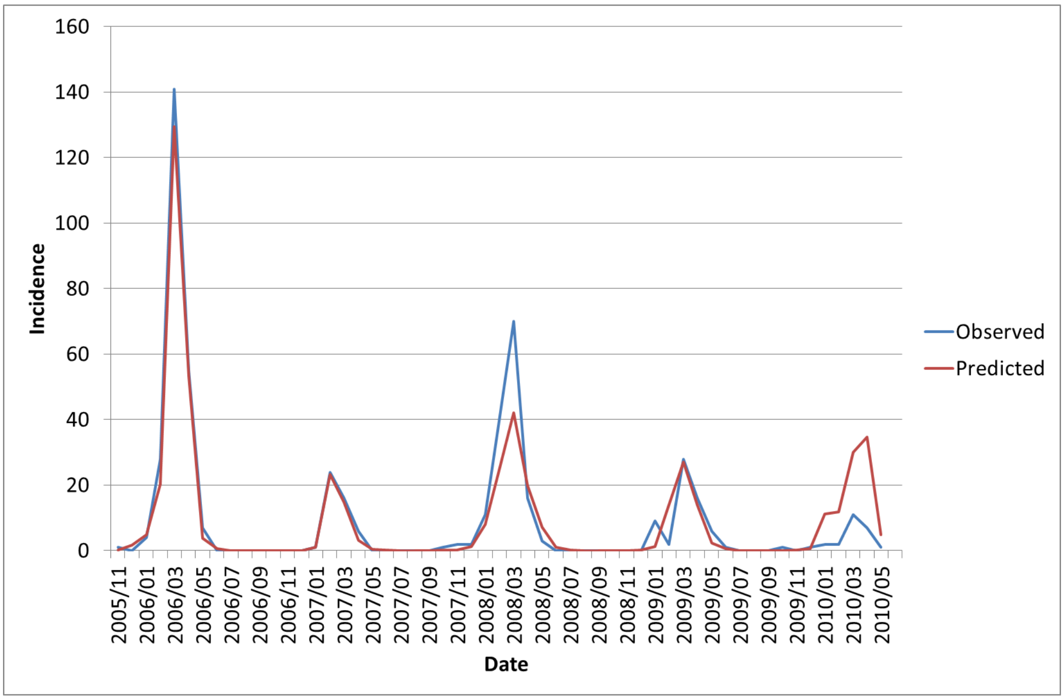

Figure 5.10: Model 2: $\log(\mu_i) = 5.7338 + 3.2434 sinyr - 0.3308.Tmax + 0.0501.Rain + 0.0413.Tmin^2 + 0.0003Rain^2 - 0.0091.Rain \times Tmin - 25.3877.SI1$

Table 5.2: Model selection process for the three models for Johannesburg incidence. Model 1 has no severity index included, Model 2 has severity index at lag 1, Model 3 at lag 2, and Model 4 includes severity index at lags 1 and 2.

	Model Information			Model Checking		Variable to be dropped		
	Log-Likelihood	Deviance	DF	Change in Deviance	$P > \chi^2_{(df)}$	Variable	Type III p-value	df
				Model 1				
1	1274.7575	220.1564	45	-	-	$Tmin \times Rain$	0.3473	1
2	1274.3212	221.0290	46	0.8726	0.3502	$Tmax \times Rain$	0.2554	1
3	1273.6851	222.3012	47	1.2722	0.2594	none		
				Model 2				
1	1297.8157	174.0399	44	-	-	$Tmax^2$	0.7500	1
2	1297.7651	174.1411	45	0.1012	0.7504	$Tmax \times Rain$	0.3720	1
3	1297.3636	174.9440	46	0.8029	0.3702	$Tmin$	0.2113	1
4	1296.5391	176.5930	47	1.6490	0.1991	none		
				Model 3				
1	1307.7342	154.2028	44	-	-	$Tmin^2$	0.6781	1
2	1307.6476	154.3760	45	0.1732	0.6773	$Tmax \times Rain$	0.3713	1
3	1307.2506	155.1700	46	0.7940	0.3729	$Rain$	0.2314	1
4	1306.5272	156.6169	47	1.4469	0.2290	none		
				Model 4				
1	1307.9046	153.862	43	-	-	$SI1$	0.5590	1
2	1307.7342	154.2028	44	0.3408	0.5594	$Tmin^2$	0.6781	1
3	1307.6476	154.3760	45	0.1732	0.6773	$Tmax \times Rain$	0.3713	1
4	1307.2506	155.1700	46	0.7940	0.3729	$Rain$	0.2314	1
5	1306.5272	156.6169	47	1.4469	0.2290	none		

Table 5.3: Parameter estimates for the three models showing severity index

Analysis Of Parameter Estimates							
Parameter	DF	Estimate	Std Error	Wald 95% CI		χ^2	Pr > χ^2
			Model 1				
Intercept	1	31.7656	5.7221	20.5504	42.9807	30.82	<.0001
$sinyr$	1	3.4912	0.3508	2.8037	4.1788	99.06	<.0001
$Tmax$	1	-2.9563	0.5086	-3.9532	-1.9594	33.78	<.0001
$Tmin$	1	1.4629	0.1896	1.0914	1.8345	59.55	<.0001
$Rain$	1	-0.0199	0.0041	-0.0280	-0.0118	23.39	<.0001
$Tmax^2$	1	0.0511	0.0103	0.0309	0.0713	24.54	<.0001
$Tmin^2$	1	-0.0468	0.0085	-0.0635	-0.0302	30.30	<.0001
$Rain^2$	1	0.0001	0.0000	0.0000	0.0001	18.01	<.0001
			Model 2				
Intercept	1	5.7338	0.8369	4.0936	7.3741	46.94	<.0001
$sinyr$	1	3.2434	0.2858	2.6832	3.8037	128.75	<.0001
$Tmax$	1	-0.3308	0.0434	-0.4159	-0.2458	58.12	<.0001
$Rain$	1	0.0501	0.0092	0.0321	0.0680	29.83	<.0001
$Tmin^2$	1	0.0413	0.0036	0.0343	0.0483	133.93	<.0001
$Rain^2$	1	0.0003	0.0000	0.0002	0.0004	83.55	<.0001
$Rain * Tmin$	1	-0.0091	0.0010	-0.0112	-0.0071	78.73	<.0001
$SI1$	1	-25.3877	3.4043	-32.0600	-18.7155	55.62	<.0001
			Model 3 & 4				
Intercept	1	18.4110	6.1355	6.3857	30.4363	9.00	0.0027
$sinyr$	1	3.2204	0.3208	2.5917	3.8492	100.79	<.0001
$Tmax$	1	-1.6940	0.5127	-2.6988	-0.6891	10.92	0.0010
$Tmin$	1	0.9094	0.0697	0.7729	1.0460	170.35	<.0001
$Tmax^2$	1	0.0230	0.0103	0.0028	0.0432	4.97	0.0258
$Rain^2$	1	0.0002	0.0000	0.0002	0.0003	65.12	<.0001
$Tmin * Rain$	1	-0.0046	0.0005	-0.0056	-0.0037	92.60	<.0001
$SI2$	1	19.0685	2.2627	14.6338	23.5032	71.02	<.0001

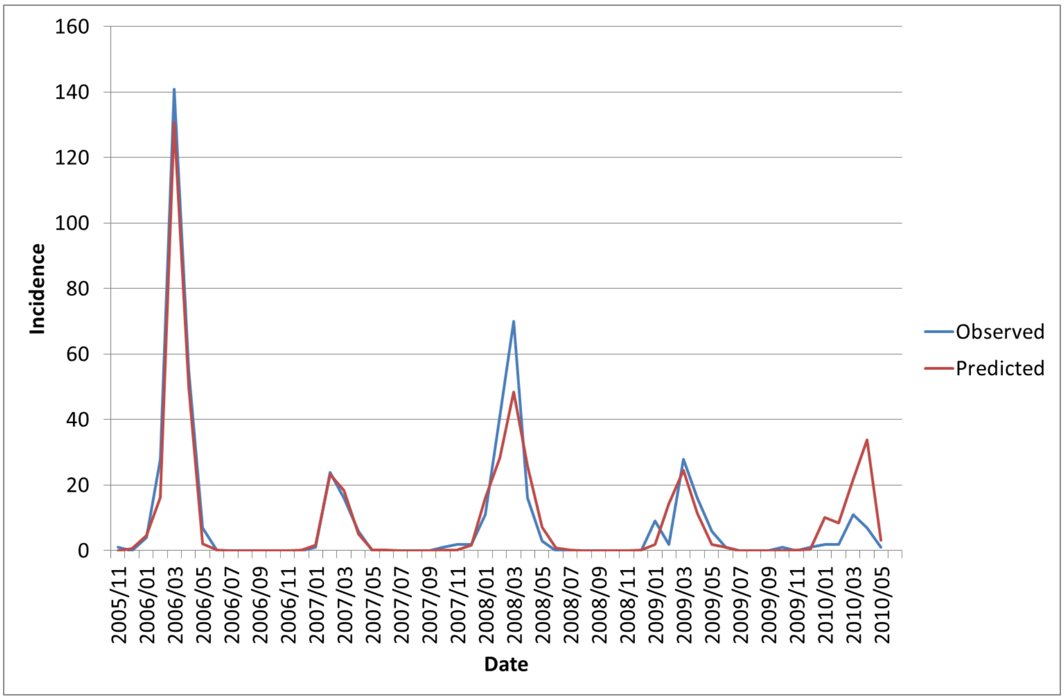

Figure 5.11: Model 3: $\log(\mu_i) = 18.4110 + 3.2204.sinyr - 1.6940.Tmax + 0.9094.Tmin + 0.0230.Tmax^2 + 0.0002.Rain^2 - 0.0046.Tmin \times Rain + 19.0685.SI2$

Chapter 6

Conclusions and Discussion

African Horse Sickness is still a major problem in South Africa, and as such requires much further work in order to understand the mechanisms which can be used to predict and ultimately to control the disease. This work has attempted to use the data available to explore some of these mechanisms. To this end both the incidence and mortality of the disease have been modelled in various ways.

Before proceeding with the modelling, however, it is important to understand the reliability of the data which is to be utilised. In the current study we know the data may be deficient by its very nature. We are well aware that not all cases are reported, particularly due to the lack of education amongst the owners of horses in the rural areas. Even when reported, the majority of cases were not sent for laboratory testing due to the cost and difficulties of administration associated with laboratory tests. The format in which the cases were reported was also not optimal as only the "primary" case in an area was recorded with accuracy, after which the additional horses presenting symptoms similar to AHS were simply recorded as the number alive and number dead. This meant that these additional cases could not be used for modelling the probability of mortality, as the explanatory variables on an individual basis (such as vaccination, presentation etc.) had not been recorded. Although it is important to keep these considerations in mind when fitting and analysing the models, however, we use this data in the knowledge that it is a measure of what is really happening in an AHS outbreak even if it is not the entire picture.

Because of the non-normality of counts data it was realised that standard general linear regression would not be adequate to model the incidence. Therefore a Generalized Linear Model with a log link was utilised. The variables chosen to explain the incidence were time and meteorological variables. Quadratic and interaction terms were also investigated in case the relationships were not strictly linear. The final model found to fit the data best included the terms $Tmin$, $Rain$, $Tmin^2$, $Tmax^2$, $Rain^2$ and $Tmin \times Rain$ to be significant, as well as a sine function of time. It was therefore evident that the incidence relied most heavily on minimum temperature and rainfall. From Figure 3.8 it is evident that incidence is highest at high minimum temperatures and with moderate rainfall. At minimum temperatures approaching freezing the predicted incidence tends towards zero, and the same with extreme high or low rainfall. We know this to be the case as *Culicoides imicola* requires temperate weather, with moistness, to reproduce. Over the range of maximum temperatures the incidence decreases since the relationship with all other variables remaining constant is $\log \mu \propto -0.0099Tmax^2$. Thus very high maximum temperatures will not favour the propagation of the disease. Although relationships of Culicoides abundance to the weather have been investigated in previous publications (Baylis, Meiswinkel and Venter, 1999), and the fact that there exists a relationship between AHS and weather is well established, it has not, to our knowledge, been quantified in this manner previously.

However, in plots of this model (Figure 3.2) it appeared as though there could be other explanatory variables which might explain more of the variation in the data. In Section 5.2 we explored refractory periods as a possible explanation for this. The fitted models showed strong evidence of AHS having a refractory period of at least two years during which the population is protected to an extent by the immunity acquired after a large outbreak. This is an entirely novel finding for AHS, although it has been discovered for other diseases such as cholera (Koelle et.al., 2005). The final length of the refractory period could not be ascertained with the short five years of data which we had available to us. With further and more accurate data, however, it will be possible to repeat this method and determine the exact length of this period. In particular this has important implications for the planning of vaccination drives and other protective measures, as it will enable us to use historical and current data to predict future outbreaks.

We also investigated as reported in Section 5.1 what role climate change may have in the incidence of the disease within South Africa. Climate change predictions were used from the IPCC MMD-A1B. These predictions were used to alter the observed weather variables from our dataset, and then the models were re-run. Although it was found that with climate change the disease may have a longer season, our predictions did not find an increase in incidence. This may be due to the fact that the predictions from the IPCC involve an overall decrease in precipitation, which may make the climate less hospitable to the vector *C. imicola*. However it is important to note that *C. imicola* is not the only capable vector species of the disease, and that climate change may bring about conditions which are favourable to other potential midge vectors.

In Chapter 3.11 the Generalized Linear Model theory was applied to the probability of mortality of the cases on an individual level. The aim of this modelling was to discover which explanatory variables had an impact on the probability that a horse would die. Type III p-values were used to determine the significance of an effect. The findings were discussed in detail in Section 3.11.1. However, it was considered important to further investigate whether the structure of the data, in particular the fact that the cases could be considered to be clustered by place, might affect the model. For these ends two further modelling techniques were used.

Both Generalised Estimating Equations in Section 4.2.3 and Generalised Linear Mixed Models in Section 4.3.6 were utilised to account for the clustering effect of “Place”. They yielded similar results to the GLM, although neither found “Province” to be a significant variable. This indicates that clustering according to place was strong and useful in providing a more parsimonious model. Generalized Linear Mixed Models were also fitted using “Place” and “Outbreak” nested in place as random effects. The differences between and relative merits of the four models were discussed in Section 4.4. All four models, however, are useful; depending on what the requirements of the model are.

The models presented in this thesis form an initial attempt to model African Horse Sickness in the South African context, although much further work is necessary if the country is truly to get this disease under control. However, it should be recognized that this constitutes a body of work from which other continents, such as Europe, can gain information and form plans of action should epizootics occur there. This is especially important since climate change guarantees that vectors of the disease will move to areas previously inhospitable to them, thereby having the ability to carry the virus to areas where it is not endemic. Together with recent advances in serotyping via a rapid diagnostic assay (Groenink, 2009), modeling work such as this would enable policy makers to form early warning systems and plan vaccination campaigns using monovalent vaccines in order to both reduce mortality from, and ultimately bring an end to, an epizootic. Furthermore it should be evident that this work has applicability not only to African Horse Sickness, but also to other vector-borne viruses of the family reoviridae, such as Bluetongue Virus and Epizootic Hemorrhagic Disease Virus in ruminants, and Equine Encephalosis Virus in horses, as they all have similar mechanisms of transfer. Other vector-borne diseases such as malaria research and control in humans can also benefit from the modelling approach suggested in the current thesis.

In conclusion, it is hoped that this work will see the beginning of much more such research into this problem, which could kickstart major advances for early warning systems, planning processes and prophylaxis that will contribute towards the control of African Horse Sickness in South Africa in the future.

Bibliography

[1] Abramowitz, M. and Stegun, I. (1964), "Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables," New York: Dover Publications.

[2] African Horse Sickness Trust (2010a), "African Horse Sickness Information Booklet"

[3] African Horse Sickness Trust (2010b), "African Horse Sickness Controlled Area"

[4] African Horse Sickness Trust (2005), "About the African Horse Sickness Trust", accessed March 2010 (Available at http://www.africanhorsesickness.co.za/about_ahs_trust.asp)

[5] Agresti, A. (2002), "Categorical Data Analysis," New Jersey: John Wiley and Sons.

[6] Baylis, M., Mellor, P. S., and Meiswinkel, R. (1999), "Horse Sickness and Enso in South Africa," *Nature*, 397, 574.

[7] Baylis M., Meiswinkel R., and Venter G.J. (1999), "A preliminary attempt to use climate data and satellite imagery to model the abundance and distribution of Culicoides imicola (Diptera: Ceratopogodinae) in southern Africa," *Journal of South African Veterinary Association*, 70, 80-89.

[8] Boorman, J., Mellor, P. S., Penn, M., and Jennings, M. (1975), "The Growth of African Horse-Sickness Virus in Embryonated Hen Eggs and the Transmission of Virus By *Culicoides Variipennis Coquillett* (Diptera, Ceratopogonidae)," *Archives of Virology*, 47, 343-349.

[9] Braverman, Y., Lirrley, J. R., Marcus, R., and Frish, K. (1985), "Seasonal Survival and Expectation of Infective Life of *Culicoides* Spp. (Diptera: Ceratopogonidae) in Israel, with Implications for Bluetongue Virus Transmission and a Comparison of the Parous Rate in *C. imicola* from Israel and Zimbabwe," *Journal of Medical Entomology*, 22, 476-484.

[10] Braverman, Y. (1989), "Control of Biting Midges *Culicoides* (Diptera: Ceratopogonidae), Vectors of Bluetongue and Inducers of Sweet Itch: A Review," *Israel journal of veterinary medicine*, 45, 124.

[11] Braverman, Y., Chizov-Ginzburg, A., and Mullens, B. A. (1999), "Mosquito Repellent Attracts *Culicoides imicola* (Diptera: Ceratopogonidae)," *Journal of Medical Entomology*, 36, 113-115.

[12] Braverman, Y. and Chizov-Ginzburg, A. (1997), "Repellency of Synthetic and Plant-Derived Preparations for *Culicoides imicola*," *Medical and Veterinary Entomology*, 11, 355-360.

[13] Breslow, N.E. and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9-25.

[14] Breslow, N. (2003), "Whither, PQL?" *UW Biostatistics Working Paper Series*, Working Paper 192 http://www.bepress.com/uwbiostat/paper192

[15] Coetzer, J. A. W., and Erasmus, B. J. (1994), "African Horse Sickness," in *Infectious Diseases of Livestock with Special Reference to Southern Africa (Vol. 1)*, eds. J. A. W. Coetzer, G. R. Thomson and R. C. Tustin, Cape Town: Oxford University Press, pp. 460-475.

[16] Cox D.R. and Hinkley D.V. (1974),"Theoretical Statistics," Chapman and Hall, London.

[17] Diggle P. J., Heagerty P., Liang K. Y., and Zeger S. L. (1994), *Analysis of Longitudinal Data (Second ed.)*, New York: Oxford University Press.

[18] Du Toit R.M. (1994) "The transmission of bluetongue and horse sickness by *Culicoides*," *Onderstepoort Journal of Veterinary Science and Animal Industry*, 19, 7-16.

[19] Hedeker, D. and Gibbons, R. D. (2006), "Longitudinal Data Analysis", Hoboken, New Jersey: John Wiley and Sons.

[20] International Panel on Climate Change (2007), "Working Group I: The Scientific Basis", accessed September 2010 (Available at http://www.ipcc.ch/ipccreports/tar/wg1/029.htm)

[21] Jenkins, A. B. (2008), "A study of the *Culicoides* (Diptera: Ceratopogodinae) vectors of African Horse Sickness to enhance current practical control measures and research methods," *Department of Animal and Poultry Science*, University of KwaZulu Natal.

[22] John, T. J. and Samuel, R. (2000), "Herd immunity and herd effect: new insights and definitions," *European Journal of Epidemiology*, 16, 601-606.

[23] Jones, R. and Foster, N. (1971), "Transovarian Transmission of Bluetongue Virus Unlikely for *Culicoides Variipennis*," *Mosquito News*, 31, 434-437.

[24] Koelle K., Rodó, X., Pascual, M., Yunus, M., and Mostafa, G. (2005), "Refractory periods and climate forcing in cholera dynamics," *Nature*, 436, 696-700.

[25] Liang, K.Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis using Generalized Linear Models," *Biometrika*, 73, 13-22.

[26] Liu Q. and Pierce D.A. (1994), "A note on Gauss-Hermite Quadrature," *Biometrika*, 81, 624-649.

[27] Lord C. C., Woolhouse M. E. J., Heesterbeck J. E. P., and Mellor P. S. (1996a), "Vector-borne diseases and the basic reproduction number: a case study of African Horse Sickness," *Journal of Medical and Veterinary Entomology*, 10, 19-28.

[28] Lord C. C., Woolhouse M. E. J., Rawlings P., and Mellor P. S. 1996b "Simulation Studies of African Horse Sickness and *Culicoides imicola* (Diptera: Ceratopogonidae)," *Journal of Medical and Veterinary Entomology*, 33, 328-338.

[29] Lord C. C., Woolhouse M. E. J., and Mellor P. S. (1997),"Simulation Studies of Vaccination Strategies in African Horse Sickness," *Vaccine*, 15, 519 - 524.

[30] McCullagh P. and Nelder J. A. (1983), *Generalized Linear Models*, London: Chapman and Hall.

[31] Meiswinkel, R., Baylis, M. and Labuschagne, K. (2000), "Stabling and the protection of horses from Culicoides bolitinos (Diptera: Ceratopogonidae), a recently identified vector of African horse sickness," *Bulletin of Entomological Research*, 90, 509-515.

[32] Meiswinkel, R. and Paweska J. T. (2003), "Evidence for a new field *Culicoides* vector of African horse sickness in South Africa," *Preventive Veterinary Medicine*, 60, 243-253.

[33] Mellor, P. S., Boned, J., Hamblin, C. and Graham, S. (1990). "Isolations of African Horse Sickness Virus from Vector Insects Made during the 1988 Epizootic in Spain," *Epidemiology and Infection*,2, 447-454.

[34] Mellor, P. S., Rawlings, P., Baylis, M. and Wellby, M. P. (1998), "Effect of temperature on African horse sickness virus infection in Culicoides," *Archives of Virology Supplement* 14, 155-163.

[35] Mellor, P. S. (1999), "African Horse Sickness: Transmission and Epidemiology," *Veterinary Research*, 24, 199-212.

[36] Mellor, P.S. and Hamblin, C., (2004),"African Horse Sickness," *Veterinary Research*, 35, 445 - 466.

[37] Mellor, P. S., Boorman, J., and Baylis, M. (2000), "*Culicoides* Biting Midges: Their Role as Arbovirus Vectors," *Annual Review of Entomology*, 45, 307-340.

[38] Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society Series A (General)*, 135, 370-384.

[39] OIE (2010, 04/01/2010). "OIE Listed diseases," Retrieved 11 August, 2010, from http://www.oie.int/eng/maladies/en_classification2010.htm?e1d7

[40] Onderstepoort Biological Products (2010), "Vaccines," Retrieved 18 October, 2010, from http://www.obpvaccines.co.za/prods/imu_horses.htm

[41] Onderstepoort Biological Products (2010), "African Horse Sickness Vaccine," Retrieved 18 October, 2010, from http://www.obpvaccines.co.za/prods/engpdf/54.pdf

[42] Pan, W. (2001), "Akaikes Information Criterion in Generalized Estimating Equations," *Biometrics*, 57, 120-125.

[43] Pawitan Y. (2001), "In All Likelihood: Statistical Modelling and Inference Using Likelihood," Oxford University Press Inc., New York.

[44] Portas, M., Boinas, F. S., Oliveira, J., Sousa, E. and Rawlings, P. (1999), "African Horse Sickness in Portugal: A Successful Eradication Programme," *Epidemiology and Infection* 123, 337-346.

[45] Sellers, R. F., Rees, W. H. G., Gillett, J. D., Boorman, J. P. T., and Rainey, R. C. (1983), "Seasonal Variations in Spread of Arthropod-Borne Disease Agents of Man and Animals: Implications for Control," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 302, 485-498.

[46] Simpkin, T. L. (2008), "Prophylactic strategies in the control of African horse sickness," *Department of Animal and Poultry Science*, University of KwaZulu-Natal.

[47] Venter, G. J., Graham, S. D., and Hamblin, C. (2000), "African Horse Sickness Epidemiology: Vector Competence of South African *Culicoides* Species for Virus Serotypes 3, 5 and 8," *Medical and Veterinary Entomology*, 14, 245-250.

[48] Veronesi, E., Venter, G. J., Labuschagne, K., Mellor, P. S., and Carpenter, S. (2009), "Life-History Parameters of *Culicoides* (Avaritia) Imicola Kieffer in the Laboratory at Different Rearing Temperatures," *Veterinary Parasitology*, 163, 370-373.

[49] Wedderburn,R.W.M., (1974) "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method", *Biometrika*, 61, 439-447.

[50] Wellby, M. P., Baylis, M., Rawlings, P., and Mellor, P. S. (1996), "Effect of Temperature on Survival and Rate of Virogenesis of African Horse Sickness Virus in *Culicoides* Variipennis Sonorensis (Diptera: Ceratopogonidae) and Its Significance

in Relation to the Epidemiology of the Disease,", *Bulletin of Entomological Research*, 86, 715-720.

[51] Wetzel, H., Nevill, E. M., and Erasmus, B. J. (1970), "Studies on the Transmission of African Horsesickness," *Onderstepoort Journal of Veterinary Research*, 37, 165-168.

[52] Wittman, E. J. and Baylis, M. (2000), "Climate Change: Effects on Culicoides-Transmitted Viruses and Implications for the UK," *The Veterinary Journal*, 160, 107-117.

[53] Wittmann E.J., Mellor P.S., and Baylis M. (2001), "Using climate data to map the potential distribution of *Culicoides imicola* (Diptera: Ceratopogonidae) in Europe," *Revue scientifique et technique (International Office of Epizootics)*, 20, 731-740.

Appendix A

Plots

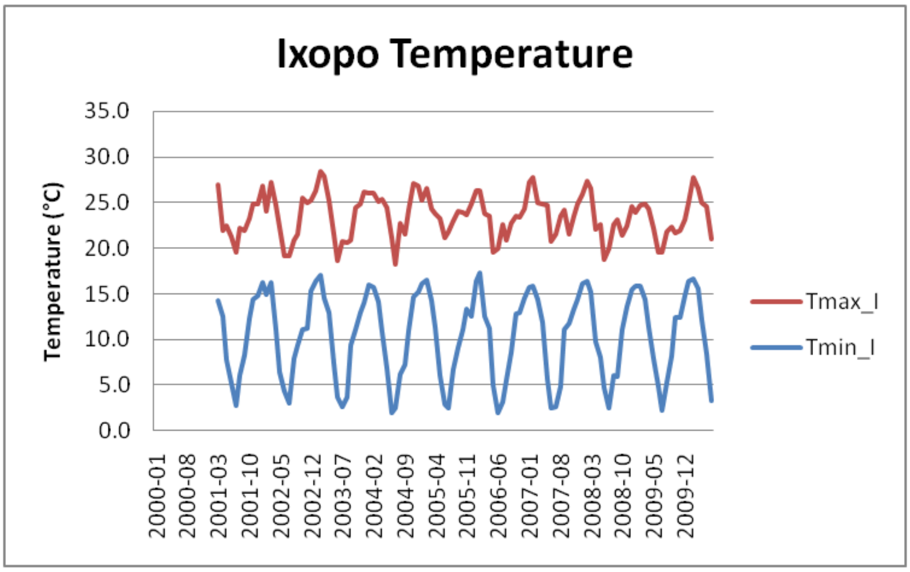

Figure A.1: Ixopo Temperature

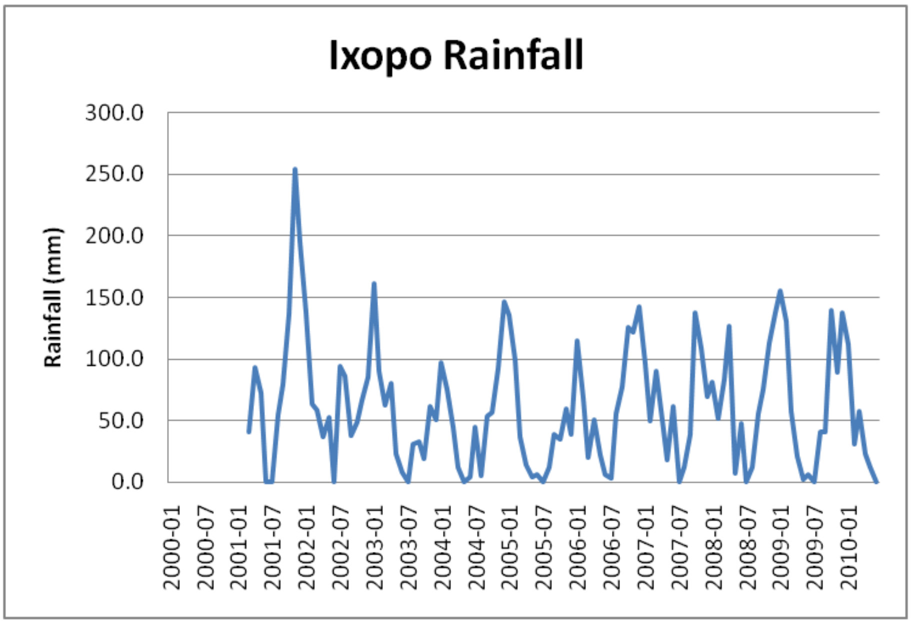

Figure A.2: Ixopo Rainfall

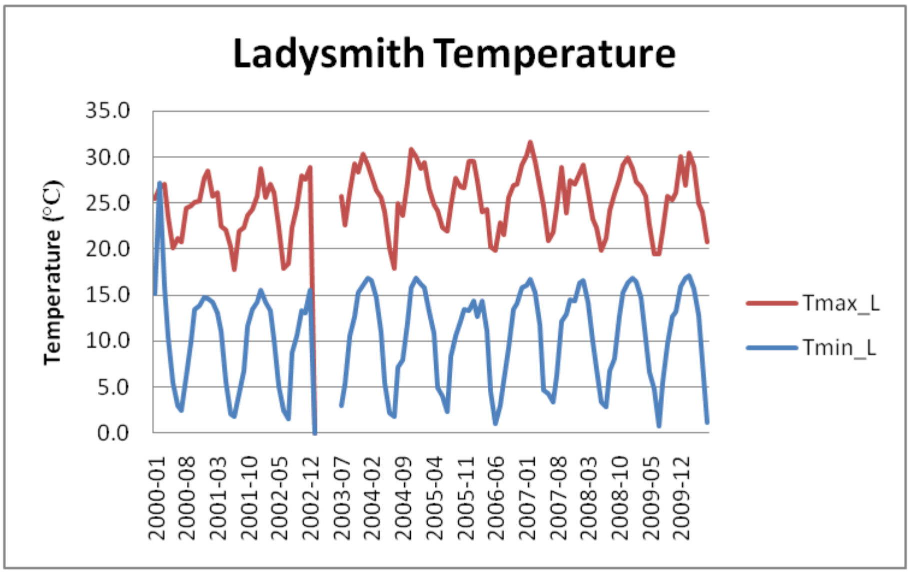

Figure A.3: Ladysmith Temperature

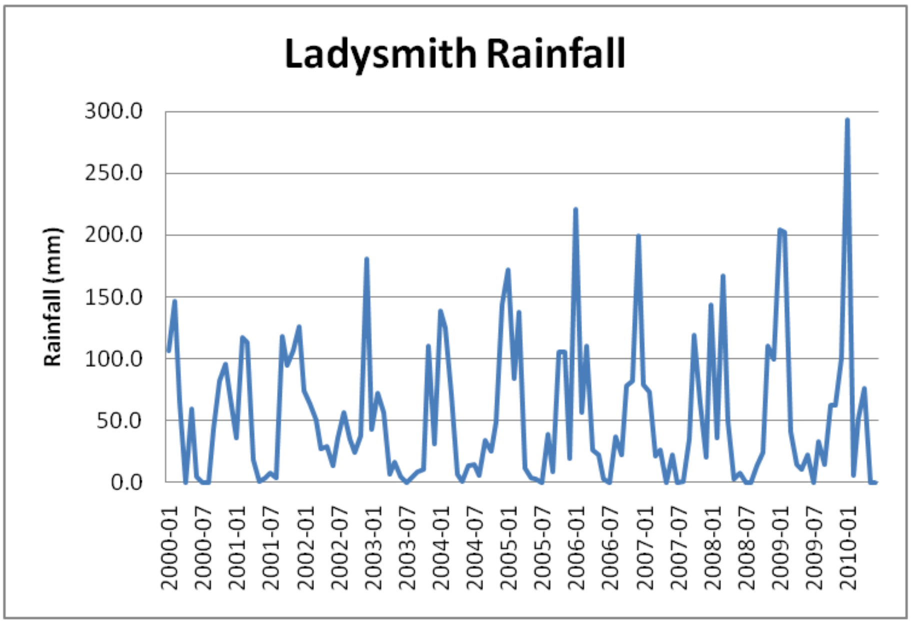

Figure A.4: Ladysmith Rainfall

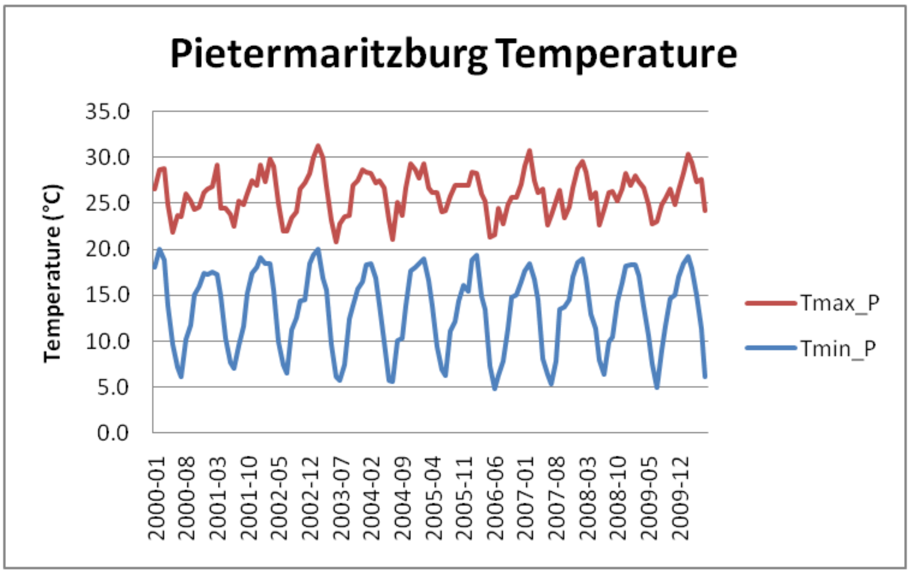

Figure A.5: Pietermaritzburg Temperature

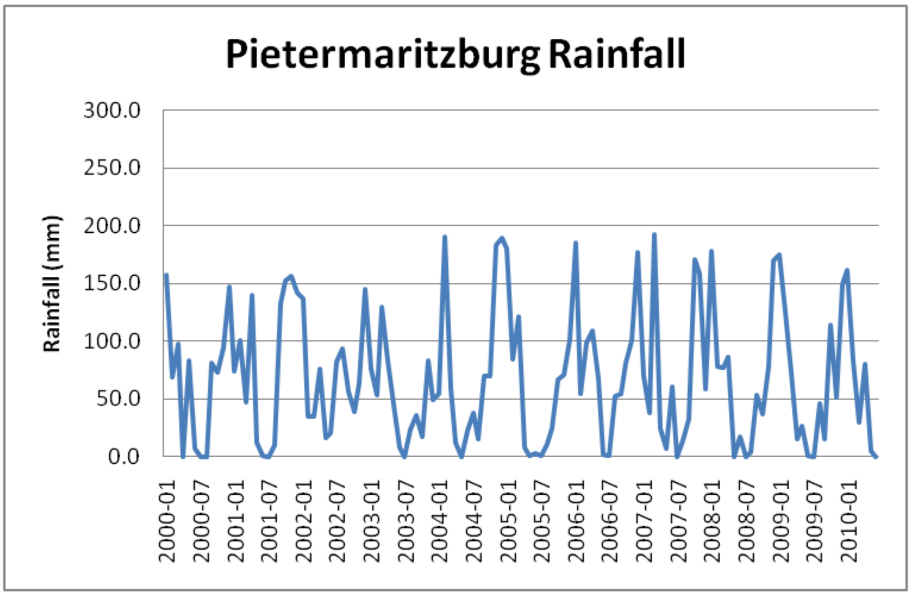

Figure A.6: Pietermaritzburg Rainfall

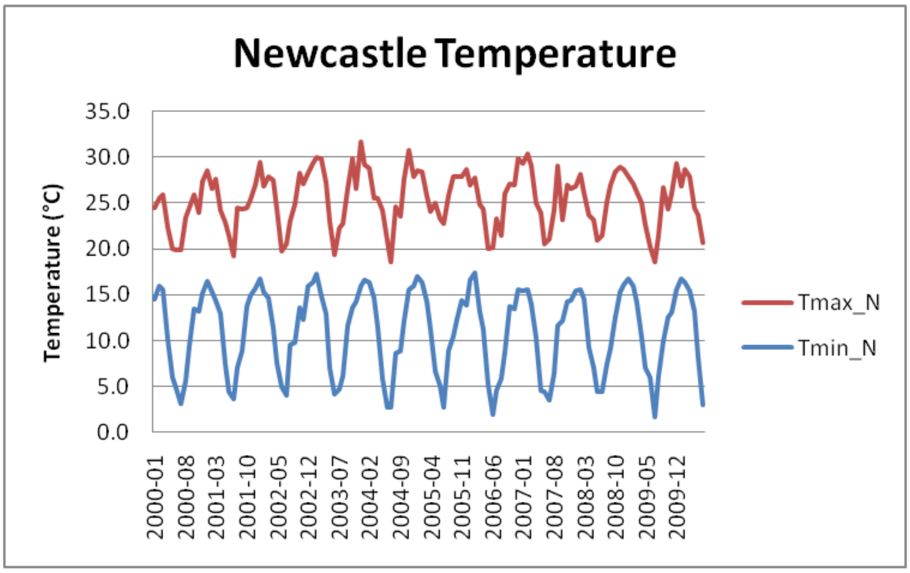

Figure A.7: Newcastle Temperature

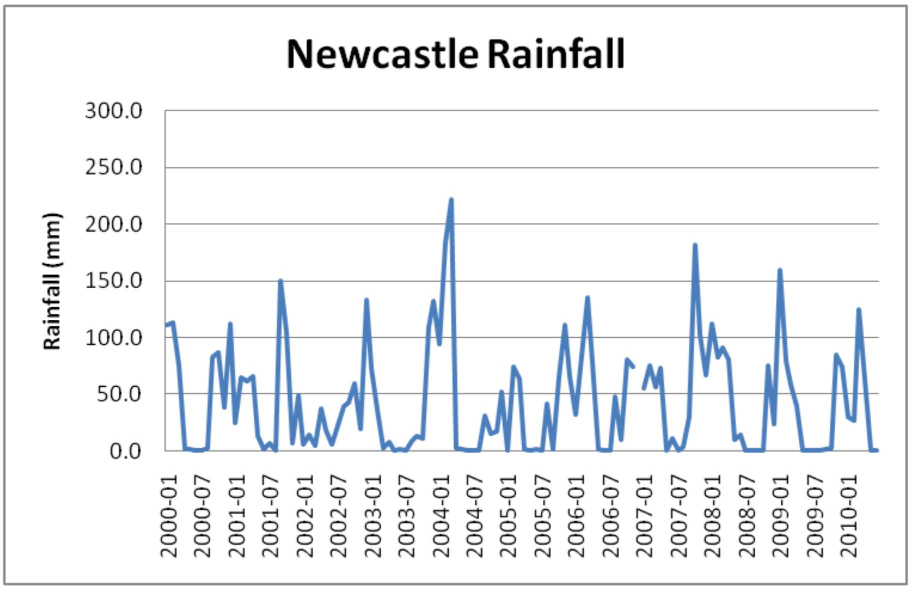

Figure A.8: Newcastle Rainfall

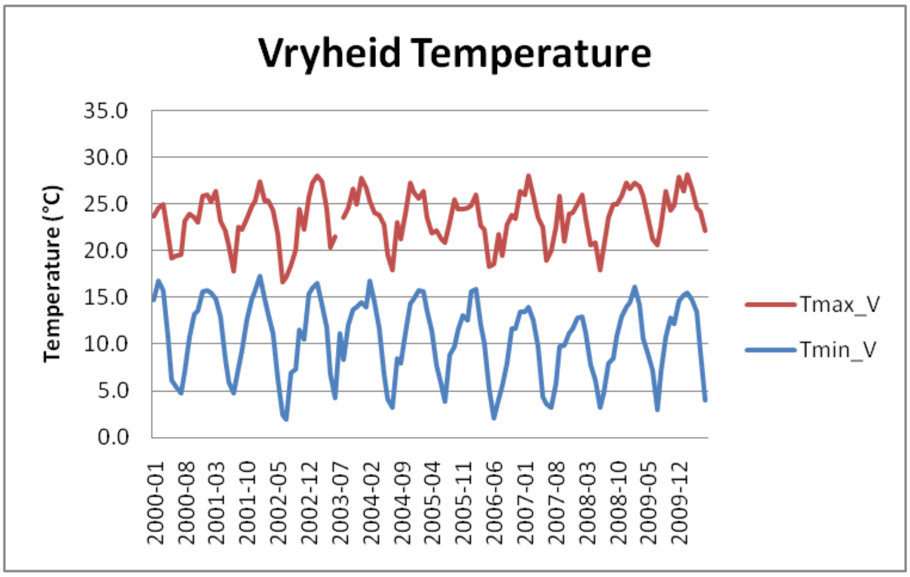

Figure A.9: Vryheid Temperature

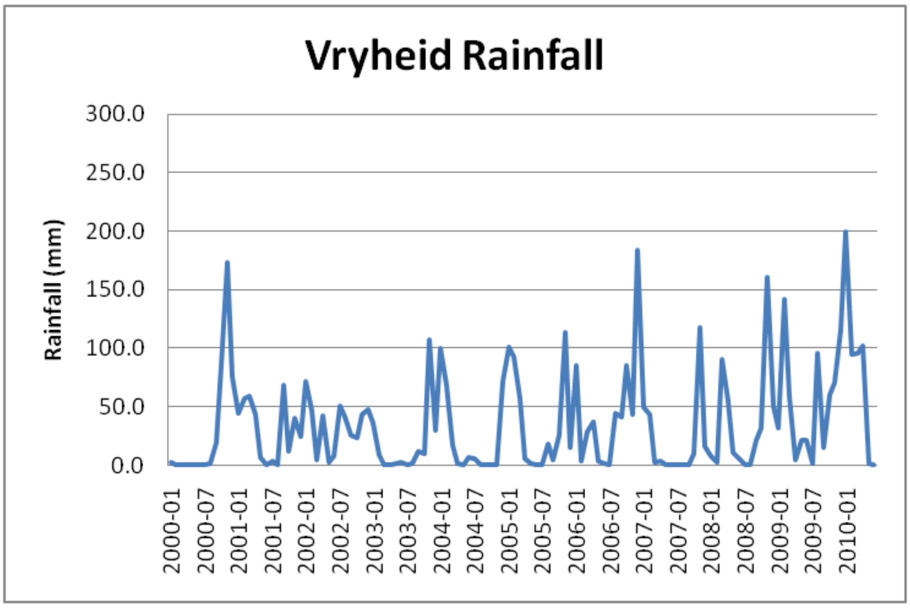

Figure A.10: Vryheid Rainfall

Appendix B

SAS Code

```
/****************** Exploratory Data Analysis *****************/
proc freq data=kzn;
tables horsestatus*province HorseStatus*classification HorseStatus*othercases
HorseStatus*vaccinated HorseStatus*presentation HorseStatus*treatment HorseStatus*stabled
HorseStatus*pesticides HorseStatus*isolation / chisq;
run;

/****************** POISSON GLM ************************/
**Calculation for sinyr term;
data kzntrig;
set kzn;
sinyr = SIN(6.28319*t);
run;

**Model without interaction effects;
proc genmod data=kzntrig;
model cases = sinyr tmax tmin rain / dist=poi link=log type3 wald;
output out=model0 p=predicted stdresdev=stdresdev ;
run;

**Model with interaction effects;
proc genmod data=kzntrig;
model cases = sinyr tmax tmin rain tmin*tmin tmax*rain / dist=poi link=log
type3 wald;
output out=model1 p=predicted stdresdev=stdresdev ;
run;
```

```
symbol1 value=dot color=red interpol=join;
symbol2 value=dot color=blue interpol=join;
title 'Predicted and Observed Cases over time';
proc gplot data=model1;
plot cases*date predicted*date / overlay legend;
run;

**model checking;
title 'Residual Plot';
symbol value=circle interpol=none;
proc gplot data=model1;
plot predicted*stdresdev;
run;
title 'QQ-Plot';
symbol value=dot color=black;
proc univariate data=model1;
qqplot stdresdev;
run;

**Model with estimating scale parameter;
proc genmod data=kzntrig;
model cases = sinyr tmax tmin rain / dscale dist=poi link=log type3 wald;
output out=model2 p=predicted stdresdev=stdresdev ;
run;
symbol1 value=dot color=red interpol=join;
symbol2 value=dot color=blue interpol=join;
title 'Predicted and Observed Cases over time';
proc gplot data=model2;
plot cases*date predicted*date / overlay legend;
run;
**model checking;
symbol value=circle interpol=none;
proc gplot data=model2;
plot predicted*stdresdev;
run;
ods html close;
symbol value=dot color=black;
proc univariate data=model2;
qqplot stdresdev;
run;
```

```
/*************** BINOMIAL GLM *******************/
proc genmod data=binom descending;
class province classification othercases vaccinated presentation treatment
stabled pesticides isolation / param=ref;
model HorseStatus = province classification othercases vaccinated presentation
treatment isolation / dist=binomial link=logit type3 wald scale=deviance aggregate=caseid;
run;

/**************** BINOMIAL GEE *********************/
proc genmod data=binom descending;
class province classification othercases vaccinated presentation treatment
stabled pesticides isolation placeID;
model HorseStatus = classification vaccinated presentation treatment / dist=bin
link = logit type3;
repeated subject = placeID / corr=exch;
output out=GEE pred=predicted;
run;

/**************** BINOMIAL GLMM ********************/
proc glimmix data=binom IC=Q noclprint=10 method=quad;
class province classification othercases vaccinated presentation treatment
stabled pesticides isolation place;
model HorseStatus (descending) = classification vaccinated presentation treatment
/ distribution=binary link=logit cl;
random place / r;
run;

proc glimmix data=binom IC=Q noclprint=10 method=quad;
class province classification othercases vaccinated presentation treatment
stabled pesticides isolation place outbreakID;
model HorseStatus (descending) = classification othercases vaccinated presentation
treatment isolation
/ distribution=binary link=logit cl;
random intercept / subject= place ;
random int / subject = outbreakID(place);
run;

/*************** POISSON GLM for JHB *********************/
data jhb;
```

```
set jhb;
sinyr = SIN(6.28319*t);
run;

Initial dataset - no severity index;
proc genmod data=jhb;
model cases = sinyr tmax*tmax tmin*tmin rain*rain tmax*rain tmin*rain / dist=poi
link=log type3 wald; output out=model2 p=predicted stdresdev=stdresdev ;
run;
symbol1 value=dot color=red interpol=join;
symbol2 value=dot color=blue interpol=join;
proc gplot data=model2;
plot cases*month predicted*month / overlay legend;
run;

One year lag for severity index (SI1);
proc genmod data=jhb;
model cases = sinyr tmax tmax*tmax tmin*tmin rain*rain tmax*rain tmin*rain
SI1 / dist=poi link=log type3 wald;
output out=model5 p=predicted stdresdev=stdresdev ;
run;
proc gplot data=model5;
plot cases*month predicted*month / overlay legend;
run;

Two years lag (SI1 and SI2);
proc genmod data=jhb;
model cases = sinyr tmax tmin*tmin rain*rain tmax*rain tmin*rain SI1 SI2 /
dist=poi link=log type3 wald;
output out=model6 p=predicted stdresdev=stdresdev ;
run;
proc gplot data=model6;
plot cases*month predicted*month / overlay legend;
run;
```