Data Mining For Exotic Pathogen Spread

David Ross

Ph.D Thesis

Department of Mathematics and Statistics
University of Strathclyde
Glasgow, U.K.

2015

Abstract

Major disease outbreaks command worldwide attention. Many recent outbreaks were caused by pathogens that were considered 'exotic' with severe implications from a health or economic standpoint. As such, there is need for an in depth examination of these threats and the means by which they might be introduced to effectively manage future risk. This thesis examines a means of identifying key emerging threats and, once identified, then modelling techniques are used to estimate the risk of introduction.

To determine the relevant exotic pathogens, data from a survey of experts were examined. In 2010 the 4th Annual Meeting of the EPIZONE network was held at which work was carried out to elicit the opinions of delegates on current and future epidemic threats to the EU. Data from this study were examined using both univariate and multivariate analytical techniques to fully explore and understand what might become an emerging threat.

This found that a particular group of zoonotic arboviruses are viewed as important potential emerging threats for Europe. Increasingly realistic and complex modelling approaches were utilised to give an increasingly accurate estimate of the risk of introduction of one of these viruses, Crimean-Congo Haemorrhagic Fever Virus (CCHFV), by means of migratory birds - a potentially key means of introduction.

Evaluating this risk must take into account not just disease related factors but also geographic factors especially the migration distance. To model this risk, spatially explicit models that correctly reflect bird migratory behaviour were used in contrast to models published previously. The approaches in this thesis show that for CCHFV there is a definite risk of introduction but it is smaller than has been estimated previously. Results also show that the bird species that should be focused on are not those intuitively identified. The migratory speed of birds is a key factor in identifying the species that represent the greatest risk of introducing CCHFV positive ticks

into Europe.

Acknowledgements

I would like to thank my supervisors Dr. Louise Kelly and Prof. George Gettinby for their guidance and high level of support. In particular the effort put in by Louise to give such consistently detailed feedback and comments on the various drafts will always be appreciated, it helped more than I can express. The support and enthusiasim provided by George was also massively important in helping me carry through the earlier stages before a plan for this project was fully developed and it will always sadden me that he never saw its completion but *Rhipicephalus appendiculatus*, the tick with brown ears, will always make me smile.

I would also like to thank Dr. David Young for agreeing to step in as a supervisor at such a late stage at what was not an easy time for many in the department.

Acknowledgement is also due to the team that collected all the expert opinion data from the EPIZONE annual meeting which was used for a large part of this thesis. Thanks is also due to Dr. William Newell formerly of AHVLA for support during the start of the thesis, Prof. Franz Conraths, Dr. Paul Gale, and Dr. Martin Hoek for providing advice and access to data.

On a personal level I would like to thank my parents for help and support throughout all my time at university. I would also like to acknowledge the support and encouragement of all my friends in particular Andrew and Catriona.

This research was funded by the Engineering and Physical Sciences Research Council, the University of Strathclyde and the Animal and Plant Health Agency.

Contents

Chapter 1

Introduction and Thesis Outline

1.1 Introduction

In the last decade there have been several news stories regarding outbreaks of disease around the world. Many of these stories are considered newsworthy due to the risk to humans (BBC, 2015, 2014; ECDC, 2014) but just as many featured outbreaks amongst animals (DEFRA and AHVLA, 2014), both wild and domestic, and the often high economic and social costs of these which had significant affects on many people's lives. The majority of these outbreaks were caused by diseases that were considered 'exotic' at the time; that is they were diseases not normally found within the country. It is obviously of benefit to try and have some manner of system in place to forewarn of when such an outbreak is starting to occur so that it can be controlled and tackled as early as possible thus minimising the damage it can cause.

Focusing on animal disease outbreaks in the United Kingdom (UK) alone, a number of exotic diseases have made incursions in recent years. These include Classical Swine Fever (CSF) in 2000 (Gibbens et al., 2000) which was a relatively small outbreak affecting only 16 farms in East Anglia. This is a potentially highly fatal and contagious disease of pigs which was generally eradicated from the UK in the 1960s (DEFRA, 2014).

In contrast, a large amount of media attention focused on the Foot-and-Mouth disease (FMD) outbreak in 2001 (Haydon et al., 2004) which was the worst outbreak

of FMD, in terms of yearly incidence and animals culled, in over 70 years (Peiso et al., 2011). This was followed by another outbreak in 2007 which, though on a much smaller scale and much more localised, still drew a large amount of media attention. The disease affects cloven hoofed animals which includes a large number of animals common to the UK such as sheep, cattle, pigs, goats and deer. While generally not fatal the disease has very serious consequences for animal welfare and can be very damaging to the economy, with over six million animals slaughtered and costs estimated at over eight billion pounds (Bourn, 2002; Haydon et al., 2004).

At an international level, 2007 also saw a major exotic disease scare with an outbreak of highly pathogenic avian influenza (specifically H5N1). A highly contagious virus that affects the respiratory, digestive and nervous systems of birds, it did not spread widely in humans but there was much focus on its potential for zoonotic transmission (Beigel et al., 2005; Claas et al., 1998; Dinh et al., 2006). Much the same happened with an outbreak of swine influenza (H1N1) in 2009 (Pawaiya et al., 2009).

The following year brought an outbreak of Bluetongue disease (Gro, 2008), a disease affecting ruminants that is spread by midges and so tends to have an element of seasonality but also has the potential to become widely distributed quickly. This affected 125 farms in the South and East of England (Gro, 2008)

Headlines were also made very recently when a man who flew into Glasgow was diagnosed with Crimean-Congo Haemorrhagic Fever Virus (CCHFV) (Atkinson et al., 2012) and died shortly after. This disease is relatively unknown in most of Europe, being more common in parts of Africa, the middle East and Russia (Sang et al., 2011; Leguenno et al., 1990; Deyde et al., 2006). Despite this, there is evidence of its spread into southern Europe, and as an exotic pathogen, it might become very important over the coming years Gale et al. (2011). In Peiso et al. (2011) a historical review of all exotic outbreaks in the UK is made and papers such as this provide a good indication of what might be a threat.

There are many potential sources of data on the introduction of diseases and viruses but they often focus on those that are endemic to a region so for exotic pathogens

there are two general options; firstly to make use of scarce data that directly relates to the risk of introduction or secondly to use more available information on potential means of introduction to create a model to estimate the risk.

1.2 Possible Data Sources that will Indicate the Likelihood and Driver of an Introduction

To select which diseases to look into it is worth starting with the list of notifiable diseases for the UK. This list was first set out in the Animal Health Act of 1981 and names a number of diseases that are considered severe enough so that potentially infected animals must be reported immediately. This provides a good starting point and from here we can look for diseases for which there has been an outbreak either in the UK or elsewhere that can be considered exotic. While looking elsewhere obviously increases the availability of data and thus would potentially allow better development and testing of methods and techniques, it will raise a number of issues that will have to be checked. These will mainly involve the sampling methodology in comparison to the UK; we would require the datasets to contain most of the same key variables and be recorded or collected in a reasonably similar manner or we would have to be able to at least take into account any major differences. Taking all of this into consideration it is possible to identify some possible pathogens to use as case studies.

There are European wide organisations that provide data on animal disease outbreaks as well as the UK's own data. While papers make good use of historical data, they also warn of gaps and scarcity of data (Peiso et al., 2011). The World Animal Health Information Database (WAHID) is an online database maintained by the World Organisation for Animal Health (OIE). It is publicly available and offers a number of ways of retrieving data on animal diseases from 2004 onwards. Its information services are split across three broad areas; by country, by disease, and by disease control measures.

The 'by country' option provides information on reported disease events, a country's

animal health situation, the numbers of animal health and veterinary personnel present, animal populations, the number of human cases of zoonoses, the country's laboratory capability, vaccination levels and vaccine production, their OIE notification history and timelines for diseases along with some time series analysis. This can be retrieved for a single country, a region or a selection of countries

Disease data are returned by disease and contain general background information on that disease, a list of all reported events whether weekly, immediate or historical, maps of disease outbreaks and distribution, country disease incidence and sanitary situation and disease timelines.

Disease control measures provides data on control measures in place either by country or by disease. In addition, there is a portal to the Handistatus database which was the forerunner of WAHID and contains data from 1996 up to 2004. This database contains monthly data on the old list A diseases (defined as those that were highly transmissible and so could rapidly and easily spread across national borders, that are of serious concern from a public health or socio-economic perspective or are of major significance in the international trade of animal or animal products (for Animal Health, 2005) and annual data on the old list A and B diseases (List B disease having a similar definition to A but without the expectation of rapid and easy spread so likely to be contained nationally (for Animal Health, 2005). This can be searched by country and disease for monthly data and by country or country and disease for annual data. In both cases information on outbreaks within a selected timeframe is returned detailing when and where an outbreak occurred (often down to a regional level) and the outbreak's affects, i.e. animals infected, culled etc.

The EPIZONE network also set up an online database, though this was not as widely accessible as the WAHID database. This database offered a large range of options for stratifying and filtering data and an automated alert system was also set up to allow experts to more easily keep track of information on outbreaks which would have been potentially very useful for the monitoring of exotic diseases. Unfortunately only the Veterinary Laboratories Agency (now the Animal and Plant Health Agency) from the UK and the Federal Research Institute for Animal Health from Germany shared

any information within the database so this severely limited its usefulness.

An alternative source of data is to make use of non-symptomatic data sources such as sales data or those offered by the development of the Internet. In recent years, there have been a number of papers on the use of search engine queries for the prediction of human disease outbreaks (Ginsberg et al., 2009; Shmueli and Burkom, 2010; Hulth et al., 2009) or grocery sales for detection of a bioterrorist attack (Goldenberg et al., 2003; Fienberg and Shmueli, 2005). In 2009 a letter was published in Nature produced by employees of Google detailing a method of detecting outbreaks of influenza through a particular form of health seeking behaviour (Ginsberg et al., 2009). Millions of search queries are submitted daily to Google by people all round the world and a method of using these queries to predict disease outbreaks was developed.

In this study, data on the 50 million most common search queries were normalised and a model was developed to investigate the likelihood of a physician visit being related to an influenza like illness being linked to these search queries. The model's explanatory variable was the log odds of the probability of a random physician visit from a particular region of the United States being related to an influenza like illness based upon data from the US Centers for Disease Control and Prevention (CDC) US Influenza Sentinel Provider Surveillance Network. The independent variables were the log odds of each search query and an automated process was used to test the fit of each independent variable separately and those that offered the best fit across all regions were used to develop a model for predicting influenza outbreaks.

The original study, having been carried out by Google, had full access to all search data. Public access to such data is more restricted. Google Trends is the public interface to Google's search data and allows data on the popularity of search terms to be retrieved. Google Insights for Search was a more detailed version of Google Trends that has since been shut down. Insights for Search in a similar way to Google Trends took a subsample of total Internet search queries across the timeframe specified and returned a scaled set of data describing the popularity of the search term across that time. The scaling was done based on the highest number of queries within the time frame so all datasets had a point where search queries were 100% and all other time

periods were scaled against this.

As a dataset, this has a number of weaknesses. First, since data are scaled as a percentage of the highest number of queries within the selected time frame, then datasets covering different time frames might not be comparable. Secondly, because only a subsample is used, then even repeatedly searching the same search term and time period can return different results. However, for human influenza and with access to the full dataset, this approach was considered such a success that Google keeps the final model running with results available to the general public, although the actual model itself has been kept confidential so as to preserve its accuracy. While Google has much greater access to facilities for processing these data, as well as more access than they grant the general public to the data, this is still of great interest.

A useful source of data, though one that can be difficult to collect, is the opinions of subject matter experts. At the 4th Annual Meeting of the EPIZONE network, an interactive question session was carried out to elicit the opinions of delegates on current and future epidemic threats to the EU. The aim of the interactive session was to identify the most threatening viruses, both now and in the future, and to identify those tools which contribute most to prediction, prevention and control of future epidemics. The output from this was a set of scores for each disease group selected calculated in part from proportions applied to the likelihood of introduction of each disease group. This was potentially a useful suorce of data for investigation of percieved threats but the nature of the data itself meant that care had to be used in selecting the correct analytical techniques.

The data for individual disease groups could be examined using many common statistical techniques but since the scoring of groups were in part dependent on how the other groups were scored then multivariate as well as univariate techniques should be used. Principal Component Analysis (PCA) was one of the earliest developed and most commonly used dimension reduction techniques. It works by taking a vector of possibly correlated random variables and uses an orthogonal transformation to convert them into a set of uncorrelated variables that are referred to as the Principal

Components (PCs). Orthogonality refers to the variables being able to vary independently, so when used in terms of statistics, it means they will be uncorrelated and if represented graphically will be perpendicular to each other. Each PC consists of a linear combination of our original variables and is formed in such a way that the first PC will explain as much of the variation in the dataset as possible. The number of PCs will be limited to as many as the original variables but it is to be hoped that the majority of information, i.e. variation, can be explained by the first few PCs. Thus, if we have a vector $\mathbf{x}$ consisting of p variables, then our principal components would be a linear function of $x_1, ..., x_p$ so if, for example, we let α represent the coefficients of each PC (so a vector of p constants) we would have for PC1:

$$\alpha_{\mathbf{1}}'\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + .. + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j$$

Plotting these PCs allows underlying patterns in the data to be investigated. To find these PCs we use either the covariance or correlation matrix of $\mathbf{x}$ and solve to find its eigenvalues and eigenvectors. For our first PC we take the largest eigenvalue of the matrix and the corresponding eigenvector consists of the coefficients that make up α_1. For our second PC we do the same with the second largest eigenvalue whose corresponding eigenvector will give us the coefficients of α_2 and so on for all p possible principal components.

One of the more well known examples of PCA is in its application to anatomical measurements and identifying the sources of variation in these for different groups and species. A common example (Jolliffe, 2002) is reproduced below in which a small dataset of 7 measurements of 28 students (15 women, 13 men) were examined to explore whether different factors are associated with the most variation between the sexes. The 7 measurements were circumferences of chest, waist, wrist and head, lengths of hand and forearm and overall height. The first PC will then be a linear combination of these measurements that has the maximum variance, with the second PC then being a linear combination, uncorrelated with the first, with the maximum possible variance and so forth. The most common way to find these linear combinations is to take the covariance (or correlation) matrix and use algebraic methods

to find the eigenvalues and associated eigenvectors. Since eigenvectors are required to be orthogonal, these linear combinations will not be correlated and taking an eigenvalue over the sum of all the eigenvalues gives the proportion of the variance associated with that PC. The results as given in Jolliffe are recreated below (Table 1.1):

Table 1.1: First three PCs of student anatomical measurements

Component number	1	2	3
		Women	
Hand	0.33	0.56	0.03
Wrist	0.26	0.62	0.11
Height	0.40	-0.44	-0.00
Forearm	0.41	-0.05	-0.55
Head	0.27	-0.19	0.80
Chest	0.45	-0.26	-0.12
Waist	0.47	0.03	-0.03
Eigenvalue	3.72	1.37	0.97
Cumulative % of total variation	53.2	72.7	86.5
		Men	
Hand	0.23	0.62	0.64
Wrist	0.29	0.53	-0.42
Height	0.43	-0.20	0.04
Forearm	0.33	-0.53	0.38
Head	0.41	-0.09	-0.51
Chest	0.44	0.08	-0.01
Waist	0.46	-0.07	0.09
Eigenvalue	4.17	1.26	0.66
Cumulative % of total variation	59.6	77.6	87.0

So, for women our largest eigenvalue is 3.72 and, as a proportion, this accounts for 53.2% of variation which will be 3.72 divided by the sum of all eigenvalues multiplied by 100%. This is then associated with an eigenvector of the form (0.33, 0.26, 0.40, 0.41, 0.27, 0.45, 0.47)' which is our first PC. As far as interpretation goes, since all

the coefficients are positive, then this is a balance of the 7 different measurements and in effect describes the overall size of the individual which would be expected, as generally a person will have directly proportional anatomical measurements, so a tall person will also have a larger chest and a wider waist and so on. The relative magnitude of the coefficients indicates the importance of that particular factor in the principal component and so in the overall variance of the data. For PC1 for women the most important factor is waist size, but it's not noticeably more important than other factors such as chest and height. Our second PC has a mix of coefficients and so is a contrast of different measurements with those having a larger magnitude being more important, in this case it is dominated by hand and wrist against height. This means that after the general size of a person (PC1), variation in anatomical measurements is mostly determined by having either large hand and wrist measurements relative to height or the opposite. The sign here is in a way arbitrary if all the coefficients were to be multiplied by minus one so as to switch the signs about the interpretation would still be the same.

Examining the rest of Table 1.1, it is seen that for both sexes the majority of variation is determined by the first three PCs and the first PC in both is a description of overall size. The second varies slightly between sexes with women being dependent on hand and wrist contrasted to height and for men hand and wrist being contrasted against forearm length. This is seen by examining the coefficients of greatest magnitude for each, and looking at PC3 the sexes differ even more.

However since the likelihood of introduction of disease group are what are referred to as compositional data; that is they consist of a vector of proportions and so have the constraint that they must sum to one, there are additional issues that must be considered. Such data also tend to display a curved relationship between variables and both of these factors mean that standard PCA can sometimes be inadequate and produce results that are of little use. This issue was first highlighted by Pearson (Pearson, 1897) in the 1800s and, generally, in relation to the correlation was communicated by considering a D-part composition $[x_1, x_2, ..., x_D]$ which is constrained to sum to one (or any other fixed value) then:

$$cov(x_1, x_1 + ... + x_D) = 0 \tag{1.1}$$

The covariance in equation (1.1) must be zero as the second element represents the sum constraint; i.e. $x_1 + ... + x_D$ will always equal one, therefore it will not vary and so the covariance will be zero.

Equation (1.1) can be expanded out using the properties of covariance:

$$cov(x_1, x_1) + cov(x_1, x_2) + ... + cov(x_1, x_D) = 0$$

$$cov(x_1, x_2) + ... + cov(x_1, x_D) = -cov(x_1, x_1) \tag{1.2}$$

$$cov(x_1, x_2) + ... + cov(x_1, x_D) = -var(x_1)$$

Then in equation (1.2) assuming that x_1 is not a constant, then the right hand side must be negative and so at least one of the left hand side covariances must be negative. Since this can be done for each x variable, then at least D elements of the covariance matrix are negative and so correlations will not be free to take all values over the (-1,1) range. A discussion of the history and state of solutions to this problem are given in Aitchison and Egozcue (2005).

To illustrate this an example is taken from a paper (Aitchison, 1982a) on this problem which uses two datasets consisting of vectors of three proportions; the relative proportions of urinary excretions of three steroid metabolites, here simply labeled 1, 2 and 3, of 37 healthy adults and the AFM compositions of 23 aphyric Skye lavas (the proportions of three oxides present in lava samples). In the paper, the datasets are used to form triangular coordinates and ternary plots are produced.

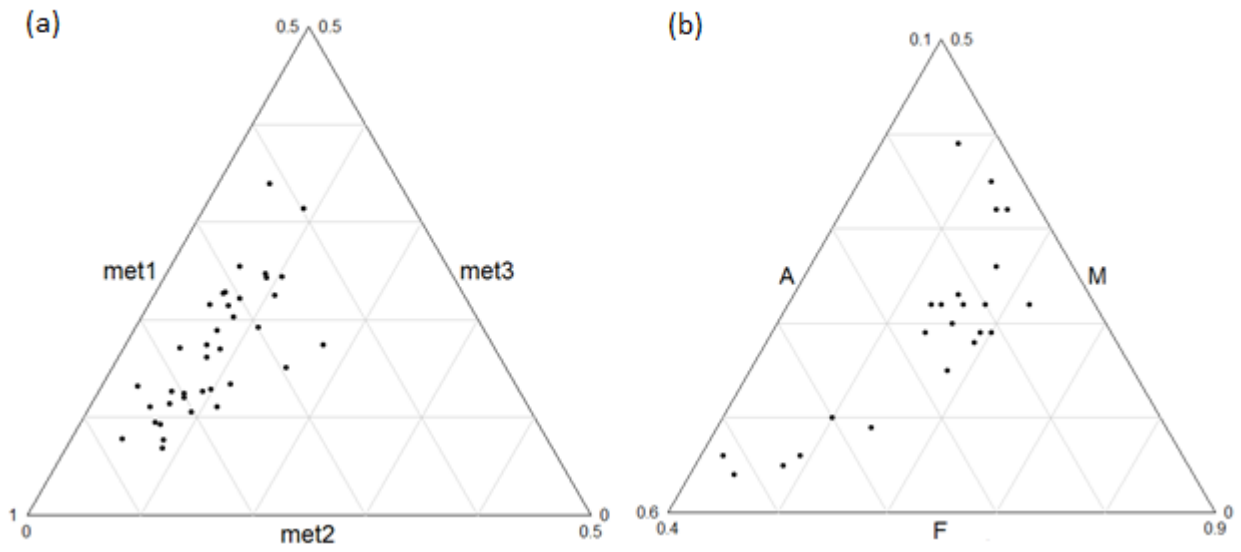

Figure 1.1: Compositional data in triangular coordinate form for (a) steroid metabolites, (b) aphyric Skye lavas, showing axes obtained by standard principal component analysis

In Figure 1.1(a) we can see a more conventional shape that is perfect for PCA where the data take an elliptical shape but in Figure 1.1(b) a non-linear relationship is clearly present meaning that fitting a first PC which takes a linear form means that it will miss out on a clear relationship that may explain much of the variation within the dataset. This is not the case for all datasets of compositional data but is held to often be the case (Aitchison, 1982a).

The other issue with this type of data is the constraint that they must sum to one, which tends to be visible as a trade off type curve in scatter plots of the variables. To illustrate this, a simple example could be considered where our data consist of just two proportions; sticking with our notation from earlier we shall call these x_1 and x_2 so taking our constraint we can write:

$$x_1 + x_2 = 1$$

Considering this equation it is clear that as x_1 increases then in order to continue to

sum to one then x_2 must decrease and producing basic points we can produce a plot showing this relationship:

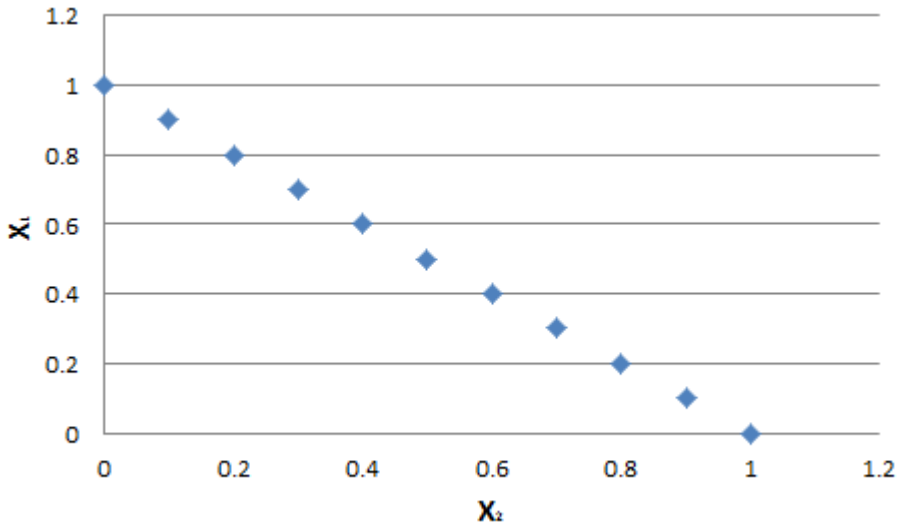

Figure 1.2: Scatter plot of x_1 against x_2 showing trade off

Figure 1.2 shows this inversely proportional relationship between x_1 and x_2 and this can cause problems with PCA as it is dependant on the correlation (or covariance) matrix and this type of relationship will put a constraint on the levels of correlation that will occur.

Many papers have been written on possible solutions to this (Aitchison, 1982a, 1999; Egozcue and Barcel, 2003) with an overall summary in Pawlowsky-Glahn and Egozcue (2006) and generally involve a transformation of the data, or some suggest the use of non-linear functions for the principal components and the technique used in this chapter was one of the earlier ones, now known as the centered log ratio transform, and involves taking the natural logarithm of each variable, a nonlinear function, and centering them before then performing principal component analysis as usual. So, for example, using the notation we used above we would have a new variable $\mathbf{y}$:

$$\mathbf{y_1} = log(x_1) - [\frac{log(x_1) + log(x_2) + .. + log(x_p)}{p}]$$

1.3 Modelling of Exotic Pathogens

When there is a lack of historical data, modelling techniques can often be used to examine the possible means of introduction and risk represented by diseases, whether these are exotic pathogens for which data are unavailable or to explore how an environmental change will affect the risk represented by an exotic or endemic disease. Many recent papers have focused on the possible influences of climate change, for example Naicker (2011); Gale et al. (2011); Gould and Higgs (2009); Baylis and Githeko (2006), which could make a significant difference in particular to vector borne pathogens where the ranges of host species may expand.

In Naicker (2011) a review of current literature was carried out and consideration was made of how changing temperatures as well as how human and other natural factors will affect the range and survivability of mosquitoes, sandflies and ticks. This could have the potential to drastically change the incidences and range of a number of vector borne diseases. This is due to warmer climates allowing many of these vector species to be active across more of the year and at higher latitudes. The paper summarises the results of a number of modelling papers focusing mostly on those concerning developing countries. There is discussion of the possibility of emerging diseases such as the Lujo virus, a haemorrhagic fever virus, but the conclusion focuses on the emergence of diseases that are currently present in Africa but have not been affecting humans. There is therefore little focus on exotic pathogens that will be introduced to a country.

More traditional epidemiological models, such as compartment models, can be used to represent the geographical movement of a pathogen and so can be used to estimate the risk of it being introduced to an area. Generally, these models describe different population groups but do not simulate the spatial aspects of spread. A discussion of this drawback, along with others, is given in White et al. (2007) followed by an approach to modelling epidemics using Cellular Automata (CA). Generally, these are two dimensional models that use an array of cells which have states that change across discrete time steps according to rules based upon the states of their neighbours. These have been used to model a number of epidemics or viruses, such as

the spread of influenza (Beauchemin et al., 2008) or Hepatitis B through cells (Xiao et al., 2006) or the spread of rabies amongst foxes (Benyoussef et al., 1999).

In contrast to the spatially explicit cellular automata approach, Gale et al. (2011) describes a Geographic Information System (GIS) model that analyses the difference between a current baseline and a future climate change scenario on the risk of CCHFV being introduced into Europe by migratory birds. The model is a GIS that uses a number of layers of 25km square cells each of which spatially covers Europe. Each layer contains different data with each cell containing the data for the geographical area it covers, these layers are then combined to calculate a level of risk for that area and the sum of all these calculations gives a level of risk for all of Europe.

The first layer describes the distribution of selected bird species across Europe; the selected species were the four most populous species considered at risk of acquiring CCHFV. This was judged by their behaviour with the four species selected being birds that migrated to Europe from sub-Saharan Africa where the virus is present and which nested and fed on the ground putting them at higher risk of becoming hosts for ticks; with ticks being the hosts for the virus. The layer was formed from two datasets: the first being data on the presence or absence of a species and the other being a total number of breeding pairs for all of Europe and these were combined to give a spatial distribution of the breeding ground of the species.

This layer was multiplied by an estimated prevalence for CCHFV positive ticks on migratory birds to give a predicted number of CCHFV positive ticks being introduced into Europe. This was combined with the second GIS layer which contained environmental data on temperature and humidity and was used to estimate the probability of the tick moulting and reaching adulthood. The final layer described the density of potential hosts for these adult ticks and a probability of successful questing combined with this layer and the previous results gave the final results of the paper. This was a geographical distribution of the number of incursions of CCHFV in host animals within 25km square cells across Europe. An alternative scenario was considered where the possible effects of climate change were taken into account, these caused an increased number of cells that held suitable environmental conditions for

successful tick moulting but also caused a decrease in the birds that migrated between Europe and sub-Saharan Africa.

The paper concludes that there will be almost no change in the number of CCHFV incursions into Europe but that climate change may result in a changing geographical distribution of these incursions. However this particular approach makes a number of statistical assumptions that are unlikely to be borne out in real life with the main assumption being that all infected ticks on a bird will stay attached throughout the entire migratory journey. To remove this assumption would require a more complex model where migratory movement is modelled.

All of this shows that with a rapidly changing world, in terms of more international movement and changing climates, exotic pathogens may be more likely to be introduced into Europe. As such, there is greater need for in depth examination of these threats. This thesis plans to examine a means of identifying key emerging threats and a number of data sources will be examined to investigate means of doing so. Once a key pathogen or pathogens is identified then modelling and estimating of the risk of introduction will be carried out. This means the thesis can be broadly broken down into two sections; firstly the identification of a key emerging threat and secondly the process of modelling its introduction.

1.4 Thesis outline

To determine which exotic pathogens might pose the greatest emerging threat, data from a survey of experts were examined. The EPIZONE network was founded in 2006 and involves a number of partner institutes from the EU, Turkey and China and has of course done work on predicting emerging threats, (Izs-ve et al., 2006) and (Kelly et al., 2013) for example. In 2010 the 4th Annual Meeting of the EPIZONE network was held, and during this, work was carried out to elicit the opinions of a large number of delegates on current and future epidemic threats to the EU.

In Chapters 2 and 3, the EPIZONE study of expert opinion (Kelly et al., 2013) will

be examined in detail to fully explore and understand the expert view on what might become an emerging threat to Europe (Figure 1.3). This will initially be done from a univariate approach in Chapter 2 before then being tackled from a multivariate perspective in Chapter 3. From this exotic pathogens that are expected to become a threat can be identified and those that seem biologically plausible for the Uk can be selected and a key emerging threat can be selected.

Once a key emerging threat has been identified, then a number of increasingly realistic and complex modelling approaches (Figure 1.3) will be utilised to estimate the risk of introduction by means of migratory birds; which was identified as a potentially key means of introduction. The first modelling approach, Chapter 4, will build upon the work done in Gale et al. (2011) by adding the effects of the migratory distance. This is an important factor as the ability of birds to introduce the virus into Europe is time dependent. The approach used here will be relatively simple and will still allow results to be easily and efficiently produced and explored but will still not be unrealistic as movement will not be explicitly modelled.

As part of the above model, analytical solutions for the risk to Europe were arrived at and were used to carry out sensitivity analysis for the various bird and tick parameters that were used in the model. For some of the key parameters, there is relatively little 'real life' data available and the parameter estimates used in Chapter 4 and in Gale et al. (2011) were based upon expert opinion. In Chapter 5, Bayesian techniques were used to combine the small amount of real data with the expert estimates in order to find a more accurate estimate that made use of all available information.

In light of the findings in Chapter 4, in Chapter 6 a more complex model will be developed where the explicit movement of bird populations will be modelled. This chapter outlines a cellular automata model for aggregated bird populations which has the disadvantage of being more time consuming to produce and run but will have less biologically unrealistic assumptions built into it although it will have birds move as populations rather than as distinct entities. This is followed by a more complex individual agent based cellular automata model in Chapter 7 and finally by the most realistic but most complex of the models: a continuous space model of bird

migration in Chapter 8. As the complexity of the model increases, a more realistic estimate of the risk of introduction of CCHFV is found but the models also become more computationally intensive.

The thesis then finishes with a discussion of the overall findings and potential further work.

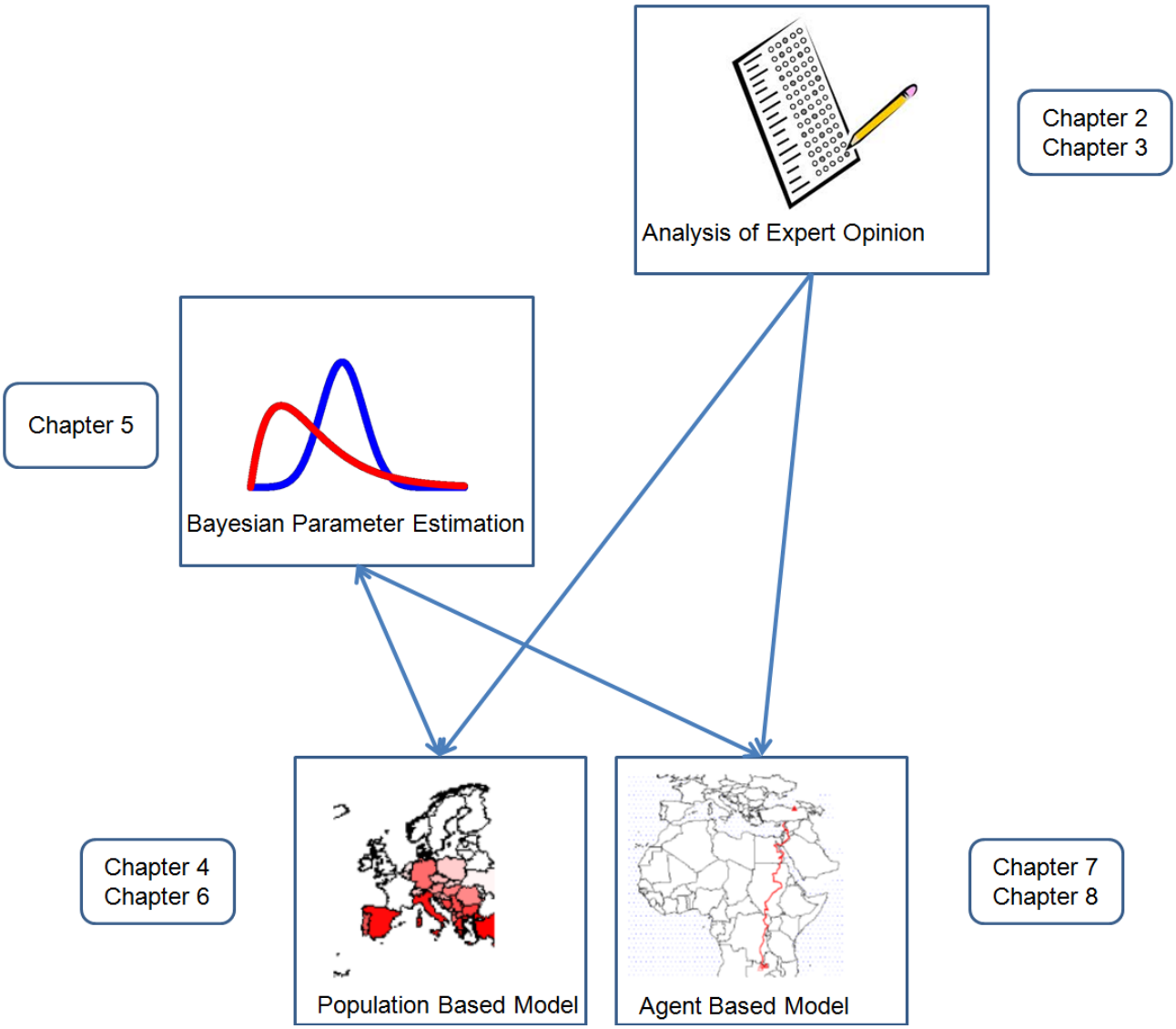

Figure 1.3: Chapter topics and interactions

Chapter 2

Univariate analysis of expert opinion data[1]

2.1 Introduction

In recent years, a number of epidemics have occurred within Europe, and many of these outbreaks have had a devastating effect on both animal health and the economy (Gro, 2008; Bourn, 2002; Haydon et al., 2004). For emerging threats, there is often a lack of 'hard data' and so expert opinion is often used as part of a risk assessment process, whether just to suggest possible threats or opinions (more qualitative data) or to provide parameter values for predictive models (quantitative data). These opinions can be used to supplement other sources of data and to aid in decision making on investments towards prevention and control.

EPIZONE is a Network of Excellence for Epizootic Disease Diagnosis and Control and was founded by the European Union (EU) in 2006. It involves 19 partner institutes from the EU, Turkey and China and was set up with the aim to work together towards the minimization of the likelihood and consequences of future epidemic threats by creating a network of scientists to improve research on the prevention, detection and control of epizootic diseases. At the 4th Annual Meeting of the EPIZONE network, an interactive question session was carried out to elicit the opinions of delegates on current and future epidemic threats to the EU. The aim of the

[1]Many of the results of this chapter were published in Kelly et al. (2013).

interactive session was to identify the most threatening viruses, both now and in the future, and to identify those tools which contribute most to prediction, prevention and control of future epidemics.

The delegates were questioned on a number of issues during two sessions which were held at different times and involved a slightly changed group of delegates. Session one covered delegates' background, the impact of disease on different areas such as economics or health, future changes of an environmental, social and economic nature and the likelihood of a disease being a threat both now and in the future (2020). The diseases considered for current and future threats were split between six disease groups: Influenza; Bluetongue (BT) and African Horse Sickness (AHS); African Swine Fever (ASF); CSF; FMD; West Nile Fever (WNF), Rift Valley fever (RVF), and CCHFV. Session two again covered the same background questions before questions covering more detail on the two most threatening diseases as selected during session one. The results of both these sessions were then examined to determine expert views on disease risk and the results were published (Kelly et al., 2013). The aim of this chapter is to perform a more detailed analysis of the data resulting from session 1 focusing on how much bias may be present amongst the expert opinions and to investigate factors that influence someone's views on what constitutes a threat.

2.2 Materials and methods

2.2.1 Questions

Interest here lies in identifying the disease groups that represent the most risk, and this is contained within the results of session one; therefore this analysis will be restricted to session one also. This interactive session was held on the final day of the 4th annual meeting of EPIZONE, held in St. Malo in November 2010. It was carried out using an interactive, hand held, electronic device which was used to select a response to each question; delegate responses were then transmitted from the device and collated in a spreadsheet and results were displayed graphically after each question. There were nine buttons on the device though most questions had fewer

possible options as answers.

After a few initial questions that were used to familiarise the delegates with the device, the delegates were then questioned on their background. The focus of the background questions and possible options are shown in Table 2.1 .

Table 2.1: Questions to describe delegates backgrounds

Question topic	Options
Background	EPIZONE, Research, Industry, Policy
Expertise	Diagnostic Development, Vaccine & anti-viral development, Surveillance & epidemiology, modelling & risk assessment
Years experience	< 10 years, $10-20$ years, >20 years
Two main diseases worked on [2]	Influenza, BT & AHS, ASF, CSF, FMD, WNF, RVF & CCHFV, Other
Region	Northern Europe, Western Europe, Eastern Europe, Southern Europe, Non-Europe

Since this meeting was an open meeting with delegates from EPIZONE partner institutes and elsewhere, the first two questions were used to record a delegates organisational area (Background) and the area they worked on (Expertise) which was broadly formed along the EPIZONE scientific themes of diagnostics, intervention strategies, risk assessment and surveillance and epidemiology. Prior to this session, the diseases for which most expertise was available were identified and the six disease groups in Table 2.1 were formed from these. BT and AHS are both non-zoonotic arboviruses sharing a common vector and so were grouped together and likewise

[2]Delegates could select up to two of the options e.g. Influenza and other or alternatively just one option, for example FMD, indicating that their line of work only involves working on one disease.

WNV, RVF & CCHFV are all zoonotic arboviruses and so were grouped into one option. For region the zones were defined using United Nations regions and a map (Figure 2.1) was shown to help clarify for the delegates the countries in each region.

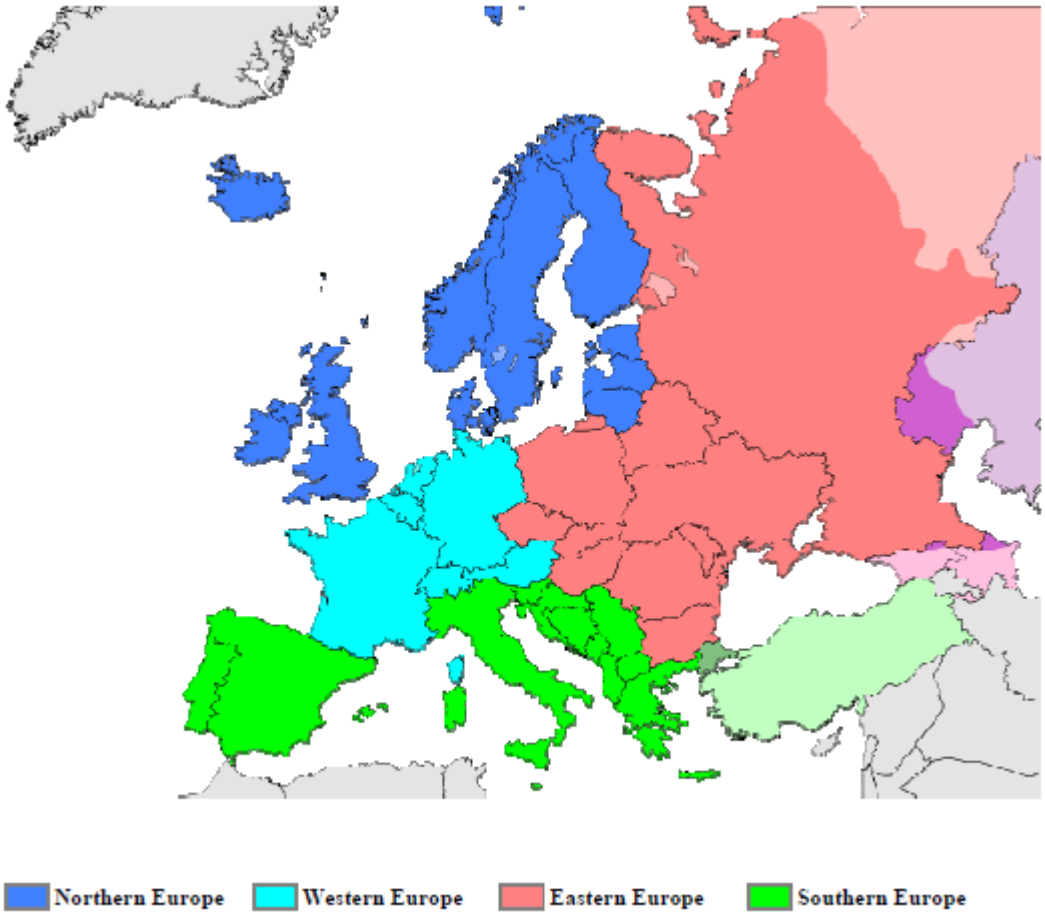

Figure 2.1: Map of European regions for delegates' origin

Following the background questions, delegates were asked a series of questions focusing on the impact of an epidemic and these are shown in Table 2.2.

Table 2.2: Questions on delegates' views on the type of impact, and their importance, and disease groups that affect each impact type

Question topic	Options
Contributes most to the threat posed by an outbreak	Impact on Human Health, Economic Impact, Impact on Animal Welfare
Contributes least to the threat posed by an outbreak	Impact on Human Health, Economic Impact, Impact on Animal Welfare
Largest impact on human health	Influenza, BT & AHS, ASF, CSF, FMD, WNF, RVF & CCHFV
Largest economical impact	Influenza, BT & AHS, ASF, CSF, FMD, WNF, RVF & CCHFV
Largest impact on animal welfare	Influenza, BT & AHS, ASF, CSF, FMD, WNF, RVF & CCHFV

The questions in this part can be broken down into two sub-sections, the first consisting of the first two questions can be used to rank the three areas of impact in terms of importance. The second subsection, consisting of the remaining three questions, then determines which disease group is most relevant for each of the 3 impact areas.

The next section focuses on current threats and asked delegates to identify which disease group is the most likely to cause a new incursion into their region, which is most likely to spread if it did get introduced and which would be most likely to persist if introduced. The question topics and the possible responses are shown in Table 2.3.

Table 2.3: Questions on delegates' views on the most likely disease groups to be introduced, spread and persist in a region at the current point in time

Question	Options
Most likely to cause incursion	Influenza, BT & AHS, ASF, CSF, FMD, WNF, RVF & CCHFV
Spread fastest if introduced	Influenza, BT & AHS, ASF, CSF, FMD, WNF, RVF & CCHFV
Most difficult to eradicate once introduced	Influenza, BT & AHS, ASF, CSF, FMD, WNF, RVF & CCHFV

Delegates were then questioned on future changes covering four key environmental areas that may change. Each of these were chosen as areas that could have an impact on the likelihoods assessed (Table 2.3) so, for example, from Table 2.4 the first question relates to how delegates believe the climate will be in ten years time, the length of time between the present and future timeframes, compared to the present which could have the potential to change the likelihood of a disease being introduced if, for example, a vector species becomes able to survive in the new climate. Two of the areas are about more human controlled activities, importation and farming, so strong views on these could be used to draw conclusions on possible future policies on these areas and the other two, concerning climate and wild animal populations, are concerned with the natural environment.

Table 2.4: Question on delegates' views of environmental differences between the current point in time and 2020

Question topics	Options
Climate	No significant change, More variable, Drier hotter summers, Milder wetter winters, Drier hotter summers & milder wetter winters
Importation of live animals and meat products	Large decrease, Small decrease, No significant change, Small increase, Large increase
Farming	No significant change, Decrease in farming, Increase in intensive farming, Increase in extensive farming, Increase in all types of farming
Wild populations of large mammals and birds	No significant change, Increase in large mammals, Decrease in large mammals, Increase in bird populations & migrations, Decrease in bird populations & migrations, Increase in both, Decrease in both

After considering the possible ways in which the environment could change over the next ten years (Table 2.4) delegates were then asked to reconsider the likelihood questions (Table 2.3) they were asked earlier, but this time for future threats.

2.2.2 Data Analysis

The results of the first session were transmitted directly from the hand held devices and stored in an Excel worksheet and after the session were cleaned according to two rules; firstly any delegates who had failed to answer any of the background questions

on region, background or expertise were removed from the dataset, secondly any delegates who had more than two missing answers for any of the other questions were also removed so as to have a dataset that had few missing values but without greatly reducing the number of subjects. The resulting spreadsheet contained the responses for 192 delegates.

Ranking Process

Using the results of the first two impact questions, a weighting of each of these impact areas was calculated. For this let α_1, α_2 and α_3 be defined as the number of delegates who selected each of these measures (human, animal and economic respectively) as contributing most to the impact of an epidemic. Likewise let β_1, β_2 and β_3 be the number of delegates who selected the measure as contributing the least to the impact of an epidemic and let N be the total number of delegates. We then have for $i = 1, .., 3$:

$$W_i = \frac{[3\alpha_i + \beta_i + 2(N - \alpha_i - \beta_i)]}{6N} \tag{2.1}$$

The weights W_i sum to one and can be regarded as representing the proportion of total delegate views on the importance of each of the areas of impact i.e. the larger the weighting the more important delegates think that impact measure is.

An impact score for each of the disease groups was then calculated using the responses from the remaining impact questions. For this the six virus groups are denoted using $j = 1, .., 6$ and using the i from above we have x_{ij} which is the number of delegates that selected virus group j as having the greatest impact on the impact area i if there was to be an outbreak of that virus. So with $i = 1, .., 3$ and $j = 1, .., 6$ gives.

Impact Score:

$$IS_j = W_1 x_{1j} + W_2 x_{2j} + W_3 x_{3j} \tag{2.2}$$

So for example x_{12} would be the number of delegates that selected BT & AHS as

having the greatest impact on human health. This produces a value representing the total view of delegates on the impact of a particular virus, being the sum of the number of delegates who selected a disease for an impact area multiplied by the importance of the area as represented by its weighting. A likelihood score was then calculated for each disease group using the three likelihood questions (Table 2.3), introduction, spread and difficulty to eradicate, represented by $i = 1, .., 3$. Letting y_{ij} represent the number of delegates who selected virus group j as being most likely to lead to likelihood measure i. For example y_{12} would be the number of delegates who selected BT and AHS as the most likely to be introduced into their region.

Likelihood Score:

$$LS_j = \sum_{i=1}^{3} \frac{y_{ij}}{3} \tag{2.3}$$

Each of the 3 likelihood scenarios were considered equally important which is why the 3 values of y for each are averaged i.e. the likelihood score has a denominator of 3.

This leaves us with two measures, one of the perceived impact a disease group would have and another for the likelihood of such a disease being an issue. A final score for each virus group was then calculated by multiplying the two of these.

Score:

$$S_j = IS_j LS_j \tag{2.4}$$

This means that the final score for each disease is determined by both the impact of an epidemic and the likelihood of such an epidemic occurring. This allows a balancing of a delegates' view as simply asking for threats would most likely result in just highly likely diseases being given but if a delegate thinks a disease is very unlikely but of high impact this means it will still get a reasonably high score. A similar score was calculated for the future using the same impact score but calculating a future likelihood score (FLS) using the likelihood questions for 2020 rather than the current likelihood questions (Table 2.3) and combining these for a future score (FS).

A similar approach was followed for future threats where, instead of y_{ij} there is z_{ij}, which represents the number of delegates who selected virus group j as being most

likely to lead to likelihood measure i in 2020. That is, those delegates who selected a particular virus group for a question in Table 2.3 for the year 2020.

Future Likelihood Score:

$$FLS_j = \sum_{i=1}^{3} \frac{z_{ij}}{3} \tag{2.5}$$

Future Score:

$$S_j = IS_j FLS_j \tag{2.6}$$

The scores and future scores for each of the disease groups across all delegates were plotted in Figure 2.2:

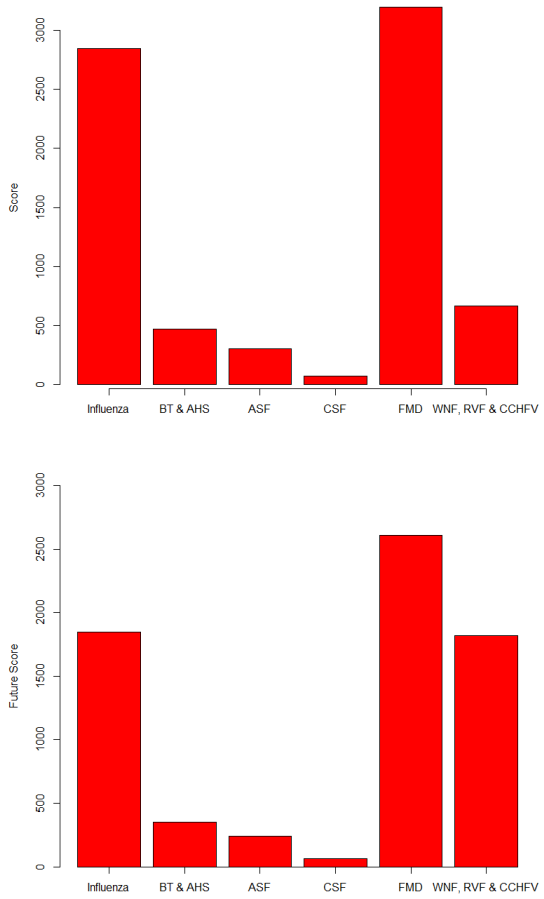

Figure 2.2: Score and Future Score for each disease group calculated from the returns from all delegates

These scores can be regarded as the level of threat a particular disease group is perceived to have by delegates, relative to others considered and using these rating criteria, and it was used to rank the different disease groups. These scores were also calculated for different subgroups formed using the Background questions so, for example, scores were calculated separately for EPIZONE, research, industry and policy groups with N being replaced by the total number of delegates within each subgroup. This was carried out to see if perceived current and future threats varied between different groups e.g. did EPIZONE delegates and industry delegates rank current threats in the same way or differently or was there a bias towards the disease groups that a delegate worked on.

Creation of new response variables

Since one of the results identified in the EPIZONE paper (Kelly et al., 2013) was of a change of opinion regarding current and future threats, new variables were formed to allow this to be investigated in more detail. The change of opinion identified was that delegates seemed to think that influenza would be less of a threat in 2020 compared to now and that WNF, RVF, and CCHFV would be more of a threat. To investigate what factors might influence delegates who changed their opinions in either of these ways, new datasets were formed using the paired current and future likelihood questions (Table 2.3). In particular, a delegate's disease choice for current and 2020 risk of incursion, spread and persistence were used to derive two sets of binary variables for each delegate indicating whether or not they had switched from Influenza and whether or not they had switched to WNF, RVF & CCHFV for each likelihood question. The final dataset then had six new indicator variables with values for each delegate, three of these marking whether a delegate changed from influenza as a current threat (for risk of incursion, spread and persistence) to a different disease group as the future threat and three for a delegate changing to the zoonotic arboviruses as a future threat e.g. if a delegate listed influenza as a current risk of incursion but not as a future then a 1 would be placed against that delegate for the appropriate new variable.

Finally, to allow investigation of how background might influence how an individual

scores a disease as a threat, an individual score was produced for each delegate for each disease using a variation of the original scoring process. As before, the first two impact questions were used to calculate an impact weighting. However, for an individual we let α_1, α_2 and α_3 be defined as an indicator (i.e. taking zero or one) of whether a delegate selected each of the measures (human, animal and economic respectively) as contributing most to the impact of an epidemic. Likewise let β_1, β_2 and β_3 be an indicator of whether a delegate selected the measure as contributing the least to the impact of an epidemic, due to the nature of the questions only one α and one β will be one and the other two will be zero. We then get:

$$W_i = \frac{[3\alpha_i + \beta_i + 2(1 - \alpha_i - \beta_i)]}{6} \tag{2.7}$$

In contrast to equation (2.1) the weights being calculated are for the proportion of an individual delegate's views on the importance of an area of impact and this is done for each delegate as opposed to the proportion of total delegate views on the importance of an area of impact. This is also why N is now replaced by 1 and α and β are indicator variables, indicating whether a individual delegate selected a particular response rather than a count of delegates who selected that response.

This produces three values for each delegate (one for each impact measure) that sum up to one and can be regarded as representing a delegates views on the importance of each of the areas of impact; i.e. the larger the weighting the more important a delegate thinks that impact measure is. However, in this case, the possible values are very restricted, if a delegate selects a different impact area for each question then the one they regard as being most important will have a weighting of $\frac{1}{2}$, the second most important will have a weighting of $\frac{1}{3}$ and the least important will have a weighting of $\frac{1}{6}$. If a delegate were to select the same impact measure as both the most and least important then all three will be equally weighted with weightings of a $\frac{1}{3}$. This could occur for a number of reasons; for example the delegate believes all three are equally important or they do not hold the necessary knowledge to be able to confidently decide between the measures.

An impact score for each of the disease groups is then calculated for each delegate

using their responses for the remaining impact questions. Again for this, the six virus groups are denoted using j and x_{1j} would be an indicator (i.e. taking the value zero or one) of whether a delegate had selected virus group j as having the greatest impact on human health if there was to be an outbreak of that virus and similarly for the other two impact measures. The impact score is then as before.

Impact Score:

$$IS_j = W_1 x_{1j} + W_2 x_{2j} + W_3 x_{3j} \tag{2.8}$$

This produces a value representing the total view of a delegate on the impact of a particular virus. Again the values are restricted, in this case having to sum to one and having possible values of $0, \frac{1}{6}, \frac{2}{6}, ..., \frac{6}{6}$.

A likelihood score is then calculated for each disease group using the three likelihood questions (so $i = 1$ to 3) and with y_{ij} representing an indicator variable of whether a delegate selected virus group j as being most likely to lead to likelihood measure i. For example $y_{12} = 1$ would mean a delegate selected BT and AHS as the most likely to be introduced into their region.

Likelihood Score:

$$LS_j = \sum_{i=1}^{3} \frac{y_{ij}}{3} \tag{2.9}$$

Clearly from the equation this is restricted to the values of 0, $\frac{1}{3}$, $\frac{2}{3}$ or 1.

The individual score for each delegate by virus group is then calculated using the impact and likelihood scores and a similar modification is done for the individual future score.

Individual Score:

$$S_j = IS_j LS_j \tag{2.10}$$

Given that the possible values for the Likelihood Score are restricted to 0, $\frac{1}{3}$, $\frac{2}{3}$ or 1 and the Impact Score to 0, $\frac{1}{6}$, $\frac{2}{6}$,..., $\frac{6}{6}$ then clearly the possible values for the Individual Score (and Individual Future Score) will be restricted to the values of $0, \frac{1}{18}$,

$\frac{2}{18}, \ldots, \frac{18}{18}$.

This is technically a discrete variable but, since it takes a fairly large number of values across a short range, it could be regarded as quasi continuous between zero and one.

This dataset of individual scores had the two policy delegates removed from it, due to the small number of delegates from this background and the fact that therefore there won't be much of a range of opinion it would be likely to produce spurious results. The final set of individual scores for both now and the future are displayed graphically in Figure 2.3.

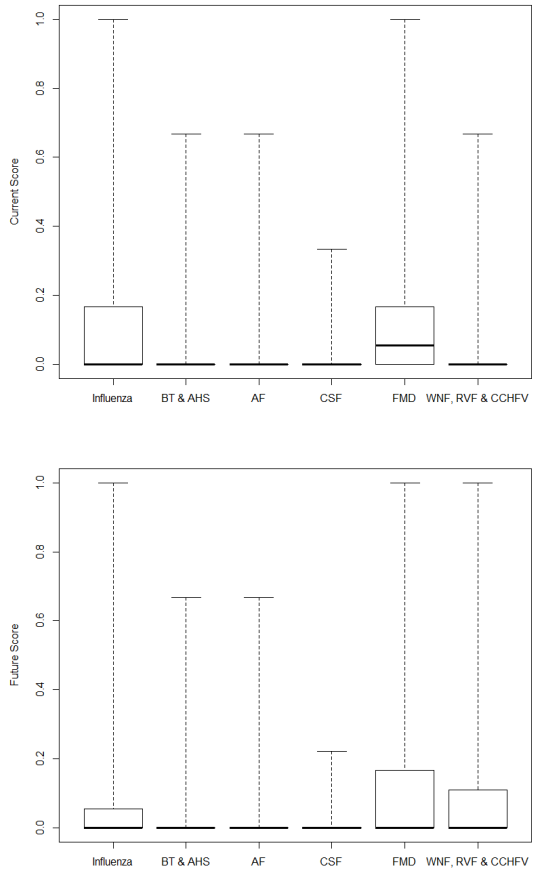

Figure 2.3: Individual Current Score and Future Score for each disease group calculated from the returns from all delegates

A dataset was also created where any score greater than zero was replaced with a one i.e.

$$S_j > 0 => S_j = 1$$
$$S_j = 0 => S_j = 0$$

This binary response to individual scores was to allow a logistic regression approach to be taken and would have the interpretation of simply whether a delegate thought a disease group was a threat or not with nothing to distinguish between level of threat.

2.2.3 Analysis

2.2.4 Total Scores & Rankings

In the EPIZONE report (Kelly et al., 2013), tables of the disease groups ranked using the Score and Future Score broken down into subgroups using the background, expertise, disease worked upon and region questions described previously were presented. These tables of results were statistically analysed using Spearman's rank correlation. Spearman's rank correlation is a non-parametric test of dependence between two variables and was used to test the similarity between the rankings of our six diseases when viewed by sub groups. This produced a correlation coefficient that represents the strength and direction of the relationship between the two sets of rankings. The closer to one the stronger the sense of agreement, the closer to negative one the more opposite the views (i.e. if ranks were completely inverted) and zero would indicate there is no link at all.

One factor ANOVA (analysis of variance) was used to examine the total scores for each disease group broken down into subgroups to investigate if there was a significant level of variation between scores for different subgroups. The most basic interpretation of this is to test whether the means of different groups are equal. This was carried out on each future and current score for each disease by subgroup with our null hypothesis being that there is no difference in how each group has scored this disease with the alternative hypothesis being that scoring does change across

groups. Tukey's range test (with overall significance level of 5%) was used to test each group against the others and further check for differences.

2.2.5 Individual Scores

A number of analyses were carried out on the individual scores with the focus being on whether or not the background of delegates could be used to explain their scoring behaviour. However, before this, an examination of what combination of disease groups delegates scored was carried out. Since an individual delegate only has three questions in both the impact and likelihood sections in which they specify the disease groups they perceive as a threat, this then places a limitation on the number of disease groups they can score. If a different disease group is specified for each of the three impact questions and the same three disease groups are specified for the likelihood then those three disease groups will receive a score. If due to either specifying the same disease group for multiple questions or putting a different combination of disease groups for the impact questions as opposed to the likelihood questions then possibly two, one or zero disease groups may be scored for a delegate. Therefore the maximum number of possible final score combinations for each delegate are:

$$\binom{6}{3} + \binom{6}{2} + \binom{6}{1} = 41 \tag{2.11}$$

With the first term in equation 2.11 being the possible disease combinations for a delegate scoring three different diseases, the second term being where only two diseases are scored and the final term being the options for a delegate scoring only one disease group. To investigate the combinations of disease groups for current and future individual scores were counted and examined.

To begin examining the individual scores, a general linear modelling approach was taken that depended on the quasi-continuous nature of the individual score. This was done as analysing it this way allowed a greater level of variation between delegates despite the fact that it stretched the assumptions required, since the data are not truly continuous but only approximately so. This approach attempts to identify an underlying mathematical model which explains the level of variation affecting delegates responses. The model is represented as

$$\begin{aligned} Y_j &= \theta B + e \\ &= \theta_0 + \theta_1 B_1 + \theta_2 B_2 + \theta_3 B_3 + \theta_4 B_4 + e \end{aligned} \tag{2.12}$$

In equation (2.12) Y is the individual current or future score for a delegate for a particular disease group and B is a set of categorical variables for a delegate's response to the background questions (Table 2.1) relating to their background, area of expertise, whether they have worked on the disease group in question and their region.

The aim is then to identify the factors that contribute most towards this controlled variation; i.e. so which of the background questions seems to contribute the most variation. The general linear model approach uses least squares regression to estimate the coefficients (θ) in the model and the size and significance of these indicate the importance of the particular background factor. A general linear model was fitted for all current and future scores for each disease group with the explanatory variables being all of the background questions with the exception of years worked which was removed due to the number of missing values that were present.

A forward stepwise selection procedure was used as follows. Univariate analyses were run for each factor and factors were selected for inclusion in the initial model using a significance level of 0.25, before then being used in the stepwise procedure with a significance level of 0.05, which included main effects and two factor interactions. Once this preliminary model was set up the terms removed during the univariate stage were added back in separately to check if they had now become significant. Once a final model was found, Tukey's range test was used to test each subgroup of delegates against the other and check for differences and factor and interaction plots were produced to explore the behaviour of the model.

To analyse whether or not a delegate regarded a disease group as a threat, a logistic regression approach was used to analyse the binary response data for individual

scores. This is an adaptation of the general linear model, equation (2.12), where the dependent variable, Y, is binary so can be only zero or one. It makes use of maximum likelihood estimation to estimate the values of θ. However, due to separation issues the standard logistic regression approach could not be used. The separation issue occurred when all delegates from a particular subgroup scored or didn't score a disease, for example all delegates whose area of expertise was modelling & risk assessment did not score ASF as a current threat. The problem here is that since delegates who view the disease as a threat (or not as the case may be) all have the same value for a factor then they can be perfectly separated i.e. a perfect predictor can be found (Albert and Anderson, 1984). In this case, our maximum likelihood estimate will not exist and will produce an infinite estimate for the parameter value or in many softwares a large estimate and very large standard error.

There are various approaches that can be used to handle the issue of data separation (Heinze and Schemper, 2002) and two of them were used here. Firstly, there is what is referred to as an *ad hoc* adjustment which is relatively straightforward and involves changing a single delegate's response so as to remove separation. To continue the example, a single delegate from a modelling & risk assessment area of expertise would have their response changed so as to score ASF as a threat.

Secondly, there is the use of penalized maximum likelihood where a modified version of a delegate's score is used to create and test the model (Heinze and Schemper, 2002). The score function is based on the distribution of the log likelihood and for penalized maximum likelihood this is replaced with a penalized log likelihood:

$$logL(B)^* = logL(\beta) + \frac{1}{2}log|I(\beta)| \tag{2.13}$$

Here I(β) is the Fisher information matrix, the negative of the second derivative of the log likelihood. This new log likelihood is biased away from zero and so is biased away from infinite parameters being produced.

There are then two methods of calculating confidence intervals for the point estimates produced, penalized profile likelihood confidence intervals (an alternative iterative

approach using the log likelihood) or the more traditional Wald intervals. The latter are more efficient to calculate and less computationally intensive; however, are still affected by separation though will not be effectively infinite since they are based on a finite form of point estimate as opposed to the intervals that would be found using a standard logistic regression approach. Therefore, the former are used here since they are not affected by separation; however, the Wald intervals for the final models were also calculated and are included in appendix A.

There are other potential approaches for handling separation, two of the main ones being omission of the problem group or variable from the model and exact logistic regression. The first of these is to be avoided where possible as it removes much information from our model and dataset and does not allow for adjusting effects on the other factors by the one removed. The second is unfortunately computationally intensive and is only appropriate for very small datasets and for fitting simpler models.

The same selection procedure was used as for our general linear models and was carried out separately for the *ad hoc* and penalized maximum likelihood methods.

2.2.6 Examining delegates' changes of opinion

For the current and future likelihood questions, interest centered on examining the relationship between what disease delegates viewed as a threat now and those they saw as threats in 2020. Separate tests were run for each disease, in effect creating a series of two by two tables, e.g. the number of delegates who selected influenza as the greatest threat in terms of an incursion now and the number of delegates who selected another disease group (so the sum of the other 5 groups) as the greatest threat in terms of an incursion now against influenza in 2020 and all other disease groups in 2020

To compare the scores, standard contingency table tests were not applicable since these were dichotomous variables, that is paired outcomes, rather than two independent variables; i.e. we have present threats selected with delegates then being

made aware of how other delegates voted before selecting future threats. As such, the McNemar test was used to determine if the proportions of delegates choosing a particular disease changed depending on the time frame.

For this test, our null hypothesis would be that as many delegates are as likely to switch from the disease of interest as are likely to switch to it when selecting their current and future threats with the alternative hypothesis being that delegates are more likely to change their opinions in a particular direction.

H_0: Delegates equally likely to change their opinion in either direction
H_a: Delegates change their opinion in a particular direction

The equation for the test statistic itself is:

$$\chi^2 = \frac{(|n_1 - n_2| - 1)^2}{n_1 + n_2}, df = 1 \tag{2.14}$$

Where n_1 and n_2 are the counts of delegates who have changed their mind; i.e. from the disease group of interest as a current threat to another disease group for a future threat or *vice versa*. Equation (2.14) is the standard test statistic used in the Minitab statistical software and includes a correction for continuity which has little effect unless $n_1 + n_2 < 25$ (Edwards, 1948)

As mentioned previously some interest revolved round the changing of delegates opinions of influenza and the zoonotic arboviruses and a new dataset was formed to represent this. To analyse this, a logistic regression approach was taken with the explanatory variables being the new indicator variables described above and a univariate analysis of this against each of the responses for the Future Changes questions. Again, the issue of separation occurred and the same approach was taken of producing models using both an *ad hoc* adjustment and penalized maximum likelihoods. Variables here were selected based upon both a p-value of less than 0.05 being present and examination of the confidence interval i.e. if it does not span one, then it was accepted as significant.

2.3 Results

2.3.1 Total Scores & Rankings

The results for Spearman's rank correlation coefficient were broken down by subgroups and current and future rankings. The first division of the delegates was by background and they were split in to four groups. For example the ranking of the diseases for EPIZONE delegates and Research delegates were used to calculate the coefficient and also a p-value which tests the null hypothesis that there is no link in how these two background groups rank the disease groups in terms of threat. For interpretation in this case we can see from Table 2.5 that the agreement between Research and EPIZONE is perfect as can be seen by visually examining the data since both groups gave the same rank to every disease. Agreement between EPIZONE and Industry and Research and Industry is very strong; however, there is strong evidence of low agreement between Policy and the other groups, altogether the low number of Policy delegates means that this result must be treated with some scepticism.

Table 2.5: Spearman's rank correlation coefficients for current disease scoring calculated when delegates are split by Background

Present	EPIZONE	Research	Industry	Policy
EPIZONE	-	1.000	0.886(p=0.019)	-0.091(p=0.864)
Research	-	-	0.886(p=0.019)	-0.091(p=0.864)
Industry	-	-	-	-0.152(p=0.774)
Policy	-	-	-	-

Examining the future rankings, Table 2.6, we see much the same pattern though the values have dropped slightly meaning some disease groups have been ranked slightly differently amongst groups but generally EPIZONE, Research and Industry delegates agree and the negative correlations show the policy delegates disagree with individuals from all other backgrounds.

Table 2.6: Spearman's rank correlation coefficients for future disease scoring calculated when delegates are split by Background

2020	EPIZONE	Research	Industry	Policy
EPIZONE	-	0.943(p=0.005)	0.886(p=0.019)	-0.135(p=0.798)
Research	-	-	0.714(p=0.111)	-0.101(p=0.848)
Industry	-	-	-	-0.372(p=0.468)
Policy	-	-	-	-

Splitting the delegates by area of expertise results in a strong level of agreement across all sub groups both now and in the future (see Tables 2.7 and 2.8). This is evidenced by the tables containing positive high values for all combinations.

Table 2.7: Spearman's rank correlation coefficients for current disease scoring calculated when delegates are split by Expertise

Present	Diagnostic development	Vaccine & anti-viral development	Surveillance & epidemiology	modelling & risk assessment
Diagnostic development	-	0.943 (p=0.005)	0.943 (p=0.005)	0.928 (p=0.008)
Vaccine & anti-viral development	-	-	1.000	0.870 (p=0.024)
Surveillance & epidemiology	-	-	-	0.870 (p=0.024)
modelling & risk assessment	-	-	-	-

Table 2.8: Spearman's rank correlation coefficients for future disease scoring calculated when delegates are split by Expertise

2020	Diagnostic development	Vaccine & anti-viral development	Surveillance & epidemiology	modelling & risk assessment
Diagnostic development	-	0.886 (p=0.019)	0.943 (p=0.005)	0.899 (p=0.015)
Vaccine & anti-viral development	-	-	0.943 (p=0.005)	0.928 (p=0.008)
Surveillance & epidemiology	-	-	-	0.986 (p=0.000)
modelling & risk assessment	-	-	-	-

When delegates were split by the diseases they work on there was a greater level of variation (Tables 2.9 and 2.10). As far as the present is concerned, it seems that those delegates who work on ASF and CSF seem less in agreement with delegates who have worked on all other diseases. This can be seen in the lower positive correlations for these 2 disease groups groups against each of the others. However, it should be noted that while there is no strong agreement there is not complete disagreement; that is, complete disagreement would be represented by a value of zero indicating no correlation and, in this case, while the values are low they are still greater than zero. There appears to be reasonably good agreement amongst the other disease groups but, when we move to the future, the levels of agreement drop amongst most groups maybe indicating uncertainty about the future or certain groups of delegates who work on emerging threats moving emphasis to a different group which ties into the EPIZONE paper (Kelly et al., 2013) where it was noted that the delegates who worked on arboviruses (WNF, RVF, CCHFV and BT & AHS) both changed their rankings of WNF, RVF & CCHFV.

Table 2.9: Spearman's rank correlation coefficients for current disease scoring calculated when delegates are split by Disease Worked On

Present	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV	Other
Influenza	-	0.943 (p=0.005)	0.371 (p=0.468)	0.486 (p=0.329)	0.886 (p=0.019)	0.943 (p=0.005)	1.000
BT & AHS	-	-	0.314 (p=0.544)	0.429 (p=0.397)	0.943 (p=0.005)	0.886 (p=0.019)	0.943 (p=0.005)
ASF	-	-	-	0.657 (p=0.156)	0.371 (p=0.468)	0.429 (p=0.397)	0.371 (p=0.468)
CSF	-	-	-	-	0.486 (p=0.329)	0.543 (p=0.266)	0.486 (p=0.329)
FMD	-	-	-	-	-	0.943 (p=0.005)	0.886 (p=0.019)
WNF, RVF & CCHFV	-	-	-	-	-	-	0.943 (p=0.005)
Other	-	-	-	-	-	-	-

Table 2.10: Spearman's rank correlation coefficients for future disease scoring calculated when delegates are split by Disease Worked On

2020	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV	Other
Influenza	-	0.600 (p=0.208)	0.429 (p=0.397)	0.657 (p=0.156)	0.771 (p=0.072)	0.600 (p=0.208)	0.829 (p=0.042)
BT & AHS	-	-	0.371 (p=0.468)	0.657 (p=0.156)	0.829 (p=0.042)	1.000	0.771 (p=0.072)
ASF	-	-	-	0.771 (p=0.072)	0.371 (p=0.468)	0.371 (p=0.468)	0.257 (p=0.623)
CSF	-	-	-	-	0.600 (p=0.208)	0.657 (p=0.156)	0.429 (p=0.397)
FMD	-	-	-	-	-	0.829 (p=0.042)	0.943 (p=0.005)
WNF, RVF & CCHFV	-	-	-	-	-	-	0.771 (p=0.072)
Other	-	-	-	-	-	-	-

Interestingly, the opposite can be seen when viewing delegates by region; in this case agreement seems to improve for the future. For the present, delegates from Western and Northern Europe showed strong agreement and the recent BT epidemic (Carrasco et al., 2010) may have contributed to this and views on the emerging threat posed by WNF, RVF & CCHFV may be what causes the increased agreement about the future.

Table 2.11: Spearman's rank correlation coefficients for current disease scoring calculated when delegates are split by Region

Present	North	West	East	South	Non
North	-	0.943 (p=0.005)	0.600 (p=0.208)	0.600 (p=0.208)	0.657 (p=0.156)
West	-	-	0.543 (p=0.266)	0.714 (p=0.111)	0.714 (p=0.111)
East	-	-	-	0.657 (p=0.156)	0.600 (p=0.208)
South	-	-	-	-	0.714 (p=0.111)
Non	-	-	-	-	-

Table 2.12: Spearman's rank correlation coefficients for future disease scoring calculated when delegates are split by Region

2020	North	West	East	South	Non
North	-	0.943 (p=0.005)	0.771 (p=0.072)	0.771 (p=0.072)	0.771 (p=0.072)
West	-	-	0.657 (p=0.156)	0.886 (p=0.019)	0.829 (p=0.042)
East	-	-	-	0.714 (p=0.111)	0.657 (p=0.156)
South	-	-	-	-	0.714 (p=0.111)
Non	-	-	-	-	-

It is also worth noting that there were never any noticeably high negative correlations in any of the comparisons suggesting that while there might be some disagreement about the rankings there was never any group that held completely contrary views to the others.

For the one factor ANOVA a number of potential factors were flagged as being significant in explaining the level of variation between subgroups. When total scores were broken down by disease worked on a significant result was found for all disease groups and times indicating that there is a statistically significant difference whereby delegates who work on a disease are more likely to rate it a threat. This indicates that there may be a bias amongst delegates towards a disease that they themselves work on.

When carrying out the same analysis for delegates broken down by expertise, there is no significant result for any disease indicating that there is agreement across these groups for all diseases. Subgroups formed by background only resulted in one significant result and that was for CSF in the present where there was a statistically significant difference between delegates from industry and from research with delegates from industry as a whole deeming it more of a threat.

By region, there is a difference amongst delegates when it comes to scoring two of the diseases. The first is Foot-and-mouth disease which is seen as more of a threat outside of Europe though the level of difference becomes less noticeable when looking at the future as opposed to the present. The other disease group is the zoonotic arboviruses where the pattern is a little less clear. Regarding current threats what can be concluded is that delegates from the South regard it as significantly more of a threat than delegates from the West and for the future the same delegates regard it as significantly more of a threat than delegates from the North.

2.3.2 Individual Scores

Combinations

Once the individual scores were calculated for each delegate a total was taken of all the final combinations of diseases scored for both the present and the future. So each delegate had the potential to score between zero, if a delegate selected completely different disease groups for impact and likelihood, and three diseases and the five most common combinations of disease groups scored are shown in Table 2.13 for current delegate perception and Table 2.14 for delegate future perception.:

Table 2.13: Current Disease Score Combinations

Present scores Disease combination	Number of delegates
Foot-and-mouth disease	38
Influenza	32
Influenza & Foot-and-mouth disease	29
None	19
BT & AHS & Foot-and-mouth disease	10

Table 2.14: Future Disease Score Combinations

Future scores Disease combination	Number of delegates
Foot-and-mouth disease	33
None	30
Foot-and-mouth disease & WNF, RVF & CCHFV	23
Influenza	22
Influenza & Foot-and-mouth disease	16

So, for example, 38 of the delegates scored only Foot-and-mouth disease as a threat to be focused on currently (Table 2.13) and 10 scored both BT & AHS and Foot-and-mouth disease. A couple of points can be noted from these combinations, firstly the two highest scoring disease groups from the original report (Influenza and Foot-and-mouth disease) dominate both tables as might be expected. Secondly, a smaller proportion of delegates are covered by the top five combinations in the future and the number who scored no disease group has increased both of which could reflect a greater level of disagreement or uncertainty on an individual level i.e. the disease a delegate thinks will have a great impact are not those they think are likely. This is also supported by the fact that for current scores there was a total of 20 combinations, with only the 5 most popular shown in Table 2.13. In contrast, for future scores this went up to 24 combinations of disease groups indicating a more diverse view amongst delegates of what will constitute a threat in the future.

General Linear Model

Analysing the individual scores of delegates using the general linear model approach resulted in one variable that was always significant and this was whether or not a delegate had worked on the disease in question with there being a negative effect if they hadn't. Using the selection procedure described for fully half of the final models (i.e. current and future scores for all six disease groups) this was the only factor to

remain in the final model, however the rest were more complex.

Table 2.15: Major Coefficients (rounded to 1 decimal place) for General Linear Model for Influenza Scores

Factor	Coefficient	p-value
Worked on Influenza	0.2	p<0.001
Delegate from North Europe	0.2	p<0.001
Delegate from the East Europe	0.2	
Delegate from the West Europe	0.1	
Non-European Delegate	0.1	
Delegate worked on Influenza & from North Europe	0.3	p<0.001
Delegate worked on Influenza & from East Europe	0.3	
Delegate worked on Influenza & non-European delegate	0.1	
Delegate does not work on Influenza & from North Europe	0.1	
Delegate does not work on Influenza & from West Europe	0.1	
Delegate worked on Influenza & from West Europe	0.1	
Delegate worked on Influenza & from South Europe	0.1	

As regards influenza as a threat in the present, region was also included in the final model (Table 2.15) with it being present as a main effect and an interaction between it and whether a delegate had worked on influenza, see Figure 2.4 which is a main effects and interaction plot showing the mean level for each factor and interaction. That is, the coefficient associated with a particular level of a factor, for example delegates from the North of Europe regarded influenza as more of a threat with those from the North who work on the disease tending to score it even higher. This could possibly be due to the high concentration of cases in Iceland during the 2009 Influenza pandemic (Sigmundsdottir et al., 2010). Those delegates from South EU and from outside of Europe who don't work on the disease had the opposite opinion and scored it lower.

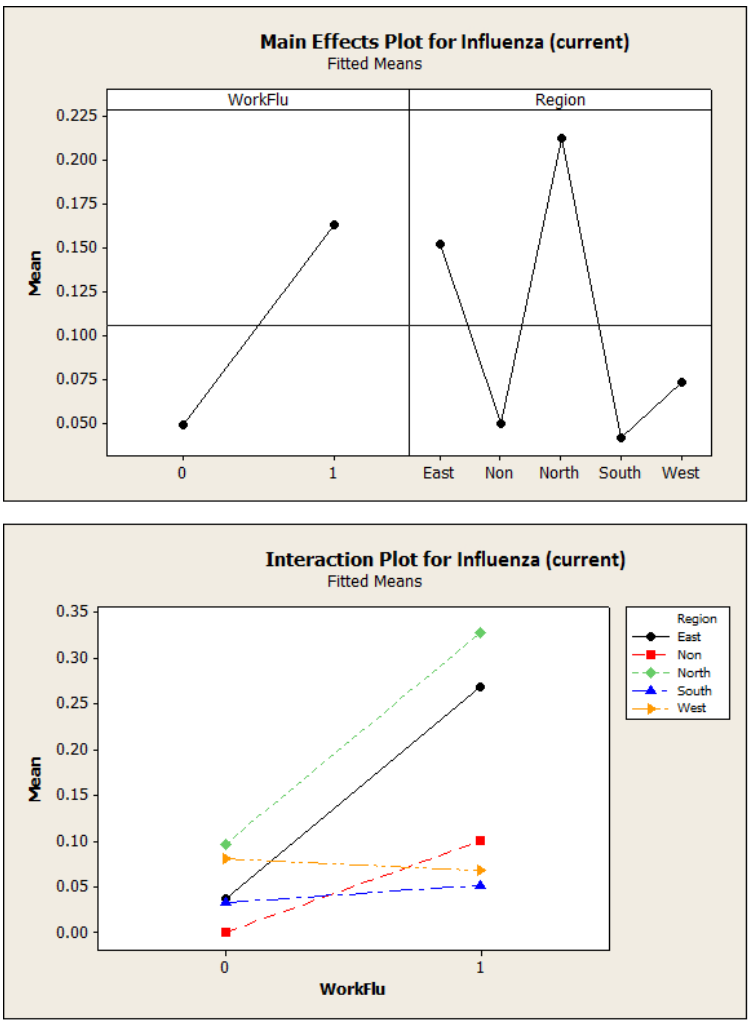

Figure 2.4: Factor & Interaction Plots: Influenza (Current)

For the scoring of BT & AHS as a future threat (Table 2.16) everything was a little less clear. Northern delegates tended to score the disease group higher (Main effects plot in Figure 2.5) which is rather surprising seeing as both these diseases are historically more of a threat in the other European regions, although outbreaks of BT in 2007/8 and a smaller outbreak in 2009 did go as far north as Norway so, possibly, since it is less common but has recently appeared, the reaction to it in the north was more extreme so it is viewed as more of a future threat.

Table 2.16: Major Coefficients for General Linear Model for BT & AHS Future Scores

Factor	Coefficient	p-value
Delegate worked on BT & AHS	0.1	p<0.001
Delegate from North Europe	0.1	0.057
Delegate worked in Research & from North Europe	0.1	0.012
Delegate worked in Research & from North Europe	0.1	

In addition to this, the interaction between background and region (Interaction plot in Figure 2.5) was also significant and again here it is Northern delegates that stand out; however, the fact that Northern EPIZONE delegates do not view these diseases as being as big a future threat is of interest. A search of the Experts database on the EPIZONE website (http://www.epizone-eu.net/epizone/experts.aspx) only returns a single expert on BT associated with an EPIZONE partner in Northern Europe and none for any region for AHS and so possibly, since there is less expertise in this subject area, there is a tendency to score it slightly lower than the other backgrounds. Of course, this is merely conjecture and, since many more Northern EPIZONE delegates indicated that they have worked on the disease group, it is likely to be flawed and to investigate it fully, more information would be needed about all the delegates who attended the conference.

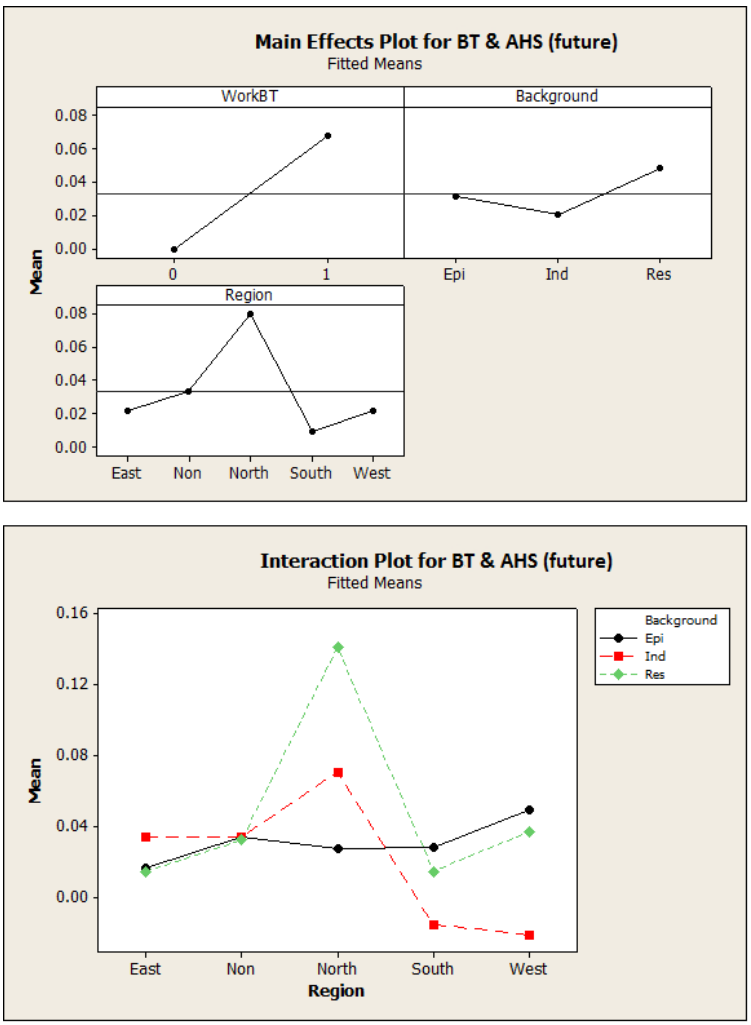

Figure 2.5: Factor & Interaction Plots: BT & AHS (Future)

For ASF in the present, with the main coefficients in Table 2.17, those delegates from East Europe were flagged as more likely to score it a threat than those from the North. This could be to do with the proximity of the Trans-Caucasus region and Russia where the disease is endemic in areas (Sánchez-Vizcaíno et al., 2012).

Table 2.17: Major Coefficients for General Linear Model for ASF Scores

Factor	Coefficient	p-value
Delegate worked on ASF	0.1	p<0.001
Delegate from East Europe	0.1	0.032

Table 2.18: Major Coefficients for General Linear Model for CSF Scores

Factor	Coefficient	p-value
Delegate worked on CSF	0.1	p<0.001
Delegate worked in Industry	0.1	p<0.001
Delegate worked on CSF & in Industry	0.2	p<0.001

As a present threat, CSF had a final model (Table 2.18) consisting of whether a delegate worked on the disease and a delegate's background and it was deemed more important by industry delegates and especially by industry delegates who work on the disease, the much higher value for industry in Figures 2.6.

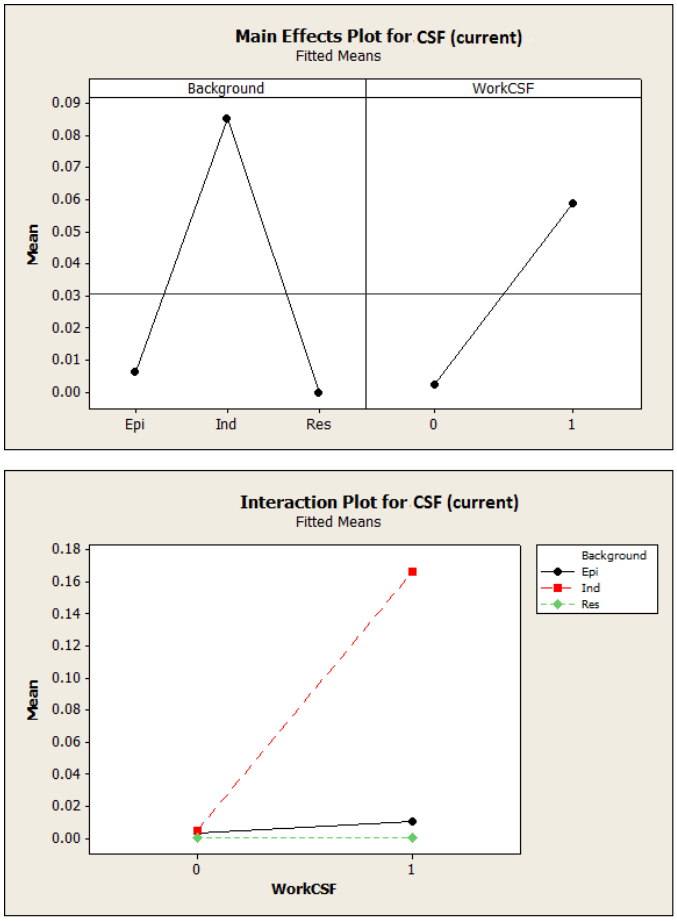

Figure 2.6: Factor & Interaction Plots: CSF (Current)

This pattern was found earlier when looking at total scores so was worth investi-

gating further. Firstly the result must be viewed with some caution as firstly only five delegates scored CSF as a threat in the present and only two of those delegates were from an industry background. These two delegates did, however, have other factors in common, both were from the west of Europe and both worked in the area of diagnostic development. A possible reason for this might be that the history of outbreaks of CSF in the Netherlands and Belgium (Boender et al., 2008) may have influenced their views of it as a current threat; however there were a number of other industry delegates from the West who did not view it as such a threat, so this is possibly simply due to the small sample size of delegates who scored this disease.

Foot-and-mouth disease (Tables 2.19 and 2.20) had what at first appeared to be more complex models (containing multiple terms), but much of the interpretation was quite straightforward. As both a present and future threat, delegates from outside of Europe, and especially those who work on the disease, regard it as more of a threat. Since some of the non-European delegates were from China, and this disease is endemic in parts of Asia, then this is perhaps not surprising. The final model for foot-and-mouth disease as a current threat, Table 2.19, also included background as a factor with delegates from an industry background scoring it less of a threat. This could perhaps be explained by the fact that almost all industry delegates working on this disease work in vaccine & anti-viral development and so possibly view it as something more easily countered and contained.

Table 2.19: Major Coefficients for General Linear Model for FMD Scores

Factor	Coefficient	p-value
Worked on FMD	0.3	p<0.001
Did not work on FMD	0.1	
Delegate from North Europe	0.1	p<0.001
Delegate from the South Europe	0.1	
Delegate from the West Europe	0.1	
Non-European Delegate	0.5	
Delegate worked on FMD & non-European delegate	0.9	p<0.001
Delegate worked on FMD & from West Europe	0.1	
Delegate worked on FMD & from North Europe	0.1	
Delegate worked on FMD & from South Europe	0.1	
Delegate does not work on FMD & non-European delegate	0.1	
Delegate does not work on FMD & from South Europe	0.1	
Delegate does not work on FMD & from North Europe	0.1	
Delegate does not work on FMD & from West Europe	0.1	
Delegate works in Research	0.2	0.029
Delegate works in EPIZONE	0.2	
Delegate works in Industry	0.1	

Table 2.20: Major Coefficients for General Linear Model for FMD Future Scores

Factor	Coefficient	p-value
Worked on FMD	0.2	p<0.001
Did not work on FMD	0.1	
Delegate from North Europe	0.1	p<0.001
Delegate from the South Europe	0.1	
Delegate from the West Europe	0.1	
Non-European Delegate	0.3	
Delegate worked on FMD & non-European delegate	0.6	p<0.001
Delegate worked on FMD & from West Europe	0.1	
Delegate worked on FMD & from North Europe	0.2	
Delegate worked on FMD & from South Europe	0.2	
Delegate does not work on FMD & non-European delegate	0.1	
Delegate does not work on FMD & from West Europe	0.1	
Delegate does not work on FMD & from North Europe	0.1	
Delegate does not work on FMD & from South Europe	0.1	

Logistic Regression

Analysing the same variables using logistic regression with the binary dependant variable being whether or not a delegate scored a disease resulted in more straightforward models. Again, as in previous models, a delegate who worked on a disease was significantly different and in almost all cases far more likely to score the disease. There was, however, one exception which was the odds of scoring WNF, RVF & CCHFV as a future threat being unaffected by whether or not a delegate had

worked on it.

As per the separation issue discussed previously, two methods were used when conducting these logistic regressions and both sets of results are included, though in all cases, with the exception of CSF, both methods resulted in the same model with factors having similar behaviours if slightly different odds ratios.

The results here are in the form of odds ratios and so are interpreted as how much more likely a delegate is to score a disease if they are in that group in comparison to the control or referent group, which in Table 2.21 would be delegates who do not work on Influenza and Southern delegates. As such, those odds ratios that are most different from 1 will be more significant as they will indicate a proportional difference in the effect of that group against the referent group: if the odds ratio was 1 it would indicate both the group and referent group have the same association. So, the first odds ratio in the table is for delegates who worked on flu (using the *ad hoc* approach) and suggests that those delegates are 2.66 times more likely to score Influenza as a current threat in comparison to those delegates who do not work on Influenza.

For Influenza (Table 2.21) Northern delegates were much more likely than delegates from any other region to award a score to the disease. As a future threat, this is even more pronounced.

Table 2.21: Logistic Regression: Individual Score (Influenza)

Influenza Factor	Current		Future (2020)	
	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)
Worked on Flu	2.66 (1.35, 5.22)	2.69 (1.39,5.32)	3.93 (1.89, 8.17)	4.09 (2.00, 8.61)
Did not work on Flu (Reference)				
Region (East)	1.29 (0.30, 5.54)	1.31 (0.30, 5.25)	3.57 (0.68, 18.61)	2.26 (0.41, 11.85)
Region (Non)	1.67 (0.43, 6.54)	1.22 (0.29, 4.67)	2.27 (0.41, 12.50)	2.28 (0.42, 11.67)
Region (North)	5.07 (1.93, 13.35)	4.86 (1.94, 13.11)	6.20 (1.79, 21.51)	5.71 (1.86, 21.22)
Region (West)	2.40 (0.96, 6.01)	2.32 (0.97, 5.93)	3.50 (1.05, 11.68)	3.23 (1.09, 11.51)
Region (South) (Reference)				

For BT and AHS (Table 2.22) and ASF (Table 2.23) there was no other significant factor than whether or not a delegate worked on the disease. The odds ratios (Tables 2.22 and 2.23) are very high and the fact that these are two of the lower ranked diseases may indicate something of interest for example that delegate bias is the only thing making someone more likely to score these diseases.

Table 2.22: Logistic Regression: Individual Score (BT & AHS)

BT & AHS Factor	Current		Future (2020)	
	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)
Worked on BT & AHS	6.23 (2.61, 14.85)	5.57 (2.41, 13.62)	5.86 (2.24, 15.30)	5.13 (2.05, 13.88)
Did not work on BT & AHS (Reference)				

Table 2.23: Logistic Regression: Individual Score (ASF)

ASF Factor	Current		Future (2020)	
	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)
Worked on ASF	7.62 (2.77, 20.92)	7.50 (2.75, 20.30)	5.83 (2.11, 16.13)	6.25 (2.24, 17.02)
Did not work on ASF (Reference)				

CSF has a delegates' background and whether or not they have worked on the disease as determining factors for its current threat status (Table 2.24) whilst having similar behaviour to ASF and BT & AHS for the future, where a delegate having worked on the disease is the only significant factor. The NA here represents the fact that the stepwise selection procedure was done separately for each approach, so, for example, the *ad hoc* approach for CSF as a current threat found no significant factors and it is worth noting that the confidence intervals for the penalized maximum likelihood approach span 1 for all factors in the current model indicating that there is no reason to reject H_0 of no significant difference between the factor levels.

Table 2.24: Logistic Regression: Individual Score (CSF)

CSF Factor	Current		Future (2020)	
	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)
Worked on CSF	NA	5.19 (0.79, 31.08)	9.52 (2.37, 38.28)	11.80 (2.93, 54.18)
Did not work on CSF (Reference)				
Background (Industry)	NA	3.41 (0.52, 20.58)	NA	NA
Background (Research)	NA	0.22 (0.00, 2.36)	NA	NA
Background (EPIZONE) (Reference)				

Foot-and-mouth disease only had one significant result other than whether or not a delegate had worked on it and this was that those from an industry background were less likely to score this disease, with industry being the referent group in Table 2.25 and the other backgrounds having higher odds ratios in comparison. This concurs with the general linear model and again may be due to many of these delegates' working backgrounds, where those from industry who worked on it work on vaccine & anti-viral development and so possibly view it as something that can be dealt with more easily.

Table 2.25: Logistic Regression: Individual Score (Foot-and-mouth Disease)

FMD Factor	Current		Future (2020)	
	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)
Worked on FMD	2.42 (0.93, 6.31)	2.31 (0.94, 6.15)	2.77 (1.05, 7.30)	2.63 (1.06, 7.09)
Did not work on FMD (Reference)				
Background (EPIZONE)	2.62 (1.01, 6.80)	2.50 (1.02, 6.61)	3.00 (1.11, 8.10)	2.82 (1.12, 7.85)
Background (Research)	3.90 (1.43, 10.60)	3.69 (1.43, 10.23)	4.59 (1.62, 12.95)	4.28 (1.63, 12.44)
Background (Industry) (Reference)				

Finally for WNF, RVF & CCHFV, asides from what was noted earlier regarding working on the disease group, which may indicate a more open less biased view of this disease as a future threat, the other factor of note was that delegates from the South were more likely to score this as a future threat (Table 2.26). Since there have been some outbreaks of West Nile in Southern Europe (Sambri et al., 2013) and CCHFV has appeared in areas nearby (Foley-Fisher et al., 2012; Midilli et al., 2009), then as a future threat, it would make sense that this disease group would be perceived as more threatening by Southern delegates.

Table 2.26: Logistic Regression: Individual Score (WNF, RVF, CCHFV)

WNF, RVF & CCHFV Factor	Current		Future (2020)	
	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)	Odds Ratio (*ad hoc* approach)	Odds Ratio (PML)
Worked on WNF, RVF & CCHFV	5.01 (2.01, 12.46)	4.95 (2.00, 12.20)	NA	NA
Did not work on WNF, RVF & CCHFV (Reference)				
Region(East)	NA	NA	1.43 (0.26, 7.91)	1.60 (0.27, 7.15)
Region(Non)	NA	NA	2.86 (0.70, 11.63)	2.89 (0.71, 10.99)
Region(South)	NA	NA	7.56 (2.69, 21.23)	7.12 (2.69, 20.55)
Region(West)	NA	NA	2.56 (0.99, 6.61)	2.45 (1.01, 6.55)
Region (North) (Reference)				

As mentioned earlier, there were two possible approaches to calculating the confidence intervals for the penalized maximum likelihood estimates and the results of this are included in Appendix A; however, since only one case of a separated variable remained in any of the final models the majority of these intervals are very similar, the exception was for CSF, for which delegates of a research background did not score the disease, but even in this case, we can see from Table 2.27 that these do not vary hugely so in this case either method seems to be suitable.

Table 2.27: Logistic Regression: Individual Score Confidence Intervals (CSF)

CSF Factor	Current		Future (2020)	
	CIs (Wald)	CIs (Profile Penalized)	CIs (Wald)	CIs (Penalized Profile)
Worked on CSF	(0.95, 28.51)	(0.79, 31.08)	(2.83, 49.22)	(2.93, 54.18)
Did not work on CSF (Reference)				
Background (Industry)	(0.58, 19.96)	(0.52, 20.58)	NA	NA
Background (Research)	(0.01, 4.04)	(0.00, 2.36)	NA	NA
Background (EPIZONE) (Reference)				

2.3.3 Changes of Opinion

McNemars test which was used to examine the results of those delegates who have changed their minds, flagged up several significant results and the p-values are presented in Table 2.28 with the null hypothesis being that delegates are equally likely to change their mind to or from a disease group and the alternative being that there is a significant change of opinion in a particular direction.

Table 2.28: p-values from McNemars test of delegates changing to/from a particular disease group for the likelihood to be introduced, spread and persist

Disease	Measure	p-value
Influenza	Incursion	**0.003**
	Spread	**0.005**
	Persistence	0.307
BT & AHS	Incursion	0.059
	Spread	0.481
	Persistence	**0.019**
ASF	Incursion	0.362
	Spread	1.000
	Persistence	0.057
CSF	Incursion	0.774
	Spread	1.000
	Persistence	1.000
FMD	Incursion	**p<0.001**
	Spread	1.000
	Persistence	**0.023**
WNF, RVF & CCHFV	Incursion	**p<0.001**
	Spread	**p<0.001**
	Persistence	**p<0.001**

The p-values below the significance level of 0.05 in Table 2.28 indicate a significant shift amongst delegate opinions regarding the threat posed by Influenza, Foot-and-mouth disease and WNF, RVF & CCHFV. This means a large number of delegates

have changed their views on these disease groups between the current point in time and the year 2020. So as to further examine the shift in opinions for these diseases, the relevant counts were produced and shown in Table 2.29. These are the number of delegates who have changed from and to this disease; the two counts which are greyed out were those which were not significant in Table 2.28, that is had a p-value above 0.05. This agrees with some of the earlier analysis (Kelly et al., 2013) which indicated a shift from influenza and Foot-and-mouth disease and a very strong move towards WNF, RVF & CCHFV indicating that many delegates see this as a serious emerging threat. For example 37 delegates changed from viewing Influenza as the most likely disease to be introduced at the moment to a different disease group for 2020 (Table 2.29) and 64 delegates changed from a different disease group to viewing WNF, RVF & CCHFV as the most likely to be introduced in 2020.

Table 2.29: Number of delegates whose opinion on the most likely disease group to be introduced, spread and persist changed to or from a particular disease group between now and 2020

Disease	Measure	Changed From	Changed To
Influenza	Incursion	37	15
	Spread	28	10
	Persist	15	9
FMD	Incursion	19	2
	Spread	19	18
	Persist	18	6
WNF	Incursion	2	64
	Spread	3	27
	Persist	5	43

The logistic regressions on the indicators of whether or not a delegate changed from influenza resulted in three potentially significant results. The first of these involved the risk of influenza being introduced into a delegate's region and a delegate's opinions on how farming would change between now and 2020. The explanatory variable would be if a delegate who viewed influenza as the greatest threat for a current incursion thought a different disease was more of a threat in 2020 compared to now

and the odds ratios are shown in Table 2.30.

Table 2.30: Logistic Regression: Changed from Influenza (Incursion: Farming)

Incursion	Changed from influenza	
	Odds Ratio	Odds Ratio
Farming	(Wald)	(PML)
Decrease	0.89 (0.35, 2.26)	0.88 (0.35, 2.25)
Increase(intensive)	0.26 (0.07, 0.94)	0.28 (0.08, 0.92)
Increase(extensive)	1.73 (0.46, 6.51)	1.74 (0.46, 6.24)
Increase(both)	0.67 (0.16, 2.86)	0.72 (0.16, 2.71)
No significant change (Reference)		

Similarly the risk of influenza spreading once introduced into a delegates region and a delegates opinions on how farming would change between now and 2020 was also found to be significant and the odds ratios are shown in Table 2.31.

Table 2.31: Logistic Regression: Changed from Influenza (Spread: Farming)

Spread	Changed from influenza	
	Odds Ratio	Odds Ratio
Farming	(Wald)	(PML)
Decrease	0.72 (0.27, 1.97)	0.72 (0.27, 1.96)
Increase(intensive)	0.15 (0.03, 0.74)	0.18 (0.03, 0.69)
Increase(extensive)	0.60 (0.11, 3.27)	0.69 (0.12, 2.99)
Increase(both)	0.78 (0.18, 3.39)	0.84 (0.18, 3.21)
No significant change (Reference)		

The factor that is significant here, having a confidence interval that does not span 1, is an increase in intensive farming (with particular emphasis put on poultry and pigs) which in relation to livestock is when the level of livestock is very high in relation to land area, the most famous example of this might be the concept of battery farming with chickens (Kelly et al., 2013; Fraser, 2008). Since two of the most

wide scale outbreaks of Influenza in recent years were avian and swine varieties and concentrated populations of these animals would potentially offer a breeding ground for this disease, so it seems reasonable that delegates who believe there will be such a change in farming would be unlikely to view Influenza as less of a threat in the future, with an odds ratio of 0.15 for Wald and 0.18 for PML showing they are much less likely to change their opinion in comparison to the delegates who think there will be no change in farming.

In addition to changes in farming, changes in the populations of wild animals (Table 2.32) also had a possible effect on delegates opinions on the risk of influenza being introduced. With those who thought that the wild bird population would increase being 3.83 times more likely to think that Influenza would be less of a threat.

Table 2.32: Logistic Regression: Changed from Influenza (Incursion: Wild Animal Populations)

Incur	Changed from influenza	
	Odds Ratio	Odds Ratio
Wild populations	(Wald)	(PML)
Decrease(bird)	1.28 (0.13, 12.64)	0.45 (0.00, 4.46)
Decrease(mammal)	0.43 (0.05, 3.65)	0.59 (0.06, 2.91)
Decrease(both)	0.73 (0.23, 2.32)	0.76 (0.23, 2.26)
Increase(bird)	3.83 (1.09, 13.44)	3.75 (1.09, 12.84)
Increase(mammal)	1.53 (0.49, 4.77)	1.56 (0.49, 4.64)
Increase(both)	2.19 (0.78, 6.15)	2.17 (0.78, 6.01)
No significant change (Reference)		

This result does not, however, make intuitive sense, if a delegate believed there is likely to be an increase in the population of wild birds this would imply an increased way of influenza being introduced into the region and so an incursion would be more likely. This combined with the fact that the low end of the confidence interval for this value is close to one means that this result should be viewed with caution and further investigation as to why delegates scored this way would need to be carried out.

Table 2.33: Logistic Regression: Changed to WNF, RVF, CCHFV (Incursion: Farming)

Incursion	Changed to zoonotic arboviruses	
	Odds Ratio	Odds Ratio
Farming	(Wald)	(PML)
Decrease	0.60 (0.27, 1.33)	0.60 (0.27, 1.33)
Increase(intensive)	0.32 (0.13, 0.81)	0.33 (0.13, 0.81)
Increase(extensive)	0.59 (0.16, 2.09)	0.61 (0.17, 2.04)
Increase(both)	0.23 (0.06, 0.92)	0.25 (0.06, 0.89)
No significant change (Reference)		

Table 2.34: Logistic Regression: Changed to WNF, RVF, CCHFV (Difficult to eradicate: Wild Animal Populations)

Persist	Changed to zoonotic arboviruses	
	Odds Ratio	Odds Ratio
Wild populations	(Wald)	(PML)
Decrease(bird)	0.77 (0.08, 7.40)	0.27 (0.00, 2.64)
Decrease(mammal)	0.92 (0.22, 3.79)	1.00 (0.23, 3.58)
Decrease(both)	1.16 (0.47, 2.88)	1.17 (0.47, 2.85)
Increase(bird)	0.24 (0.03, 1.96)	0.10 (0.00, 0.86)
Increase(mammal)	0.40 (0.11, 1.52)	0.45 (0.11, 1.44)
Increase(both)	1.53 (0.59, 3.99)	1.53 (0.11, 1.44)
No significant change (Reference)		

For both factors that were found to be significant for the binary variable indicating a delegate had changed to scoring the zoonotic arbovirus group as a future threat, Tables 2.33 and 2.34, again the results do not make intuitive sense and so would require further investigation. The key result (those where confidence interval does not include one) from Table 2.33 is that delegates who believe there will be an increase in intensive or both types of farming are less likely to change to the zoonotic arbovirus

group as a threat in terms of a possible incursion. Since birds can be important for two of the diseases in this group, this result would not seem to make sense. From Table 2.34 the key result is that delegates who think there will be an increase in wild bird populations are less likely (odds ratios of 0.24 (Wald) or 0.1 (PML)) to change to the zoonotic arbovirus group as a threat in terms of persisting in their region. For the Wald value the confidence interval includes one and so is not quite as significant but again since an increase in wild birds should have the opposite effect this result may be simply coincidence.

2.4 Discussion

In this chapter the results of a survey of expert opinion on current and future threats were analysed using a number of techniques to identify what delegates regarded as current and emerging threats and what factors might explain their choices. This was done by examining the group scoring of delegates and the number of delegates who changed their opinions to identify current and emerging threats as well as by deriving an individual scoring approach and examining the explanatory variables behind this using a general linear model and logistic regression approach to explore what might cause a delegate to perceive a particular disease group as a current or emerging threat.

Examining delegates' total scores by backgrounds and expertise does not suggest much of a difference although policy delegates have widely differing opinions to other groups. However, this could be due to the relatively small number of policy delegates present in the sample.

The lack of significant factors in the one factor ANOVAs on total scores supports the idea that background or expertise is not particularly important in determining an expert's opinion in this case. While these variables do appear as significant factors in the general linear models (Figures 2.5, 2.6) they tend to be in instances in which there are fewer delegates. The sole exception to background was for Foot-and-mouth disease for which a large number of delegates do score the disease and for the general linear model and the logistic regression model a delegates' background is found to be significantly associated.

Stronger delegate bias is found when results are examined on whether a delegate has worked on a disease group or the region a delegate is from. The diseases a delegate works on was found to be significant in almost all cases and regional significance was found in many of the results. For example, total scores (Tables 2.11 and 2.12) have lower levels of correlation amongst regions indicating less agreement and a delegate's region was found to be significant in the general linear models and logistic regressions for Influenza and the zoonotic arbovirus groups. This could suggest 2 forms of opinion bias: firstly, that a delegate is more likely to believe that a disease group that threatens their region is equivalent to one that threatens all of Europe. This is understandable, especially since regional media and academic interest might focus on these groups. The second form of bias is that a delegate who works on a disease will obviously believe that they are working on something worthwhile and to support this belief the disease group must be important and so be either a current or emerging threat and the belief will influence their scoring.

An additional trend that runs through most of the results is that delegates are less certain about the future than the present. This can be seen by stronger correlations in total scores for current disease rankings than in future disease rankings and the greater number of none scoring combinations in Table 2.14 in comparison to Table 2.13. This makes intuitive sense since any future prediction is always going to be less certain than an opinion given on the present state of things; however, also related to the difference between current and future views is the increasing significance of the zoonotic arbovirus group. This disease group does not feature at all in the most common combinations of current disease groups perceived as a threat by delegates (Table 2.13) suggesting it is not seen as much of a current threat but, in combination with Foot-and-mouth disease, is viewed as a future threat by a large number of delegates (Table 2.14). The highly significant results in the testing of changes of delegate opinions for the incursion, spread and persistence of this disease group (Tables 2.28 and 2.29) show that these diseases are viewed as a much more serious threat in the future possibly at the expense of more traditional threats, such as Influenza and Foot-and-mouth disease.

The lack of certainty in an explanation for these future shifts highlights an issue with this approach. The differing scoring behaviours of delegates towards diseases are useful in highlighting potential biases that may appear when using expert opinion to identify future threats. However, examining disease groups individually does not properly examine such data and the scoring combinations is quite a simplistic approach to viewing all scores. If we look only at how a delegate scores a single group then the information present in the contrasts and potential tradeoffs between disease groups is missed. For example, the nature of the scoring system in this survey means that to score a disease group highly a delegate must score others lowly.

Overall in this chapter the analysis supports what was found in the original paper (Kelly et al., 2013), that experts believe that particular zoonotic arboviruses will become a threat and highlights a number of key factors that might explain this result. However, the approaches in this chapter do not adequately take into account one of the key factors in an expert's opinion and that is to score one disease group a delegate must not score others. Therefore, a univariate approach where just one disease group is modeled ignores a key factor. In the next chapter, in an attempt to examine these relationships, a multivariate approach will be utilised. The individual scores for all disease groups will be analysed at once and the differences between groups and current and future perceived threats will be identified.

Chapter 3

Multivariate analysis of expert opinion data

3.1 Introduction

In the previous chapter, a dataset of expert opinions gathered at an interactive question session at the 4th annual meeting of the EPIZONE network was analysed to examine what might determine expert's opinions on current and future disease threats. As part of this analysis, an individual score was formed for each of the six disease groups for each delegate and this score was used along with background information to explore what factors might be associated with a delegates views. This approach found that a delegates' views seemed to be strongly linked to the region they came from and the disease group that they worked in. Importantly, it identified a particular disease group, WNF, RVF & CCHFV, as an emerging threat. However, the univariate approach where each disease group was examined on its own means that potential patterns and factors that might be behind a delegate's total scoring choice for all six disease groups were missed. To examine their scoring behaviour as a whole, so how delegates score all 6 disease groups as the dependant variable, we make use of multivariate analysis techniques.

3.2 Aim

The aim of this chapter is to use the scores derived for the delegates for all six diseases and look into possible explanations of the overall scoring behaviour for a delegate as well as trying to identify patterns between the scores.

For this, it was desirable to reduce our six dimensional dataset (a dimension for each disease group score calculated in Equation 2.10, i.e. a delegate would have a score for influenza and a score for BT & AHS and a score for each of the other four disease groups) down to something simpler, that is the six variables down to one or two new variables that still contain as much information as possible from the originals. There are a few dimension reduction techniques available and, in this case, Principal Component Analysis was used.

As a secondary consideration, and for a deeper examination of delegate's behaviour, the likelihood and impact scores used to calculate the individual scores of each delegate are also looked at along with an examination of the possible issues brought about by analysing compositional data (data that consists of proportions).

As an alternative approach, cluster analysis was also used to examine the delegate disease group scores and the results from both methods were compared.

3.3 Methods

3.3.1 Dataset

The initial dataset used was detailed in the previous chapter and consists of the responses of 190 delegates (for consistency, the two policy delegates continued to be left out). The key variables in the dataset are those covering each delegate's background and the individual scores, both current and future, calculated for each disease group for each delegate. A full description of how these were derived is in the previous chapter but suffice to note that the score is found by multiplying a delegate's impact score (taking values of $\frac{j}{6}$ with j=1,..,6) and likelihood score (taking values of $\frac{i}{3}$ with

i=1,..,3) and so is technically a discrete variable (with values $\frac{i*j}{18}$ with i=1,..,3 and with j=1,..,6) but since it takes a fairly large number of values across a short range, it could be regarded as quasi continuous between zero and one.

3.3.2 Principal Component Analysis

Principal Component Analysis (PCA) was one of the earliest developed and most commonly used dimension reduction techniques. It works by taking a vector of possibly correlated random variables and uses an orthogonal transformation to convert them into a set of uncorrelated variables that are referred to as the Principal Components (PCs). Orthogonality refers to the variables being able to vary independently, so when used in terms of statistics, it means they will be uncorrelated and if represented graphically will be perpendicular to each other. Each PC consists of a linear combination of our original variables and is formed in such a way that the first PC will explain as much of the variation in the dataset as possible. The number of PCs will be limited to as many as the original variables but it is to be hoped that the majority of information, i.e. variation, can be explained by the first few PCs. Thus, if we have a vector $\mathbf{x}$ consisting of p variables, then our principal components would be a linear function of $x_1, ..., x_p$ so if, for example, we let α represent the coefficients of each PC (so a vector of p constants) we would have for PC1:

$$\alpha_{\mathbf{1}}'\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + .. + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j$$

To find these PCs we use either the covariance or correlation matrix of $\mathbf{x}$ and solve to find its eigenvalues and eigenvectors. For our first PC we take the largest eigenvalue of the matrix and the corresponding eigenvector consists of the coefficients that make up α_1. For our second PC we do the same with the second largest eigenvalue whose corresponding eigenvector will give us the coefficients of α_2 and so on for all p possible principal components.

In most applications of PCA, the correlation matrix is used as this has the effect of canceling out the influence of differences in the scale and units of different variables;

i.e. if the covariance matrix were to be used, then greater weight would be given to the larger variables, and so when using the correlation matrix, we find the PCs for a standardised version of x. For example, if there were variables measuring something of similar size but one variable had millimeters as a scale and the other meters then the latter would be given more weight and a spurious result may be found where it is the dominant factor in the PC. In the case studied in this chapter, however, a stronger argument can be made for using the covariance matrix. Firstly, the units are the same, as all the variables here are measures of a delegate's opinion on a disease group and, secondly, it makes more intuitive sense for the disease groups that are scored higher to be given more weight. This could be thought of as the more highly scored diseases will be likely to cover a greater range of values and so will also be likely to contribute more to variation within the dataset.

PCA was carried out for the current and future scores and the first 2 PCs would then explain the most variation in the data. Plotting these PCs allowed underlying patterns in the data to be investigated and marking points on these plots, coloured by a delegate's background factors (Table 2.1), allowed us to see whether a background factor contributed to the pattern. Analysis of variance (ANOVA) was used for each PC and background question to provide a more numerical approach to examining the same affects. PCA was also carried out for the impact, likelihood and future likelihood scores and for these a variant of this method was also investigated. This was necessary as these are what are referred to as compositional data; that is they consist of a vector of proportions and so have the constraint that they must sum to one. Such data also tend to display a curved relationship between variables and both of these factors mean that standard PCA can sometimes be inadequate and produce results that are of little use.

A transformation of the data was used in this chapter, known as the centered log ratio transform, and involves taking the natural logarithm of each variable, a nonlinear function, and centering them before then performing principal component analysis as usual. So, for example, using the notation we used above we would have a new variable $\mathbf{y}$:

$$\mathbf{y_1} = log(x_1) - [\frac{log(x_1) + log(x_2) + .. + log(x_p)}{p}]$$

There is some debate about this type of approach, however, and one of the key issues for this example is that it cannot be used whenever any of the proportions are zero, which is very common in this case, and so the results for both the transformed and non-transformed data were found and investigated. For the zero values in the dataset, these were replaced with half of the smallest non-zero value as was suggested in the paper from which the method was taken (Aitchison, 1982a).

To examine the principal components, the coefficients of the three principal components associated with the largest eigenvalues were considered and scatter plots of the top two components were produced with and without groups (broken down by background, expertise, region and disease worked on) marked down. To further examine the role of background as a determinant of variability, one factor ANOVAs were performed on the top two principal component scores to see what background factors might contribute the most to variability in opinions and the results were discussed with comparisons to the graphical results. The most basic interpretation of this is to test whether the means of PC1 and PC2 for different groups are equal. Tukey's range test (with overall significance level of 5%) was used to test each group against the others and further check for differences.

3.3.3 Cluster Analysis

Cluster analysis is a common exploratory technique used in data analysis to group observations or variables by similarity. It is used to form natural groups which in theory should represent an underlying pattern or set of rules within the data. These groups can be perfectly distinct, so an individual only falls within one group, or overlapping where an individual could be classified as belonging to multiple groups.

Within cluster analysis there are two broad approaches, hierarchical and non-hierarchical (sometimes referred to as k-means or partitioning) (Witten et al., 2011). Hierarchical clustering as the name suggests attempts to form a hierarchy of clusters and

can be agglomerative (a bottom up approach where each instance starts as its own cluster and pairs of clusters are merged as we move up the hierarchy) or divisive (a top down approach where all instances start in one cluster and splits are performed as we move down the hierarchy). The output from this kind of approach is often presented as a dendrogram with axes consisting of variables or instances against a measure of similarity.

Non-hierarchical clustering has no such hierarchy of clusters but instead clusters are formed round a distribution, central vector or level of density within multi-dimensional space (with dimensions equal to the number of variables). One of the more common non-hierarchical methods is k-means clustering; in this technique the user specifies in advance how many clusters are required, denoted k. Then k points are randomly selected and become our cluster centroids, all remaining points are then assigned to a cluster based on the closest Euclidean distance. Once this is complete a new centroid for each cluster is calculated from the mean of all instances within that cluster and all points are assigned to these new clusters. This process is repeated until the composition of clusters stabilise. Centroids can then be used to classify clusters by examining the larger coefficients in their compositions, for example one centroid may be determined by a contrast between two of the variables or a balance of three, such interpretation is very similar to that for the PCs resulting from PCA.

The number of clusters can be selected based on some prior knowledge, i.e. expert opinion, or can be selected by examining the within cluster sum of square errors. This is calculated by summing the squared error for each point against its centroid. This value is calculated for the largest estimated number of clusters, often slightly above the number of variables, and the percentage difference between the sum of square errors for a number of clusters n and n+1 is examined. The number of clusters that results in the smallest percentage increase in error is then selected.

To examine the individual scores data, k-means clustering will be used and will be expected to produce results in agreement with those produced using principal components analysis.

3.4 Results

3.4.1 Principal Component Analysis

Disease Group Scores

Analysing the disease scores, using no transformations given that the final scores cannot be assumed to be compositional data, Table 3.1 shows the eigenvalues and proportions of variation for each PC for the delegate's current disease scores using the covariance matrix so as not to standardise the disease groups. Table 3.2 shows the same results produced for the delegates future disease scores.

Table 3.1: Eigenanalysis of Covariance Matrix of Current Scores

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.029	0.020	0.008	0.006	0.004	0.001
Proportion of variation	0.425	0.292	0.121	0.089	0.060	0.014
Cumulative Proportion of variation	0.425	0.717	0.838	0.927	0.986	1.000

Table 3.2: Eigenanalysis of Covariance Matrix of Future Scores

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.023	0.017	0.014	0.004	0.004	0.001
Proportion of variation	0.343	0.257	0.208	0.117	0.060	0.015
Cumulative Proportion of variation	0.343	0.600	0.808	0.925	0.985	1.000

Examining the eigenvalues and the respective proportions of variability they represent we can see that in both cases the first three principal components explain over 80% of the variation in the dataset. We can see, however, that the proportion of variability is more skewed towards fewer principal components as regards current disease groups, this is illustrated below in Figure 3.1:

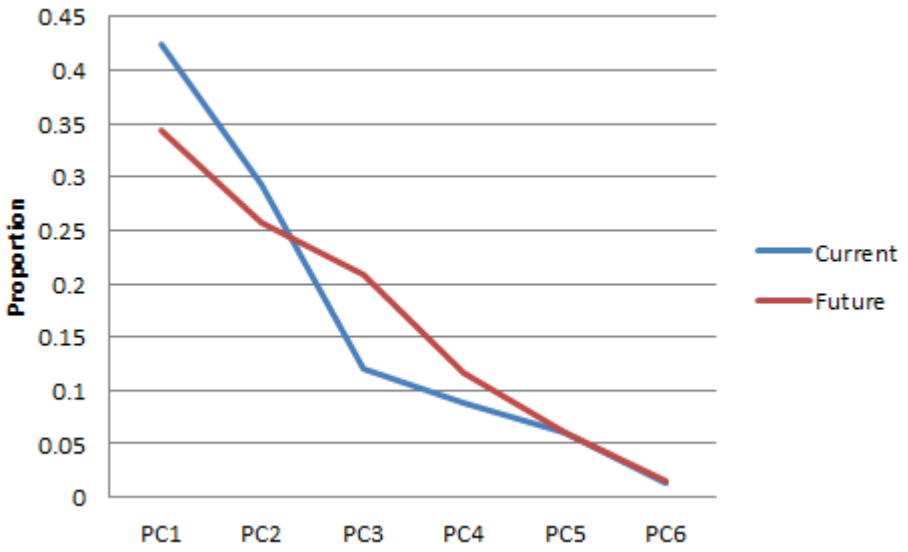

Figure 3.1: Principal Component Proportions

That is, as can be seen in Figure 3.1, the first two or three principal components explain a greater part of the variation in current opinions as opposed to future opinions where the variation is more spread out amongst components. This could be interpreted as more factors or groups having an effect in the future in comparison to the present. This reinforces what was discussed in the previous chapter where current opinion was dominated by two disease groups (influenza and foot-and-mouth disease) but for the future the balance shifted and while these two groups remained important the zoonotic arbovirus group also started to have an influence.

Table 3.3: Principal Component Coefficients (Current Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	0.673	0.024	0.016	0.014	-0.738	-0.031
PC2	-0.712	0.189	0.092	0.001	-0.648	0.173
PC3	-0.010	-0.808	0.109	0.005	-0.056	0.173
PC4	0.163	0.497	-0.331	-0.026	0.125	0.775
PC5	0.117	0.251	0.932	-0.044	0.127	0.190
PC6	0.001	0.028	0.031	0.999	0.020	0.026

Examining the coefficients of the top three principal components for the current scores we can see from Table 3.3 that the first component is dominated by a contrast between influenza (large positive coefficient) and foot-and-mouth disease (large negative coefficient) meaning that the majority of variation amongst delegates is caused by the differing opinions on these diseases. This is to be expected, since these were considered the two most important diseases and it means delegates were likely to have had strong opinions towards them (so as to give a high enough score in both the impact and likelihood questions) and, since to score one disease high you must score another low then a contrast between these makes a lot of sense.

The second principal component, and so the second biggest contributor to variation in the data, is strongly determined by a balance of the two key disease groups from the first component (both with large negative coefficients) with a weak contrast against the two arbovirus groups (much smaller positive coefficients). This means that most of the variation in current scoring is created by delegates selecting either Influenza or Foot-and-mouth disease with the majority of the remaining variation being those delegates who select both.

Finally the third principal component is a contrast between the two arbovirus groups with much greater importance being attached to BT & AHS but explains much less of the variation than the first two.

Table 3.4: Principal Component Coefficients (Future Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	-0.886	0.043	0.004	0.007	0.354	0.296
PC2	0.046	-0.046	-0.052	-0.008	0.708	-0.702
PC3	-0.391	0.445	0.068	0.029	-0.534	-0.599
PC4	-0.227	-0.876	0.224	0.053	-0.264	0.240
PC5	-0.084	-0.166	-0.969	0.075	-0.133	-0.057
PC6	0.036	0.046	0.059	0.995	0.043	0.026

Unlike the current scores, looking at Table 3.4, the first principal component for future scores is not a simple contrast between two disease groups but is instead dominated by a balance of foot-and-mouth and the zoonotic arbovirus group (both positive large coefficients) against influenza (much larger negative coefficient); however, the second principal component is then mainly a contrast between foot-and-mouth disease and the zoonotic arbovirus group. This means that, if we look at the first two principal components together, then we have a contrast between the three disease groups that were already identified by delegates as the most important in regards to future threats. The third principal component is then all three of these groups contrasted against the non-zoonotic arbovirus group (three large negative and a single larger positive coefficient).

A further point worth mentioning is that for both the current and future scores the last two PCs have much the same structure. PC5 is dominated by ASF and PC6 by CSF; what this indicates is that these two diseases are where there is very little variation and are near constant amongst delegates i.e. delegates don't really change their opinions on them.

To further show the interaction amongst disease groups as a factor in explaining variation, the two primary principal components were calculated for each delegate and were plotted against each other. So, for example, the current scores for a delegate's disease groups would be multiplied by the PC1 coefficients in Table 3.3 and

added to give a PC1 value for that delegate. The equivalent would be done for PC2 and then these two values would give the coordinate for that delegate. Viewing the components graphically in Figure 3.2 gives a better and more straightforward demonstration of how the important disease groups for scoring interact.

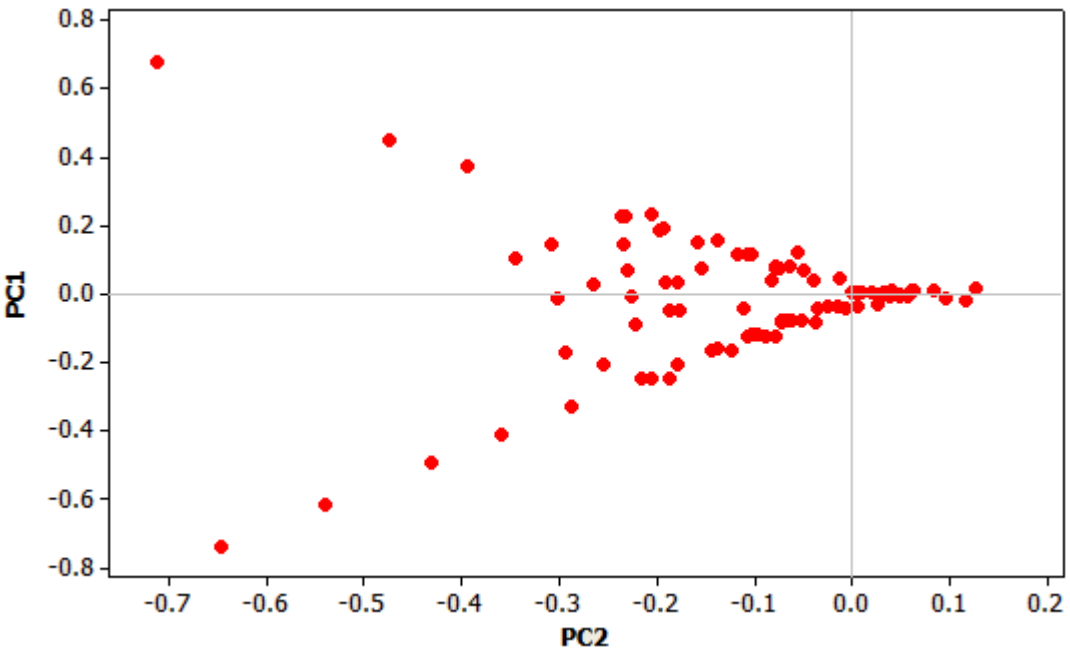

Figure 3.2: Scatter plot of Principal Components 1 & 2 for Current Scores

For our current scores, Figure 3.2, we can see that there are two 'prongs' of behaviour, the points in the top left of the graph indicate those delegates who scored influenza very highly and foot-and-mouth disease as much less of a threat and the points in the lower left are the delegates who did the opposite. Both these sets of points have high scores in principal component one meaning they contribute a lot to the level of variation in the dataset. These points also have low scores in principal component 2 which reinforces the idea that some delegates lean towards viewing one of these two diseases as a threat to the exclusion of all other disease groups.

As a delegate's views on either of these two become less strong, we move in towards the origin of the graph and lower scores for both components. There are a few delegates who didn't view influenza or foot-and-mouth as a key threat but did rate one

of the arbovirus groups and these are the points to the right of the axis, but, as can be seen, this behaviour does not dominate the dataset.

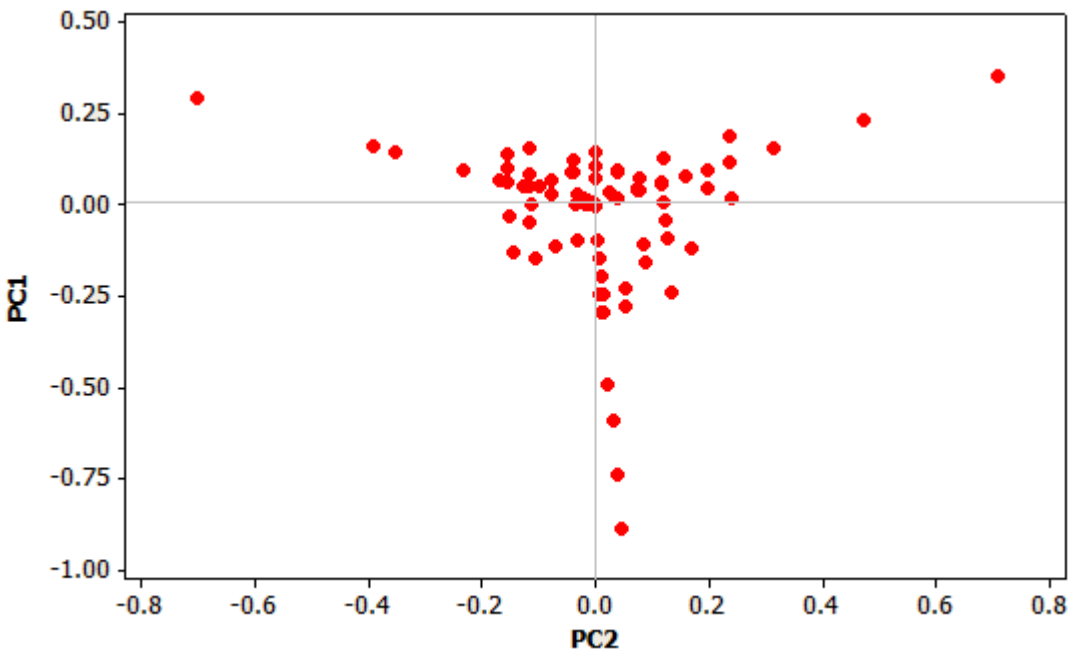

Figure 3.3: Scatter plot of Principal Components 1 & 2 for Future Scores

A similar idea can be seen in Figure 3.3 for the future scores although here there are three distinct types of behaviour displayed. The point on the top left of the plot indicates delegates who scored zoonotic arboviruses as the greatest threat, points on the top right are delegates who scored foot-and-mouth as the greatest threat and, finally, delegates in the lower part of the plot are those who regarded influenza as the most threatening. As we move in towards the origin from any of these three areas we are moving through the delegates whose opinions towards any one particular disease group were less strong and so the level of variation in their scores was lower.

This change in behaviour, shown by the change in shape between Figure 3.2 and Figure 3.3, is similar to the change identified in both the original report (Kelly et al., 2013) and in the previous chapter where the diseases of the zoonotic arbovirus group were perceived by delegates as being more threatening in the future; that is, an emerging threat.

To further explore whether there is anything more behind these patterns, the plots were reproduced with the points marked according to the various background questions delegates were asked and, for a more numerical approach, one factor ANOVA and Tukey's pairwise comparison were produced for each of the first two principal components for each background question to see whether there was a significant difference between group means. For example, for the actual values of PC1 for current scores (the values on the y-axis in Figures 3.2) is there a significant difference between the values for delegates whose background was Research, Policy, Industry or EPIZONE. The factors that were found to be significant, had a p-value below 0.05.

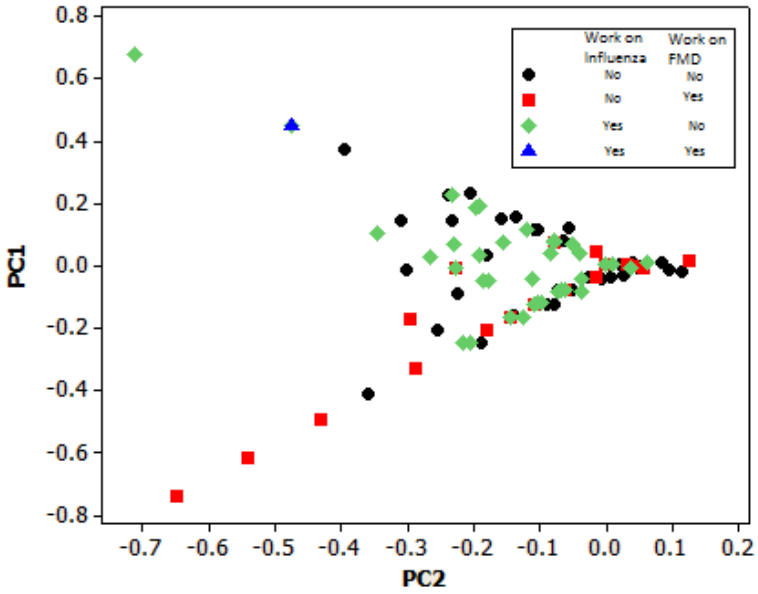

Figure 3.4: Scatter plot of Principal Components 1 & 2 for Current Scores with delegates status regarding working on Influenza and Foot-and-mouth disease

For current scores, in Figure 3.4, it can clearly be seen that delegates who work on foot-and-mouth disease and not on influenza account for nearly all the more extreme negative values in principal component one which are the delegates who score foot-and-mouth disease as the greatest threat so this is in agreement with many of the results from the previous chapter. For the other more extreme points, it is a bit less clear cut: all the delegates there did work on influenza but one also worked on

foot-and-mouth disease. As we move in towards the origin there are more delegates who didn't work on either of these diseases.

Table 3.5: ANOVA of current PC1 against Worked on Influenza

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on Influenza	1	0.19352	0.19352	0.19352	6.86	0.10
Error	188	5.30720	5.30720	0.02823		
Total	189	5.50071				

Table 3.6: ANOVA of current PC2 against Worked on Influenza

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on Influenza	1	0.04269	0.04269	0.04269	2.15	0.144
Error	188	3.73051	3.73051	0.01984		
Total	189	3.77320				

Table 3.7: ANOVA of current PC1 against Worked on FMD

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on FMD	1	0.47070	0.47070	0.47070	17.59	0.000
Error	188	5.03001	5.03001	0.02676		
Total	189	5.50071				

Table 3.8: ANOVA of current PC2 against Worked on FMD

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on FMD	1	0.12730	0.12730	0.12730	6.56	0.011
Error	188	3.64590	3.64590	0.01939		
Total	189	3.77320				

For delegates who had worked on foot-and-mouth disease, their mean was lower for both principal components, which would result in them being placed in the lower left corner of our plot, Figure 3.4, and lending more statistical evidence to the idea

that there is a distinct pattern here are the low p-values in Tables 3.7 and 3.8. For delegates who had worked on influenza (Tables 3.5 and 3.6) the p-values were still not significant but were small, suggesting that there is a bit more of a difference between those who have and have not worked on influenza.

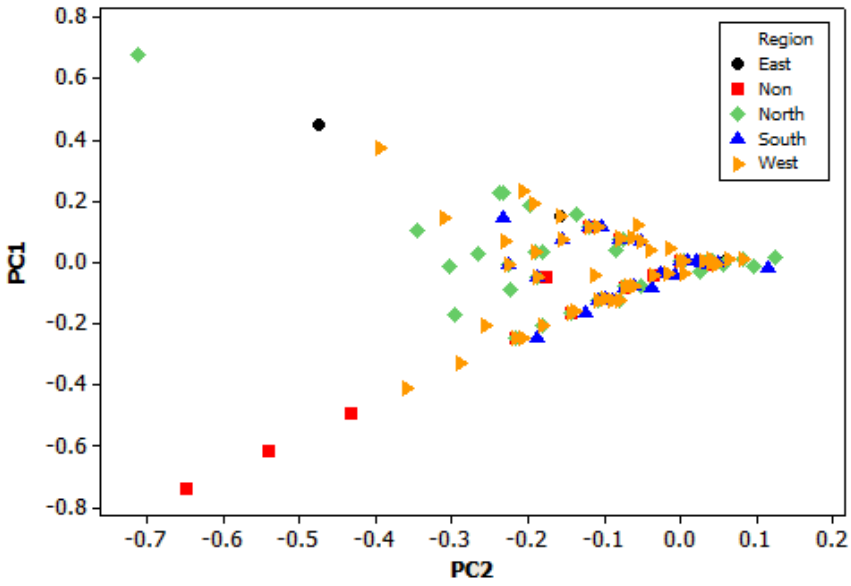

Figure 3.5: Scatter plot of Principal Components 1 & 2 for Current Scores with delegates region

Table 3.9: ANOVA of current PC1 against Region

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Region	4	0.75481	0.75481	0.18870	7.36	0.000
Error	185	4.74590	4.74590	0.02565		
Total	189	5.50071				

Table 3.10: ANOVA of current PC2 against Region

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Region	4	0.31468	0.31468	0.07867	4.21	0.003
Error	185	3.45852	3.45852	0.01869		
Total	189	3.77320				

Similarly, when plotting according to region, in Figure 3.5, it is the delegates who scored foot-and-mouth disease highly who have the most distinct pattern with all of them coming from outside the EU; again this is in accord with results from the previous chapter and is likely to be due to the fact that many of the non-EU delegates were from China where this disease is endemic in certain areas: checking the World Animal Health Information database run by the OIE displays foot-and-mouth being present in China since 2005, which is as long as the database has been running (for Animal Health , OIE). The more extreme point for high influenza scores is from a delegate from the North; however, since the other less extreme points in that area are from different areas it can't be regarded as strong a pattern.

From the ANOVA approach for region, the only strong conclusion was that non-European delegates were much likely to have a lower score for principal component one (highly significant p-value and lower mean along with different group for pairwise comparison) and that they were likely to score lower for component two than southern or western delegates. This is quite surprising when looking at the plot, Figure 3.5, as there is some overlap between these regional groups.

As with our current scores, it was the disease worked on and region that gave the strongest indication of a pattern for the future. Firstly in Figure 3.6 for delegates who worked on only one of the three disease groups, they tended to dominate the more extreme values and those delegates who worked on none of the three were predominantly spread evenly about the origin meaning these delegates contributed less to the dataset variability.

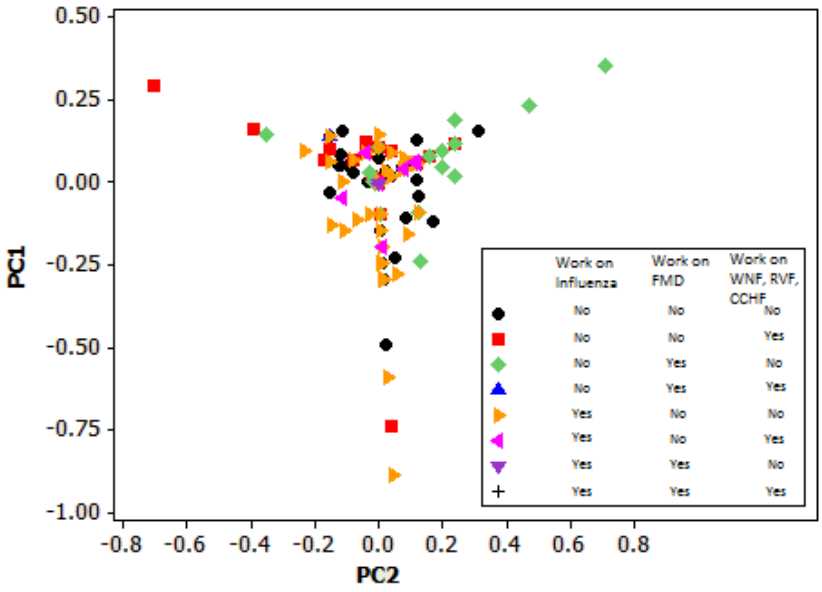

Figure 3.6: Scatter plot of Principal Components 1 & 2 for Future Scores with delegates status regarding working on Influenza, Foot-and-mouth disease and zoonotic arbovirus group

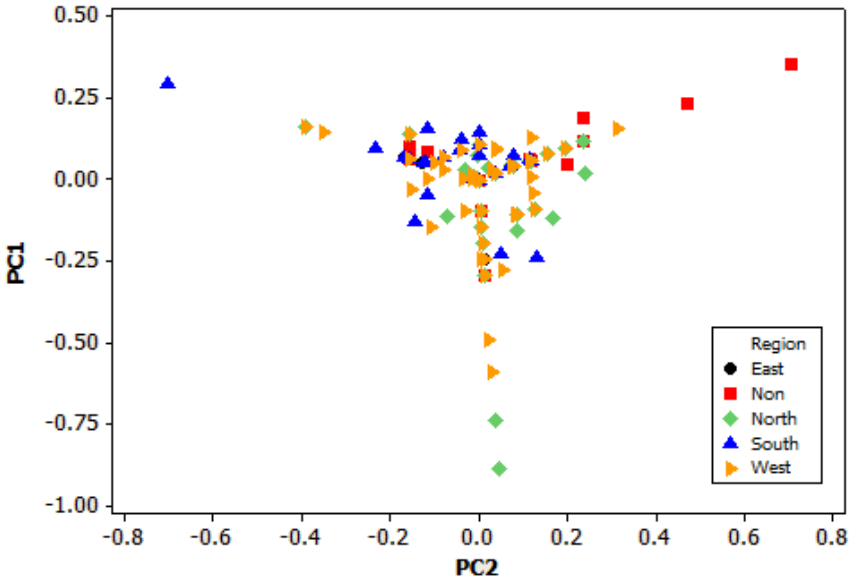

Figure 3.7: Scatter plot of Principal Components 1 & 2 for Future Scores with delegates region

When region is marked on the plot, Figure 3.7, we can see that the delegates who are

more strongly inclined to view influenza as a threat come from the North, a region where there had been outbreaks in recent years and where it would be considered likely for there to be outbreaks again, and delegates who are more strongly inclined to view foot-and-mouth as a threat come from outside the EU and this could be attributed to the same reasoning as for the similar behaviour in current scores. For the zoonotic arboviruses, it is less clear, although a case could be made for the more extreme views coming from delegates from the South, an area where these are regarded as an emerging threat. Since, under the region breakdown that was used, Corsica is counted as West, this may explain why some western delegates also have more extreme views towards the zoonotic arbovirus group since WNF has some history there. Also, since much of central Europe is listed as West, then this would also explain why they view influenza as more of a threat, for the same reasons as the northern delegates.

Table 3.11: ANOVA of future PC1 against Worked on Influenza

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on Influenza	1	0.24009	0.24009	0.24009	11.11	0.001
Error	188	4.06213	4.06213	0.02161		
Total	189	4.30223				

Table 3.12: ANOVA of future PC2 against Worked on Influenza

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on Influenza	1	0.00566	0.00566	0.00566	0.33	0.566
Error	188	3.22638	3.22638	0.01716		
Total	189	3.23204				

For the delegates who had worked on influenza, the mean for principal component one was lower than those who had not, and nonsignificant for principal component two, meaning they were down in the lower part of our plot, Figure 3.6, and, as such, a more distinct group.

Table 3.13: ANOVA of future PC1 against Worked on FMD

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on FMD	1	0.11959	0.11959	0.11959	5.38	0.022
Error	188	4.18264	4.18264	0.02225		
Total	189	4.30223				

Table 3.14: ANOVA of future PC2 against Worked on FMD

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on FMD	1	0.17335	0.17335	0.17335	10.66	0.001
Error	188	3.05869	3.05869	0.01627		
Total	189	3.23204				

Delegates who worked on foot-and-mouth disease had a higher mean for both principal components meaning they were at the top right of the plot, Figure 3.6, and, as was the case for influenza, they can be regarded as distinct.

Table 3.15: ANOVA of future PC1 against Worked on WNF, RVF & CCHFV

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on WNF, RVF & CCHFV	1	0.05896	0.05896	0.05896	2.61	0.108
Error	188	4.24327	4.24327	0.02257		
Total	189	4.30223				

Table 3.16: ANOVA of future PC2 against Worked on WNF, RVF & CCHFV

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Worked on WNF, RVF & CCHFV	1	0.12543	0.12543	0.12543	7.59	0.006
Error	188	3.10661	3.10661	0.01652		
Total	189	3.23204				

Table 3.17: ANOVA of future PC1 against Region

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Region	4	0.24156	0.24156	0.06039	2.75	0.030
Error	185	4.06067	4.06067	0.02195		
Total	189	4.30223				

Table 3.18: ANOVA of future PC2 against Region

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Region	4	0.32332	0.32332	0.08083	5.14	0.001
Error	185	2.90873	2.90873	0.01572		
Total	189	3.23204				

For region, it was significant for principal component one but highly significant (p=0001) for principal component two, though in terms of significantly different groups, all we can say is that those from outside Europe had a higher mean than western and southern delegates and southern delegates had a lower mean than northern or non-European delegates.

The other background section questions concerned the background of a delegate with the options of being a member of an EPIZONE organisation, research or industry and a delegate's area of expertise and the PC plots with these marked, Figures 3.8 and 3.10 for current scores and Figures 3.9 and 3.11 for future scores.

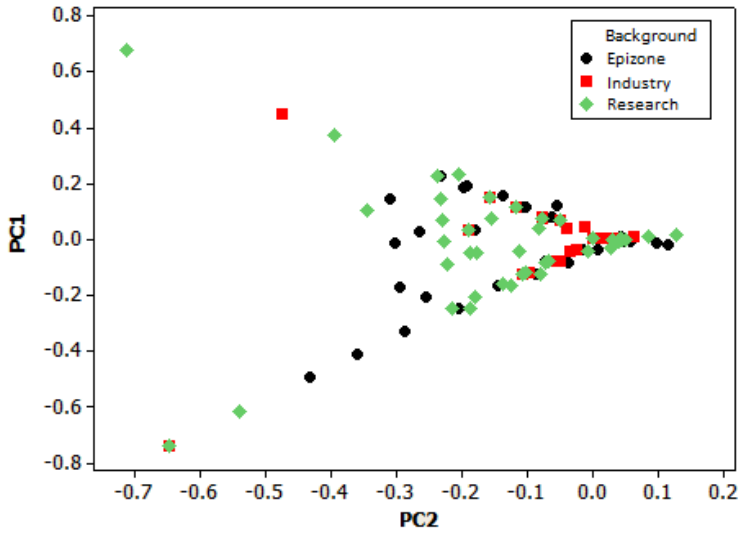

Figure 3.8: Scatter plot of Principal Components 1 & 2 for Current Scores with delegates' background

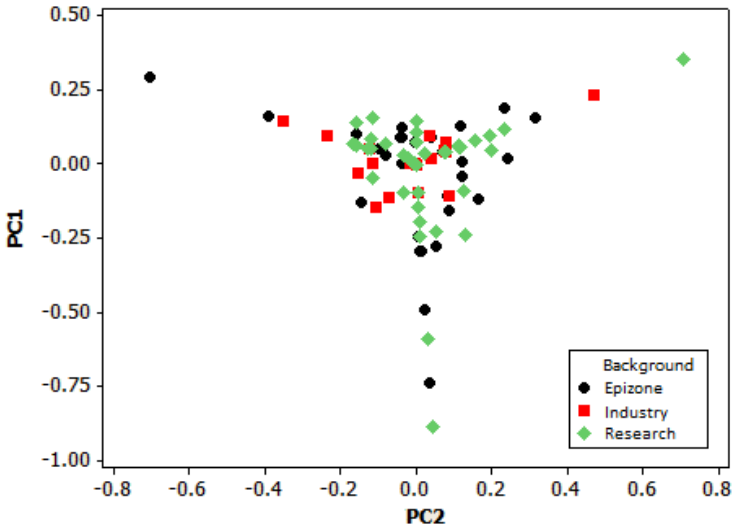

Figure 3.9: Scatter plot of Principal Components 1 & 2 for Future Scores with delegates' background

In neither the present (Figure 3.8) or future (Figure 3.9) does a delegate's background seem to suggest having a significant link with their scoring behaviour.

Table 3.19: ANOVA of current PC1 against Background

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Background	2	0.01643	0.01643	0.00822	0.28	0.756
Error	187	5.48428	5.48428	0.02933		
Total	189	5.50071				

Table 3.20: ANOVA of current PC2 against Background

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Background	2	0.02875	0.02875	0.01437	0.72	0.489
Error	187	3.74445	3.74445	0.02002		
Total	189	3.77320				

Table 3.21: ANOVA of future PC1 against Background

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Background	2	0.04218	0.04218	0.02109	0.93	0.398
Error	187	4.26005	4.26005	0.02278		
Total	189	4.30223				

Table 3.22: ANOVA of future PC2 against Background

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Background	2	0.05450	0.05450	0.02725	1.60	0.204
Error	187	3.17754	3.17754	0.01699		
Total	189	3.23204				

For current or future scores, neither of the PCs seemed to be significantly different between backgrounds, so an industry delegate varied in scoring in the same way as an EPIZONE delegate. This is the same as was found in Figure 3.8.

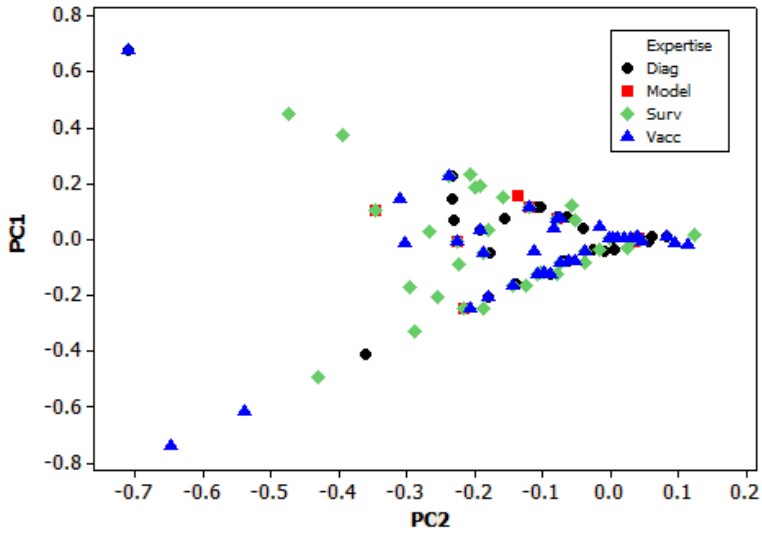

Figure 3.10: Scatter plot of Principal Components 1 & 2 for Current Scores with delegates expertise

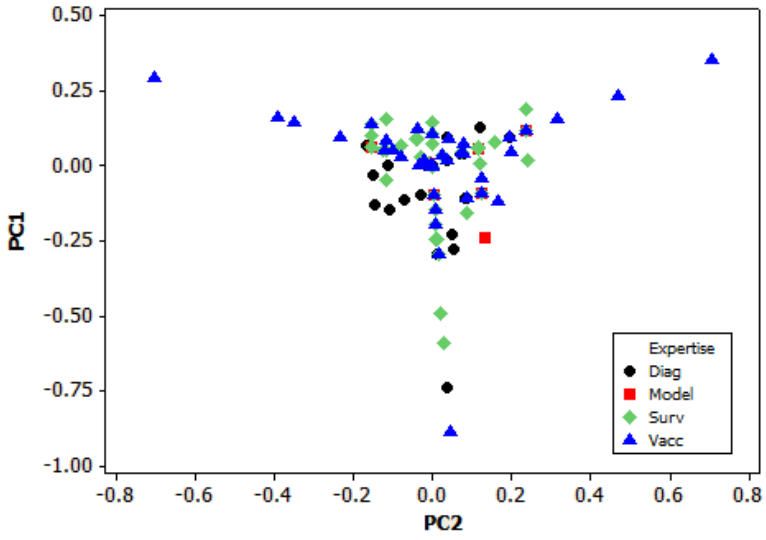

Figure 3.11: Scatter plot of Principal Components 1 & 2 for Future Scores with delegates expertise

Table 3.23: ANOVA of current PC1 against Expertise

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Expertise	3	0.09540	0.09540	0.03180	1.09	0.353
Error	186	5.40531	5.40531	0.02906		
Total	189	5.50071				

Table 3.24: ANOVA of current PC2 against Expertise

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Expertise	3	0.01298	0.01298	0.00433	0.21	0.887
Error	186	3.76022	3.76022	0.02022		
Total	189	3.77320				

Table 3.25: ANOVA of future PC1 against Expertise

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Expertise	3	0.03099	0.03099	0.01033	0.45	0.718
Error	186	4.27124	4.27124	0.02296		
Total	189	4.30223				

Table 3.26: ANOVA of future PC2 against Expertise

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Expertise	3	0.02250	0.02250	0.00750	0.43	0.728
Error	186	3.20954	3.20954	0.01726		
Total	189	3.23204				

Similarly for a delegate's expertise, there does not appear to be a pattern in the plots (Figures 3.10 and 3.11) and testing numerically (Tables 3.23, 3.24, 3.25 and 3.26) shows there is no difference in the means of the PCs between the different levels.

Unlike Figures 3.4, 3.5, 3.6 and 3.7, there is no clear pattern here, though from Figures 3.10 and 3.11, a case could be made for arguing that delegates whose expertise lies in vaccinations tend to give more varied values.

Transformation

In his main paper of compositional data Aitchison (1982b) describes the idea of perturbation and the production of a perturbed composition. This is produced by multiplying the elements of two compositions and then transforming the values to fit the correct constraints. The individual scores are produced using the Impact and Likelihood scores which are true compositions. Therefore, if the individual scores were to be modified to obey the same constraint, then a perturbed composition would be produced which would be viable for the transformation discussed earlier.

There are many zeros present in the data for individual scores and the solution of using half of the smallest non-zero value is used as suggested in Aitchison (1982a) although some more advanced techniques have since been suggested (Aitchison and Kay, 2003).

Initial interest is on whether the results are substantially different between the true compositional data that have been transformed and those produced previously, and looking at the components that are produced for the transformed scores (Tables 3.27 and 3.28) we can highlight any difference.

Table 3.27: Eigenanalysis of Covariance Matrix (Transformed Current Scores)

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.231	0.151	0.088	0.052	0.021	0.000
Proportion of Variation	0.426	0.279	0.162	0.095	0.038	0.000
Cumulative of Variation	0.426	0.705	0.867	0.962	1.000	1.000

Comparing Table 3.27 with Table 3.1 and Table 3.28 with Table 3.2 will allow any changes in the level of importance of components to be highlighted.

Table 3.28: Eigenanalysis of Covariance Matrix (Transformed Future Scores)

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.203	0.157	0.117	0.059	0.030	0.000
Proportion of Variation	0.359	0.277	0.207	0.104	0.053	0.000
Cumulative of Variation	0.359	0.636	0.843	0.947	1.000	1.000

Examining the eigenvalues and the respective proportions of variability they represent we can see that in both cases there is not a great deal of difference between the original and transformed variables in the amount of variation explained by principal components. While this shows that the importance of each component is relatively unchanged, the structure of these components may have changed radically and so the coefficients for the original and transformed variables must be compared.

Table 3.29: Principal Component Coefficients (Transformed Current Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	0.722	0.022	0.021	0.027	-0.682	-0.110
PC2	0.517	-0.373	-0.238	-0.076	0.593	-0.423
PC3	0.071	-0.763	0.117	-0.008	-0.047	0.630
PC4	-0.137	-0.276	0.815	0.120	-0.048	-0.473
PC5	0.145	0.187	0.313	-0.901	0.108	0.149
PC6	0.408	0.408	0.408	0.408	0.408	0.408

Table 3.30: Principal Component Coefficients (Transformed Future Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	0.784	0.015	0.041	0.021	-0.509	-0.353
PC2	0.178	-0.094	-0.110	-0.033	0.717	-0.658
PC3	-0.0375	0.736	0.170	0.150	-0.210	-0.471
PC4	0.200	0.501	-0.763	-0.258	0.095	0.226
PC5	0.081	0.175	0.457	-0.862	0.089	0.060
PC6	0.408	0.408	0.408	0.408	0.408	0.408

Examining the coefficients of the top three principal components for the transformed current scores we can see from Table 3.29 that similarly to Table 3.3 the first component is dominated by a contrast between influenza, with a large positive coefficient, and foot-and-mouth disease, with a large negative coefficient. Again, similarly, the second principal component is a slightly weaker contrast between the two diseases of our first principal component against the two arbovirus groups. The third principal component is a contrast between the two arbovirus groups meaning that the structure of all three of the most important principal components is the same between the transformed and original scores.

Contrasting Tables 3.4 and 3.30 there are a few differences. For the first principal component under the transformed variables the signs of the two main coefficients have reversed; however, this makes no real difference to the interpretation since it is still a contrast between the balance of foot-and-mouth and the zoonotic arbovirus group against influenza.

The second principal component has the same structure in both tables as does the third. However, in the case of the third principal component, it is still a contrast of influenza, zoonotic arbovirus group and foot-and-mouth against the non-zoonotic arbovirus group but under the transformed variables, more importance is placed on the zoonotic arbovirus group and less on influenza and foot-and-mouth disease.

Overall the transformation makes almost no difference to final results and so in this case, where it is debatable as to whether the data are truly compositional or not, then the results using the original individual scores seem perfectly valid.

3.4.2 Principal Component Analysis of Impact and Likelihood Scores

The initial analysis of the impact, likelihood and future likelihood scores was undertaken using the same principal component approach as for the disease group scores (Tables 3.1, 3.2, 3.3 and 3.4) before then examining them using the alternative log centered approach.

Table 3.31: Eigenanalysis of Covariance Matrix (Impact Scores)

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.108	0.073	0.310	0.027	0.026	0.001
Proportion of Variation	0.406	0.275	0.117	0.103	0.096	0.003
Cumulative Proportion of Variation	0.406	0.681	0.798	0.901	0.997	1.000

For impact scores, as can be seen in Table 3.31, the first three principal components again explain the large majority of the variation and so their coefficients were examined:

Table 3.32: Principal Component Coefficients (Impact Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	-0.828	-0.025	0.067	0.070	0.185	0.519
PC2	-0.033	-0.176	-0.185	-0.153	0.893	-0.336
PC3	-0.369	0.189	0.544	0.290	-0.065	-0.666

From Table 3.32 it can be seen that the first principal component is dominated by a

contrast between influenza and the zoonotic arboviruses along with a weaker contribution from foot-and-mouth disease. At first this might seem surprising since earlier results always featured strong influences from both influenza and foot-and-mouth but since this relates to the impact of the disease and the other two groups here tend to have higher fatality rates, then it is perhaps to be expected. The second component then brings in the importance of foot-and-mouth disease with it being contrasted against a balance of almost all other groups with the exception of influenza simply meaning that foot-and-mouth disease is a secondary source of variation for delegates who view the other diseases as less important in terms of impact. The third principal component is then a contrast between influenza and the zoonotic arbovirus group against non-zoonotic arboviruses and the two swine fevers. The fact that this third component contains so many groups and delegates are restricted on the number they can score means it may be less important even though it has a reasonable proportion of the data variability assigned to it.

Table 3.33: Eigenanalysis of Covariance Matrix (Current Likelihood Scores)

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.090	0.076	0.0581	0.042	0.013	0.001
Proportion of Variation	0.321	0.273	0.206	0.151	0.047	0.002
Cumulative Proportion of Variation	0.321	0.594	0.800	0.951	0.998	1.000

Moving on to likelihood scores in Table 3.33 there is less of a jump between the first and second principal components meaning they are both almost equally important in explaining variance.

Table 3.34: Principal Component Coefficients (Current Likelihood Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	-0.277	-0.629	0.198	-0.025	0.698	0.022
PC2	-0.657	0.638	-0.037	-0.061	0.329	-0.219
PC3	-0.482	-0.070	-0.018	0.018	-0.275	0.829

Our first principal component in Table 3.34 is dominated, surprisingly, by a contrast between foot-and-mouth disease and the non-zoonotic arbovirus group which has tended not to feature strongly in any results. It is, however, slightly balanced with influenza that also has a reasonably large coefficient with the same sign. The second component rearranges the same three disease groups with influenza against the non-zoonotic arbovirus group balanced weakly with foot-and-mouth disease; however, this time the other arbovirus group also has an affect being balanced with influenza. Finally, the third principal component contrasts the zoonotic arboviruses against both influenza and foot-and-mouth disease. The fact that all three components consist of rearrangements of these groups and there is less of a difference between proportions of variability for the three components mean much of the variation in delegates' likelihood scores is based upon the opinions of these groups which then equally means the two remaining groups have a high level of consistency amongst delegate opinions on their likelihood as a threat; that is neither of the swine fever groups will contribute much to variation so will be likely to never be scored unusually.

Table 3.35: Eigenanalysis of Covariance Matrix (Future Likelihood Scores)

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.122	0.075	0.058	0.042	0.013	0.000
Proportion of Variation	0.394	0.241	0.186	0.136	0.041	0.001
Cumulative Proportion of Variation	0.394	0.635	0.821	0.957	0.999	1.000

As far as future likelihood scores are concerned, again the first three principal components, as seen in Table 3.35, explain over 80% of the variability and there is more of a difference between them in terms of proportion.

Table 3.36: Principal Component Coefficients (Future Likelihood Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	0.220	0.425	0.023	0.035	0.151	-0.864
PC2	0.718	-0.678	0.052	0.025	0.024	-0.144
PC3	0.458	0.403	-0.209	-0.063	-0.741	0.175

As with the current likelihood scores, the same four disease groups appear again, Table 3.36, indicating little variability is contributed by the two swine fevers. The first component is the zoonotic arbovirus group contrasted against influenza and non-zoonotic arboviruses, with the second then being a contrast between the non-zoonotic arboviruses and influenza meaning between the two we have a lot of the variability dependant on the interplay between these three groups. The third principal component is then foot-and-mouth disease contrasted against influenza and non-zoonotic arboviruses.

Transformation

The data were then transformed according to the method detailed earlier and scatter plots created of each score combination both before and after, Figures 3.12 and 3.13:

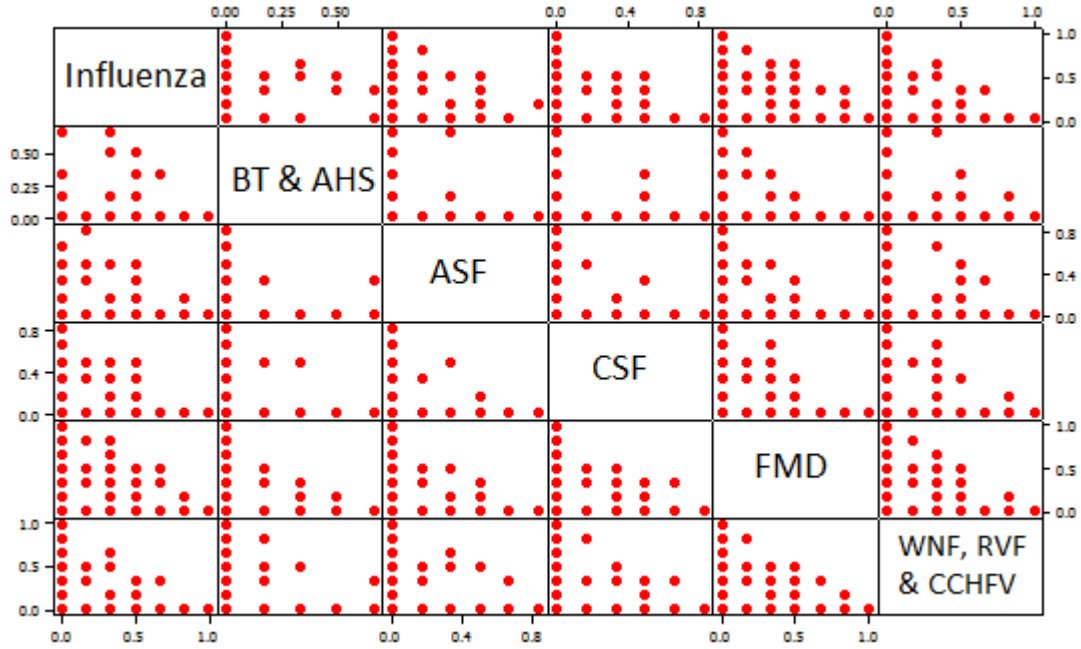

Figure 3.12: Scatter plots of Impact Scores for each combination of disease groups

This series of plots (Figure 3.12) shows the impact scores (Equation (2.8)) for each pair of disease groups plotted against each other with each cell representing a disease combination. So for example the top row of plots will be the impact score for influenza (with the influenza scores on the y axis) against each of the other disease groups (the relevant impact score plotted against the x axis), the second row will be BT & AHS against each of the others. While not as visible as the example earlier (Figure 1.2) it can be seen in Figure 3.12 that the values are restricted to the lower left of the plot, towards the origin; i.e. there is a tendency towards a trade off between variables; e.g. the high values for the disease on the x-axis have low values on the y-axis.

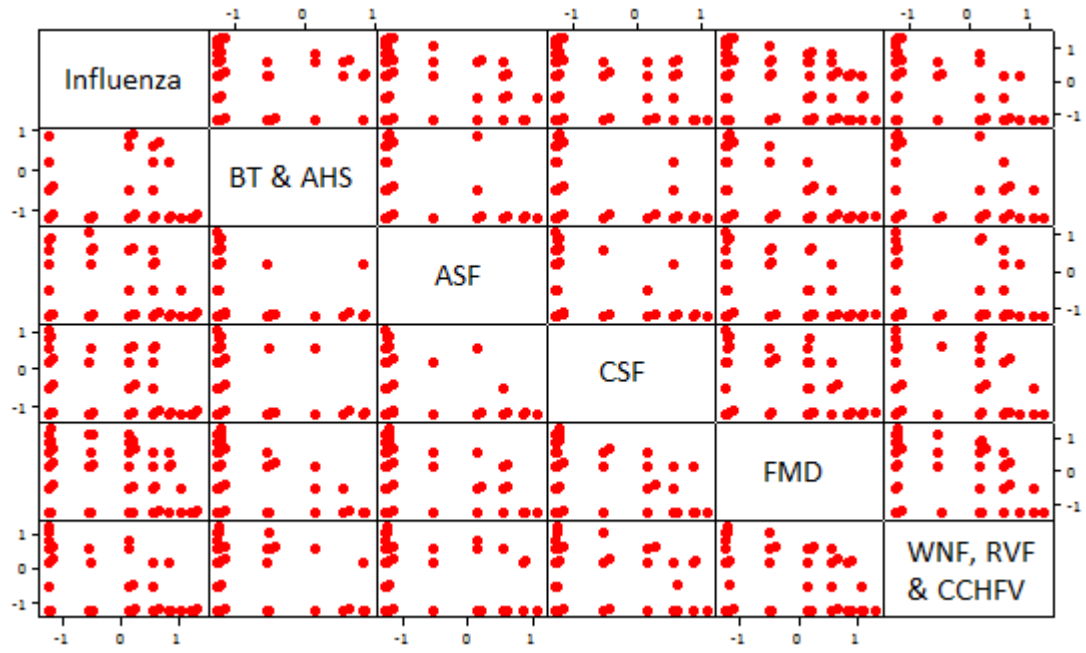

Figure 3.13: Scatter plot of Transformed Impact Scores

In Figure 3.13 the impact scores for each pair of disease groups has been plotted but the data have been transformed and the plots have clearly changed with the data points being much more dispersed i.e. the points no longer seem to be confined to the lower left section of the plot.

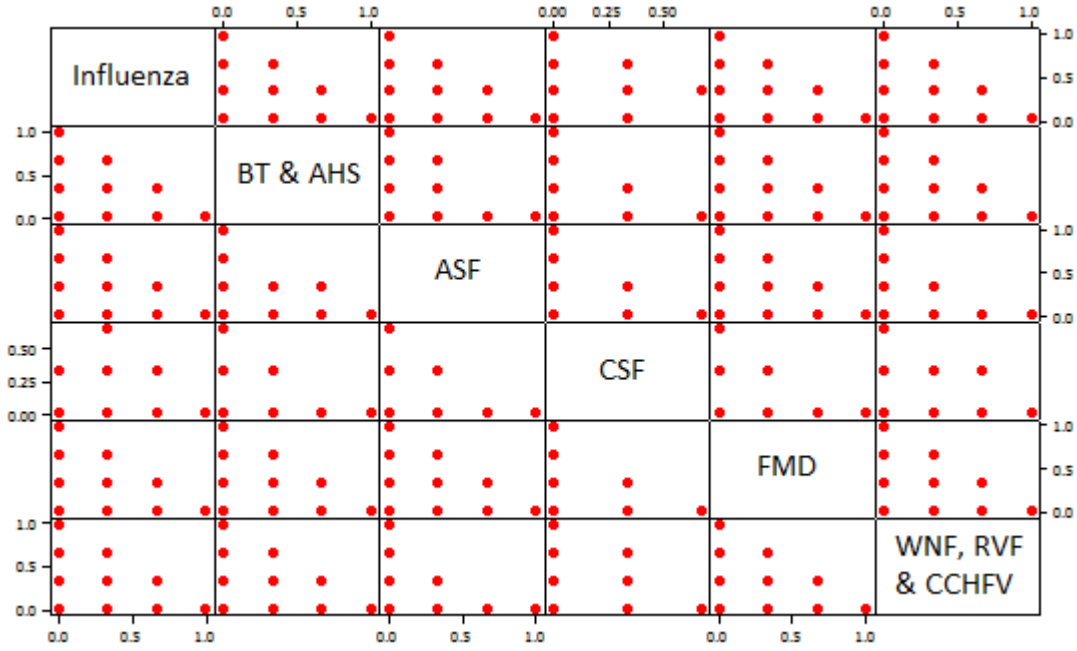

Figure 3.14: Scatter plot of Likelihood Scores

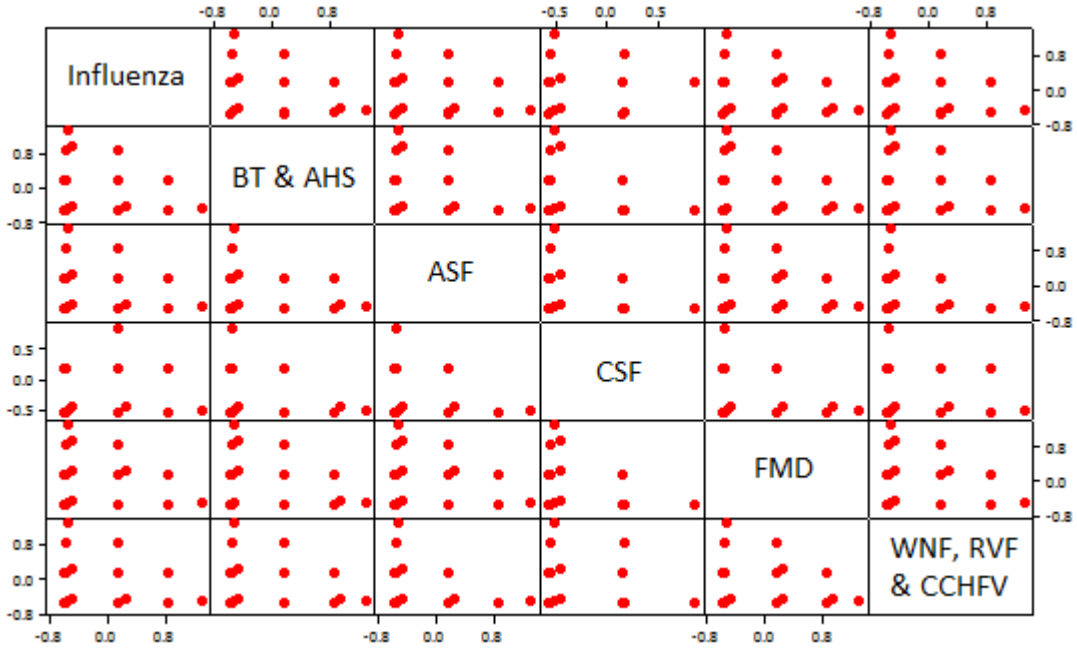

Figure 3.15: Scatter plot of Transformed Likelihood Scores

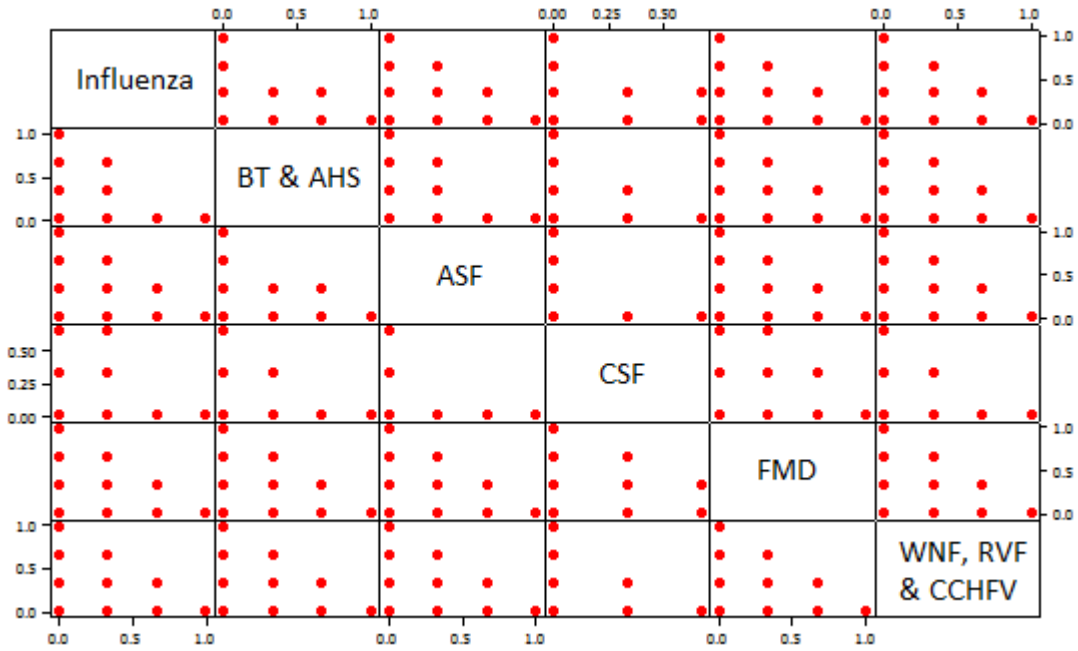

Figure 3.16: Scatter plot of Future Likelihood Scores

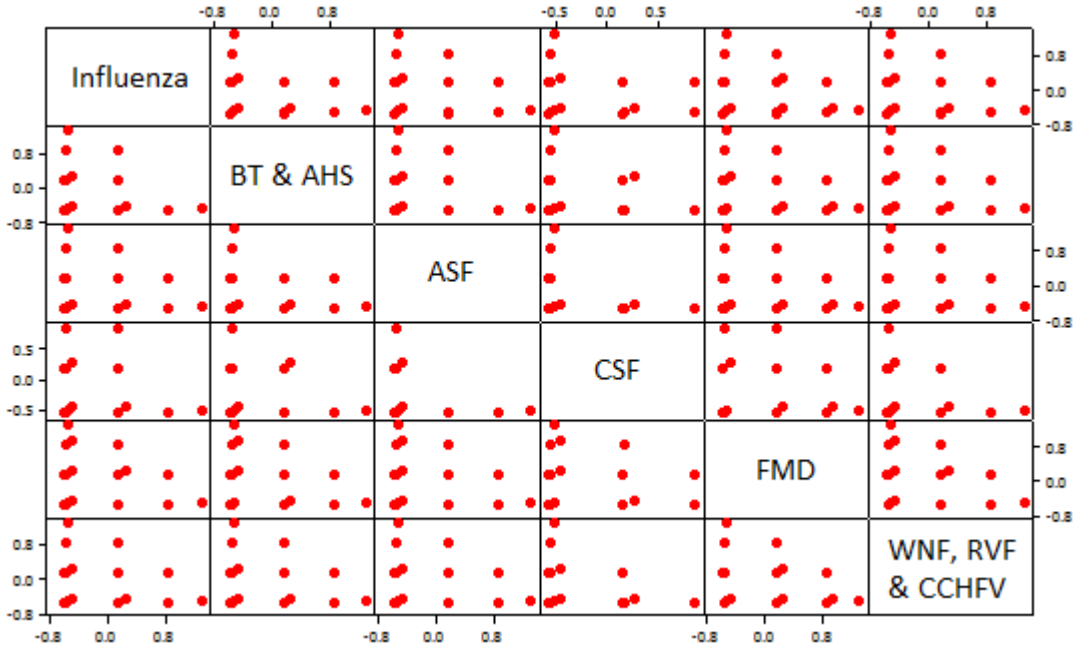

Figure 3.17: Scatter plot of Transformed Future Likelihood Scores

For our likelihood and future likelihood scores, Figures 3.14 and 3.16, again the trade

off type relationship can clearly be seen. However, once transformed, Figures 3.15 and 3.17, it is not clear that this relationship has been removed, a number of the scatter plots continue to display the same kind of trade off behaviour, that is, the values are all in the lower left corner of the plots and suggest an indirectly proportional relationship between variables.

To check for the curved shape described in Aitchison (1982a) is more difficult; the example shown earlier used a triangular plot, (Figure 1.1), but was conveniently on a dataset of only three proportions. The delegates scored six different disease groups, so a similar plot cannot be produced to display the interaction of all proportions. Taking three scores at a time, such a plot can be produced, though to check all combinations would require sixty such plots; for example the impact scores for influenza, foot-and-mouth disease and the zoonotic arbovirus group, Figure 3.18.

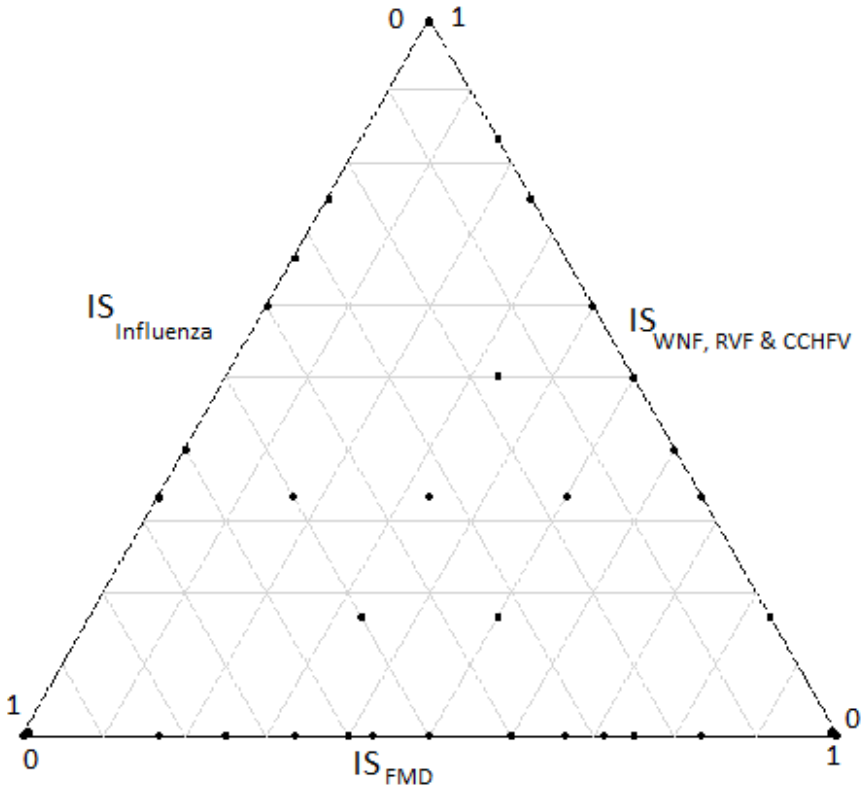

Figure 3.18: Ternary plot Impact Scores for Influenza, Foot-and-mouth disease and WNF, RVF & CCHFV

This plot shows the value of delegates' impact score for influenza on the left axis, their impact score for foot-and-mouth disease on the bottom axis and their impact score for WNF, RVF & CCHFV on the right axis. For example, the point at the very top of the plot represents a delegate that scored WNF, RVF & CCHFV for all impact measures so has an impact score of 1 for that disease group and of 0 for the rest. A curved shape like that expected is not visible but neither is the preferred elliptical shape (Aitchison, 1982a) and much of this difference in behaviour is due to the nature of the scores. Each delegate, and so each row of the dataset, will have at least three zeros present (which is why many of the points lie along an axis in Figure 3.18) and while the individual scores may be regarded as quasi-continuous, the argument for treating the impact and especially the likelihood and future likelihood scores the same way is much weaker. This may indicate that even if it is acceptable to analyse these data using principal component analysis then, for this particular example, the common type of transformation may not be suitable due to the large number of zero values. Nonetheless the analysis was carried out and the results found for the impact, current likelihood and future likelihood scores.

Table 3.37: Eigenanalysis of Covariance Matrix (Transformed Impact Scores)

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	1.183	0.806	0.373	0.330	0.312	0.037
Proportion of Variation	0.389	0.265	0.123	0.109	0.103	0.012
Cumulative Proportion of Variation	0.389	0.654	0.777	0.885	0.988	1.000

Table 3.38: Eigenanalysis of Covariance Matrix (Transformed Current Likelihood Scores)

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.384	0.329	0.248	0.187	0.062	0.000
Proportion of Variation	0.317	0.272	0.205	0.155	0.051	0.000
Cumulative Proportion of Variation	0.317	0.589	0.794	0.949	1.000	1.000

Table 3.39: Eigenanalysis of Covariance Matrix (Transformed Future Likelihood Scores)

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	0.493	0.310	0.248	0.185	0.061	0.000
Proportion of variation	0.380	0.239	0.192	0.143	0.047	0.000
Cumulative Proportion of variation	0.380	0.619	0.810	0.953	1.000	1.000

As far as explaining variation, it can be seen that generally the transformation results in more of a spread across the principal components in comparison to the untransformed results (Table 3.31 compared to 3.37, Table 3.33 compared to 3.38 and Table 3.35 compared to 3.39); i.e. the first two principal components now explain less and the other components now tend to explain slightly more of the dataset variation.

Table 3.40: Principal Component Coefficients (Transformed Impact Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	0.743	0.025	-0.091	-0.094	0.008	-0.656
PC2	-0.152	-0.224	-0.201	-0.161	0.920	-0.118
PC3	-0.111	-0.484	0.829	-0.024	0.007	-0.256

For our impact scores, the interpretation of the first two principal components (Table 3.40) remains much the same as their untransformed counterparts; however, the third principal component has changed quite radically. It is dominated by ASF contrasted against the two arbovirus groups and so is much simpler than it was before now only really involving three disease groups. However, the fact that it is heavily dependant on ASF makes the result rather questionable, since this was generally a very low scored or unscored disease it would not be a likely candidate to contribute much to variation.

Table 3.41: Principal Component Coefficients (Transformed Current Likelihood Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	-0.310	-0.616	0.236	-0.027	0.683	0.032
PC2	-0.657	0.651	-0.034	-0.061	0.308	-0.210
PC3	0.463	0.062	0.049	-0.024	0.287	-0.835

The first principal component for the transformed current likelihood score, Table 3.41, is similar to the original though there is a bit more of a contribution from influenza and ASF. The second component remains effectively unchanged; although the signs of the coefficients have switched for the third component, it has not in fact changed, as the magnitude of the coefficients are much the same and the contrasts (the diseases that have a negative sign as opposed to those diseases with a positive sign) are the same as before.

Table 3.42: Principal Component Coefficients (Transformed Future Likelihood Scores)

Variable	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV
PC1	0.202	0.464	0.005	0.035	0.146	-0.849
PC2	-0.722	0.663	-0.055	-0.032	-0.038	0.183
PC3	-0.440	-0.372	0.152	0.056	0.783	-0.170

All three principal components for the transformed future likelihood scores in Table 3.42 are reasonably similar to the untransformed original, there has been some switching round of signs and there is some slight difference in coefficients for the third principal component with the non-zoonotic arbovirus group and ASF now mattering less than they did before.

Overall, despite there being some change in some of the principal components the fact that the curved shape remains after transformation and many of the components being relatively unchanged suggests that this method of transformation is either unnecessary or ineffectual and some alternative method should be found.

3.4.3 Cluster Analysis

The first requirement is to determine the number of clusters that should be produced. This can be determined by examining the within cluster sum of squares for different number of clusters.

Table 3.43: Within cluster sum of squares for current scores

Number of clusters	Sum of squares	% Change
7	7.93	na
6	9.60	21.03
5	10.05	4.66
4	12.23	21.67

Table 3.43 shows a rapid drop in percentage change of within cluster sum of squares when moving from six clusters to five. In contrast, for the future scores (Table 3.44) the drop is less extreme and coincides with a move from seven to six clusters.

Table 3.44: Within cluster sum of squares for future scores

Number of clusters	Sum of squares	% Change
8	8.03	na
7	8.64	7.66
6	9.03	4.54
5	10.22	13.08
4	13.07	27.96

The coefficients of each cluster centroid describes the relationships between scores

for each disease group.

Table 3.45: Cluster centroids for current scores

Cluster	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV	% of delegates
1	0.0327	0.0076	0.0038	0	0.2298	0.0099	38.42
2	0.125	0	0	0.1944	0.0278	0	2.11
3	0.1736	0.0056	0.016	0.0007	0.0188	0	42.11
4	0.0214	0.3162	0	0	0.0171	0	6.84
5	0.0111	0.0028	0.0778	0	0.0306	0.2167	10.52

As far as current threats are perceived, there are five distinct groups. Of these five, the majority of delegates are contained within two with almost all delegates contained within clusters one and three. Cluster three has a centroid mostly determined by the score of influenza with small contributions from some of the other disease groups. Cluster one's centroid is defined by foot-and-mouth disease with contributions from other groups. The third most important cluster, by percentage of delegates, is mostly dependent on the zoonotic arbovirus group but contains far less of the delegates than the top two.

Table 3.46: Cluster centroids for future scores

Cluster	Influenza	BT & AHS	ASF	CSF	FMD	WNF, RVF & CCHFV	% of delegates
1	0.0069	0.0046	0.0035	0	0.2384	0.044	25.26
2	0.02	0.0044	0.0067	0	0.0489	0.2978	13.16
3	0.0106	0.0062	0.0088	0.0194	0.0247	0.0062	33.16
4	0.2994	0.0015	0.0062	0	0.0509	0.0108	18.95
5	0	0.3611	0	0	0.0185	0.0093	6.32
6	0	0	0.3148	0	0	0.1019	3.16

Table 3.46 shows that the number of delegates are less focused than for current

scores (Table 3.45). The most important cluster, containing the greatest number of delegates, contains less than the second most important for the current scores. This cluster, cluster three, is dependent on the scoring of foot-and-mouth disease, CSF and influenza. In the second most important cluster (cluster one), there is a strong dependency on foot-and-mouth disease with some contribution from the zoonotic arbovirus group.

The third and fourth most significant clusters are dependent on influenza and the zoonotic arbovirus groups respectively.

3.5 Discussion

Principal component analysis and cluster analysis were used here to examine the scoring behaviour of delegates from a multivariate perspective. This was considered important as each disease group was not scored independently of the others and so how a delegate scored one group had a direct influence on the others.

Despite this, many of the same patterns were found as for the univariate approaches. Evidence of regional bias and a delegate's bias towards a disease they work on continued to be found.

Of greater interest is the repeated pattern of less certainty in future risk as evidenced by the increased number of principal components (Figure 3.1) and clusters (Table 3.43 contrasted against Table 3.44) required to explain the data. Again, this makes an intuitive sense; any opinion about the future should intrinsically be less certain than an opinion about the present, so it would be expected to still be present in the results. This future change also reinforces the increased future importance of the zoonotic arbovirus group that was highlighted in previous studies (Gale et al., 2010) and in Chapter 2. This is shown in the structure of the more significant principal components and clusters for future scores where large coefficients are associated with this disease group as opposed to current score principal components and clusters where it tends not to feature as much. It can also be seen graphically in the

plots of principal components one and two for the past and future which are in the main dependent on the Influenza, Foot-and-mouth disease and zoonotic arbovirus groups. Here, the difference from the current threats, where focus is on Influenza and Foot-and-mouth disease, to a situation where all three seem equally important is graphically illustrated.

In this and the previous chapter, expert opinion was analysed using both multivariate and univariate and group and individual approaches in order to identify what diseases might emerge as a threat and which therefore can be regarded as exotic threats. All of these approaches result in the same disease groups being identified. Both now and in the future, delegates view influenza and foot-and-mouth disease to be a threat but the key emerging threat in coming years is seen to be WNF, RVF & CCHFV. All of these disease are zoonotic arboviruses and are vector borne. The following chapters will focus on quantifying and exploring the risk represented by these diseases.

Chapter 4

Non-spatially explicit mathematical model of the risk of CCHFV incursion via migratory birds

4.1 Introduction

At the 2010 annual EPIZONE meeting, a survey was carried out of over 150 experts on current and emerging disease threats (Kelly et al., 2013). The experts were questioned on their current views and their expectations for 2020 about the likelihood of incursion, spread and persistence of six different disease groups. The data from this survey were analysed in chapters 2 and 3 using both univariate and multivariate techniques and all the evidence suggested that the disease group that was viewed by the experts as the most significant emerging threat was the zoonotic arbovirus group. This group consists of diseases that are zoonotic, that is capable of being transmitted from other animal species to humans, and are vector borne, so transmitted by one or more arthropod species. Three specific diseases of this class were considered; WNV, RVF and CCHFV only one of which has a potential vector present throughout Europe - CCHFV. One of the potential means of introduction of this disease is through *Hyalomma marginatum* ticks brought in upon migrating birds and so there is value in understanding and exploring this in order to better understand the risks.

In this chapter, this risk will be explored. In particular, a geographical information system approach will be used to assess the risk of CCHFV positive ticks being introduced into Europe from Africa (Palomar et al., 2013; Jameson et al., 2012). A large number of bird species migrate annually between sub-Saharan Africa, where CCHFV is present in tick species, and Europe (Kirby et al., 2008; Dorst, 1962) and many of these species have feeding or nesting habits that put them at risk of becoming hosts for ticks, thereby creating a means for the disease to be introduced into Europe.

4.2 Crimean-Congo Haemorrhagic Fever Virus

Found throughout much of Africa, Asia and Europe, CCHFV is a fatal viral infection. It is a member of the Nairovirus genus, of the family Bunyaviridae (Deyde et al., 2006), and gives its name to one of the seven sub-groups this family is often divided into (i.e. the Crimean-Congo hemorrhagic fever serogroup consists of CCHFV and HAZARA virus). All of these viruses are tick borne and can have human or animal hosts (Turell, 2007). CCHFV causes a severe disease in humans with a reported mortality rate of 30% (World Health Organisation, 2012) and is the most geographically widespread of the medically significant tick borne viruses. Figure 4.1 shows the geographic distribution of both the virus and its main vector and illustrates that, so far, the distribution of the virus seems to be curtailed by the geographic distribution of the tick. The grey box at the top of Figure 4.1 is the 50° North latitude line and neither the tick nor virus are established beyond this point (Formenty et al., 2007). However, over the course of the last few years, a number of CCHFV positive ticks have been found in many countries north of this line (Hasle, 2010) and the virus has potentially established itself in some hitherto virus free countries south of the line (Foley-Fisher et al., 2012).

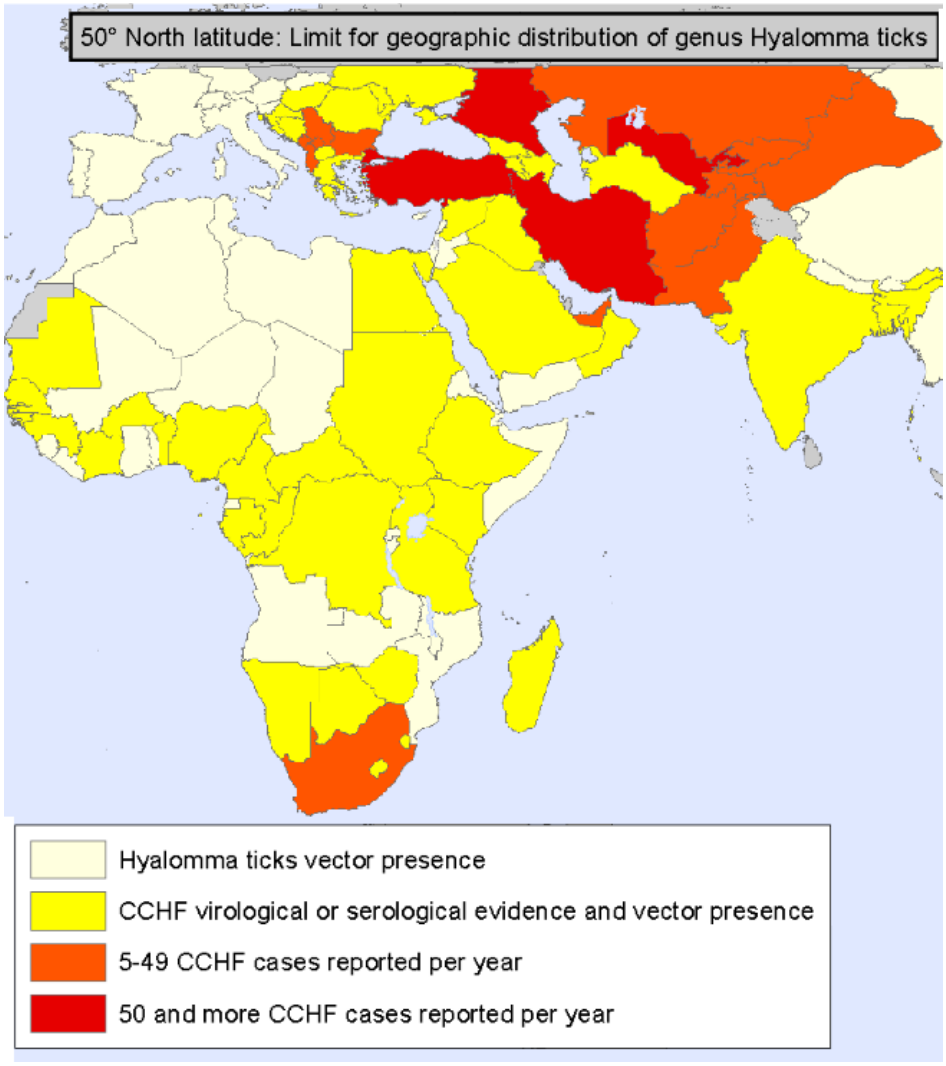

Figure 4.1: Geographic distribution of CCHFV and vector (WHO 2008)

Humans can contract the virus through a number of routes including direct contact with infected blood or tissue; this would include infected meat products or a nosocomial (hospital acquired) infection (Randolph and Rodgers, 2007). However, the most important route is via a tick bite, as not only are ticks a source for human or animal infection but they are also important in the disease lifecycle. The long lifespan of the tick means they can act as both a vector, introducing the virus to other animals, and a reservoir; in other words a long term host.

The continued existence of the disease is strongly linked to the presence of its vector

as can be seen in Figure 4.1. While many countries where this vector is present (marked in cream) do not have reported cases of CCHFV, it is yet to become established in any country where this vector is not present (marked in grey) and all those where there is evidence for its presence or reported cases (marked in yellow or red respectively) it is associated with this vector. It relies on a number of stages of transmission within tick populations and Figure 4.2 shows the transmission of the disease within the tick lifecycle:

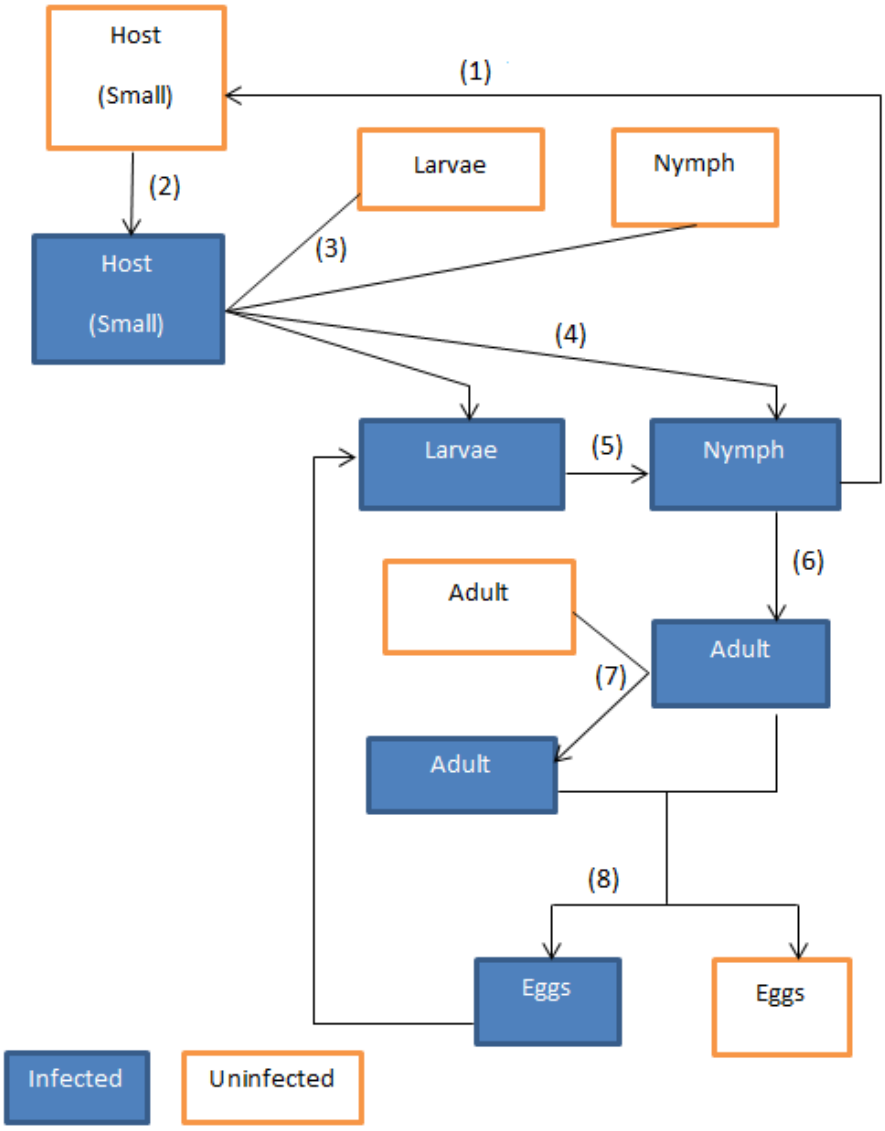

Figure 4.2: CCHFV transmission within the tick lifecycle

Following through Figure 4.2 in order, at step (1) we have an infected larvae or nymph feeding on an uninfected small host (generally any small vertebrate species; e.g. rodents or ground feeding birds) and so transmitting the infection (2). At step (3) uninfected ticks feed on the infected host and contract the virus (4). At stages (5) and (6) transtadial transmission is considered, that is, the tick carries the virus between lifecycle stages, so an infected larvae becomes an infected nymph and an infected nymph becomes an infected adult. In stage (7) an infected tick reproduces with an uninfected tick and this tick contracts the virus. This is known as venereal transmission. In addition, at stage (7), adult tick to adult tick infection can occur through a phenomenon known as co-feeding where the ticks never come into contact but the virus is still transmitted from one to the other through the bloodstream of a shared host. Stage (8) involves vertical transmission whereby the infected female lays her eggs some of which carry the virus and so hatch into infected larvae (Turell, 2007).

Birds are not as important as other small animals in the lifecycle of the disease (Turell, 2007). They are hosts of ticks so are important for the tick lifecycle but the majority of birds are refractory for CCHFV. This means they can never be viraemic and so generally cannot play a role in transmitting the virus. The one exception to this is co-feeding where an uninfected tick feeding on the same host as an infected tick can acquire the virus without the host ever becoming infected (Turell, 2007). Despite not being important as part of the disease lifecycle, birds are potentially very important in the spread of the disease as they are capable of, and in fact regularly, migrate long distances and so can carry infected ticks into regions where CCHFV is not present and so introduce it.

4.3 Selection of Bird Species

There are over 300 species of birds listed for Europe, some of which are resident throughout the year and some of which migrate between Europe, to breed, and warmer locations, to winter. Therefore, it is necessary to have some criteria by which to identify species to focus on as a means of introduction of infected ticks. Firstly, we are interested in species that move between Sub-Saharan Africa and Eu-

rope, i.e. between an area where the virus and vector are present and our area of interest. Each year, a large number of bird species do this, wintering in Africa where the temperature is more comfortable and food is more readily available, and migrating North to Europe in the spring in order to return to breeding grounds. Secondly, we are interested in species whose behaviour puts them at risk of coming into contact with ticks such as those species who spend time on the ground or in low undergrowth (Hoogstraal, 1979; Gale et al., 2011). For example the Chaffinch (*Fringilla coelebs*) is likely the most populous species in Europe (Papazoglou et al., 2004); however, it is not truly migratory and where limited migration does take place, it does not move overly far South and certainly not into Sub-Saharan Africa. Similarly, the Barn Swallow (*Hirundo rustica*), which does migrate from Europe to Sub-Saharan Africa, but feeds and nests above the ground, would not be considered further. Examining the hundred most populous European Species (Papazoglou et al., 2004) and eliminating those that fail to meet these criteria results in the species in Table 4.1.

Table 4.1: Bird species that meet the criteria for possible means of introduction of infected ticks from Africa to Europe with population data from (Papazoglou et al., 2004)

Species	Behaviour	Population Size (Breeding Pairs)
Willow Warbler	Ground Feeding/Nesting	49000000
Tree Pipit	Ground Feeding/Nesting	16000000
Garden Warbler	Low Feeding/Nesting	13000000
Common Whitethroat	Ground Feeding/Low Nesting	10000000
Turtle Dove	Ground Feeding/Low Nesting	2600000
Sedge Warbler	Low Feeding/Nesting (damp)	2500000
Marsh Warbler	Ground Feeding/Low Nesting (damp)	2500000
Northern Wheatear	Ground Feeding/Nesting	1700000
Common Quail	Ground Feeding/Nesting	1300000
Hoopoe	Ground Feeding/High Nesting	980000
Thrush Nightingale	Ground Feeding/Low Nesting	860000
Ortolan Bunting	Ground Feeding/Nesting	700000
Grasshopper Warbler	Low Feeding/Nesting (damp)	670000
Iberian Chiffchaff	Low Feeding/Nesting	530000

From Table 4.1, it can be seen that the Willow Warbler (*Phylloscopus trochilus*) is the most common true migratory bird in Europe, showing obligate migration (Newton, 2011; Dorst, 1962), (it is outnumbered by the Chaffinch, House Sparrow, Common blackbird and European Robin, but none of these have true migratory behaviour) and could thus be considered an important candidate for modelling purposes. Much of it's behaviour further suggests it as the most interesting candidate. Firstly, it winters in sub-Saharan Africa (as do all species in Table 4.1) in areas where *H. marginatum*, the primary CCHFV vector is present. Secondly, it is a ground nesting and ground feeding bird meaning it is more likely to come into contact with the vector, as opposed to other migratory bird species whose nesting and feeding habits mean they are less likely to come into contact with ticks.

To consider just one species of bird would, however, be unrealistic and it would be a mistake to ignore species that may be less populous but might turn out to be more important for the transportation of ticks e.g they are wider ranging or migrate at a faster rate, meaning that ticks could be carried further. Returning to Table 4.1 and using a more strict set of criteria to examine the Behaviour column so that only species whose behaviour maximises their risk of tick contact are accepted, that is, those that both feed and nest on the ground, we are left with five species to be modeled in this chapter as shown in Table 4.2:

Table 4.2: Bird Species most likely to be a means of introduction of infected ticks from Africa to Europe

Species	Population
Willow Warbler	49000000
Tree Pipit	16000000
Northern Wheatear	1700000
Common Quail	1300000
Ortolan Bunting	700000

4.4 Avian Migration

4.4.1 Population Distribution

Obtaining accurate population data for birds can be difficult and a number of surveying methods are used by ornithologists to obtain estimates. These include observatories collecting data on bird passage, normally along the main migratory routes, either by visual or radar observation. Alternatively, more time consuming but detailed approaches such as spot mapping, a labour intensive method where a bird's territory is mapped out allowing a guide to bird densities to be estimated based on this and the overall area e.g. if a bird's territory is approximately $10m^2$ then it can

be estimated that ten birds may be found in an area of $100m^2$. Similarly, a transect count where an observer moves along a fixed path and counts occurrences (in this case breeding birds) and distance from the path and uses this to arrive at an estimated distribution of birds within an area and so an overall population.

All these methods are labour intensive and involve mainly human measurement and, thus, there tends to be a high level of variation in population estimates. For this study, population data were taken from Birdlife International, a global partnership of conservation organisations who operate in over 100 countries and have made available estimates of breeding pairs for all common European Species for each European country based on 2004 estimates. These data are given as an estimated range of breeding pairs for each country which is then tripled to give an estimate of the true number of birds. This was undertaken as estimating breeding pairs neglects birds that will be unable to breed that year and so underestimates the true population. Such birds would be too old, too young or unable to find a mate. Some papers (Fuller, 2009) would suggest quadrupling numbers of breeding pairs, but since the breeding pair estimates are taken from Birdlife International and their approach, shown in any of their species factsheets, is to triple this number, then that is what was done here.

For many of these countries, bird populations can be present in only a small part of the country's territory. Birdlife International maintains spatial plots of the presence and absence of a species with a different polygon for the wintering, breeding and resident populations.

A plot of the European breeding distribution of the Willow Warbler based on data from Birdlife International can be seen in Figure 4.5. Examining a country like Romania demonstrates the importance of not simply uniformly distributing a population across the territory, as in the case of the Willow Warbler population in Romania they are present in only a small part of the country.

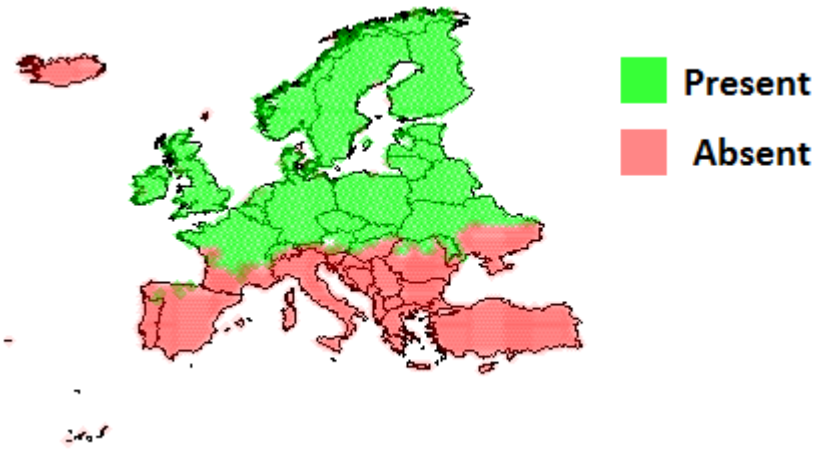

Figure 4.3: European breeding distribution of the Willow Warbler

4.4.2 Arrival Dates

Like many migratory birds, the behaviour of the Willow Warbler when returning to its nesting grounds tends to have a level of predictability, with the dates of arrival across Europe on average being constant from year to year (Dorst, 1962). These dates are gathered each year by observatories and professional or amateur ornithologists and the average arrival date for locations across Europe is calculated. This is the average day on which that species tends to arrive in that region. Lines are then produced by joining locations which have the same average arrival dates for returning Willow Warblers. For example the South of France, Southern Italy and the West of Turkey all have the Willow Warbler normally first arriving in these regions about the 15th of March. These lines are known as isochronal lines. Figure 4.4, taken from Dorst (1962), shows the isochronal lines for the Willow Warbler.

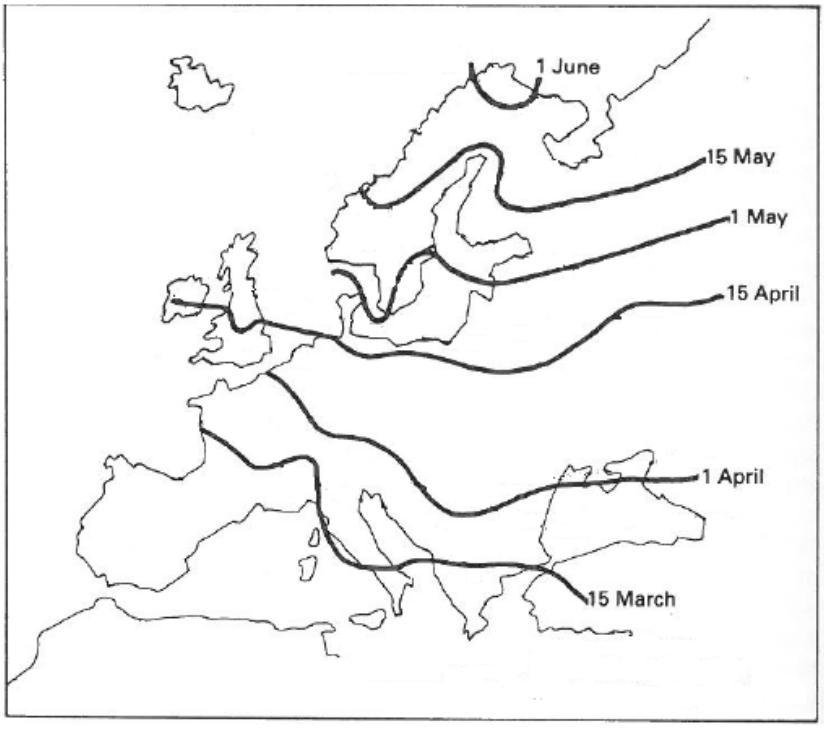

Figure 4.4: Isochronal lines for the Willow Warbler (Dorst, 1962)

Unlike for the Willow Warbler, isochronal lines for the other species in Table 4.2, or indeed for many other bird species, have not been produced by ornithologists. However, there is considered to be a strong level of spatial correlation between those isochronal lines that have been constructed and isothermal lines (Dorst, 1962). Isothermal lines are similar in idea and are formed by joining locations where a particular temperature 'arrives' on the same date. As an example, those for 9° C can be seen in Figure 4.5:

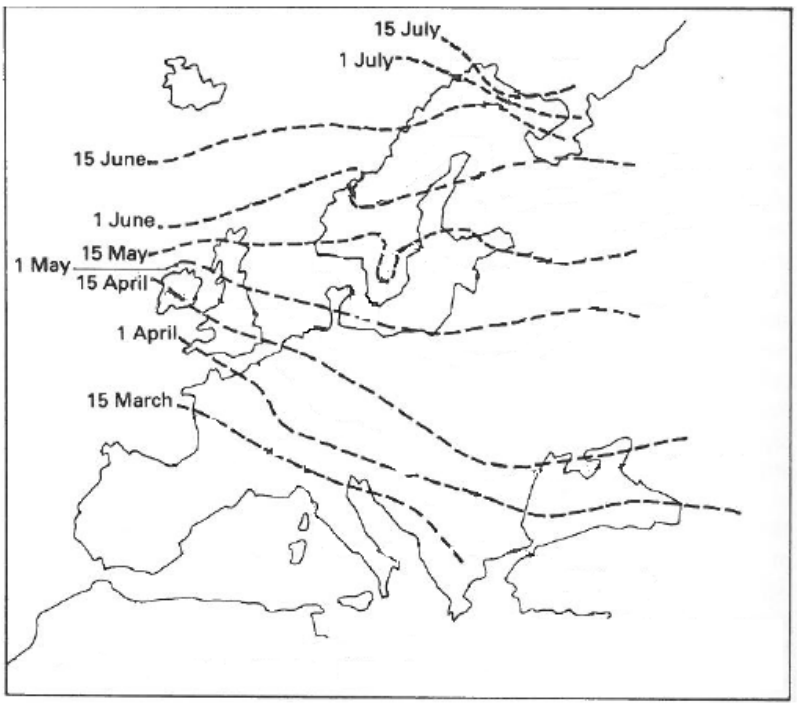

Figure 4.5: Isothermal Lines for 9°C (Dorst, 1962)

The level of correlation between the data for the Willow Warbler (Figure 4.4) and the 9°C isothermal lines was investigated using the raster spatial package in the statistical programming language R. Taking Figures 4.4 and 4.5, the isochronal lines and isothermal lines were interpolated out to the edges of the plots to form zones. Figure 4.6 shows the extended isochronal lines for the Willow Warbler. Each zone was given a numerical value, starting from zero for the southern most zone and increasing by one for each day's difference based on the arrival dates for each subsequent zone. So, for example, from Figure 4.5, there would be a difference of 14 between the two most northern zones as there is a fortnight difference in their arrival dates. Both plots were rasterised (where a 2D plane is broken up into cells) and any cell that fell within a zone would have the value associated with that zone.

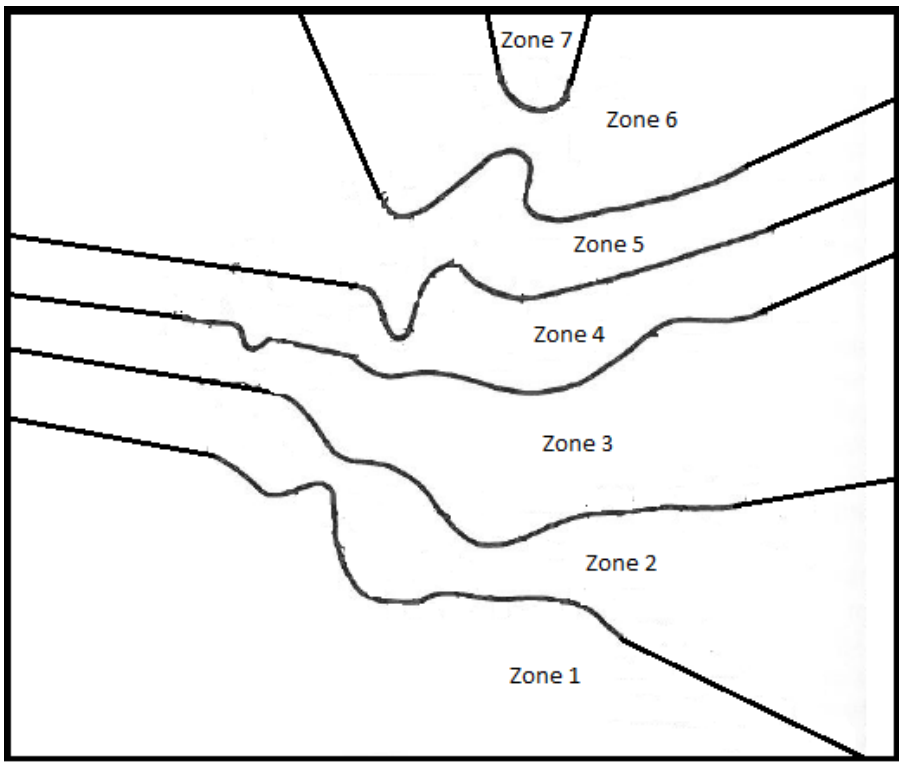

Figure 4.6: Interpolated Isochronal zones for the willow warbler

An equation very similar to the standard Pearson correlation coefficient equation was used to calculate the level of correlation between these two sets of raster data. This is used so as to compute a single correlation coefficient for the two raster layers as opposed to the standard correlation coefficient equation which is designed to assess correlation between two vectors, while the data here consist of two matrices, and so the standard approach would result in a coefficient for each column of the matrix. The equation used is shown in (4.1).

$$r_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\Sigma_{i=1}^{m}\Sigma_{j=1}^{n}(x_{i,j} - \mu_{\boldsymbol{X}})(y_{i,j} - \mu_{\boldsymbol{Y}})}{(mn-1)\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{Y}}} \tag{4.1}$$

$$\boldsymbol{X} = [x_{i.j}]_{m\times n}$$

$$\boldsymbol{Y} = [y_{i.j}]_{m\times n}$$

In equation (4.1) $r_{\boldsymbol{X},\boldsymbol{Y}}$ is the sample correlation and gives a measure of the linear dependence between our two matrices $\mathbf{X}$ and $\mathbf{Y}$ which must be of the same dimension. The $\mathbf{X}$ and $\mathbf{Y}$ matrices contain the rasterised images of the isothermal and isochronal lines with $x_{i,j}$ being the value associated with the zone that cell i,j is within, $\mu_{\boldsymbol{X}}$ and $\mu_{\boldsymbol{Y}}$ being the mean of the elements and $\sigma_{\boldsymbol{X}}$ and $\sigma_{\boldsymbol{Y}}$ their standard deviation. The values of the cells in the matrix will generally be higher for the lower rows as these would be equivalent to being further North and, as we move towards row m the values of the cells will be lower as these represent the more southerly locations.

This resulted in a correlation coefficient of $r_{\boldsymbol{X},\boldsymbol{Y}} = 0.9284$ $(0.9279, 0.9290)$ and so it can be reasonably concluded that the 9° C isothermal lines are a good isochronal surrogate for those bird species for whom isochronal lines have not been constructed.

Whether these arrival dates i.e. the isochronal lines, do indeed indicate anything about the mechanisms of bird migration is a matter of some debate and Dorst (1962) does discuss the varied reasons put forward to explain them. That there is a link between isochronal lines and temperature is suggested by the level of correlation between isochronal and isothermal lines, but this could reflect a difference in departure dates for migrants rather than the mechanism of migration itself. If a species migrates at the same speed regardless of route but their departure date varies depending on the temperature of their final destination i.e. those birds whose nesting grounds reach a required temperature earlier will start their migration earlier, then the same pattern would be observed. Many ornithologists believe that the influence of the movement of temperature is simply a trigger for migration and has no affect on its mechanics with the reason for different arrivals being the availability of temperature dependent insect species that are required as a food source.

This means that even if isochronal lines were available for any of our other species of interest, the reasons discussed would suggest that these should not be taken into account when modelling.

4.4.3 Migration Flight Speed

Bird speeds, or daily migration distances can be measured using a number of different techniques. Handheld instruments such as laser based rangefinders or ornithodolites can be used (Pennycuick, 2001; Pennycuick et al., 2013). The latter instrument combines a rangefinder with an altazimuth mount (a mount designed for rotating an instrument between two perpendicular axes in this case horizontal and vertical) and readings from the mount and rangefinder are fed to a computer. It was designed to be able to allow a single observer to collect a large number of bird flight speed estimates in a short space of time (Pennycuick, 1982).

Alternatively, many studies use ringing data to estimate speeds (Yohannes et al., 2009; Cleere et al., 2000). This is where birds are trapped and rings with a unique identification number are attached to them (often on the leg or wing) before they are released to continue their migration. If the bird is later recaptured, the identification number can be used to link the original capture date and location with this new date and location of capture. The difference in distance between the two locations and the time elapsed between the two capture dates can be used to estimate the average flight speed.

The final method is that of radar (Bruderer, 1997) which has the benefit of being able to gather large amounts of data in a short space of time and allows a larger scale of study. The data from radar studies can illustrate how birds navigate using the surrounding landscape and what environmental changes cause a change in migration. The main weakness of radar is the inability to identify a particular species; in recent years work is still being done on merely correctly identifying radar targets as birds (Zaugg et al., 2008).

Although all of these methods exist, for many of these species the only data available described autumn migration. However, it is a common belief amongst ornithologists that spring migration speeds are generally faster with the limited data available suggesting an average increase in speed of 34% (Bauchinger and Klaassen, 2005), this means that spring migration speeds can be estimated for those species that only have data on autumn migration speed.

4.4.4 Orientation and Navigation

Papers examining avian migration are often interested in how birds navigate and whether stars, magnetic fields or geographical markers are used (Erni et al., 2003) and this is still an area of debate (A. Vrugt et al., 2007). A bird's ability to orient itself based upon whatever means of navigation is a very important parameter as more accurate orientation and navigation will result in a more efficient migration giving a competitive advantage. As discussed in the previous section, radar studies can provide information on the large scale movements of birds which can be examined to investigate how they navigate. Orientation studies are also carried out where some of these means of orientation are examined and manipulated in order to investigate their importance and the level of variation amongst birds orienting themselves by these means (Wiltschko and Wiltschko, 1975). In these studies, birds are often captured and placed in special cages with sensor equipped perches and a record is taken of which perches are triggered by the bird flying to them. These data will indicate the general direction that particular bird is trying to fly and then the means of orientation can either be removed, e.g. covering the cage so the stars cannot be seen, or manipulated e.g. modifying the local magnetic field.

Leaving asides the question of how accurately birds navigate the question of what they orient on is also unsure. For example, a simple clock and compass method (Mouritsen, 1998) where fixed directions are followed for a fixed period of time, but other papers argue against this form of navigation (Thorup and Rabøl, 2001) and suggest geographic markers and a measure of bird orientation and papers such as Wiltschko and Wiltschko (1975) examine more complicated methods involving magnetic fields.

4.4.5 Model Factors

The distribution of birds across Europe will affect both the migration routes taken by birds and potentially the distribution of introduced ticks and so will feature in all the

modelling approaches. Since CCHFV positive ticks being introduced by birds has a time dependency from the length of on host attachment, then migration speed will feature in all models. A bird's orientation requires birds to be modelled in greater detail, so will only feature in the more advanced models used in later chapters.

4.5 Method

4.5.1 Background

In chapter 1 the work of Gale et al. (2011) was reviewed. This paper made use of a geographic information system (GIS) approach to estimate the risk of livestock in Europe becoming infected by CCHFV via CCHFV positive ticks introduced by migratory birds. This approach consists of using multiple layers of geographic and spatial data to analyse a problem. The model in Gale et al. (2011) focused on four bird species and as far as risk of introduction is concerned a single GIS layer was used for each species. These layers consisted of grids of 25km square cells for Europe and contained a binary value indicating the presence or absence of that species within the geographical region represented by that cell.

The population for each bird species for all of Europe was evenly distributed amongst the cells with a present value for that species. This resulted in a new GIS layer which contained 25km squared cells containing species number for each geographical region. So, from two original layers, one containing species population values for parts of Europe by country and another containing species presence/absence data by grid cells, a single layer is produced for each species containing population in grid cells. These layers were multiplied by a prevalence rate for CCHFV positive ticks on migrating birds, which was estimated from a number of papers and represents the average number of CCHFV positive larval or nymphal ticks on birds in sub-Saharan Africa. This gives a number of ticks deposited per cell, all of which were summed to produce a total number of deposited ticks in Europe.

This can be represented mathematically using the notation from the paper:

$$N_{PosTicks}[i,j] = N_{birds}(i,j) \times p_{prev} \times Mean_{tick}$$

$$N_{PosTicks} = \Sigma_{i=1}^{m}\Sigma_{j=1}^{n} N_{birds}(i,j) \times p_{prev} \times Mean_{tick} \tag{4.2}$$

In equation (4.2) the values m and n represent the dimensions of the grid, $N_{birds}[i,j]$ is the total number of birds in grid cell i,j and is derived as described above. p_{prev} is the proportion of immature *H. marginatum* nymphs which are CCHFV positive and is assumed after reviewing multiple papers (see 4.2), $Mean_{tick}$ is the average number of *H. marginatum* nymphs per migrant bird and is derived from information in Molin et al. (2011).

As discussed in chapter 1, this takes no account of the distance of the cell from sub-Saharan Africa where the birds start their migration, so two countries with similar breeding populations of bird species will have the same estimated number of deposited ticks regardless of the country's geographical position and the resulting migration distance from sub-Saharan Africa.

4.5.2 Geographic Information System

In this chapter, a similar approach to Gale et al. (2011) will be used. Their model will be extended in an attempt to further reflect the spatial element of migration. The increased distance birds have to fly to reach some more northern European countries compared to the Mediterranean countries means that there is a much smaller chance of ticks reaching those countries. This is due to the finite duration of on host feeding time for the Hyalomma tick (Estrada-Peña et al., 2011; Caporale, 2009; Gale et al., 2011; EFSA Panel on Animal and Welfare (AHAW), 2010).

Rather than use a grid approach, as in Gale et al. (2011), the number of deposited ticks was calculated for a series of polygons, in this case the countries of Europe. To incorporate the spatial element, an expression of the distance between the breeding

grounds of a country and the African wintering grounds was needed. Points were sampled in the breeding grounds of a country and across the wintering grounds in sub-Saharan Africa. The Euclidean distance for each pair of points allows a distribution to be formed of the distance a bird might have to fly to reach its nesting ground. Dividing by the speed of the bird species this can be converted into a distribution for the number of days of migration. The proportion of this distribution ($p_{distance_{country,birdspecies}}$) that lies under the maximum on host attachment time for ticks represents the probability that a bird will reach it's breeding ground in that country before any ticks that are feeding on it finish and detach. This value, when multiplied by the number of birds per country ($N_{country,birdspecies}$) and by the prevalence rate of CCHFV positive ticks per bird (average ticks per bird (μ_{tick}) multiplied by CCHFV prevalence in ticks (ρ_{prev})) gives an estimate of the potential number of CCHFV positive tick incursions per European country with the distance flown taken into account.

This can be represented mathematically as in equation (4.3):

$$T_{country,\,bird\ species} = \rho_{prev} \times \mu_{tick} \times N_{country,\,bird\ species} \times p_{distance_{country,\,bird\ species}} \quad (4.3)$$

$$p_{distance_{country,\,bird\ species}} = Pr\left(D_{country,\,bird\ species} \leq a \times v_{bird\ species}\right) \quad (4.4)$$

In equation (4.3) the values ρ_{prev} and μ_{tick} are equivalent to p_{prev} and $Mean_{tick}$ in equation (4.2) and thus came from Gale et al. (2011). However, in contrast to equation (4.2), in equation (4.3) the number of ticks introduced is calculated by country and species and so N varies dependent on this. That is, in equation (4.2) $N_{birds}[i,j]$ represents the sum of the breeding populations for all five species of interest in the 25 km square region i,j. In equation (4.3) $N_{country,\,bird\ species}$ represents the breeding population for a particular species and country. This is required as the main difference between the equations is in the term $p_{distance_{country,\,bird\ species}}$ and since this is dependent on species then N must be too. This is the proportion of a bird species migrating to that country that will do so in less time than tick on host attachment. So $Pr\left(D_{country,\,bird\ species} \leq a \times v_{bird\ species}\right)$ with $D_{country,\,bird\ species}$ being a distribu-

tion describing the time taken for migration, $v_{bird\ species}$ being the migration speed for a particular species and a being the maximum on host attachment time. Thus, the right hand side of the inequality is the number of days migrating multiplied by the distance traveled each day giving the total distance a bird can travel carrying a CCHFV positive tick.

To create distributions for the distance from each bird species' wintering ground to each breeding country of interest in the EU ($p_{distance_{country,\ bird\ species}}$) spatial libraries within the statistical programming language R were used along with two distinct GIS layers. These offered various methods of spatial sampling with defined polygons, in this case an EU country and a list of African countries. The first GIS layer consisted of polygons of the countries of the world, the second layer was one of a set of polygons detailing the distribution of the bird species and was based upon data from Birdlife International (BirdLife International, 2012a). Since the species distribution did not contain political boundaries, the initial points were sampled for an EU country then those that fitted into the bird species distribution map were kept and the rest discarded.

To test which points fitted into the species distribution map, the coordinates from one layer were plotted on another. All GIS layers require a map projection method which is used to convert a 3-dimensional surface into a plane. There are numerous methods based on the initial shape assumed for the earth, the methods of recording positions on its surface and the means by which it is transformed dimensionally and differences between these methods can mean a transformation is often necessary to plot information from one layer onto another. For the two layers used here, there is a slight difference in projection method in that our world map uses GRS80 (geodetic reference system 1980 - geodetic meaning it relates to the measurement of the Earth's surface and GRS being a coordinate system and points of reference for locations on this surface) and the Birdlife layers make use of WGS84 (world geodetic system 1984) which is a slightly refined version of GRS80 (Clarke, 2003). Since the second layer is simply a higher precision version of the first then there is no need for any transformation to ensure compatibility between the two and the points from our country map can be plotted directly onto the species distribution map.

The spatial sampling technique used was based upon a hexagonal grid rather than random sampling so as to ensure equal coverage of all possible departure and destination points rather than risk the points being too heavily distributed in one part of an area. For example, the common quail departs from a number of countries in sub-Saharan Africa, so a set of hexagonal gridded points was sampled (Figure 4.7 blue points). If this had been sampled randomly a greater proportion may have been sampled in e.g. Angola and other more southern countries, thus skewing the distribution of distances. The points can be seen to be based upon a hexagonal lattice taken over the polygons representing the countries in which the common quail spends the winter months. Points were also sampled for a European country, in this case the UK (Figure 4.7 red points).

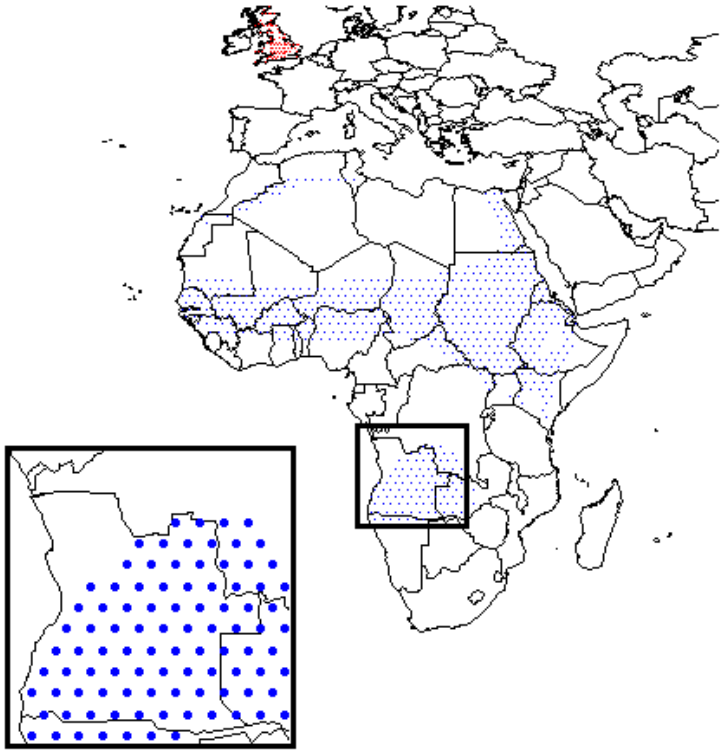

Figure 4.7: Sampled points in wintering grounds and UK

Spatial sampling in R works by taking the defined polygon, in this case often a country, and distributing points within it (Bivand et al., 2008). It starts by get-

ting the bounding box for the polygon or polygons, that is a rectangular space that contains the polygon so for a country or part of a country it will be defined using the most northerly, southerly, easterly and westerly point. Points can then be either distributed in a random or non-random way within this space. The points that fall within the polygon itself will be kept. This often means that the number of points recorded will be less than requested. For example if 100 regularly distributed spatial points were to be found for the UK then only 88 points may be returned. Within the bounding box the points would be distributed on a rectangular grid starting from a preset offset from the upper left corner of the bounding box. Points can also be distributed on a hexagonal grid or randomly whereby a random uniform distribution would be used to select an x and y coordinate within the bounding box. The number of attempts to find an acceptable spatial point within a bounding box, i.e. one that falls within the polygon, can be increased meaning that a number of points closer to what is required will be returned and for awkwardly shaped polygons a result is more likely. This is often useful when there is a need to map all areas of a country and a part of it has a very irregular shape.

The Euclidean distance between each combination of the European country (UK in Figure 4.7) and African points, so between every combination of blue and red point, was calculated and these gave a distribution for the distance between the wintering grounds and that country. These were in terms of latitude and longitude and so were converted back to kilometers and divided by the average daily flight distance for the bird species. The resulting distribution is in terms of number of days of migration and the proportion under 26 days (maximum on-host attachment time (Estrada-Peña et al., 2011; Caporale, 2009; Gale et al., 2011; EFSA Panel on Animal and Welfare (AHAW), 2010)) was calculated. Figure 4.8 shows this distribution for the UK and the common quail.

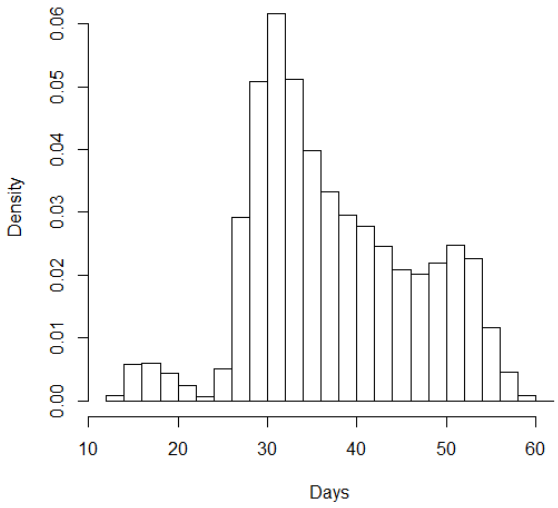

Figure 4.8: Histogram of the number of day's flight from wintering grounds to the UK for the common quail $\left(\frac{D_{UK,\,Quail}}{v_{Quail}}\right)$

As can be seen from Figure 4.8, only a very small proportion of the distribution lies under 26 days. This can be contrasted with Figure 4.9 which shows the equivalent distribution for Italy, a more southern country.

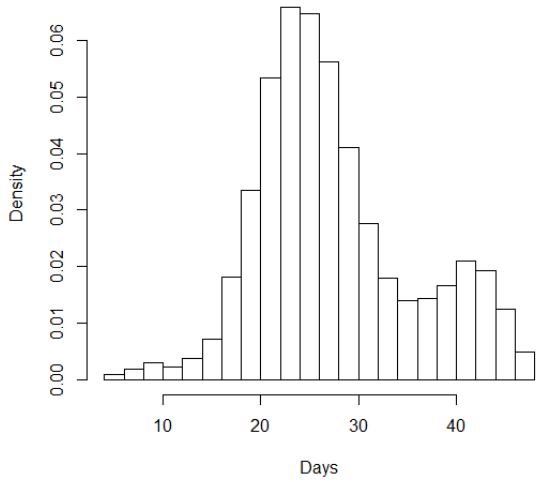

Figure 4.9: Histogram of the number of day's flight from wintering grounds to Italy for the common quail $\left(\frac{D_{Italy,\,Quail}}{v_{Quail}}\right)$

For Italy (Figure 4.9), a much greater proportion of the distribution lies under 26 days meaning a greater number of birds would reach Italy than the UK before their attached ticks finished feeding and detached.

Doing this for each European country gave a series of proportions which could be multiplied by the prevalence rate of CCHFV positive ticks on migrating birds and by the number of birds for that country giving an estimate of the number of infected ticks that will still be attached to their host when they reach their breeding ground.

Unfortunately, the distributions of distance for the majority of countries do not follow or resemble any standard distribution; however, doing the same for all European countries at once (Figure 4.10) gives results that generally resemble the Normal distribution.

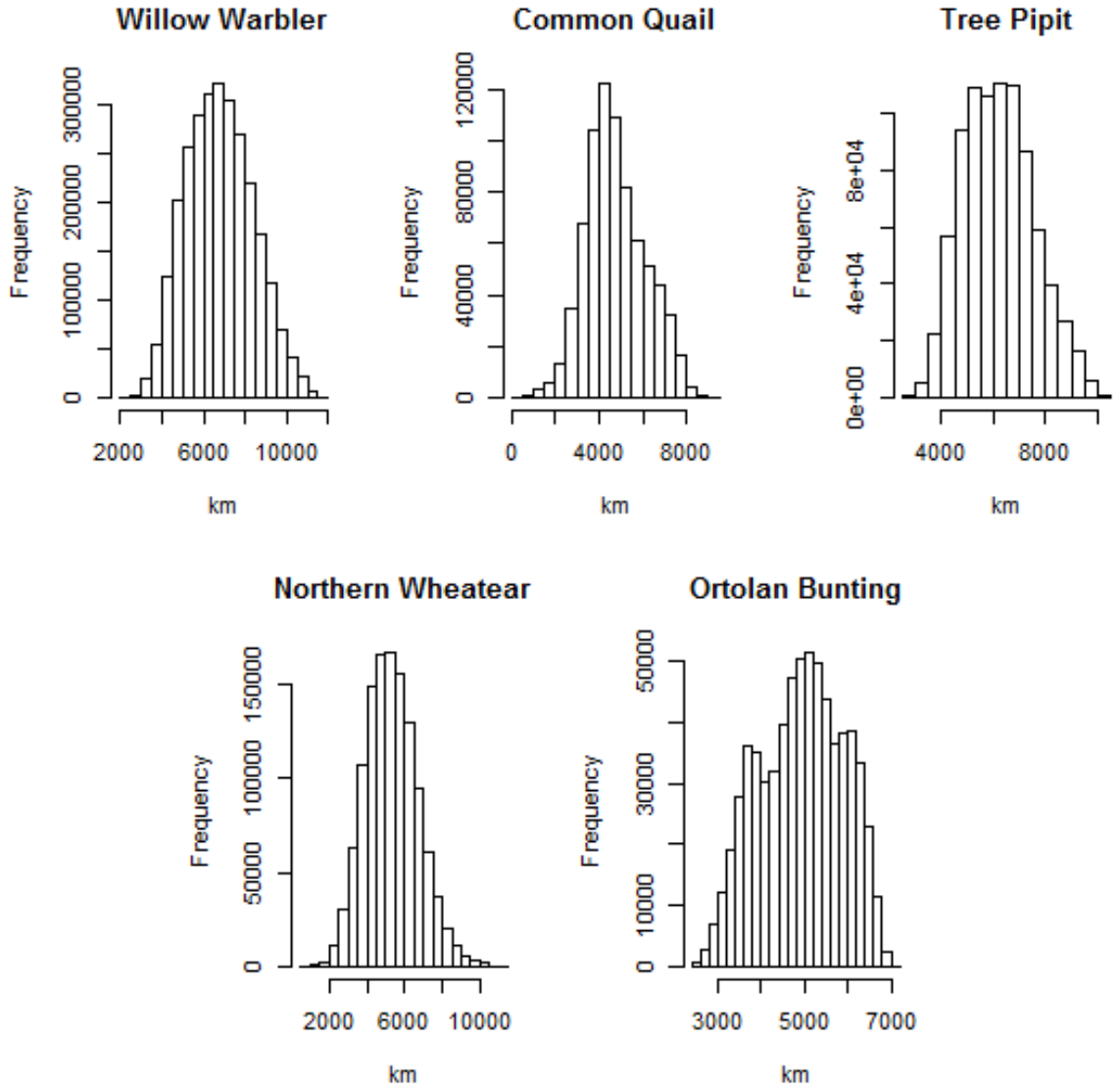

Figure 4.10: Distribution of migration distances for each bird species between wintering grounds and European breeding grounds ($D_{Europe,\, bird\ species}$)

The Normality of these distributions can be judged in part from the histograms in Figure 4.10, where they look approximately Normal, though the distribution for the Ortolan Bunting could be argued to have two peaks. Q-Q plots can be used to compare each of these against a theoretical Normal and gives a better graphical indication of normality (Figure 4.11).

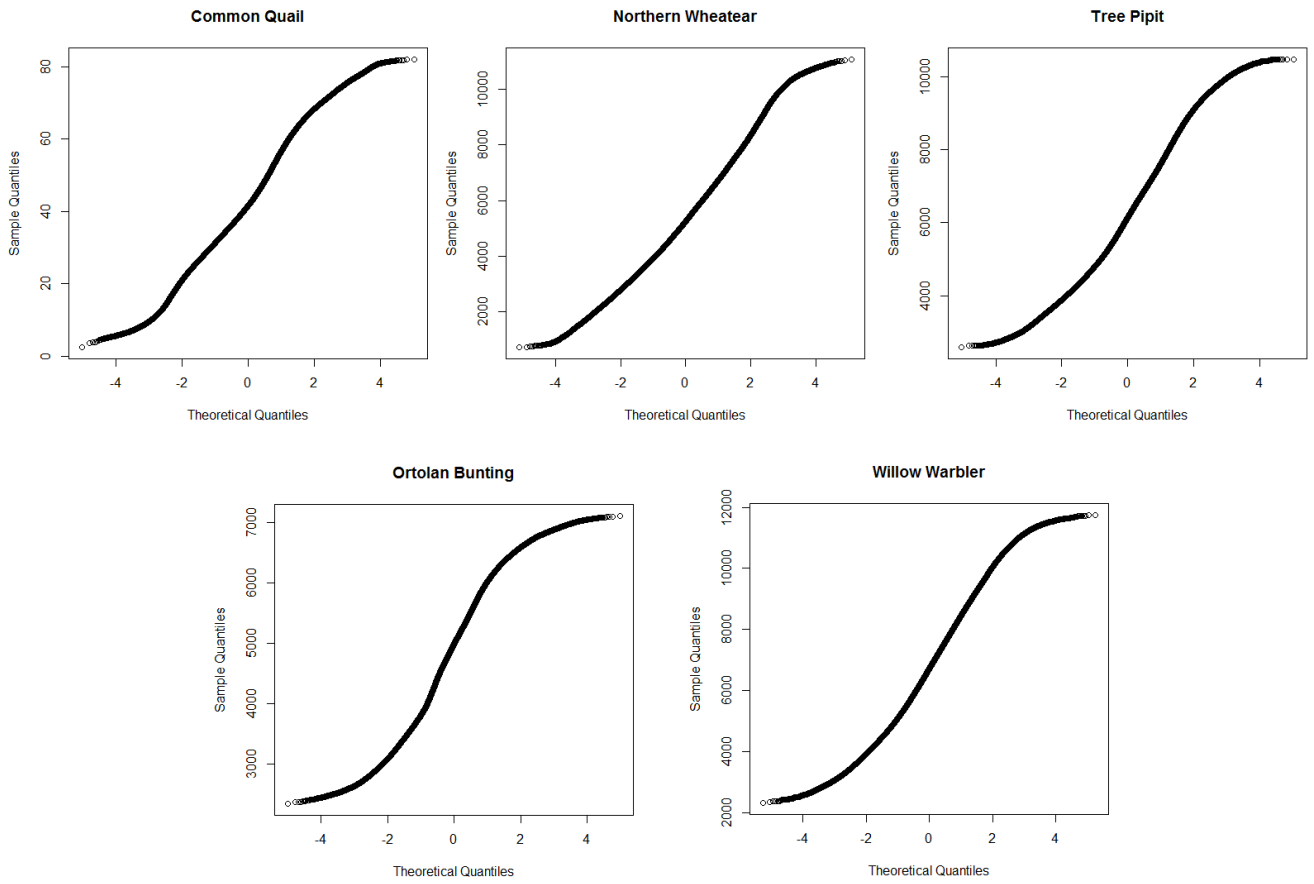

Figure 4.11: QQ-plots for distribution of migration distances for each bird species between wintering grounds and European breeding grounds

None of the plots in Figure 4.11 have the straight line indicative of a perfect fit to a Normal distribution. The plot for all 5 species gives an approximate fit for the central part of the plot with the left side being above and the right side falling below the line. This suggests there are short tails at both ends of the distribution. Intuitively, this might be expected as the Normal distributions are infinite in both directions while the distributions of migration distances must be bound by the minimum and maximum distances between the wintering and breeding areas which are both measurable finite quantities.

Calculating the mean and standard deviation for the distance distributions for Europe allows a Normal distribution to be formed which will approximate the true

distribution. The distributions of bird migration speeds ($v_{bird\ species}$), CCHFV prevalence amongst ticks (ρ_{prev}) and population numbers ($N_{Europe,\ bird\ species}$) are assumed to be uniform between the estimated minimum and maximum and thus there is a set of standard distributions from which sampling can be carried out. The results of this will be a distribution of CCHFV positive tick incursions under all possible parameter values.

Under these same assumptions, since equations can be found for each of these distributions, then analytical solutions for all of Europe will be produced and used for sensitivity analysis to examine the affects of varying parameter values and the contributions towards overall variation. These same distributions will then be used to produce simulated results for Europe.

Since defined distributions cannot be assumed for each country then only simulated results can be produced for these with the number of CCHFV infected ticks being simulated using the distance distribution for each species and Europe country, along with the distributions of migration speed, prevalence and bird populations.

4.5.3 Approaches

There will thus be three separate approaches: two based around Europe as a single entity, and one based around individual countries. The first approach will form analytical solutions for all of Europe using the possible ranges for speed along with the derived distance distributions and the attachment time of ticks to calculate the proportion of birds that can reach Europe. This will be combined with the species populations and prevalence rate to calculate the total number of CCHFV positive ticks introduced. This approach will also be used to perform sensitivity analysis on the parameters.

The second approach will use the single distance distribution for all of Europe and a single species population distribution for all of Europe along with uniform distributions across the ranges of the other parameters to simulate results for the total

number of CCHFV positive ticks introduced into Europe.

The final approach will use distinct distance distributions and species population distributions for each country along with the same distributions for the other parameters to estimate the total number of CCHFV positive ticks introduced to each European country and multiple runs will be used for each to try and cover the full sample space.

Parameter estimates

Parameter values are in Table 4.3 with bird speeds, or daily migration distances, being taken from a number of sources. A summary of the daily migration speeds for a large number of bird species (Payevsky, 2013) was used for estimates of the majority of the species being modeled. However, as discussed, for many of these species the only data available described autumn migration and so an increase in speed of 34% (Bauchinger and Klaassen, 2005) was applied. For species migration speeds, the maximum and minimum estimates were taken with the maximum estimates for those from autumn migrations being increased by 34% to represent the full possible range of migration speeds. This covers the lowest estimated speed with no multiplier to cover the possibility that there is not a change in migration speeds between spring and autumn.

Table 4.3: Estimates of parameter values

Parameter	Description	Value(s)	Simulation Distribution	Reference
v_{Quail}	Average distance (km) covered per day during spring migration by the Common Quail	150-160	U(150,160)	(Perennou, 2011)
$v_{Warbler}$	Average distance (km) covered per day during spring migration by the Willow Warbler	62-114	U(62,114)	(Payevsky, 2013)
$v_{Wheatear}$	Average distance (km) covered per day during spring migration by the Northern Wheatear	110-147	U(110,147)	(Payevsky, 2013)
v_{Pipit}	Average distance (km) covered per day during spring migration by the Tree Pipit	57-106	U(57,106)	(Payevsky, 2013)
$v_{Bunting}$	Average distance (km) covered per day during spring migration by the Ortolan Bunting	181-243	U(181,243)	(Payevsky, 2013)
a	Tick Length of Attachment	26		
μ_{tick}	Mean number of ticks on migrating birds	0.049		(Gale et al., 2011)
ρ_{prev}	Prevalence of CCHFV in *Hyalomma* ticks	0.0001-0.058	U(0.0001,0.058)	(Gale et al., 2011; Palomar et al., 2013)

The length of time of attachment (Table 4.3) is commonly given as being between 12 and 26 days (Gale et al., 2011; Jameson et al., 2012; Caporale, 2009; EFSA Panel on Animal and Welfare (AHAW), 2010). The prevalence of CCHFV infected ticks (ρ_{prev}) is less certain, with papers suggesting a range of values some based on review of tick data in Africa and expert opinion (Gale et al., 2011) and others on data collected from migrating birds in Morocco (Palomar et al., 2013) and at Mediterranean observatories (Lindeborg et al., 2012). The latter paper used a relatively small sample of birds, and so an estimated range of prevalence is based on the figures from the two former papers.

The distributions for D were created as described earlier for each European country and also for Europe as a whole (Table 4.4).

Table 4.4: Parameters of the distributions of migration distances for all Europe for each species ($D_{Europe,\, bird\ species}$)

Species	Mean (km) ($\mu_{bird\ species}$)	Standard Deviation (km) ($\sigma_{bird\ species}$)
Willow Warbler	6772	1593
Common Quail	4763	1356
Tree Pipit	6202	1352
Northern Wheatear	5310	1398
Ortolan Bunting	4928	968

As discussed previously, obtaining accurate population data for birds can be difficult. For this study, population data (Table 4.5) were taken from Birdlife International and consisted of estimates of breeding pairs for all common European Species for each European country based on 2004 estimates. The maximum and minimum estimate was taken for each country and then tripled to give an estimate of the true number of birds. This was undertaken as most of the methods described previously are used to count breeding pairs and so neglect birds that will be unable to breed

that year and Birdlife International suggest tripling to produce estimates of the true population.

Certain more Eastern countries where birds often winter in Asia rather than Africa were removed from this list, e.g. Russia and Georgia. This was done to keep the area being modeled to a smaller size and focus on central Europe. It also means that birds that might winter in Asia will not be included, which is desirable as CCHFV introductions into Europe are believed to be from Africa rather than Asia (Lindeborg et al., 2012) so focus should be on birds that winter there. As a result, the estimated number of birds over all of Europe will be lower.

Table 4.5: Population range estimates for each European country for each species ($N_{country,\ bird\ species}$)

Country	Common Quail	Willow Warbler	Tree Pipit	Northern Wheatear	Ortolan Bunting
Andorra	18-27	0-0	2250-3000	7500-12000	12-30
Albania	2400-3450	0-150	3000-6000	15000-30000	3000-6000
Austria	15000-30000	60000-120000	105000-210000	13500-27000	45-75
Bosnia-Herzegovina	0-0	0-0	0-0	255-267	0-0
Belgium	7200-12150	60000-300000	20700-33000	84-93	0-0
Bulgaria	24000-34500	0-15	30000-54000	60000-180000	75000-150000
Belarus	45000-67500	2850000-3300000	2400000-3600000	150000-210000	7500-12000
Switzerland	4500-5250	12000-24000	150000-210000	60000-90000	300-450
Cyprus	3000-7500	0-0	0-0	0-0	0-0
Czech Republic	15000-22500	1500000-3000000	1500000-3000000	600-1200	300-600
Germany	36000-66000	5100000-8400000	1500000-2640000	21000-39000	16800-21000
Denmark	600-1200	1200000-1800000	30000-150000	3000-6000	0-0
Estonia	30-90	2400000-6000000	1350000-2400000	30000-60000	6000-12000
Spain	960000-1132500	150-750	900000-1200000	978000-1083000	600000-675000
Finland	30-165	21000000-33000000	3900000-5100000	450000-600000	90000-150000
France	300000-900000	4500000-22500000	750000-3000000	45000-135000	30000-120000
Guernsey	0-0	0-0	0-0	0-0	0-0
Gibraltar	0-0	0-0	0-0	0-0	0-0
Greece	6000-10500	30-300	1200-2400	90000-300000	60000-150000
Croatia	30000-37500	150-300	30000-150000	15000-18000	3000-15000
Hungary	210000-246000	162000-345000	390000-705000	84000-171000	30-45
Isle of Man	0-0	0-0	0-0	0-0	0-0
Ireland	0-30	1500000-3000000	0-0	7500-30000	0-0
Iceland	0-0	0-0	0-0	30000-150000	0-0
Italy	15000-37500	0-0	120000-240000	300000-600000	12000-48000
Jersey	0-0	0-0	0-0	0-0	0-0
Liechtenstein	15-39	90-180	210-360	9-15	0-0
Lithuania	3000-4500	1200000-1800000	900000-1500000	15000-30000	600-2400
Luxembourg	30-54	24000-36000	15000-18000	15-30	0-0
Latvia	60-780	1500000-1800000	1500000-2700000	30000-90000	1500-6000
Monaco	0-0	0-0	0-0	0-0	0-0
Moldova	10500-12000	3000-4500	45000-60000	10500-12000	13500-15000
Netherlands	6000-12750	1350000-1650000	105000-135000	1800-2400	0-0
Macedonia	6000-7500	0-0	6000-9000	30000-90000	9000-30000
Malta	3-6	0-0	0-0	0-0	0-0
Montenegro	0-0	0-0	0-0	0-0	0-15
Norway	150-525	6000000-30000000	3000000-6000000	1500000-3000000	450-465
Poland	300000-375000	3000000-6000000	1500000-2400000	60000-150000	450000-900000
Portugal	15000-82500	0-0	150-300	1500-15000	1500-7500
Romania	480000-570000	180000-255000	1800000-2550000	675000-1029000	375000-765000
Serbia	30000-37500	0-0	105000-150000	33000-48000	13500-19500
Sweden	30-75	30000000-48000000	9000000-21000000	300000-1500000	6000-21000
Slovenia	3000-4500	600-900	60000-90000	900-1500	600-900
Slovakia	6000-12000	1200000-1800000	600000-1200000	18000-27000	0-15
Turkey	900000-1650000	0-0	30000-90000	6000000-24000000	9000000-30000000
Ukraine	300000-390000	1260000-2280000	2550000-3720000	420000-510000	174000-201000
UK	15-684	6375000-6375000	223200-223200	157500-157500	0-0
EU	3733581-5774775	92437020-181792095	34621710-64549260	11613663-34405005	10949637-33328995

4.6 Results

4.6.1 CCHFV Incursions to all of Europe

Analytical Results

Due to the fact that the distributions for D for each species across all European countries (Figure 4.10) are being assumed to be Normal, then analytical rather than simulated solutions can be reached. So $D_{Europe,\, bird\ species}$ will be a Normal distribution:

$$D_{Europe,\, bird\ species} \sim \mathcal{N}(\mu_{bird\ species}, \sigma^2_{bird\ species})$$

and will have a probability density function (Equation 4.5) describing the probability of different distances of migration (d) for individual species and defined by the means ($\mu_{bird\ species}$) and standard deviations ($\sigma_{bird\ species}$) in Table (4.4).

$$f(d) = \frac{1}{\sigma_{bird\ species}\sqrt{2\pi}} e^{\frac{(d-\mu_{bird\ species})^2}{2\sigma^2_{bird\ species}}} \tag{4.5}$$

Referring back to equations (4.4), it is the cumulative distribution function of $D_{Europe,\, bird\ species}$ that is of interest for the random variable $a \times v_{bird\ species}$ with the minimum, mean and maximum value of migratory speed. These were calculated using the built in functions in R. To calculate the number of ticks introduced into Europe by each species, these cumulative distribution functions would be multiplied by the mean, minimum and maximum values (where appropriate) of prevalence (ρ_{prev}), mean ticks per bird (μ_{ticks}) and species population numbers ($N_{Europe,\, bird\ species}$) from Table 4.5. Table 4.6 gives the minimum, average and maximum predicted introduced CCHFV positive ticks.

Table 4.6: Minimum and maximum introduced CCHFV positive ticks by species

Species	Min number of introduced ticks	Average number of introduced ticks	Max number of introduced ticks
Ortolan Bunting	22	22906	87143
Northern Wheatear	2	2604	13689
Common Quail	5	2420	7330
Willow Warbler	0	477	4345
Tree Pipit	0	89	991
Total	29	28496	113498

Sensitivity Analysis

There is a large difference in the maximum and minimum values for each species in Table 4.6, so it is of interest to examine the contribution of each model parameter to this overall variation. Equation (4.3) is multiplicative and so a change in almost any of the parameters will result in a direct proportional change in the value of T; e.g. doubling μ_{tick} would result in T doubling. Differentiating with respect to the parameter would give the rate of change for that parameter. However $p_{distance_{country,\ bird\ species}}$ is a function of parameter values and so its behaviour will be different. So, for example, a twofold increase in migration speed ($v_{bird\ species}$) will not result in T doubling.

The Normal distributions for $D_{Europe,\ bird\ species}$ can be examined but with d defined by $av_{bird\ species}$. The resulting cumulative distribution has the form:

$$F(av_{bird\ species}) = \frac{1}{2}\left[1 + erf\left(\frac{av_{bird\ species} - \mu_{bird\ species}}{\sqrt{2\sigma^2_{bird\ species}}}\right)\right] \tag{4.6}$$

$$erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt \tag{4.7}$$

In equation (4.6) erf is the error function that has the standard form shown in equation (4.7) and substituting this into equation (4.6) results in equation (4.8).

$$F(av_{bird\ species}) = \frac{1}{2}\left[1 + \frac{2}{\sqrt{\pi}}\int_0^{\frac{av_{bird\ species}-\mu_{bird\ species}}{\sqrt{2\sigma^2_{bird\ species}}}} e^{-t^2}dt\right] \tag{4.8}$$

However, the integral present in the error function cannot be evaluated directly and so its Taylor series expansion (equation (4.9)) is used instead to give an approximation.

$$erf(z) = \frac{2}{\sqrt{\pi}}\sum_{n=0}^{\infty}\frac{(-1)^n z^{2n+1}}{n!(2n+1)} = \frac{2}{\sqrt{\pi}}\left(z - \frac{z^3}{3} + \frac{z^5}{10} - \frac{z^7}{42} + \frac{z^9}{216} - ...\right) \tag{4.9}$$

Replacing z with $\frac{av-\mu_{bird\ species}}{\sqrt{2\sigma^2_{bird\ species}}}$ before substituting back into equation (4.6) results in equation (4.10). For ease of reading, the subscripts denoting bird species migration speed are not being used; i.e. v denotes $v_{bird\ species}$.

$$\begin{aligned} F(av) &= \frac{1}{2}\left[1 + \frac{2}{\sqrt{\pi}}\sum_{n=0}^{\infty}\frac{(-1)^n\left(\frac{av-\mu_{bird\ species}}{\sqrt{2\sigma^2_{bird\ species}}}\right)^{2n+1}}{n!(2n+1)}\right] \\ &= \frac{1}{2}\left[1 + \frac{2}{\sqrt{\pi}}\sum_{n=0}^{\infty}\frac{(-1)^n(av-\mu_{bird\ species})^{2n+1}}{n!(2n+1)(2\sigma^2_{bird\ species})^n\sqrt{2\sigma^2_{bird\ species}}}\right] \end{aligned} \tag{4.10}$$

Equation (4.10) can then be partially differentiated in terms of a or $v_{bird\ species}$ to give the rate of change of $p_{distance_{Europe,\ bird\ species}}$ in respect to each of them.

$$\begin{aligned}
\frac{\partial p_{distance_{Europe,\ bird\ species}}}{\partial a} &= \frac{v}{\sqrt{2\pi\sigma^2_{bird\ species}}} \sum_{n=0}^{\infty} \frac{(-1)^n (av - \mu_{bird\ species})^{2n}}{n!(2\sigma^2_{bird\ species})^n} \\
&= \frac{v}{\sqrt{2\pi\sigma^2_{bird\ species}}} e^{-\left(\frac{av-\mu}{\sqrt{2\sigma^2}}\right)^2} \\
\frac{\partial p_{distance_{Europe,\ bird\ species}}}{\partial v} &= \frac{a}{\sqrt{2\pi\sigma^2_{bird\ species}}} \sum_{n=0}^{\infty} \frac{(-1)^n (av - \mu_{bird\ species})^{2n}}{n!(2\sigma^2_{bird\ species})^n} \\
&= \frac{v}{\sqrt{2\pi\sigma^2_{bird\ species}}} e^{-\left(\frac{av-\mu}{\sqrt{2\sigma^2}}\right)^2}
\end{aligned} \tag{4.11}$$

The rate of change of T can now be found for each variable for each bird species. These are shown in equations 4.12.

$$\begin{aligned}
\frac{dT_{Europe,\ bird\ species}}{d\rho_{prev}} &= \mu_{tick} \times N_{Europe,\ bird\ species} \times p_{distance_{Europe,\ bird\ species}} \\
\frac{dT_{Europe,\ bird\ species}}{d\mu_{tick}} &= \rho_{prev} \times N_{Europe,\ bird\ species} \times p_{distance_{Europe,\ bird\ species}} \\
\frac{dT_{Europe,\ bird\ species}}{dN_{Europe,\ bird\ species}} &= \rho_{prev} \times \mu_{tick} \times p_{distance_{Europe,\ bird\ species}} \\
\frac{dT_{Europe,\ bird\ species}}{da} &= \rho_{prev} \times \mu_{tick} \times N_{country,\ bird\ species} \times \frac{\partial p_{distance_{Europe,\ bird\ species}}}{\partial a} \\
\frac{dT_{Europe,\ bird\ species}}{dv_{bird\ species}} &= \rho_{prev} \times \mu_{tick} \times N_{country,\ bird\ species} \times \frac{\partial p_{distance_{Europe,\ bird\ species}}}{\partial v_{bird\ species}}
\end{aligned} \tag{4.12}$$

Holding all parameters at the minimum values in Table 4.3 and using equations (4.12), a value for the rate of change can be found for each variable for each bird species. Since they are based upon fixed geographical points the means and standard deviations for $p_{distance_{Europe,\ bird\ species}}$ will not vary. The resulting rates of change are shown in Table 4.7.

Table 4.7: Rate of change in number of CCHFV positive ticks introduced into Europe with respect to each parameter taken from minimum estimated values (rounded to 2 decimal places)

Species	ρ_{prev}	μ_{tick}	$N_{Europe,\ bird\ species}$	a	$v_{bird\ species}$
Willow Warbler	2714.88	5.54	0.00	0.04	0.02
Common Quail	47977.06	97.91	0.00	0.66	0.11
Tree Pipit	407.99	0.83	0.00	0.01	0.00
Northern Wheatear	22673.79	46.27	0.00	0.38	0.09
Ortolan Bunting	219604.1	448.17	0.00	3.90	0.56

From Table 4.7 it can be seen that for all species the level of risk of CCHFV positive ticks being introduced is far more sensitive to parameters relating to the ticks themselves rather than the birds. The greatest sensitivity is associated with a change in prevalence, followed by the mean number of attached ticks and then the length of on-host attachment and the sensitivity of these parameters are noticeably larger than those associated with the bird itself, that is the population size and migration speed. The estimated species population is relatively insensitive and would require a large change to have any effect on risk of CCHFV introduction. The Tree Pipit, which represented the lowest risk amongst the species in Table 4.6 is also the least sensitive to changes in parameter values. The fastest species, and so that with the greatest range, the Ortolan Bunting, is most sensitive to those parameters that affect the possible range; that is the migration speed and length of attachment.

Simulation Results

The distributions of D can be assumed to be approximately Normal as discussed previously. Means and standard deviations can be estimated for each distribution (Table 4.4) and sampling from migration speeds ($v_{species}$) and taking the length of attachment (a), the probability of a bird arriving within the maximum on host attachment time for ticks can be found. Combined with sampling from CCHFV prevalence amongst ticks (ρ_{prev}), the mean number of ticks on migrating birds (μ_{tick}) and sampling from population numbers ($N_{Europe,\, bird\ species}$) and inputting these into equation (4.3) results in an estimated number of introduced CCHFV positive ticks. Repeated sampling results in the following distributions for CCHFV positive tick incursions into Europe.

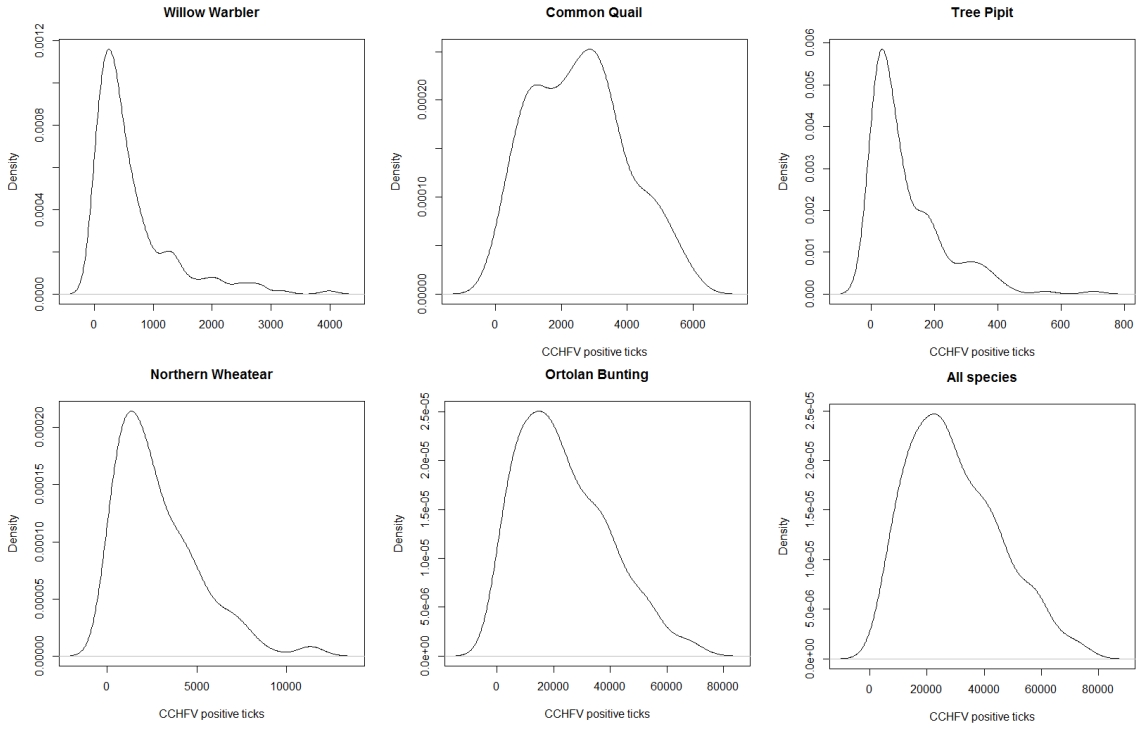

Figure 4.12: Distributions of CCHFV positive tick incursions into Europe for each and all bird species

Due to the skewness of many of these distributions, Table 4.8 gives the minimum, median and maximum predicted introduced CCHFV positive ticks for each bird species and for all bird species for Europe.

Table 4.8: Minimum, median and maximum number of introduced CCHFV positive ticks introduced into Europe, by species

Species	Minimum number of CCHFV positive ticks introduced	Median number of CCHFV positive ticks introduced	Maximum number of CCHFV positive ticks introduced
Ortolan Bunting	163	19093	76603
Common Quail	23	2373	6890
Northern Wheatear	18	2035	9863
Willow Warbler	7	334	3299
Tree Pipit	1	64	612
All	211	23900	97267

The results in Table 4.8 can be examined graphically (Figure 4.13) to illustrate the different contributions to total risk made by each bird species.

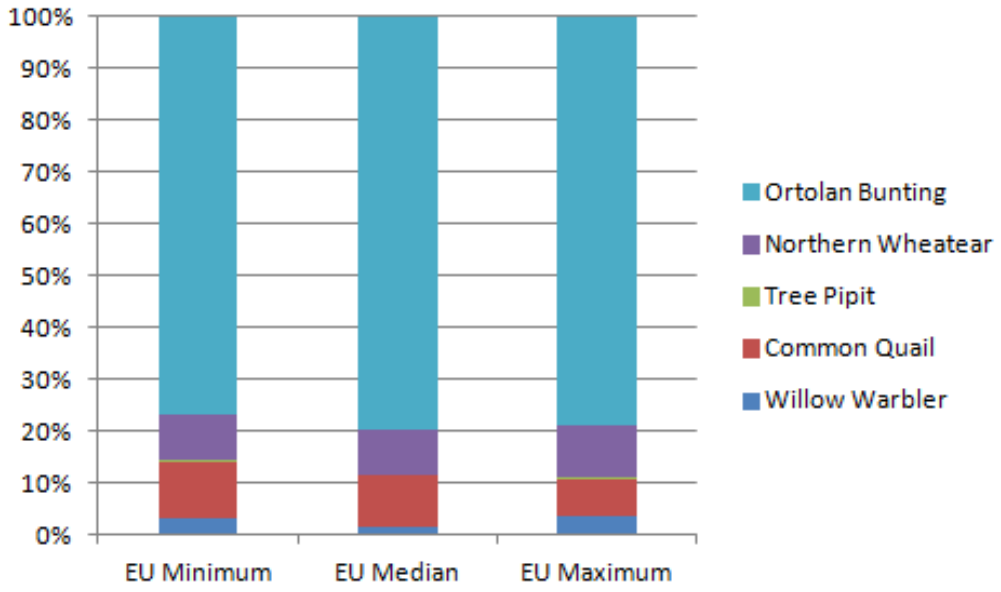

Figure 4.13: Percentage of CCHFV positive ticks introduced by each bird species in terms of the maximum, median and minimum results

The proportion of CCHFV positive ticks seems to remain consistent with a far greater proportion being introduced by the Ortolan Bunting as opposed to any other species. There is some variation in the risk represented by the Common Quail and the Willow Warbler and the Tree pipit represents almost no risk relative to the other species despite being the second most populous of the species examined.

4.6.2 CCHFV Incursions by Individual Country

Simulating the number of CCHFV positive ticks introduced by the migration of birds into each European country separately and aggregating these gives an alternative estimate of the risk to Europe. This total number of ticks introduced into Europe allows an examination of the importance of each bird species and a comparison with the analytical results above.

Table 4.9: Summary statistics of introduced CCHFV positive ticks by species for Europe

Species	Mean number of introduced ticks	Standard deviation	Median number of introduced ticks	Minimum number of introduced ticks	Maximum number of introduced ticks
Ortolan Bunting	29848	19267	26763	836	75478
Northern Wheatear	5376	3762	4619	159	17112
Common Quail	2732	1526	2657	109	5699
Tree Pipit	0	1	0	0	9
Willow Warbler	0	0	0	0	0

The first result to note in Table 4.9 is that the more detailed approach, where individual countries are modeled, has resulted in the most populous species, the Willow Warbler, representing absolutely no risk of introducing CCHFV. Additionally, the second most populous species, the tree pipit, introduced no CCHFV positive ticks in the vast majority of the simulations. The species that posed the greatest risk was that with the fastest flight speed, the Ortolan Bunting, and so has the greatest migration range possible within the on host attachment time of *H. Marginatum*.

Examining the species proportion of total introduced ticks can highlight the importance of the different species (Figure 4.14) or, in the case of the Willow Warbler and Tree Pipit, the lack of importance, with a negligible percentage between both species.

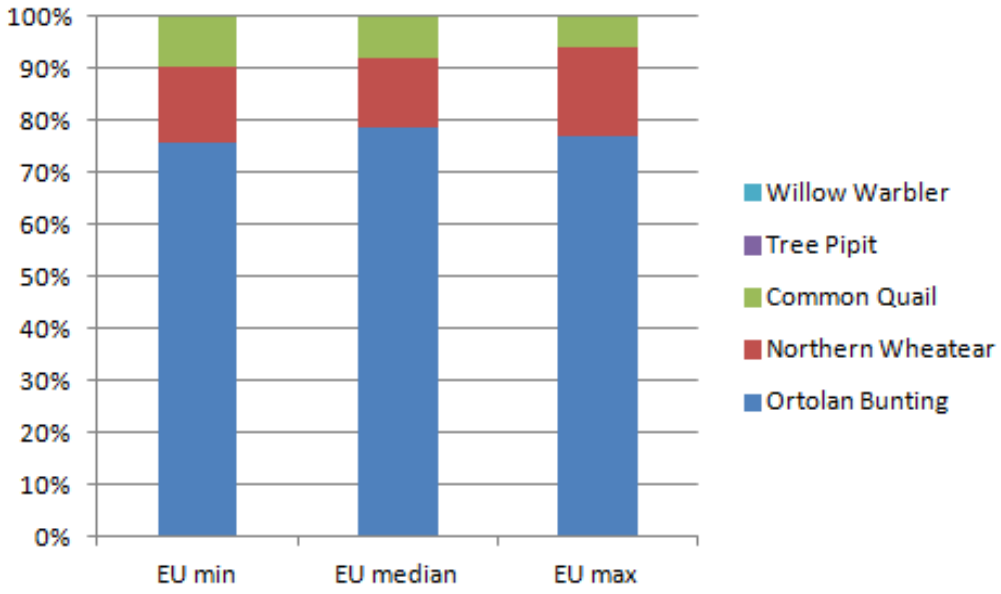

Figure 4.14: Percentage of CCHFV ticks introduced by each bird species under maximum, median and minimum results

Examining the number of introduced CCHFV positive ticks by countries gives an idea about what geographical areas surveillance efforts might be focused on. Taking the five countries with the greatest number of minimum incursions, median incursions and maximum incursions (Table 4.10) we can see the geographical distribution of risk.

Table 4.10: Top five countries most at risk of introduced CCHFV positive ticks for all bird species as ranked by minimum, median and maximum risk

Minimum number of introduced ticks	Median number of introduced ticks	Maximum number of introduced ticks
Turkey (881)	Turkey (28039)	Turkey (83255)
Spain (55)	Spain (1667)	Spain (3806)
Romania (55)	Romania (1191)	Romania (2955)
Poland (28)	Poland (895)	Poland (2471)
Ukraine (18)	Ukraine (413)	France (997)

Examining by individual country shows a distinctly limited geographic distribution, as a proportion of the total number of introduced CCHFV positive ticks, the five most at-risk countries are representative of the vast majority of incursions (over 90%).

The most at-risk countries for each species can also be examined in order to see if there is a different geographical distribution of risk for each bird species. Figure 4.15 shows the top five countries where there is the highest median number of introduced ticks; for two of the species, this was zero for all countries and so results for these species have not been plotted.

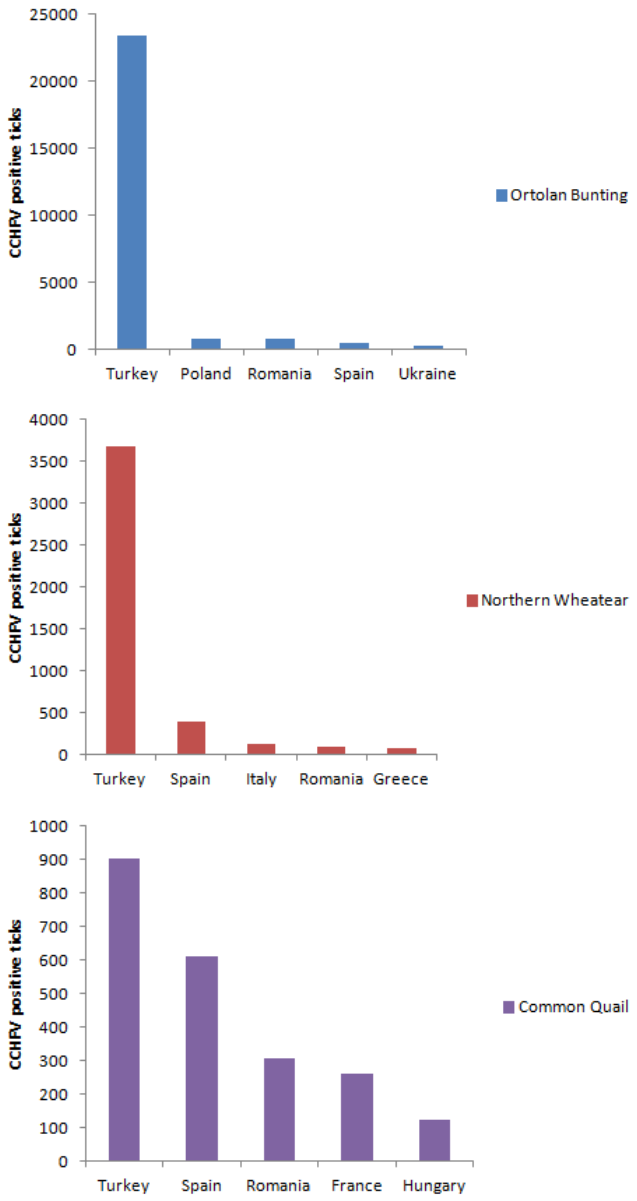

Figure 4.15: Top five countries most at risk of introduced CCHFV ticks for each bird species as ranked by median risk

4.7 Discussion

Adapting the linear model from Gale et al. (2011) in order to introduce a spatial distance element and to have a non-uniform distribution of birds resulted in three sets of results. The first makes use of an assumed Normal distribution for the migra-

tion distance between wintering grounds in sub-Saharan Africa and Europe and a uniform distribution of bird populations to simulate results. The second makes use of the same assumptions to arrive at analytical solutions and finally separate non-standard distributions of distance are developed for each country and non-uniform distributions of birds across Europe are used to simulate results separately for each country.

Comparing the European simulated (Table 4.8) and analytical results (Table 4.6), the results are broadly similar with the analytical approach producing a slightly wider range. This, of course, might not be the case if a larger number of iterations were used in the simulation approach as there would be a greater chance of extreme values being chosen for all parameters in a single run.

One of the main points of interest is that the species that represents the most risk is the Ortolan Bunting, which was not one of the species modeled in the original paper (Gale et al., 2011) which used only the top four bird species that were judged to be the greatest risk for introducing CCHFV positive ticks. This means that it may be more important to look at those species with a greater potential migratory range within our attachment time, rather than looking merely at those with the largest population. Put simply, many of the less populous species may represent a much greater risk of introducing CCHFV compared to the more populous species and so focusing on them might give a better overall indication of the potential risk.

The sensitivity of either of these approaches to the different parameters is dependent on which bird species is being discussed, but a couple of general conclusions can be drawn. Firstly, it seems very clear that the number of CCHFV positive ticks is much more influential than the number of birds (large values in Table 4.7 as opposed to values that were effectively zero for population) though it is worth noting that the absolute change in estimates of bird numbers will be much larger than either the mean number of ticks per bird or the prevalence of CCHFV positive ticks. In Table 4.7, the values associated with population are so low that they are rounded to zero; however, with a large enough change in population there could still be a small change in risk although this would be insignificant compared to a much smaller change in

prevalence. However, examining Tables 4.3 and 4.5 the biggest proportional change between our lowest and largest estimates is for prevalence, so it is this that our model is most sensitive to. This makes intuitive sense as, since prevalence is very small, then it would require a large increase in the number of birds to result in a small change in the number of CCHFV positive ticks, but since N for all of Europe is very large, then a small change in prevalence will have a large effect on the number of CCHFV positive ticks introduced.

The parameters relating to distance, migration speed ($v_{bird\ species}$) and length of attachment (a), show more variation between different bird species. Using either of these parameters, our five bird species can be ranked by their level of sensitivity from most to least with the Ortolan Bunting being the most sensitive to a change followed by the Common Quail, the Northern Wheatear, the Willow Warbler and the Tree Pipit being the least sensitive to a change in migration speed or attachment time. Examining Table 4.3, it can be seen that this is the same order found if the bird species are ranked in terms of their minimum estimated speed. Both these parameters are used to form $p_{distance_{Europe,\ bird\ species}}$ which, in this case, takes the form of the cumulative distribution of the Normal distribution which is sigmoidal and so the rate of change will be relatively lower at the tails of the distribution. Since the point taken on this distribution is equal to the migration speed multiplied by attachment, $av_{bird\ species}$, then a bird species with a much lower migration flight speed will be much further left on this distribution where the gradient is smaller and so less sensitive to any changes.

The greater sensitivity to changes in migration speed and attachment time for the Ortolan Bunting may also be in part due to the fact that the geographical distribution for this species is much more focused i.e. they winter and breed across a much smaller area of Europe. As such, the distribution describing their flight distance (Figure 4.10) has greater kurtosis and so increasing x results in more of an increase in $Pr(X \leq x)$ than in a less peaked distribution.

Comparing the results for all of Europe (Table 4.8 and Table 4.6) with those of Europe where each country is calculated independently (Table 4.9), the first thing

to note is that the Willow Warbler and Tree Pipit are now virtually no threat in terms of the introduction of CCHFV positive ticks. Examining Table 4.5 for both these species, a large part of their European population breeds in Scandinavia, so the greatest distance from their wintering grounds. Relaxing the assumption of uniform distribution of birds across Europe now means that for some species the larger part of their populations will coincide with the upper part of the European distance distributions (Figure 4.10) and so not contribute to introducing CCHFV positive ticks. This means that, under this model, the two most populous of our bird species are in fact the least important.

For some of the other bird species, the opposite is true. The Northern Wheatear now introduces more CCHFV positive ticks. In Table 4.5 a large part of the Northern Wheatear population can be seen to breed in Turkey which is relatively far south and so closer to the winter feeding grounds. This means that a greater part of the population coincides with the lower part of the European distance distributions (Figure 4.10) and so a smaller proportion of the population will be associated with the upper part where the distance is greater than can be covered. These differences suggest that the assumption of a uniform population, as in Gale et al. (2011) or the first simulated approach, to be deeply flawed. Modelling the countries individually not only results in a more detailed geographic distribution that examines where risk might be focused but also results in a far more accurate estimate of the overall risk of introduction.

Examining the countries with the greatest number of introduced CCHFV positive ticks (Table 4.10) and the plots of most at risk countries for those species that do introduce CCHFV positive ticks (Figure 4.15), then Turkey is consistently most at risk. For two of these species, the Northern Wheatear and the Ortolan Bunting, Turkey has the largest European population (Tables 4.5) and is a southern country so close to the wintering grounds in sub-Saharan Africa meaning its likely well within the 26 day (on-host attachment time) flight range of migrating birds. The Common Quail has a large Turkish population as well, though slightly smaller than Spain, and so it is unsurprising that it is at a far greater level of risk. For the separate species, it is interesting that Spain is not at the most risk for the Common Quail despite having

the largest population and this is presumably to do with having a generally greater required flight distance than Turkey ($\bar{D}_{Turkey,\, Common\ Quail} < \bar{D}_{Spain,\, Common\ Quail}$). A more evenly distributed European population for the Common Quail can also be seen by the smoother results between countries as opposed to the Northern Wheatear or Ortolan Bunting which are very much focused in Turkey and have results that reflect this.

In terms of a single European approach (as in the first two approaches in this chapter) or a separate country approach (the third and final approach) there are a number of results that could be argued to agree, but an equal number that appear very different. This is likely to be due in part to the fact that, while a single distance distribution for all of Europe is fine as an approximation, having a single population figure for Europe is not. To assume a single European population like this is to assume that the species is spatially uniformly distributed across Europe. This means an unrealistic distance of migration will be assumed for several species i.e. those species that live only in the North will have a much lower distribution of distance than they should and so represent a much greater threat than they should. Risk will also be more uniformly distributed especially if the actual migration itself is not taken account of, as in Gale et al. (2011), but even if the migration is examined this will still have an affect. It can be concluded that a simplistic analytical approach is not suitable.

The level of difference in levels of risk in both the analytical and simulated results indicate the level of uncertainty. The magnitude of difference between maximum and minimum estimated risk means that tackling many of the parameter uncertainties especially for prevalence of CCHFV amongst larval and nymphal ticks on migratory birds would be very important in estimating the risk represented by CCHFV

Leaving aside the uncertainty in prevalence rates, some species have a much larger difference in maximum and minimum predicted migration speeds but in Payevsky (2013) they also have the most data available and so this level of variation in migratory speed may be the norm and for our other species the more concise range of speeds is simply due to lack of data.

Results by country are, in part, consistent with what little real data there are in that CCHFV is present in Turkey and may have been introduced by migratory birds. It has also recently been found to be present in Spain and Portugal so it is consistent with real data that Spain has the second highest level of introduced CCHFV positive ticks.

Results from this modelling approach must be viewed with some caution. Despite improving on the approach in Gale et al. (2011), there are still a couple of major weaknesses with this approach.

Firstly, while the distance distributions take account of the possibility of ticks detaching from their host before reaching their breeding grounds, it does not allow for those ticks that detach in another European country. For example, a bird that breeds in France may pass through and deposit a CCHFV positive tick in Spain and this would not be counted. This can be contrasted with the approach from Gale et al. (2011) where all ticks stayed on host and so these ticks would be counted but so additionally would those ticks that would have detached outside of Europe.

Secondly, the distance distributions make use of the Euclidean distance which is a straight line between the two points. For avian migration, this is not appropriate as birds will take an indirect course in order to avoid obstacles. In particular many species will change their route to avoid, where possible, flying over open water. This means for migrating birds flying between Africa and Europe they tend to follow three flyways that avoid the most open parts of the Mediterranean. This would results in a greater flight distance and so change our distributions.

The next modelling approach will make use of a more spatially explicit model where these two issues will be addressed and the results generated can be compared and contrasted with those above. Before this, however, an examination of the estimate of CCHFV prevalence will be carried out. This is due to the importance of this parameter as shown by the sensitivity analysis and the fact that real life data do not seem to agree with estimates in previous papers. A Bayesian approach will be

used to examine these two data sources and to create a better estimate of prevalence based upon both of them.

Chapter 5

Bayesian statistical estimate of the prevalence rate of CCHFV in sub-Saharan ticks

5.1 Introduction

In the previous chapter a Geographic Information System approach was used to model the number of CCHFV infected ticks introduced into Europe by five different species of migratory birds. As part of this modelling process, an analysis of the sensitivity of the model to its different parameters was carried out. It was found that across all five species the number of introduced infected ticks was most sensitive to a change in the prevalence of CCHFV amongst ticks.

Of the two estimates of CCHFV prevalence in *H. marginatum* used in Chapter 4 the first is based on an expert opinion (Gale et al., 2011), which is in turn based on data of CCHFV prevalence amongst adult ticks in Africa, and the second is taken from one of two recent studies testing ticks on migratory birds. In this chapter, a Bayesian approach will be used to combine these estimates using the expert opinion as a prior and the two studies as data points. This will mean that no single source of information is being used and that the small amount of real data available will not be ignored, as in Gale et al. (2011), but neither will such a small sample be used on its own to estimate prevalence.

5.2 Methods

Bayesian inference is a technique based on Bayes' theorem (Equation (5.1)) and can be used to take new data into account to improve a previous estimate of a parameter (Vose, 2008).

The approach can be broken down into three parts, containing x which is a vector or matrix of data and θ a vector or matrix containing the parameters of a potential model to explain x, with the first being a prior ($\pi(\theta)$) that represents all that is known about our parameter and represents the current belief or view held on its value. Secondly an appropriate likelihood function ($l(x|\theta)$) for the observed data and finally the calculation of the posterior ($f(\theta|x)$) which is the revised estimate of the parameter. As seen in Equation (5.1) the posterior is calculated by multiplying the prior by the likelihood and the denominator is used to normalise this so the area under the curve for the posterior is equal to one.

$$f(\theta|x) = \frac{\pi(\theta)l(x|\theta)}{\int \pi(\theta)l(x|\theta)d\theta} \tag{5.1}$$

The prior, the initial state of information, is often uncertain so rather than simply being a point estimate is instead represented as a probability density function. The likelihood function measures the probability of observing the data (x) for a particular value of θ and the posterior is the description of our knowledge of the parameter after the data have been taken into account. If the data are considered very unlikely to have randomly occurred, given our prior beliefs, then our posterior beliefs will be radically different from the prior. If, however, the data are very likely to have occurred randomly, given our prior distribution, then the posterior distribution will more closely resemble the prior.

5.2.1 Priors

In this case our prior, the current belief on CCHFV prevalence in larval and nymphal ticks, is taken from Gale et al. (2011) and consists of a single estimated level of prevalence for CCHFV amongst ticks. This single point estimate must be turned into a distribution that represents our prior knowledge and a number of approaches can be used. These approaches can be split by the type of prior used and broadly speaking there are three types (Vose, 2008). The first is an informative prior that reflects the initial view including the particular distribution chosen for the prior as well as the parameter values. For many studies involving proportions, such as prevalence, a beta distribution is used (Enøe et al., 2000) as it is bounded by zero and one and can, by manipulation of its parameters, portray a large number of shapes.

In Enøe et al. (2000) a technique is outlined to form suitable parameters for a beta distribution by making use of an estimated most likely value (mode or in some case a mean) and an upper estimate (95% percentile) and/or lower (5% percentile value). The equation of the mode for the beta distribution (Equation 5.2) is dependent on the two shape parameters, a and b, that determine the beta distribution.

$$Mo = \frac{a-1}{a+b-2} \tag{5.2}$$

This can be rearranged to give an equation for the shape parameter a (Equation 5.3).

$$a = \frac{1+Mo(b-1)}{1-Mo} \tag{5.3}$$

The value of the mode can then be substituted and for a given value of the shape parameter b then the parameter a can be found. A software program such as R can be used to simulate random draws from a beta distribution with different values of b and the equivalent a and the 5% and 95% percentiles can be found. These can be compared to the upper and/or lower estimate and an iterative approach can be used to find appropriate parameter values to represent the prior.

The second form of prior is the conjugate prior. This is a distribution chosen so

that the posterior distribution is in the same family as the prior distribution. This is done in the main for computational convenience.

The final form of prior is the noninformative prior distribution. These represent minimal existing belief or knowledge and are generally flat or diffuse and give very little information about the parameter of interest and these can often be difficult to construct (Gelman et al., 1998).

To investigate the sensitivity of results and to ensure a better model fit is found, a prior of each type will be used. In order to easily form a noninformative prior, the same distributions as for the informative prior will be used, but with parameter values changed in order to give a more diffuse shape.

To compare the different prior choices, the deviance information criterion (DIC) will be used. This was designed to select a model based on fit and complexity (Spiegelhalter et al., 2002) and takes the form:

$$DIC = -2ln[P(y|\bar{\theta})] + 2p \tag{5.4}$$

In equation (5.4) $\bar{\theta}$ is the posterior mean(s) of θ and p is the posterior mean of the deviance minus the deviance of the posterior means and should be approximately the true number of independent parameters (Spiegelhalter et al., 2002). A higher likelihood ($[P(y|\bar{\theta})$) will result in a smaller value and this indicates a better fit and a smaller p indicates a more parsimonious model and so the smaller the value of DIC the better the model.

5.2.2 Likelihood and Data

The likelihood function for the prevalence of CCHFV amongst larval and nymphal ticks in sub-Saharan Africa is set as a Normal distribution. This distribution was chosen as it is unimodal and many natural occurrences follow this distribution, so it seems a reasonable choice.

The data for this analysis are taken from studies of ticks removed from captured migratory birds during their northwards migration.

5.3 Parameters

The informative prior for our estimate is taken from Gale et al. (2011) and consists of a single estimated prevalence for CCHFV amongst larval and nymphal ticks. This value is very small and, since prevalence must be bounded by zero and one, then a distribution with a positive skew would be most appropriate. A beta distribution with $b > a$ will give this shape. This single point estimate can be used as the most likely value for the technique in Enøe et al. (2000). An upper and/or lower estimate is then needed and since Gale et al. make a case for this estimated larval/nymphal prevalence being less than the estimated CCHFV prevalence for adult ticks, then the value for adult tick prevalence is used as the upper estimate (Table 5.1). Values of the two shape parameters were then found by filling the most likely value into equation (5.3), selecting values for the shape parameter b and calculating the value of the shape parameter b. For this beta distribution, the 95% percentile can be found and compared to the CCHFV prevalence for adult ticks and from this comparison a new value of b can be selected and the process repeated until a distribution is found that correctly reflects the estimated parameters, Mo and θ_U (see Table 5.1), and this process produced the following distribution:

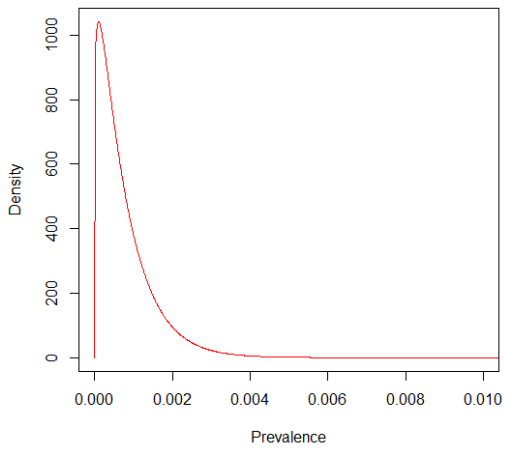

Figure 5.1: Informative prior distribution for mean value of prevalence of CCHFV amongst ticks in sub-Saharan Africa

Figure 5.1 is the prior distribution for the mean prevalence of CCHFV amongst larval and nymphal ticks in sub-Saharan Africa. Prior information about how certain we are of this value is needed, but there is no available information on how much this might vary. It is based on expert opinion and the reputation and certainty of the expert could be used to define a level of uncertainty, but, in this case there is little in the original paper to convey such a measure. Therefore, for the standard deviation of CCHFV prevalence a uniform prior covering the entire sample space is used.

The noninformative prior was created in a simplistic fashion by taking the informative prior and changing the parameters of the beta distribution so as to flatten the distribution.

The data for this analysis are taken from Lindeborg et al. (2012) and Palomar et al. (2013); both of these papers were studies of ticks removed from captured migratory birds during their northwards migration in Capri and Antikythira and in Morocco respectively.

Table 5.1: Estimates of parameter values

Description	Value(s)	Reference
Mode of distribution of mean prevalence of CCHFV amongst sub-Saharan ticks (Mo)	0.0001	(Gale et al., 2011)
Upper limit of mean prevalence of CCHFV amongst sub-Saharan ticks (θ_U)	0.0027	(Gale et al., 2011)
Informative Prior distribution of mean prevalence of CCHFV amongst sub-Saharan ticks	$Beta(1.14995, 1500)$	
Informative Prior distribution of standard deviation of prevalence of CCHFV amongst sub-Saharan ticks	$U(0, 1)$	
Noninformative Prior distribution of mean prevalence of CCHFV amongst sub-Saharan ticks	$Beta(1, 1)$	
Noninformative Prior distribution of standard deviation of prevalence of CCHFV amongst sub-Saharan ticks	$U(0, 1)$	
Data on mean number of infected ticks removed from captured migratory birds	0.057692308, 0.004563709	(Lindeborg et al., 2012), (Palomar et al., 2013)

5.3.1 Posterior

To arrive at a posterior distribution for all of the priors and data in Table 5.1, WinBUGS (Windows Bayesian inference Using Gibbs Sampler) software was used. This

makes use of a Bayesian technique known as the Gibbs Sampler which is a Markov Chain Monte Carlo (MCMC) method used for posterior simulation. This method is used to find a sequence of observations from a multivariate distribution from which direct sampling is tricky. It achieves this by sampling from a conditional distribution instead, Our posterior $f(\theta|x)$ in Equation 5.1 is often not easy to draw from when θ is more complex and, if independent samples cannot be drawn to allow a standard Monte Carlo approach then instead slightly dependent draws can be taken. For example if θ consists of 3 separate values, then initial values of θ are selected, regarded as θ^0 which will contain θ_1^0, θ_2^0 and θ_3^0 as starting estimates for each of the 3 values.

A random draw for θ_1^1 can then be taken from $p(\theta_1|\theta_2^0, \theta_3^0)$.

This new estimate of θ_1^1 can then be used to find a random draw for θ_2^1 from $p(\theta_2|\theta_1^1, \theta_3^0)$.

Similarly, for a estimate of θ_3^1 as a random draw from $p(\theta_3|\theta_1^1, \theta_2^1)$.

These steps can then be repeated S times to yield a set of S draws for each value of θ, a number of these draws, the default in WinBUGs is 1000 which was sufficient here, can be discarded as 'burn in' to eliminate the effect of the selection of the starting values; plots of the coefficient estimates can be used to visually inspect for convergence. A weak law of large numbers is then invoked to show that the estimates of θ tend to the true values as S goes to infinity.

5.4 Results

Running the analysis for each of the 3 priors (Table 5.1) in WinBUGS the following posterior parameters were found:

Table 5.2: Posterior parameter values

Description	Mean	Standard deviation
Posterior mean prevalence of CCHFV amongst sub-Saharan ticks (Informative Prior distribution)	0.0007528	0.0007067
Posterior standard deviation of prevalence of CCHFV amongst sub-Saharan ticks (Informative Prior distribution)	0.1467	0.1713
Posterior mean prevalence of CCHFV amongst sub-Saharan ticks (Noninformative Prior distribution)	0.1433	0.1722
Posterior standard deviation of prevalence of CCHFV amongst sub-Saharan ticks (Noninformative Prior distribution)	0.2468	0.2452

Density plots of each of the posterior distributions can be produced by WinBUGS (Figure 5.2).

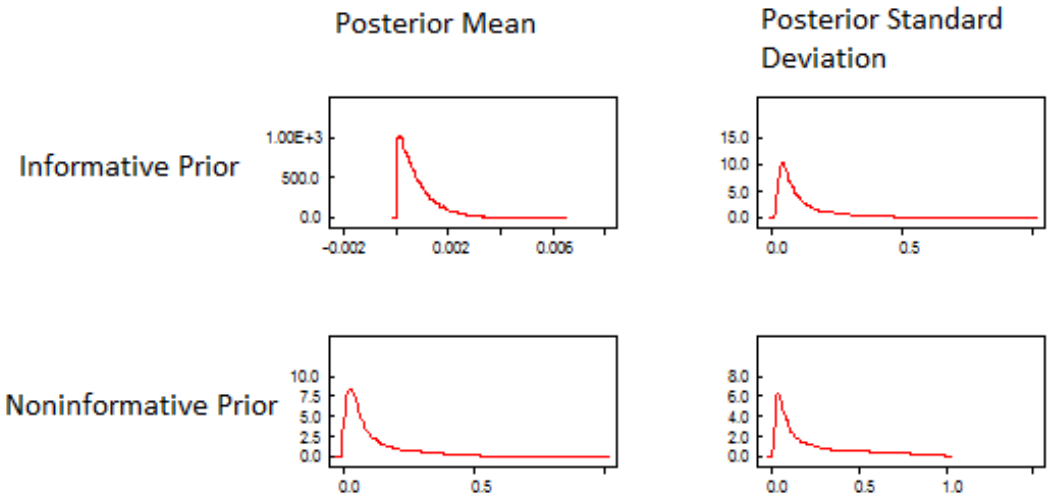

Figure 5.2: Posterior distributions of mean and standard deviation of prevalence of CCHFV amongst ticks in sub-Saharan Africa formed from each prior

In Figure 5.2, the posterior distributions produced using the informative prior has yielded a much tighter distribution compared to the non-informative prior and this is to be expected since it reflects a more certain prior. Since, in the informative prior, more weight is given to the expert opinion on CCHFV prevalence in ticks the resulting mean in Table 5.2 is much closer to the estimated value in (Gale et al., 2011) than the result from either of the other 2 priors.

To test which of these priors was the most suitable and which set of results might therefore be regarded as the most useful, the deviance information criterion was produced for each.

Table 5.3: Estimates of parameter values

Description	DIC
DIC for posterior using informative prior	-5.580
DIC for posterior using noninformative prior	-4.427

Table 5.3 indicates that the informative prior should be regarded as yielding the most useful set of results and the noninformative yields the least. This means that using all available information i.e. the available data and the expert opinion from Gale et al. (2011) produces the better posterior.

5.5 Discussion

In Table 5.2, the mean prevalence of CCHFV amongst sub-Saharan ticks has increased in all cases from the mean in the priors. This is expected, as all the data points are above the value used to create the priors and so the posterior distribution would be centered more in this direction. The greatest movement between prior and posterior estimates of prevalence is in the case of the uninformative prior. This reflects the strength of the priors; the prior representing the strongest belief (the informative prior) has the smallest difference between prior and posterior mean value and the prior representing the weakest belief (the noninformative) has the greatest.

In Table 5.2 and Figure 5.2 the prevalence estimate produced by using the informative prior has a much narrower distribution than the noninformative prior. This combined with the lower deviance information criterion (Table 5.3) show that the informative prior was the most appropriate approach in this case. This also indicates that the expert judgment used for the prevalence value in Gale et al. (2011) produced an accurate estimation of CCHFV prevalence since the prior most closely based upon this was found to be most appropriate.

As such, the posterior mean CCHFV prevalence amongst sub-Saharan ticks produced using the informative prior will be used in the next chapter where a spatially explicit model of migration will be developed. The estimate of CCHFV prevalence taken from Gale et al. (2011) will continue to be used as the lower estimate and this new estimate of prevalence, which is higher than that in Gale et al. (2011), will be used as the upper estimate. This will yield more realistic estimates that make full use of all available data whether that be expert opinion or the scarce real life data that have been collected.

Chapter 6

Population level cellular automata for the risk of incursion of CCHFV via migratory birds

6.1 Introduction

In Chapter 4, a GIS approach based upon an initial model from Gale et al. (2011) was expanded upon to take into account distance of migration. This was used to explore the risk of CCHFV positive ticks being introduced by migrating birds flying from sub-Saharan Africa to Europe. A GIS modelling approach had two main weaknesses: firstly, it did not allow for ticks detaching from their host in any part of Europe the host passed through and, secondly, it made use of the Euclidean distance when estimating the likelihood of a host reaching its destination within the on host attachment period. Both these assumptions are unrealistic as birds do not take a direct path from wintering to breeding grounds and ticks can detach from their host during this journey.

In this chapter, a more spatially explicit modelling approach will be used to explore the risk of CCHFV positive ticks being introduced to Europe. In particular, an aggregated population cellular automata model of birds' migratory behaviour will be constructed and used to assess the risk of tick introduction.

6.2 Methods

6.2.1 Birds Species

The species of birds will be the same as those discussed in Chapter 4 (Table 4.2). That is, the Willow Warbler, the Northern Wheatear, the Common Quail, the Tree Pipit and the Ortolan Bunting. These are the five most populous species that migrate between sub-Saharan Africa where CCHFV is present and Europe and which have both nesting and feeding behaviours that make them vulnerable to becoming hosts for ticks.

6.2.2 Cellular Automata

To model the movement of birds and estimate the risk it represents, a spatially explicit model is needed. The simplest of these are known as cellular automata and are grid based, so the movement at each timestep is discrete (Navier-stokes, 1985). The model created here describes the movement of birds from Africa to Europe over discrete time steps of length one day and is implemented in the statistical programming language R. The geographical area including Africa and Europe is represented by a m x n lattice with each cell on the lattice covering an area based on the average daily speed of a particular bird species; e.g for the Common Quail that has an average daily speed of 150km per day then each cell is 150x150km^2 as opposed to the Willow Warbler where the average daily speed is only 62km per day and so cells would be 62x62km^2.

In many applications of these models each cell's value is determined using a differential equation with parameters taken from surrounding cells (White et al., 2007) and many such models have been used to investigate disease spread (Estrada-Peña et al., 2012; Beauchemin et al., 2008; Xiao et al., 2006; Benyoussef et al., 1999). Since our model does not involve disease dynamics but instead involves the movement of agents whose numbers should not vary, this would not be appropriate and all our equations will be based round the movement of birds and make use of population and geographical data from surrounding cells to govern movement. The simplest ap-

proach is to model this from a population perspective so we model the movement of a group of birds within each cell as opposed to modelling from an agent perspective where we would model behaviour for each individual bird.

The model makes use of a number of matrices all of size m x n. Firstly, there is the matrix B_t, which will represent all birds that are undergoing migration, i.e. are migrating in flight, and will represent their geographical positions at time t. Each cell $B_t[x, y]$ will contain the total number of birds present in that area (in flight) at time t with y being equivalent to a longitudinal position and x a latitudinal one and so there will be a single group of birds per cell. In order to cover the required geographical region, these will vary from longitude -15° (the leftmost edge of y=1) up to 57.78° (the rightmost edge of y=n) and from latitude -46.98° (the bottommost edge of x=1) up to 71.16° (the topmost edge of x=m).

In addition, there will be 2 matrices to record the birds that have landed in available breeding space (BL_t) and ticks that have been deposited (T_t), again both at time t.

The dynamics of the lattice model, that is, the movement of birds through B_t, are also governed by two m x n matrices P_t and G which will both have the same dimensions as B_t with cells representing the same geographical region. The matrix P_t represents the available space for breeding populations for each cell of the lattice. These values will change following settlement and so at time t will represent the maximum breeding capacity minus the number of birds that have settled there in all time steps less than t. This means that each element of P will potentially decrease every timestep and that $P_{t-1}[x, y] \geq P_t[x, y]$.

As the available population in a cell decreases, the number of birds landed in that cell will increase by an equal value.

$$P_{t-1}[x, y] - P_t[x, y] = BL_t[x, y] - BL_{t-1}[x, y] \tag{6.1}$$

This is true for all time steps and thus we can rewrite equation (6.1) as:

$$-(P_t[x, y] - P_0[x, y]) = BL_t[x, y] - BL_0[x, y] \tag{6.2}$$

Proof:

$$P_{t-1}[x,y] - P_t[x,y] = BL_t[x,y] - BL_{t-1}[x,y]$$

$$\Rightarrow P_t[x,y] = P_{t-1}[x,y] - BL_t[x,y] + BL_{t-1}[x,y]$$

$$\Rightarrow P_t[x,y] = [P_{t-2}[x,y] - BL_{t-1}[x,y] + BL_{t-2}[x,y]] - BL_t[x,y] + BL_{t-1}[x,y]$$

$$\Rightarrow P_t[x,y] = P_{t-2}[x,y] - BL_t[x,y] + BL_{t-2}[x,y]$$

$$\ldots$$

$$\Rightarrow P_t[x,y] = P_0[x,y] - BL_t[x,y] + BL_0[x,y] \tag{6.3}$$

Additionally, since no birds will be landed at timestep zero (so $BL_0[x,y] = 0$) then equation (6.3) tells us that the available population for any cell is equal to the total number of birds that a cell can support ($P_0[x,y]$) minus all the birds that have landed up to time t.

Behaviour

Matrix P

Movement of the in flight bird population through B_t is determined by evaluating the total available breeding space in all neighbouring cells. That is, for cell $B[x,y]$ all cells that share an edge or vertex with $P[x,y]$, so all cells directly above, below or beside and also those diagonally touching (the Moore neighbourhood (Navier-stokes, 1985; White et al., 2007)). This means the majority of cells will have eight neighbours. The in flight bird population for that cell is divided by this total available breeding space as in equation 6.4.

$$R_t[x,y] = \frac{B_t[x,y]}{\left(\sum_{i=-1}^{1}\sum_{j=-1}^{1} P_t[x+i,y+j]\right) - P_t[x,y]} \tag{6.4}$$

$R_t[x,y]$ is therefore the ratio of the number of birds in flight in a cell against the amount of breeding space available to those birds within the next timestep.

B_{t+1}, the number of birds in flight in the next timestep, is found by evaluating the sum of all neighbouring ratios and multiplying them by the corresponding P_t for that cell. This is shown in equation 6.5.

$$B_{t+1}[x,y] = \left(\left(\sum_i \sum_j R_t[x+i,y+j]\right) - R_t[x,y]\right) * P_t[x,y] \quad \begin{array}{l} x=1 \quad i=0 \text{ to } 1 \\ x=m \quad i=-1 \text{ to } 0 \\ 1<x<m \quad i=-1 \text{ to } 1 \\ y=1 \quad j=0 \text{ to } 1 \\ y=n \quad j=-1 \text{ to } 0 \\ 1<y<n \quad j=-1 \text{ to } 1 \end{array} \tag{6.5}$$

For both equation (6.4) and equation (6.5) this is the case for the majority of cells. For cells where $x = 1$ then i will go from zero to one, for $x = m$ then i goes from negative one to zero and equivalently with j when $y = 0$ and $y = n$.

As an example, consider the following simple 3x3 matrices:

$$B_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad P_0 = \begin{bmatrix} 10 & 10 & 0 \\ 10 & 0 & 15 \\ 25 & 20 & 10 \end{bmatrix}$$

If we number the rows and columns starting from the top left cell then matrix B_0 tells us that there are 50 birds inflight over the area represented by $B_0[2,2]$. In addition we know that there is breeding space available for 10 birds in each of cells $P_0[1,1]$; $P_0[1,2]$; $P_0[2,1]$ and $P_0[3,3]$; breeding space available for 15 birds in cell $P_0[2,3]$; for 20 birds in cell $P_0[3,2]$ and for 25 birds in cell $P_0[3,1]$.

R_0 is then:

$$R_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{50}{100} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Our matrix B for the next timestep can then be calculated as:

$$B_1 = \begin{bmatrix} 10 * 0.5 & 10 * 0.5 & 0 * 0.5 \\ 10 * 0.5 & 0 & 15 * 0.5 \\ 25 * 0.5 & 20 * 0.5 & 10 * 0.5 \end{bmatrix}$$

This means each cell will receive a number of birds proportionate to its level of breeding space in comparison to other neighbouring cells. This can be thought of as considering the potential for birds to land in the following timestep and defining movement based on this. This means direction of travel for birds will be based only upon availability of resource (in this case breeding space) in neighbouring cells This is carried out simultaneously for each cell at each timestep.

After each timestep, the values in the matrices P and B will be rounded to the nearest integer. This will result in a level of rounding error which means that the number of birds in the system can vary slightly between timesteps. This is a disadvantage of this aggregated population approach but is necessary as having partial birds in a cell is physically impossible in a real world sense but will also cause modelling issues.

Matrix G

General bird movement when there is not breeding ground nearby is determined by the matrix G. It was designed to reflect two particular behaviours of bird movement; first, that birds will generally have a northbound direction of movement (during spring migration) and, secondly, they tend to avoid open water. This means that G will contain increasing values as we move from the bottom to the top of the matrix i.e. cells in row one will be greater than those in row two, and that any cell that contains only water shall have a value of zero.

In the case of there being no available breeding space, that is all neighbouring cells have P=0, then matrix G is used in a similar fashion to that described in our exam-

ple. For cells with nearby breeding space, a neighbouring cell has P>0, movement will be determined as described in the previous section and G will not be used.

Matrix BL

Once all the cells of B have been updated with new bird population numbers, birds will land to fill up the available space. This is represented by the number of birds in a cell reducing by the associated P_t e.g. if 20 birds move into a cell with a value of $P_t[x, y]$=10, $P_{t+1}[x, y]$ will then be zero and the number of in flight birds will have reduced to 10. The landed bird population (BL[x,y]) will increase to reflect this.

This can be illustrated by a simple example where we have a 3 x 3 matrix and start with birds in cell $B_0[3, 2]$ and available breeding space to the northwest, north and northeast and with no birds on the ground:

$$B_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 100 & 0 \end{bmatrix} BL_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} P_0 = \begin{bmatrix} 0 & 0 & 0 \\ 10 & 25 & 15 \\ 0 & 0 & 0 \end{bmatrix}$$

Birds in cell $B_0[3, 2]$ then split proportionally to move into these neighbouring cells:

$$B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 20 & 50 & 30 \\ 0 & 0 & 0 \end{bmatrix}$$

To complete the first timestep birds land to take up the available breeding space leaving us with:

$$B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 10 & 25 & 15 \\ 0 & 0 & 0 \end{bmatrix} BL_1 = \begin{bmatrix} 0 & 0 & 0 \\ 10 & 25 & 15 \\ 0 & 0 & 0 \end{bmatrix} P_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Since all cells are updated simultaneously then a single cell of B_1 will probably acquire birds from multiple cells of B_0. Once all movement for that timestep has been

determined and all bird related matrices are fully updated, the birds in flight (B_t) and birds landed (BL_t) matrices are summed and multiplied by the prevalence rate for CCHFV positive ticks on that particular bird species (ρ) divided by the number of days on host (a), resulting in the number of ticks that will drop off that day which is added to the previous timestep's total.

$$T_t[x, y] = \frac{(B_t[x, y] + BL_t[x, y]) * \rho}{a} + T_{t-1}[x, y] \tag{6.6}$$

ρ in this case makes no assumptions about the distribution of ticks amongst a population i.e. all the ticks introduced to a region could come from multiple birds or a single bird within that population. Since this model is based round aggregated populations it is simply the expected number of ticks per bird.

Equation 6.6 uses birds in flight (B_t) as well as birds landed as there is the possibility of birds depositing ticks during migration as well as when they have landed. Birds will make regular stopovers to rest and feed (Dorst, 1962) and so ticks can be deposited at this time meaning that birds that are listed as in flight can introduce a tick to the area they are currently passing through. The division by a represents the proportion of ticks that will be ready to detach at that point in time; for example if we know that ticks could have between 1 or 2 days of attachment left then we would expect half of them to detach each day.

The model will run for a number of timesteps equal to the maximum on host attachment time of ticks, so for a timesteps.

Matrix formation

The matrix P_0 was created by regularly distributing points across a country, removing those that don't fall within a species' breeding grounds, and forming the m by n grid and counting the number of points for each cell. Figure 6.1 illustrates this for the UK.

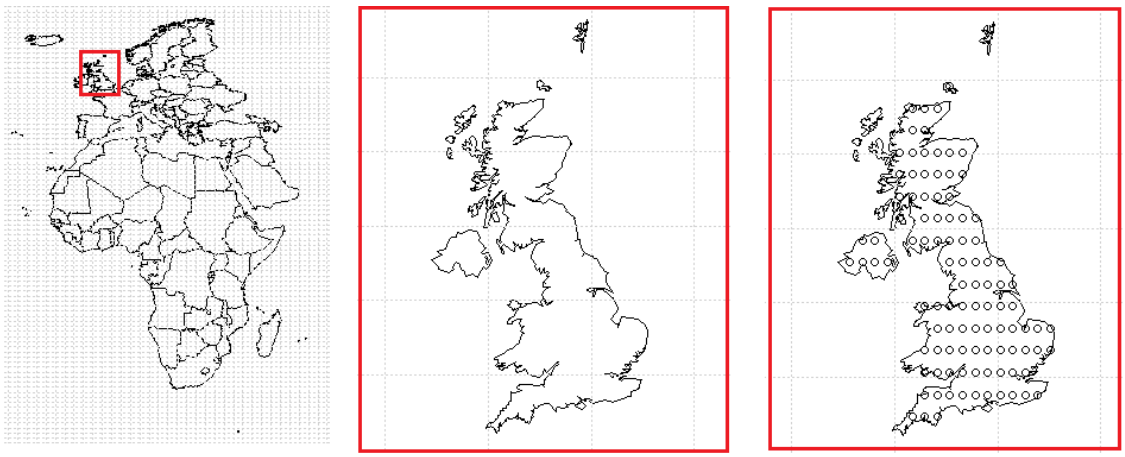

Figure 6.1: Illustration of point sampling approach for the UK that was used to derive values for the matrix P

The count for each cell was divided by the total number of points to give a percentage of the country that lay within a cell and this was multiplied by the total number of birds listed for that country taken from bird population data produced by Birdlife International (Burfield and van Bommel, 2004). Referring to the example in Figure 6.1, the cell containing Northern Ireland contains six percent (6 of 100) of the assigned points and so this cell will be given six percent of the bird population associated with the UK and Northern Ireland. This was done for each European country, and thus a $(m \times n)$ matrix was created for each country, and the results were summed to give a single matrix (P_0) describing the total breeding space for each bird population.

The second matrix (G), representing the general migration behaviour of birds to be used when there is no nearby breeding ground, was used to reflect two aspects of bird movement. Firstly, birds will generally have a northbound direction of movement and, secondly, they tend to avoid open water. It was formed by having an increasing value as birds move north up the grid and any cell that is entirely water is initially given a value of zero to represent the difficulty in crossing this area.

To form the increasing values the simple approach of counting down from m for each row (see below) was not possible as, when calculating the proportions of birds moving into a cell, it would be required to have a bird being equally likely to move

in a direction regardless of where it currently is and this would not happen with this approach. By this we mean that a bird should be equally likely to move north irrespective of where it is assuming there is no neighbouring population space and/or water. This problem can be illustrated by examining the below matrix:

$$G = \begin{bmatrix} 4 & 4 & 4 \\ 3 & 3 & 3 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$$

Then considering the case of a starting population of migratory birds in cell 2,2 ($B_{0,A}$) compared to cell 3,2 ($B_{0,B}$).

$$B_{0,A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} B_{0,B} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

In this case, the proportion of birds in cell $B_{0,A}[2,2]$ that would move directly north would be the value associated with cell G[1,2] divided by the sum of the gradient value for all neighbouring cells (the same approach used in the example after equation 6.5). This would need to be equal to that for $B_{0,B}$ which would be the value for cell G[2,2] divided by the sum of all neighbouring cells:

$$\frac{4}{3*4+2*3+3*2} \neq \frac{3}{3*3+2*2+3*1} \tag{6.7}$$

$$\frac{1}{6} \neq \frac{3}{16}$$

This is clearly not true and the resulting $B_{1,A}$ and $B_{1,B}$ show this difference:

$$B_{0,A} = \begin{bmatrix} 17 & 17 & 17 \\ 13 & 0 & 13 \\ 8 & 8 & 8 \\ 0 & 0 & 0 \end{bmatrix} B_{0,B} = \begin{bmatrix} 0 & 0 & 0 \\ 19 & 19 & 19 \\ 13 & 0 & 13 \\ 6 & 6 & 6 \end{bmatrix}$$

In both cases, as discussed earlier, there is a slight change in the total number of migratory birds present in B due to rounding error; meaning that there is an additional bird in each matrix compared to the starting population.

If we use i to represent the row number and a scalar c to represent the proportion of birds that will move directly north then:

$$\frac{m+1-i+1}{3(m+1-i+1)+2(m+1-i)+3(m+1-i-1)} = c \tag{6.8}$$

Equation (6.8) is the ratio of the value of G in the cell directly north against that of all surrounding cells minus three times the value in the row above (for north, northwest and northeast), two times the value for the current row (for east and west) and three times the value of the row below (for south, southeast and southwest).

To simplify equation (6.8) the value $m+1-i$ can be replaced with k and it can be rearranged to give:

$$\begin{aligned} \frac{k+1}{8k} &= c \\ \Rightarrow \frac{1}{8}(1+k^{-1}) &= c \end{aligned} \tag{6.9}$$

In this case, c clearly decreases as k increases; this would mean that a bird's preference to move north would decrease the further north it moves and this would not be the case. An alternative would be a linearly increasing exponentiation with a fixed base and powers increasing by one as we move up the row, taking the same example as above and using a base of two we have:

$$G = \begin{bmatrix} 2^4 & 2^4 & 2^4 \\ 2^3 & 2^3 & 2^3 \\ 2^2 & 2^2 & 2^2 \\ 2^1 & 2^1 & 2^2 \end{bmatrix}$$

In this case, the proportion of birds in cell $B_{0,A}[2,2]$ that would move directly north would again be the value associated with cell G[1,2] divided by the sum of the gradient value for all neighbouring cells. This would need to be equal to that for $B_{0,B}$ which would be the value for cell G[2,2] divided by the sum of all neighbouring cells:

$$\frac{16}{3*16+2*8+3*4} = \frac{8}{3*8+2*4+3*2} \tag{6.10}$$

$$\frac{8}{38} = \frac{8}{38}$$

In this case, we satisfy this requirement and the resulting $B_{1,A}$ and $B_{1,B}$ show this:

$$B_{0,A} = \begin{bmatrix} 21 & 21 & 21 \\ 11 & 0 & 11 \\ 5 & 5 & 5 \\ 0 & 0 & 0 \end{bmatrix} B_{0,B} = \begin{bmatrix} 0 & 0 & 0 \\ 21 & 21 & 21 \\ 11 & 0 & 11 \\ 5 & 5 & 5 \end{bmatrix}$$

Again using i to represent the row number and c to represent the proportion of birds that will move north and using b to represent the base we can write:

$$\frac{b^{i+1}}{3b^{m+1-i+1} + 2b^{m+1-i} + 3b^{m+1-i-1}} = c \tag{6.11}$$

Simplifying by replacing the term $m + 1 - i$ with k:

$$\Rightarrow \frac{b^{k+1}}{3b^{k+1} + 2b^k + 3b^{k-1}} = c$$

$$\Rightarrow \frac{b^k * b}{3b^k * b + 2b^k + 3b^k * b^{-1}} = c \tag{6.12}$$

$$\Rightarrow \frac{b}{3b + 2 + 3b^{-1}} = c$$

In this case a bird is equally likely to move in a northerly direction independent of its location; this is proved by the fact that equation (6.12) is independent of row numbers in its final form. Since the proportion moving north is now independent of location as in equation (6.12) and increases with b then we can determine how much of the bird population of a cell will move north. If, for example, we desire 25% of a cell population to move north then c=0.25:

$$\frac{b}{3b + 2 + 3b^{-1}} = 0.25$$

$$b^2 - 2b - 3 = 0 \tag{6.13}$$

$$b = -1, 3$$

To avoid oscillating values (values moving between positive and negative) in matrix G caused by having a value of b<0 then the positive solution for b would always be chosen.

To find suitable values for c, papers on bird orientation were used. A wrapped Normal distribution with a fixed mean (chosen to represent north) and a level of standard deviation ($\sigma_{orientation}$) taken from Batschelet (1981) were used for the baseline value and from Erni et al. (2003) for a stronger level of orientation. This distribution was sampled from and the cosine and sine of the resulting angles were calculated to give the equivalent movement in the x and y planes. These were then rounded to the

nearest integer to represent the cell that each bird would move to. The proportion that moved north (c) could then be found and be used in equation (6.12).

The initial starting positions of the birds, B_0, were based upon maps taken from Birdlife International (BirdLife International, 2012b). Firstly, a subset of African countries was taken, then spatial points were randomly distributed amongst them. To minimise computing time, 2000 points were taken and these were used to give a proportion of each species for each grid cell before birds were assigned proportionally to these starting grid locations.

Parameter estimates

Many of the values used in chapter 4 were used again here, in particular the migratory speeds and the population number per country for each bird species. However, the prevalence rate of CCHFV positive ticks on migrating birds was informed by the work in chapter 5.

The orientation of migrating birds, which is required for the formation of the matrix G, were taken from the sources discussed (Batschelet, 1981; Erni et al., 2003).

Table 6.1: Estimates of parameter values

Parameter	Description	Value(s)	Reference
v_{Quail}	Average distance (km) covered per day during spring migration by the common quail	150, 160	(Perennou, 2011)
$v_{Warbler}$	Average distance (km) covered per day during spring migration by the willow warbler	62, 114	(Payevsky, 2013)
$v_{Wheatear}$	Average distance (km) covered per day during spring migration by the northern wheatear	110, 147	(Payevsky, 2013)
v_{Pipit}	Average distance (km) covered per day during spring migration by the tree pipit	57, 106	(Payevsky, 2013)
$v_{Bunting}$	Average distance (km) covered per day during spring migration by the ortolan bunting	181, 243	(Payevsky, 2013)
a	Tick Length of Attachment	26	(Gale et al., 2011)
μ_{tick}	Mean number of ticks on migrating birds	0.049	(Gale et al., 2011)
ρ_{prev}	Prevalence of CCHFV in *Hyalomma* ticks	0.0001, 0.0007528	(Gale et al., 2011), (Lindeborg et al., 2012), (Palomar et al., 2013)
$\sigma_{orientation}$	Standard deviation of direction of bird migration	54°, 30°	(Batschelet, 1981), (Erni et al., 2003)

Number of runs

Since the model can be sensitive to the random initial distribution of the migrating birds, then the simulation will have to be run multiple times. To determine the number of runs required for convergence, the model was run 50 times using the minimum values from Table 6.1 and the standard error of the number of ticks deposited in Europe was taken across increasing numbers of runs (Figure 6.2). The standard error continues decreasing as the number of runs increases but it decreases at a greater rate initially, up to approximately 40 runs. Despite the fact that more runs always leads to more accurate results, as our results converge to the real value, this has to be balanced against efficiency in producing these results and so a value of 50 runs was selected.

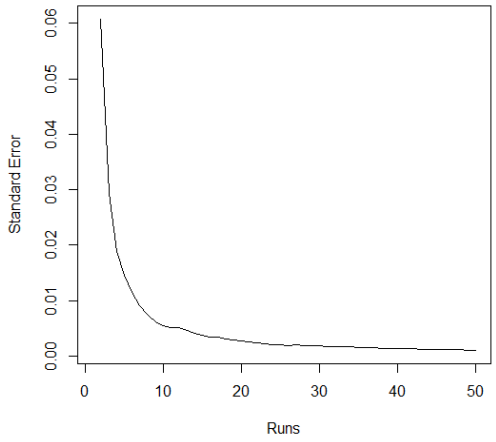

Figure 6.2: Standard error for number of introduced CCHFV positive ticks into Europe for all bird species under standard parameters

6.2.3 Scenarios

To investigate the sensitivity of the model to its parameters, a number of distinct scenarios were considered. Firstly, the model was run with the baseline parameters (Table 6.1) and the number of CCHFV infected ticks per country was calculated for each run. This is the baseline scenario and was referred to as scenario one. The

values used are the lower of those in Table 4.5 and those in Table 6.1.

This was then repeated with an increase in tick prevalence to the value calculated in Chapter 5 to see if an increase in CCHFV prevalence amongst ticks had an affect on the risk of the virus being imported into European countries via migrating birds. This was referred to as scenario two. It is worth noting, when interpreting results, that because of how the model works, this change is also equivalent to a proportional increase in the mean number of ticks per bird ($\mu_{bird\ species}$).

The model was run with the values from Table 6.1 and the higher population estimates from Table 4.5 which was referred to as scenario three. Scenario four used the lower population estimates (Table 4.5) and increased estimates of migration speeds but with all other parameters remaining unchanged (Tables 6.1).

The final scenario, scenario five, used a higher level of orientation amongst migrating birds. This is reflected by a lower level of standard deviation of their direction of migration (the smaller value of $\sigma_{orientation}$ in Table 6.1).

6.3 Results

6.3.1 CCHFV Incursions to all of Europe

Examining the number of ticks being introduced into Europe we can plot the expected numbers under each of our scenarios (Figure 6.3):

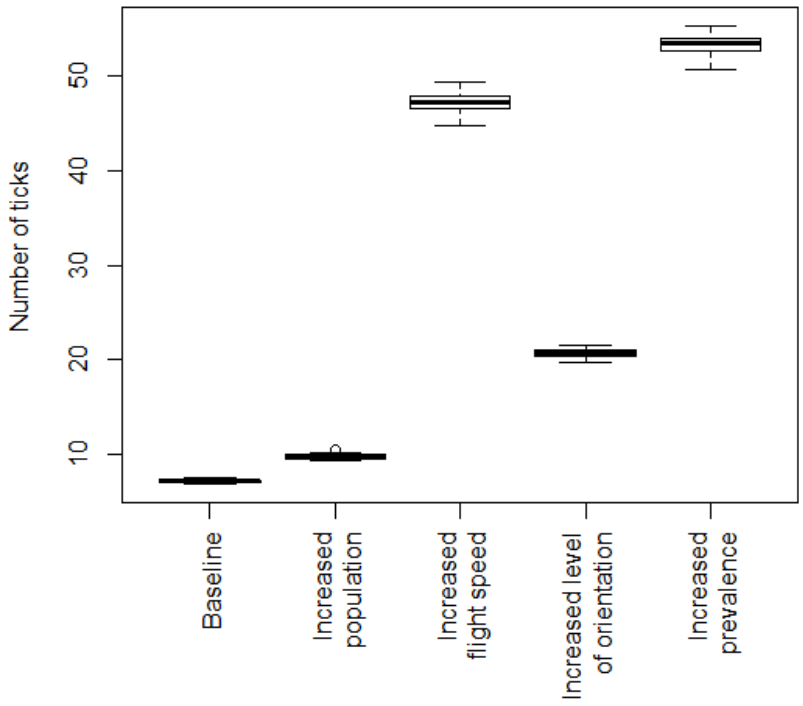

Figure 6.3: Boxplot of number of ticks introduced into Europe by all bird species under each scenario

As was found in Chapter 4, the level of sensitivity to changes in prevalence is greater than for the other parameters. From Figure 6.3, it can be seen that the level of variation in risk and the overall level of risk represented by the increased level of CCHFV prevalence is the highest for our scenarios. In contrast to the results in Chapter 4, there is not as marked a difference between the sensitivity of the model to this parameter and the other parameters notably increasing the speed of migration. In addition, it can be seen that, as expected, the baseline scenario represents the lowest risk as well as the highest level of consistency in results, but a small number of CCHFV positive ticks are still being introduced into Europe. The level of risk increases slightly with an increase in bird populations, but is more sensitive to a change in orientation.

Summary statistics of these results can be found in Table 6.2, ordered by increasing mean, which illustrate the much higher risk and level of variance present in scenario 2.

Table 6.2: Summary statistics of number of ticks introduced into Europe under each scenario

Scenario title (scenario number)	Mean	Standard deviation	Lower confidence interval	Upper confidence interval
Baseline (1)	7.22	0.15	6.94	7.51
Increased Population (3)	9.81	0.26	9.3	10.31
Increased Orientation (5)	20.67	0.39	19.9	21.43
Increased Flight Speed (4)	47.20	1.09	45.06	49.33
Increased Prevalence (2)	53.25	1.12	51.05	55.44

To inspect whether the geographic distribution of risk changes between scenarios, plots can be produced (Figure 6.4). Risk is very much not uniformly distributed amongst the at risk countries, and plotting as a heat map, where a country's colour was determined by the number of ticks introduced, resulted in only a few countries being visibly different from zero. Instead, in Figure 6.4, to stop a single country's risk meaning the risk for others is almost indistinguishable from no risk, then ordinal rankings of the mean number of CCHFV introduced ticks will be used to determine a country's colour.

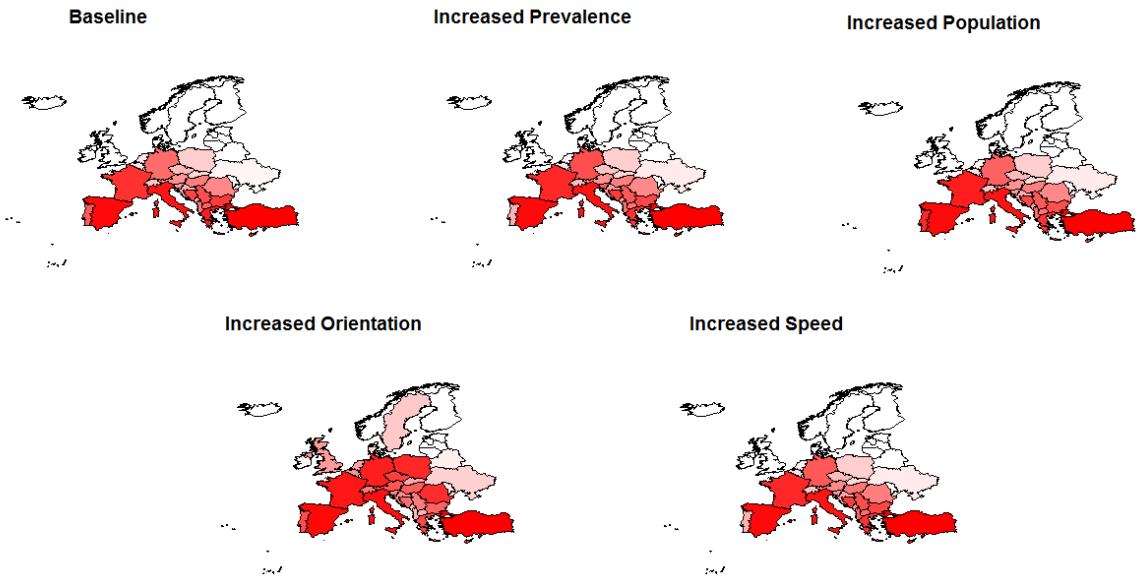

Figure 6.4: Geographic map of risk by rank; most at risk country (red) shading to no risk (white)

While there are some slight differences in Figure 6.4, in particular France under the increased prevalence scenario, generally the geographic pattern of risk is fairly similar under each scenario.

6.3.2 CCHFV Incursions by Individual Bird Species

Means and standard deviations were derived for all of Europe for each bird species to investigate differences in sensitivity to parameters between species and to examine the varying level of risk represented by each. In Table 6.3 it can be seen that the Ortolan Bunting represents the greatest risk under all of the scenarios. In addition, under a number of the scenarios, some of the species represent no risk.

Table 6.3: Summary statistics of number of ticks introduced into Europe under each scenario for each bird species

	Baseline (1)		Increased Level of Orientation (5)		Increased Population (3)		Increased Prevalence (2)		Increased Migration Speeds (4)	
Species	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Ortolan Bunting	5.46	0.11	15.42	0.33	5.48	0.09	40.89	0.93	13.19	0.78
Tree Pipit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00
Common Quail	0.92	0.06	2.57	0.10	2.01	0.13	6.89	0.37	0.84	0.06
Willow Warbler	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01
Northern Wheatear	0.85	0.08	2.68	0.14	2.32	0.20	5.47	0.52	3.27	0.13

This again means that it is the bird species with the fastest migration flight speed rather than the greatest population that introduces the most CCHFV positive ticks. This is similar to the results in Chapter 4, likewise is the fact that the two most populous of the bird species, the Willow Warbler and the Tree Pipit, contribute almost nothing to the risk of CCHFV positive ticks being introduced into Europe and it is only when their migratory flight speed is increased that they introduce any CCHFV positive ticks at all.

6.3.3 CCHFV Incursions by Individual Country

As in Chapter 4, it is worth looking at the most at-risk countries for each case to see if there is a geographical area that is more at risk where surveillance efforts might be focused. Table 6.4 lists the five most at risk countries for CCHFV positive ticks being introduced by all bird species for each scenario.

Table 6.4: Mean number of ticks for top five most at risk countries for all bird species and for each scenario

Baseline (1)		Increased Population (3)		Increased Level of Orientation (5)		Increased Prevalence (2)		Increased Migration Speeds (4)	
Country	Mean	Country	Mean	Country	Mean	Country	Mean	Country	Mean
Turkey	5.25	Turkey	6.47	Turkey	13.30	Turkey	38.20	Turkey	35.66
Spain	0.95	Spain	2.05	Spain	2.82	Spain	7.40	Spain	4.27
Italy	0.42	Italy	0.56	Italy	1.23	Italy	3.39	Italy	3.14
Greece	0.14	France	0.15	France	0.77	Cyprus	1.08	Cyprus	1.08
Cyprus	0.12	Portugal	0.14	Germany	0.76	France	0.90	France	0.89

Examining Table 6.4, there are three countries that consistently occur although not always in the same rank; however, Turkey, the most at risk country under every scenario, has a much greater level of introduced CCHFV positive ticks than any other country. Plotting the results in Table 6.4 yields Figure 6.5 which shows that there is a rapid fall in risk after the most at-risk country.

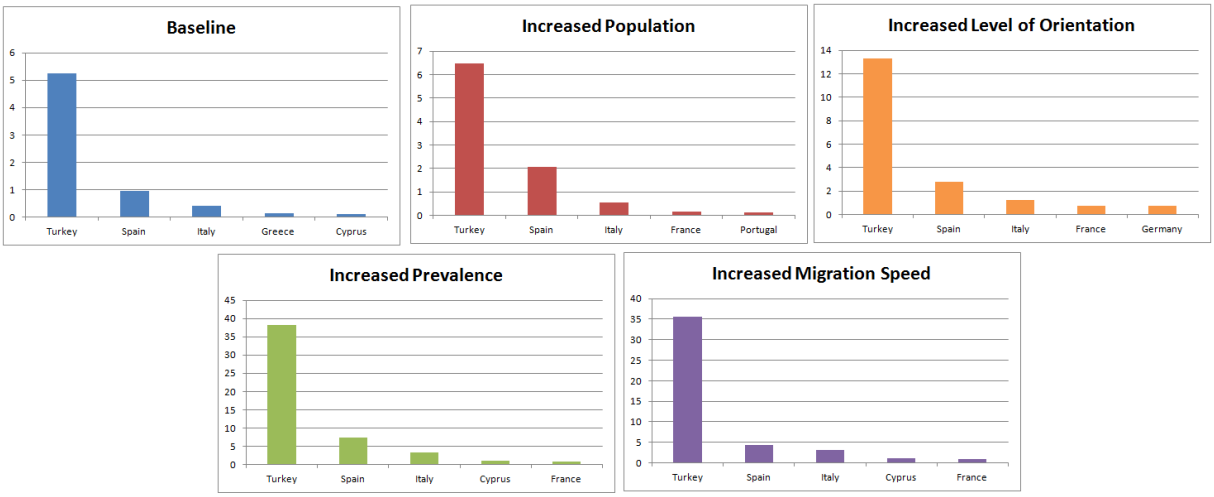

Figure 6.5: Bar charts of level of tick incursion for top five most at risk countries for all bird species and for each scenario

The scale of differences between the top five at risk countries seems to be the same for each scenario and to further test whether this is the case, the proportion of risk for each country based on only the top five most at-risk can be plotted (Figure 6.6).

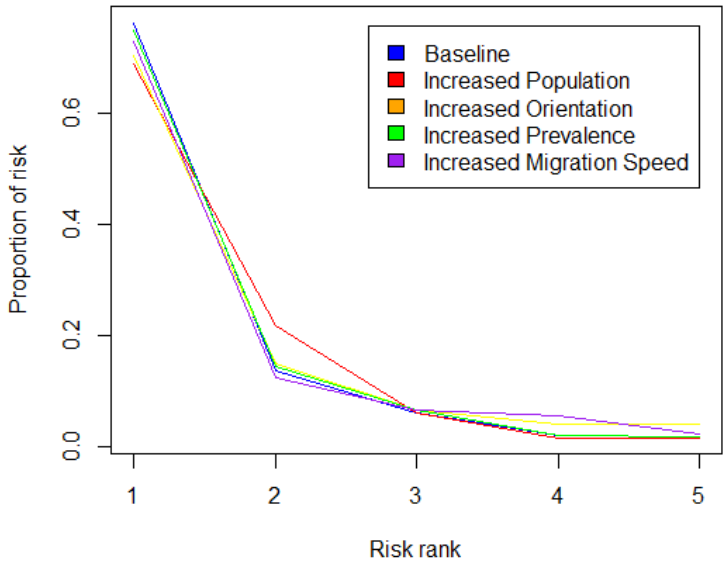

Figure 6.6: Proportions of level of tick incursion for top five most at risk countries for all bird species and for each scenario

This is the country's proportion of risk out of the total risk for the top five countries. Examining the plot, it would seem that changing the parameters does not cause much of a change as far as the proportion of risk is concerned and, under all scenarios, the top three most at risk countries are consistently the same. This could indicate that changing parameters makes very little difference to the geographical distribution of risk.

6.4 Discussion

In this chapter, a model was formed to examine the level of risk represented by migrating bird species introducing CCHFV positive ticks into Europe. After being run for the five most populous and at risk bird species, estimated numbers of CCHFV positive ticks were found for each country in Europe.

Figure 6.3 and Table 6.2 show that increasing the prevalence of CCHFV positive ticks, even by the smaller amount calculated in Chapter 5, still has far more impact

than increasing the other parameters and also results in greater variation of results. The scenarios that offer the second and third highest risk are an increased migration speed and an increased level of orientation. Both of these parameters have the effect of decreasing the number of timesteps required for a bird to reach Europe. This means that there is much the same pattern of sensitivity as there was for the GIS model (Table 4.7) where prevalence was found to have the greatest influence, followed by migration speed and with bird population numbers coming last.

The maps of risk of tick introduction (Figure 6.4) show a pattern of risk that is fairly consistent between scenarios and this is supported by Table 6.4 where under all five scenarios the top three most at-risk countries are the same. These countries are major parts of the main migration routes for birds flying between Africa and Europe so it would be expected that they be more at risk. The top two, also flagged up in Chapter 4, are both countries where CCHFV has been detected; in fact Turkey has had a large number of human cases over the last decade. For Spain, however, the presence of CCHFV has only recently been detected. It is possible then that the virus was introduced to these countries through bird migration. The repeated inclusion of Italy, where CCHFV is not believed to be present, means that this aggregated cellular automata model suggests a definite possibility of it being introduced.

While in Table 6.4 there are quite marked differences in risk between scenarios, the maps in Figure 6.4 and the proportion of risk between countries in Figure 6.6 suggest the proportional distribution of risk might be fairly similar between all five scenarios.

The risk represented by the different species in the cellular automata shows that the Ortolan Bunting represents by far the greatest level of risk (Table 6.3). Of the other species, two of them, the Tree Pipit and the Willow Warbler, embody no risk under the majority of scenarios and the two remaining species, the Common Quail and the Northern Wheatear, have means and standard deviations such that there can not be said to be a significant difference in the risk they represent. The sole scenario that is an exception to this is when there is an increased migratory flight speed. Under this scenario, the Tree Pipit and Willow Warbler have a very small risk of tick introduction, and the Northern Wheatear has a noticeably higher level

of risk than the Common Quail. For the Pipit and Warbler, this will be due to them now having the necessary flight range to reach Europe within the timeframe of tick attachment as both have quite a large increase in speed compared to the baseline scenario (over 80% above their baseline speed). The larger gap between the Northern Wheatear and Common Quail will be due in part to the fact that, of all five species, the Common Quail has the smallest proportional increase in speed, so increasing migration speed has little effect. This can be seen by comparing the results of the baseline and increased speed scenario for the Quail where the mean and standard deviation suggest there is no significant difference between the two.

In Chapter 4, full distributions were found and sampled from for each species and country in order to arrive at the final results. This means that the scenarios in this chapter are not directly comparable to those results. However, the minima found in chapter 4 should be approximate to the baseline scenario. Comparing the baseline results in Table 6.2 to those of Tables 4.8 and 4.6 or Table 6.4 to Table 4.10 show the number of expected CCHFV positive tick incursions are much lower under this model. This suggests modelling the spatial element explicitly has made a difference to the estimated risk of CCHFV positive ticks being introduced into Europe.

As far as the model is concerned, the results flagging up countries where the virus has already been found goes some way towards validating the results. However, this approach has a couple of flaws, the main one of which is behavioural; birds do not simply move northwards and settle in the first region they come across that has available space. To properly reflect this, a model would have to include a form of settling behaviour and this will be examined in the next chapter. This would affect the routes taken by birds and would have the additional affect of meaning that some of the birds who start in the northern most parts of sub-Saharan Africa will not stop in southern most Europe despite being the first birds to reach this area but will instead continue on towards their traditional nesting ground. This could have the possibility of increasing the range of ticks.

Chapter 7

Individual agent level cellular automata for the risk of incursion of CCHFV via migratory birds

7.1 Introduction

In the previous chapter, a spatially explicit model was used to simulate and estimate the risk posed by migrating birds introducing CCHFV virus positive ticks into Europe from Africa. While it was an extension of the previous simple model, it still had some areas which could be expanded.

In this chapter, an individual agent based cellular automata model will be developed. This uses a spatial grid as in the previous model but each bird, its movement and attached ticks are modeled independently. This allows a more realistic level of bird behaviour to be modeled; in particular, birds will no longer necessarily settle in the first available space and this will affect where CCHFV positive ticks will be introduced.

7.2 Methods

7.2.1 Selection of Bird Species

As in the GIS (Chapter 4) and aggregated cellular automata (Chapter 6), the bird species considered were the common quail, the ortolan bunting, the tree pipit, the willow warbler and the northern wheatear.

7.2.2 Cellular Automata

As in Chapter 6, a cellular automata type model is used to model the movement of birds and estimate the risk it represents. However, in this case, rather than model the population per grid, each agent (in this case a bird) will be modeled individually as it moves across the grid. The model describes the movement of birds from Africa to Europe over discrete time steps of length one day and will be implemented in the statistical programming language R. Again, as in the previous model, the geographical area including Africa and Europe is represented by a $m\ x\ n$ lattice with each cell on the lattice covering an area based on the average daily speed of a particular bird species; e.g for the Common Quail that has an average daily speed of 150km per day then each cell is 150x150km^2 as opposed to the Willow Warbler where the average daily speed is only 62km per day and so cells would be 62x62km^2.

The location of each bird and tick will be recorded as an individual so, unlike in the previous model, there will be no matrices B_t or T_t, instead, arrays (T and B) will be used to record information about each individual tick and bird, such as its position and number of days of attachment. However, it still makes use of a number of matrices all of size m x n. Firstly, there is the matrix P which represents the proportion of available space for breeding populations for each cell of the lattice. Each cell $P[x, y]$ will contain a number describing the breeding space of that cell as a proportion of the total breeding space with y being equivalent to a longitudinal position and x a latitudinal one. In order to cover the required geographical region these will vary from longitude -15° (the leftmost edge of y=1) up to 57.78° (the rightmost edge of y=n) and from latitude -46.98° (the bottommost edge of x=1) up

to 71.16° (the topmost edge of x=m). This matrix will be used to determine the settling behaviour of migratory birds; i.e. when they cease to migrate.

The dynamics of the lattice model, that is, the movement of the birds, are governed by an $m\ x\ n$ matrix G. As before, this will represent the general northbound direction of movement (during spring migration) and the tendency of migratory birds to avoid open water and will take the same form as in Chapter 6.

Since only those birds who are host to a CCHFV positive tick are of interest then only these birds are modeled.

Behaviour

Ticks and birds

As each bird can carry multiple ticks a negative binomial distribution is used to calculate the number of CCHFV positive ticks on each bird; this distribution is often used to model the number of parasites on a host (Bliss and Fisher, 1953; Pena-Rehbein, 2013; Shaw et al., 1998) selected as it is a discrete distribution with an unbounded positive range where the sample variance is often much larger than the sample mean. This can be sampled from in the R programming language with a number of failures set as 1 and the distribution mean set as the mean number of ticks per bird (μ_{tick}) multiplied by the prevalence of CCHFV amongst ticks (ρ_{prev}). A number of values equal to the total population size of a bird species will be drawn from this distribution and all those greater than one shall be kept and these will be the birds modeled. For identification purposes each of these birds will have a unique reference number generated.

The status of all CCHFV positive ticks will be held in an array (T) that will keep a record of the number of days on host the tick still has, the reference number for the bird the tick is attached to and the coordinates of the cell the bird is in when it either stops migrating or the ticks on host attachment reaches zero.

A similar array for migratory birds (B) will record their reference number, the cur-

rent coordinates of the cell the bird is passing through and, finally, a value describing the length of the bird's migration.

Matrix G and bird movement

Movement of the in flight bird population is determined by evaluating the values of all neighbouring cells in G. That is, all cells that share an edge or vertex, so all cells directly above, below or beside and also those diagonally touching. This is known as the Moore neighbourhood. This means the majority of cells will have eight neighbours.

These values will be transformed to give the proportion of the neighbourhood total and will represent the probability of moving in that particular direction. For example, if we consider the following simple 3x3 matrix, where north is to the top of the matrix:

$$G = \begin{bmatrix} 3 & 3 & 3 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$$

If we consider a bird who is currently in the centre cell ($G[2,2]$) then the neighbourhood total will be 16 $(3+3+3+2+2+1+1+1)$ and the probability for each direction will be the value of the cell divided by this; as shown in Table 7.1.

Table 7.1: Probability of movement for each direction

Direction	Probability
North-West	0.1875
North	0.1875
North-East	0.1875
East	0.125
South-East	0.0625
South	0.0625
South-West	0.0625
West	0.125

These probabilities can be calculated for each individual bird based on the directions of movement available to them; i.e. for geographical reasons Eastern movement might be impossible so the probability will be zero. A direction can then be randomly sampled from this categorical distribution for each bird with the outcomes being the possible directions and with the relevant probabilities being those calculated for that bird's current position, and this will be the direction the bird moves in that timestep.

The values of the matrix G will be selected so that there is a greater probability of the birds moving north; in Table 7.1, the probability regions for northerly directions are greater than those for southerly directions.

Matrix P and bird settling behaviour

Bird settling behaviour will be based around two factors the matrix P and the value describing the length of a bird's migration stored in the array B. Each cell of P will reflect the proportion of total breeding space contained within the geographical region represented by that cell. This means the sum of all values of P will be equal to one.

The value for the length of a bird's migration will be randomly sampled from the uniform distribution to give a probability between zero and one with equal likelihood.

As a bird passes through a cell (x,y), the value of $P[x, y]$ shall be subtracted from this value. Once this value ceases to be greater than zero then the bird will stop migrating and the current x and y coordinates will be recorded against the tick or ticks attached to the bird. This means the likelihood of settling will increase as the bird passes through Europe and that in geographical areas where a large proportion of the total population is known to be breed a bird is more likely to cease migrating. Additionally, it also resolves one of the weaknesses identified in the model from Chapter 6 in that birds will not stop in the first suitable region.

The model will run for a number of timesteps equal to the maximum on host attachment time of ticks; so for a timesteps.

Matrix formation

The initial steps for the creation of matrix P were as described for the aggregated cellular automata approach (Chapter 6). Points are regularly distributed across a country, those that don't fall within a species' breeding grounds are removed. An m by n grid is formed and counting the number of points for each cell and dividing by the total number of points gives a percentage of the country that lay within a cell and this was multiplied by the total number of birds listed for that country. This was done for each European country and thus a $(m \times n)$ matrix was created for each country, and the results were summed to give a single matrix (P) describing the total breeding space for the bird population. Each value of P is then divided by the sum of all P.

The second matrix (G) was produced in an identical manner to that in Chapter 6 with increasing values moving from the bottom to the top of the matrix (i.e. northwards) and with zero values for cells that contain only open water.

The initial starting positions of the birds were based upon maps of wintering locations taken from Birdlife International (BirdLife International, 2012b). To minimise computing time, 2000 points were randomly sampled from the wintering grounds for each species and these were used to produce a list of starting cells which was

randomly sampled from for each bird.

Parameter estimates

All parameter values were identical to those in Chapter 6 Table 6.1.

Number of runs

As in Chapter 6, since there is randomness present in the model, then the simulation will have to have multiple runs. To determine the number of runs required for convergence, the model was run 100 times and the standard error of the number of ticks deposited in Spain (chosen as a country that consistently had ticks deposited and so non-zero data) was taken across increasing numbers of runs. The standard error continues decreasing as the number of runs increases but it decreases at a greater rate initially, up to approximately 50 runs and so a value of 50 runs was selected for the production of results.

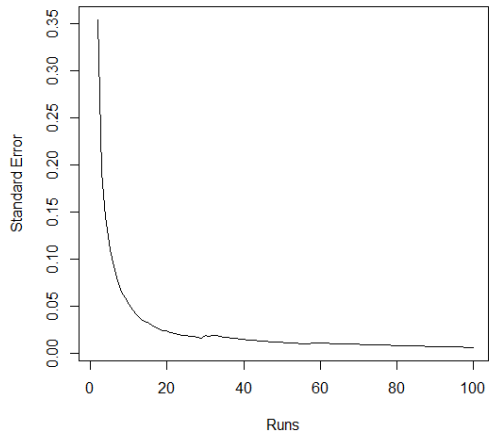

Figure 7.1: Standard error for number of ticks deposited in Spain under standard parameter values

7.2.3 Scenarios

To investigate the sensitivity of the model to its parameters, five distinct scenarios were considered. These scenarios were identical to those used in Chapter 6.

The first scenario is the baseline scenario and makes use of the lower parameter estimates in Table 4.5 and those in Table 6.1 and should represent the minimum expected level of risk of CCHFV introduction. As before, scenario two replaced the baseline prevalence of CCHFV amongst ticks with that calculated in Chapter 5 to investigate the affect of an increase in CCHFV prevalence amongst ticks.

The model was then run with the higher population estimates from Table 4.5 (scenario three); scenario four returned to the lower population estimates (Table 4.5) but with increased estimates of migration speeds. The final scenario, scenario five, used a higher level of orientation amongst migrating birds.

Since only those birds carrying a CCHFV positive tick will be modelled then the number of birds modelled will vary between the scenarios, as either the total bird population or the prevalence rate changes, the expected number of each species modelled under each scenario are summarised in Table 7.2.

Table 7.2: Expected number of birds, rounded to nearest integer, carrying CCHFV ticks to be modelled under each scenario

Species	Baseline (1), Increased Flight Speed (4), Increased Orientation (5)	Increased Prevalence (2)	Increased Population (3)
Willow Warbler	453	3410	889
Tree Pipit	170	1277	316
Northern Wheatear	57	428	176
Common Quail	18	138	38
Ortolan Bunting	54	404	163

7.3 Results

Unlike in the aggregated cellular automata, results for this model will be restricted to integer values and this restriction will affect the appearance and shape of the results. To determine the best way of summarising these results, that is, whether presenting a mean and standard deviation was appropriate, the normality of the data was inspected. A histogram of the results for Spain with a fitted Normal curve can be seen in Figure 7.2. The Shapiro Wilk normality test was also carried out, returning a p-value of $p < 0.001$.

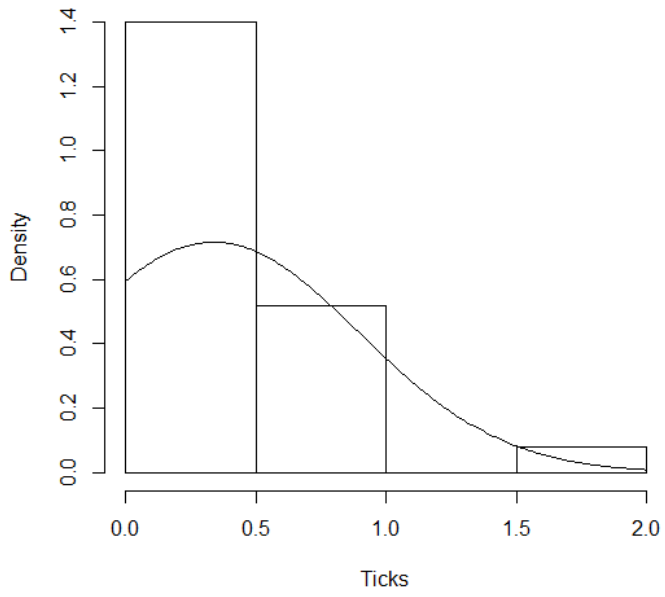

Figure 7.2: Histogram of level of tick incursion for Spain under standard parameters and average population numbers

In Figure 7.2, there very little fit between the data and the proposed Normal distribution and since the p-value from the Shapiro Wilk test is extremely significant then it is safer to assume results are not Normally distributed. Therefore, a selection of statistics will be used to summarise results.

7.3.1 CCHFV Incursions to all of Europe

Examining the number of ticks being introduced into Europe, we can plot the expected numbers under each of our scenarios:

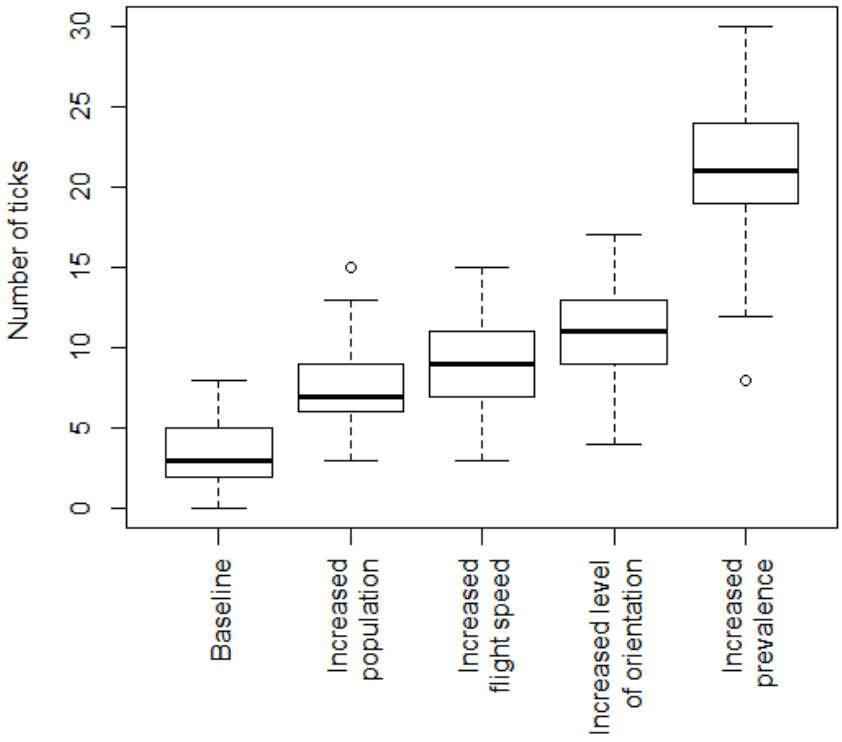

Figure 7.3: Boxplot of number of ticks introduced into Europe under each scenario

As was found in chapters 4 and 6, the level of sensitivity to changes in prevalence is greater than for the other parameters. From Figure 7.3, it can be seen that the level of variation in risk and the overall level of risk represented by the increased level of CCHFV prevalence is the highest for our scenarios. In addition, it can be seen that, as expected, the baseline scenario represents the lowest risk as well as the highest level of consistency in results, but a small number of CCHFV positive ticks are still being introduced into Europe. There is not much of a marked difference between scenarios 2, 3 and 4, indicating there may not be much of a difference in risk between these scenarios. This indicates that biological factors relating to the bird species themselves have very little impact on the risk of CCHFV positive tick introduction, and the sensitivity to each factor is similar across the estimated range of each. Maximising the speed or efficiency of migration or increasing to the maximum estimated population sizes has a similar affect on the number of CCHFV positive ticks being introduced into Europe. The single factor relating to the ticks directly, the prevalence of CCHFV, has a greater affect on the overall risk.

Summary statistics of these results can be found in Table 7.3, ordered by increasing median, which illustrate the much higher risk and level of variance present in scenario 2 with both its mean and median being almost twice those of the second most risky scenario.

Table 7.3: Summary statistics of number of ticks introduced into Europe under each scenario

Scenario title (number)	Mean	Standard deviation	Median	Range
Baseline (1)	3.6	2.05	3	8
Increased Orientation (5)	7.34	2.50	7	15
Increased Population (3)	8.9	3.05	9	15
Increased Flight Speed (4)	10.94	2.71	11	17
Increased Prevalence (2)	20.56	4.58	21	30

In Table 7.3, it is of note that almost all values of summary statistics increase as we move down the table. So, as both measures of average risk increase so too does the level of variation in risk.

To inspect whether the geographic distribution of risk changes between scenarios plots can be produced (Figure 7.4). Risk is very much not uniformly distributed amongst the at-risk countries and so to stop a single country's risk meaning the risk for others is almost indistinguishable from no risk then ordinal rankings will be used.

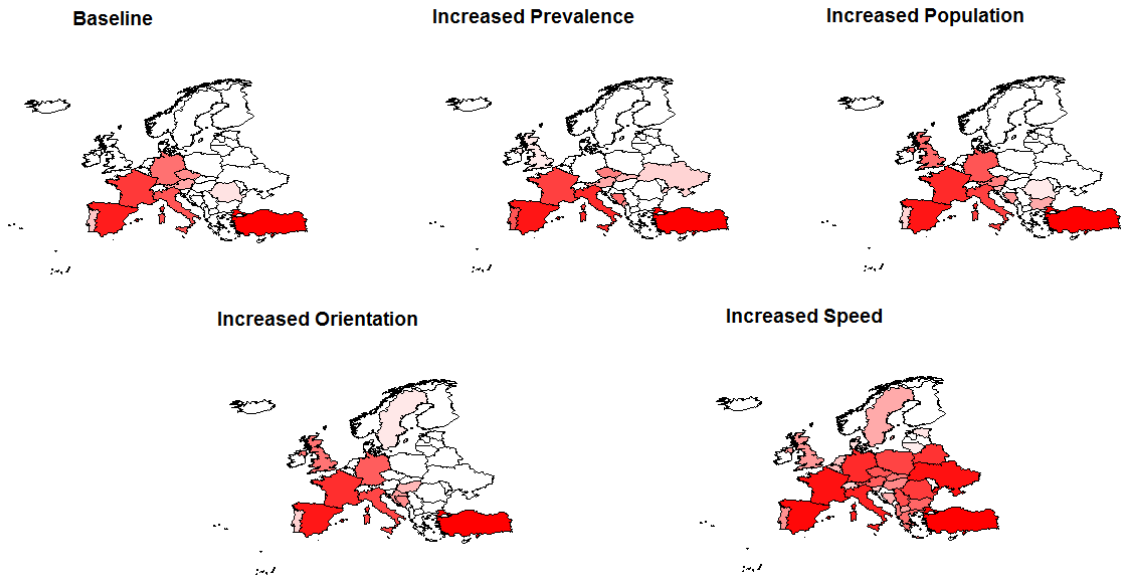

Figure 7.4: Geographic map of risk by rank; most at risk country (red) shading to no risk (white)

There are several slight differences in Figure 7.4 between scenarios 1,2,3 and 5, in particular for the UK and France. However, the most marked difference is the spatial distribution under increased migration speeds. Much more of Europe, including part of Scandinavia is at risk.

7.3.2 CCHFV Incursions by Individual Bird Species

Summary statistics can be derived for all of Europe for each bird species to investigate differences in sensitivity to parameters between species and to examine the varying level of risk represented by each. In Table 7.4, the results suggest that some of the bird species are of very little importance in introducing CCHFV positive ticks. In particular, under all scenarios the willow warbler represents no risk.

Under scenario 4 the tree pipit has a low level of risk for introducing CCHFV positive ticks but represents no risk at its lower migration speed.

The common quail and northern wheatear both have a relatively low level of intro-

duction of CCHFV positive ticks under all scenarios and have a median value of never more than three between them.

The vast majority of introduced ticks are by the ortolan bunting and this holds true under all five scenarios. Under the baseline scenario, this is still, however, a fairly small number of introduced ticks.

Table 7.4: Summary statistics of number of ticks introduced into Europe under each scenario for each bird species

	Baseline (1)				Increased Population (3)				Increased Level of Orientation (5)			
Species	Mean	Standard deviation	Median	Range	Mean	Standard deviation	Median	Range	Mean	Standard deviation	Median	Range
Ortolan Bunting	2.50	1.61	2	5	4.74	2.26	5	9	6.74	2.61	7	10
Tree Pipit	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00	0	0
Common Quail	0.80	0.97	0	3	1.58	1.13	1	4	1.22	1.39	1	6
Willow Warbler	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00	0	0
Northern Wheatear	0.30	0.65	0	3	1.02	0.96	1	4	0.94	1.02	1	4
	Increased Prevalence (2)				Increased Migration Speeds (4)							
Species	Mean	Standard deviation	Median	Range	Mean	Standard deviation	Median	Range				
Ortolan Bunting	16.32	3.86	17.00	22.00	8.52	2.50	8.5	9				
Tree Pipit	0.00	0.00	0	0	0.02	0.14	0	1				
Common Quail	1.34	2.49	0	9	1.20	1.18	1	4				
Willow Warbler	0.00	0.00	0	0	0.00	0.00	0	0				
Northern Wheatear	2.90	1.58	3	6	1.20	1.14	1	4				

7.3.3 CCHFV Incursions by Individual Country

Table 7.5 lists the five most at risk countries for CCHFV positive ticks being introduced by all bird species for each scenario to allow us to examine the most at risk countries for each case to see if there is a geographical area that is more at risk where surveillance efforts might be focused.

Table 7.5: Mean number of ticks for top five most at risk countries for all bird species and for each scenario

Baseline (1)		Increased Population (3)		Increased Level of Orientation (5)		Increased Prevalence (2)		Increased Migration Speeds (4)	
Country	Mean	Country	Mean	Country	Mean	Country	Mean	Country	Mean
Turkey	2.3	Turkey	6.12	Turkey	4.38	Turkey	14.24	Turkey	4.46
Spain	0.66	Spain	1.52	Spain	1.26	Spain	3.4	Spain	1.72
France	0.28	France	0.58	France	0.98	Italy	1.46	Ukraine	0.98
Italy	0.18	Italy	0.34	Italy	0.48	France	1.2	France	0.8
Germany	0.08	Germany	0.08	Germany	0.08	Portugal	0.06	Italy	0.52

Examining Table 7.5, Turkey and Spain, consistently are the countries with the greatest incidences of introduced CCHFV positive ticks. There are two countries, France and Italy, that consistently occur although not always in the same rank. Turkey, the most at-risk country under every scenario, has a much greater level of introduced CCHFV positive ticks than any other country. Plotting the results in Table 7.5 yields Figure 7.5 which shows that there is a rapid decrease in risk after the most at-risk country under all five scenarios.

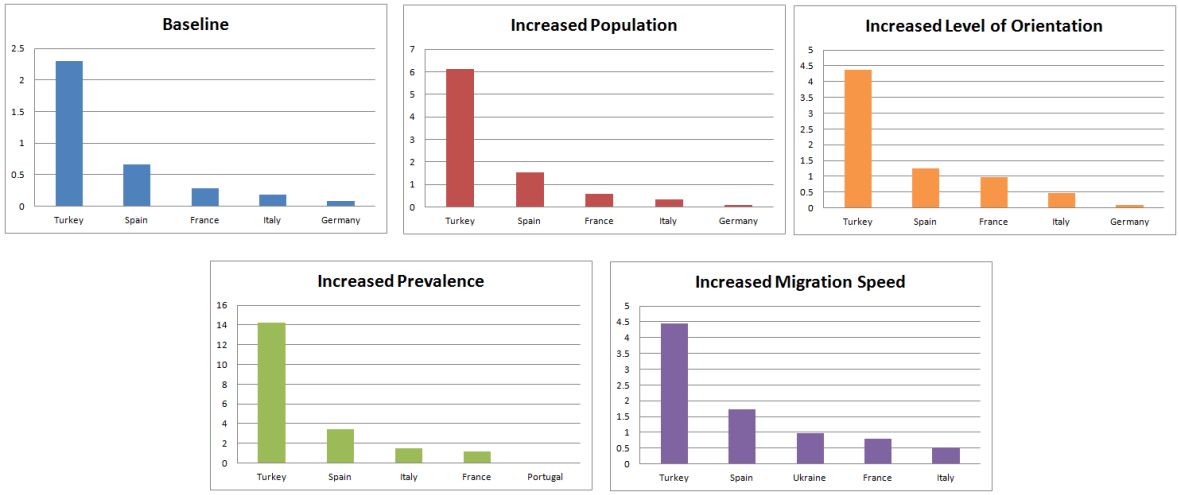

Figure 7.5: Histograms of level of tick incursion for top five most at risk countries for all bird species and for each scenario

As was seen in the comparison of the most at risk countries for the aggregated population cellular automata (Figure 6.5) the rate of reduction seems to be the same for each scenario and again this can be tested by plotting the proportion of risk for each country (Figure 7.6).

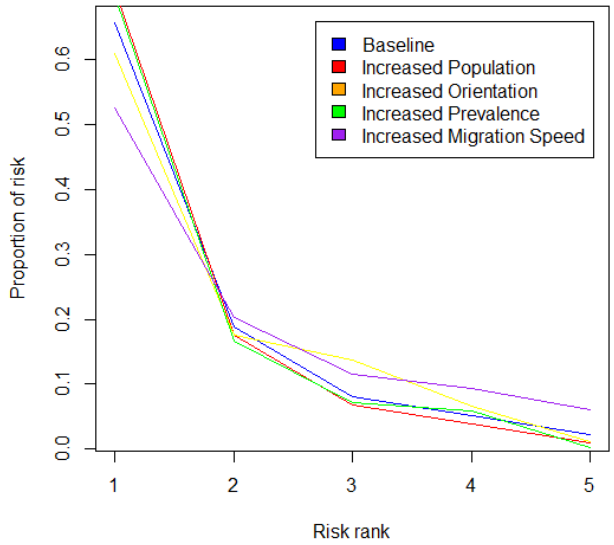

Figure 7.6: Proportions of level of tick incursion for top five most at risk countries for all bird species and for each scenario

Examining the plot, it would seem that changing the parameters does not cause much of a change. Comparing to the previous model (Figure 6.5) there does seem to be more variation between the proportion represented by the most at risk country for each scenario. Since the settling behaviour of migratory birds is no longer an automatic response then it is of additional interest whether the majority of ticks are deposited due to birds settling or total attachment time expiring and Table 7.6 gives the proportion of ticks that are deposited by settled birds per species. This is not the total proportions of ticks deposited in Europe but generally and since the results are a proportion then it wold be expected that the larger population and prevalence scenarios would be similar to the base as these simply increase the number of birds being modelled without affecting their behaviour. It is clear from Table 7.6 that total attachment time is a much more significant driver of the final location of CCHFV positive ticks given the very small proportion of birds who reach their breeding grounds before the total tick attachment time passes.

Table 7.6: Proportion of birds who reach their breeding ground before tick attachment time is complete

Species	Base Mean	Base Standard deviation	Increased Speed Mean	Increased Speed Standard deviation	Increased Orientation Mean	Increased Orientation Standard deviation
Common Quail	0.0084	0.023	0.0134	0.024	0.0113	0.022
Tree Pipit	0	0	0	0	0	0
Northern Wheatear	0.0004	0.003	0.0016	0.005	0.0015	0.005
Ortolan Bunting	0.0305	0.022	0.0796	0.012	0.0639	0.033
Willow Warbler	0	0	0	0	0	0

7.4 Discussion

In this chapter, an individual agent based model was formed to examine the level of risk represented by migrating bird species introducing CCHFV positive ticks into Europe. After being run for the five most populous and at-risk bird species, estimated numbers of CCHFV positive ticks were found for each country.

Examining the different scenarios, Figure 7.3 and Table 7.3, it is found again that increasing the prevalence of CCHFV positive ticks has a greater impact on incursions than increasing the other parameters and also results in greater variation of results. Again, as expected, the baseline scenario produces the lowest level of risk. Despite Normality not being assumed for results, there is not a large difference between the mean and median so the results don't appear to be overly skewed. The other three scenarios do not appear to be strongly different as there is a great deal of overlap in their ranges.

Comparing these results to the equivalent aggregated cellular automata results (Figure 6.3 and Table 6.2) the expected number of CCHFV positive ticks being introduced is noticeably lower and there is a greater level of variation in the results. This might be expected, as this model allows a greater range of behaviours and is not using an aggregate population approach to produce results and so behaviour would not be averaged out.

There is also less difference between scenarios when modeled from an individual agent perspective. That is, the ranges overlap for the majority of scenarios with only the baseline and increased prevalence being completely distinct, contrasted with the aggregate approach where all five scenarios are quite distinct. The pattern of scenario results, however, is also quite different; that is, for the aggregated approach the scenarios might be formed into three groups; (1) baseline and increased population, (2) increased level of orientation, (3) increased migratory flight speed and increased level of prevalence. In contrast, the larger jumps for the individual agent results are between baseline and the three middle scenarios and then between them and the top scenario. This suggests that the individual agent model is less in agreement with the

analytical sensitivities derived in Chapter 4, where the increased population should result in very little difference. This suggests that moving away from population models towards a more realistic individual agent model reveals potentially important dynamics in migration behaviour. While increasing population numbers results in more of a difference than in the population approaches, and so is more influential, it does not appear to have a markedly different affect on risk compared to changing the migratory flight speed or orientation.

The maps of risk of tick introduction (Figure 7.4) show a few different patterns of risk between scenarios and this is supported by Table 7.5, where, under all five scenarios, only the top two most at-risk countries are the same, although two of the other top countries are consistently present. These countries are major parts of the main migration routes for birds flying between Africa and Europe, so it would be expected that they be more at-risk and the top two were flagged up in Chapter 4 and 6. The greatest geographic range is associated with an increased migratory flight speed which makes intuitive sense

Again, as in the aggregated approach, there are quite marked differences in risk between the top countries for each of the scenarios (Table 7.5) but the proportion of risk between countries in Figure 7.6 suggests the proportional distribution of risk might be fairly similar between all five scenarios.

The risk represented by the different species shows that the Ortolan Bunting represents by far the greatest level of risk (Table 7.4). The Willow Warbler offers no risk of tick introduction even with an increased migration speed, in contrast to the aggregated cellular automata approach, where that scenario resulted in the species having a small level of risk. As before, increasing the flight speed and so the migration range is the only circumstance in which the Tree Pipit has a risk of entering Europe with CCHFV positive ticks still attached. Of the remaining two species, the Common Quail poses a similar level of risk to the Northern Wheatear; i.e. if mean and standard deviation is examined, there is no significant difference and medians are the same, except under an increased level of prevalence, in which the Northern Wheatear seems much more sensitive to the changed parameter. This may be due to the larger

population for the Wheatear and so an increased prevalence would result in a larger additional number of CCHFV positive tick carrying birds than it would for the Quail.

Across each species, the increased migration speed is either the second highest or in the case of the Pipit the highest level of risk; the exception to this, as in the aggregate cellular automata, is the Common Quail and again this is most likely due to the small proportional increase in flight speed that this scenario represents.

The results from the individual agent approach are repeatedly lower than in the GIS or aggregated cellular automata approach. Since the behaviours in this model can be argued to be more representative of bird behaviour, this difference in results would suggest that a general population approach even with explicit spatial modelling is not the most appropriate way to model this risk.

As in the previous chapter, the results flagging up countries where the virus is already present goes some way towards validating the results. However, this approach still contains some areas of potential improvement, the main one of which is behavioural; while birds no longer simply settle in the first available breeding space this model does not explicitly show the homing instinct birds have with birds returning to the same area to nest year after year. To properly reflect this, a model would have to include a form of homing behaviour and this will be examined in the next chapter. This will very much affect the routes taken by birds and to properly reflect this greater flexibility in movement will be needed and so continuous rather than discrete space will be used.

Chapter 8

Continuous space model for the risk of incursion of CCHFV via migratory birds

8.1 Introduction

In the previous chapter avian migration was modeled using an individual agent cellular automata approach. There are a number of weaknesses in this approach, mainly that bird movement, and more significantly settlement, is determined by carrying capacity and this is not realistic. The majority of bird species tend to return to the same breeding areas year after year (Dorst, 1962) and so modelling should reflect this with birds displaying a level of homing behaviour. Unlike the GIS model in Gale et al. (2011), many of the current models of bird migration produced by ornithologists are based in continuous space (Thorup and Rabøl, 2001; Mouritsen, 1998).

In mathematics, these kind of problems are often solved using diffusion type models or, as per the ornithological models, stochastically by making use of circular statistical distributions.

The aim of this chapter is to model bird migration in continuous two dimensional space and use this model to evaluate the risk of CCHFV positive ticks being imported into countries by migrating birds.

8.2 Continuous Space modelling

8.2.1 Particle Diffusion

Many simpler models of movement make use of random walks, the simplest of which has an individual moving along a linear space that is divided into sites labeled by integers, sites are taken as size 1 and steps are every timestep t. Moves between sites (Z) are assumed to be independent and decided by simple probabilities:

$$\begin{aligned} Pr(move\ from\ site\ i\ to\ i+1) &= p \\ Pr(move\ from\ site\ i\ to\ i-1) &= q = 1-p \end{aligned} \tag{8.1}$$

Therefore Z will be equal to one with probability p and to negative one with probability $1-p$. An individual's position at timestep t (X_t) will then be the sum of all previous steps (Z_1 to Z_t):

$$X_t = \sum_{j=1}^{t} Z_j \tag{8.2}$$

If we then replace our sites by a smaller scale, for example spaces of size δx, and take a smaller timestep, δt, then each step of this scaled random walk has a mean

$$\mu = p(+\delta x) + q(-\delta x) = (p-q)\delta x$$

and variance

$$\begin{aligned} \sigma^2 &= [p(+\delta x)^2 + q(-\delta x)^2] - \mu^2 \\ &= 4pq(\delta x)^2 \end{aligned} \tag{8.3}$$

Over a time period of length t there will be $t/\delta t$ separate timesteps. This means the individual's position $X(t)$ will have:

mean

$$\mu(t) = (p-q)t\delta x/\delta t$$

and variance

$$\sigma^2(t) = 4pqt(\delta x)^2/\delta t$$

As the size of the sites and timesteps decreases, as $\delta t \to 0$ and $\delta x \to 0$ then the mean and variance of $X(t)$ must remain finite for the process to make biological sense. If it is required for the limiting process to have a mean a and variance D^2 in unit time then it must be the case that:

$$\mu(1) = (p-q)\delta x/\delta t \to a$$

$$\sigma^2(t) = 4pq(\delta x)^2/\delta t \to D^2$$

This is satisfied by having $\delta x = D\sqrt{\delta t}$, $p = \frac{1}{2}(1 + a\sqrt{\delta t}/D)$ and $q = \frac{1}{2}(1 - a\sqrt{\delta t}/D)$.

The limiting process $X(t)$ then has a Normal distribution with a mean of at and variance D^2t of the form:

$$f(x;t) = (2\pi D^2 t)^{-\frac{1}{2}} exp\left(\frac{-(x-at)^2}{2D^2t}\right)\delta x \tag{8.4}$$

This can then be increased to cover two dimensional diffusion which makes use of the bivariate Normal distribution.

8.2.2 Simple Diffusion

The simplest form of the particle diffusion equation, in our two dimensions, simulates the dispersal of a large number of individuals from a central point and utilises the bivariate Normal distribution. In the case of no drift, i.e. no preferred direction of

motion, the equation has the following form:

$$f(x,y;t) = (2\pi D^2 t)^{-1} exp(-(x^2+y^2)/2D^2 t)$$
$$Var[X(1)] = Var[Y(1)] = D^2 \tag{8.5}$$

In many cases (i.e. in chemistry or physics) the diffusion term (D) will vary depending on the circumstances, so there will be a D_X and D_Y, but since in this case, birds should equally be able to fly in any direction then we simply have a single D. This symmetry in equation 8.5 allows us to make the following transformations:

$$x = rcos(\theta),\ y = rsin(\theta)\ and\ dxdy = rd\theta dr \tag{8.6}$$

If we then denote $b^2 = 2D^2$ we can create a polar coordinate form of equation 8.5:

$$\psi(r,\theta;t)d\theta dr = (\pi b^2 t)^{-1} exp(-r^2/b^2 t) r d\theta dr \tag{8.7}$$

Equation 8.7 would then allow birds to determine a direction of movement free from the restrictions of a grid and allows variation in the amount of distance traveled. However, if initially we allow only the direction of travel to vary, we can produce a model that relies only on a variable θ and this can be used to determine the horizontal (east-west) and vertical movement (north-south) of birds. If it is assumed that birds have no preferred direction, that is no homing instinct, then this would be modeled as follows:

$$\begin{aligned} \theta &\sim U[0,2\pi] = UC \\ X &= Cos(\theta) \\ Y &= Sin(\theta) \end{aligned} \tag{8.8}$$

This can be easily simulated by randomly sampling from θ, a circular uniform (UC) distribution, and moving our points the required horizontal and vertical distance. Doing this with 2000 individuals who all start at the origin and producing density plots we get:

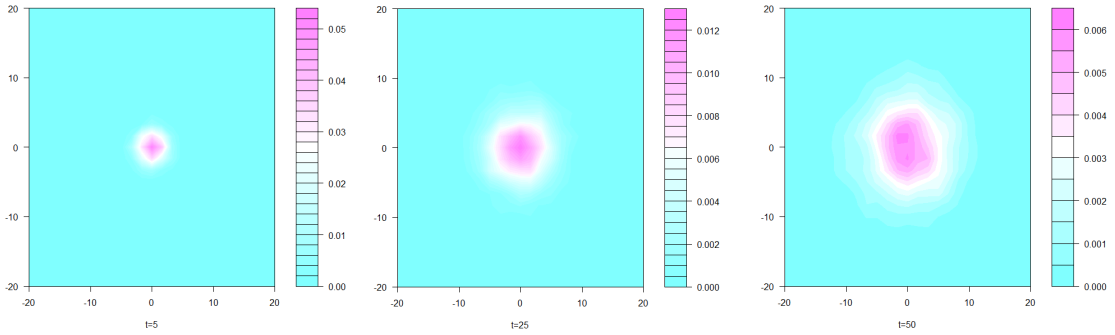

Figure 8.1: Density plot of individuals at t=5, 25, 50 under diffusion with no drift

In Figure 8.1 we can see that as time passes our individuals spread outward in a broadly symmetrical fashion and stay centered around the origin.

To examine the movement analytically though we would require the mean and variance of this distribution, and since the cosine (or sine) of a uniform distribution is not a standard distribution, this requires a transformation of the following form:

$$f_Z(z) = \left[\frac{d}{dz}g^{-1}(z)\right] f_W(g^{-1}(z))I_\eta(z) \tag{8.9}$$

Equation 8.9 allows us to find the probability distribution function for a transformed known distribution, in this case, the cosine of a uniform distribution. Taking each component of 8.9 and solving for our problem gives us equations 8.10 and 8.11.

$$\left[\frac{d}{dz}g^{-1}(z)\right] : \quad \left[\frac{d}{dx}cos^{-1}(x)\right] = \frac{1}{\sqrt{1-x^2}} \tag{8.10}$$

$$f_W(g^{-1}(z)) : \quad f_\theta(cos^{-1}(x)) = \frac{1}{2\pi} \tag{8.11}$$

This approach, however, is only applicable in the case of a one to one transformation and since across the interval 0 to 2π most values of H would occur twice then this is not applicable. If, however, our original variable, in our case θ, can be broken down into subsets so that a one to one transformation can be performed for each subset then we can replace equation 8.9 with equation 8.12 where the density is found by

summing across these subsets.

$$f_Z(z) = \sum \left[\frac{d}{dz} g_i^{-1}(z) \right] f_W(g_i^{-1}(z)) I_\eta(z) \tag{8.12}$$

Since our values between 0 and π and π and 2π are symmetrical for values of h we can simply double the results of our above component equations 8.10 and 8.11. This leaves us with a final probability distribution function for horizontal movement:

$$\begin{aligned} f_X(x) &= 2 * \frac{1}{2\pi} * \frac{1}{\sqrt{1-x^2}} \\ &= \frac{1}{\pi\sqrt{1-x^2}} \end{aligned} \tag{8.13}$$

We can then use this function to find analytical solutions for the descriptive statistics of horizontal movement for each time step:

$$\mu_X = \int_{-1}^{1} \frac{x}{\pi\sqrt{1-x^2}} dx = \left[-\frac{\sqrt{1-x^2}}{\pi} \right]_{-1}^{1} = 0 \tag{8.14}$$

$$\sigma_X^2 = \int_{-1}^{1} \frac{x^2}{\pi\sqrt{1-x^2}} dx = \left[\frac{sin^{-1}(x) - x\sqrt{1-x^2}}{2\pi} \right]_{-1}^{1} = \frac{1}{2} \tag{8.15}$$

These values then increase linearly with time. From Figures 8.1 we can clearly see that our individuals are centred around zero as predicted. For our variance, Figure 8.2, we can record the variance in the horizontal coordinates of individuals for each timestep:

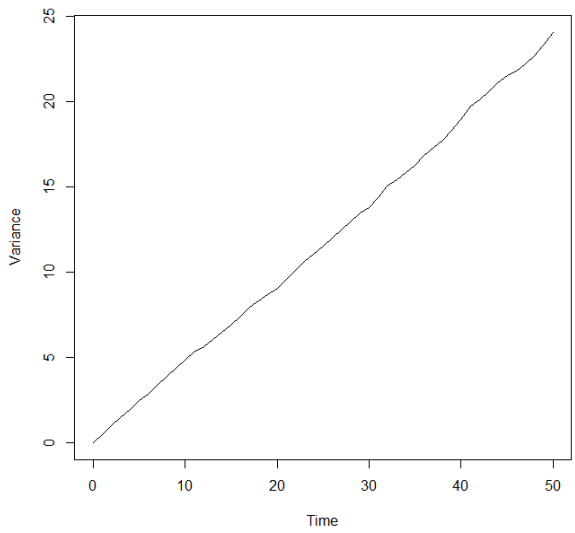

Figure 8.2: Plot of the variance in horizontal movement under diffusion with no drift against timestep

As can be seen, Figure 8.2 is very linear with a gradient equal to our values of σ_H^2

If we wish to investigate the distance $R(t)$, rather than just vertical or horizontal movement, of an individual from the origin at time t, we can integrate equation 8.7 over θ to obtain:

$$\begin{aligned} Pr(r \leq R(t) \leq r + dr) &= \int_0^{2\pi} (\pi b^2 t)^{-1} exp(-r^2/b^2 t) r d\theta dr \\ &= \frac{2r}{\pi b^2 t} exp(-r^2/b^2 t) dr \end{aligned} \tag{8.16}$$

If we then define $R(t)$ as the 'wavefront' of our model at time t, that is, the value at which we expect to find a single individual, from a population of N, at a distance further than $R(t)$ we have that:

$$\frac{1}{N} = \int_{R(t)}^{\inf} \frac{2r}{\pi b^2 t} exp(-r^2/b^2 t) dr \tag{8.17}$$

Equation 8.17 can be rearranged to give us a time dependent equation for the approximate maximum distance an individual can travel by timestep t, replacing b^2 with $2D^2$:

$$R(t) = \sqrt{2D^2\ t\ log(N)} \tag{8.18}$$

For our model without drift we found that $D^2 = \frac{1}{2}$ and so equation 8.18 becomes:

$$R(t) = \sqrt{t\ log(N)} \tag{8.19}$$

For our simulation the maximum Euclidean distance of any individual from the origin at each timestep can be recorded and these points plotted against the estimated $R(t)$:

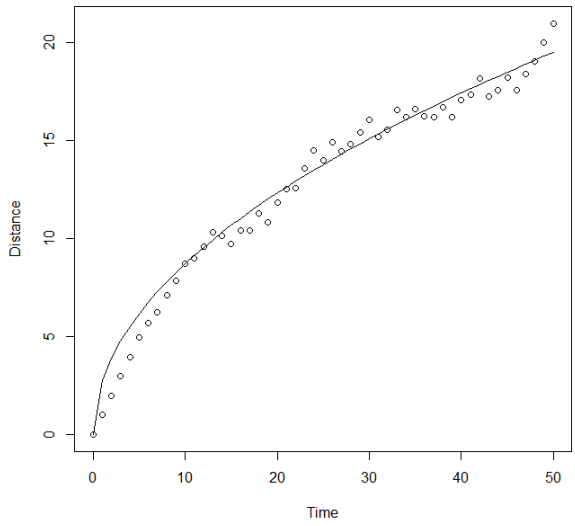

Figure 8.3: Plot of the maximum actual (points) and predicted (line) Euclidean distance of individuals from the origin under diffusion with no drift against timestep

8.2.3 Diffusion with drift

For many natural processes and for modelling bird migration this assumption of no drift is not realistic. Birds do not simply move randomly in any direction but instead have a determined direction of movement. For the particle diffusion equation, this is

often referred to as the advective term and was represented by a in equation 8.4. In circular statistics, this is modeled by one of the non-Uniform circular distributions and normally by one of the two related distributions of the wrapped/circular Normal distribution or the vonMises distribution.

This distribution has similar parameters to the Normal distribution in that there is a mean and standard deviation. The mean, in this case, is an angle between 0 and 2π and the standard deviation is a measure of the concentration of angles around this mean. The standard deviation is sometimes replaced with a concentration parameter, inversely proportional to the standard deviation, or by the mean resultant length (MRL) which gives the average distance moved as a proportion. So, for example, the MRL for the circular uniform distribution would be zero and would tend towards one as the standard deviation decreases. It is equivalent to the size of the drift term (a in equation 8.4) in the particle diffusion equation with an MRL of zero indicating that diffusion is the stronger force and an MRL of one indicating that advection is entirely dominant.

A convenient equation is given (Batschelet, 1981) to link the standard deviation and MRL, assuming the standard deviation is given in radians, then:

$$s^2 = 2(1 - MRL)$$

Taking a mean direction of 45° ($\pi/4$) and a standard deviation of 30° ($\pi/6$), the MRL can be calculated to be 0.863. Running this for 100 timesteps the mean distance from the origin can be plotted:

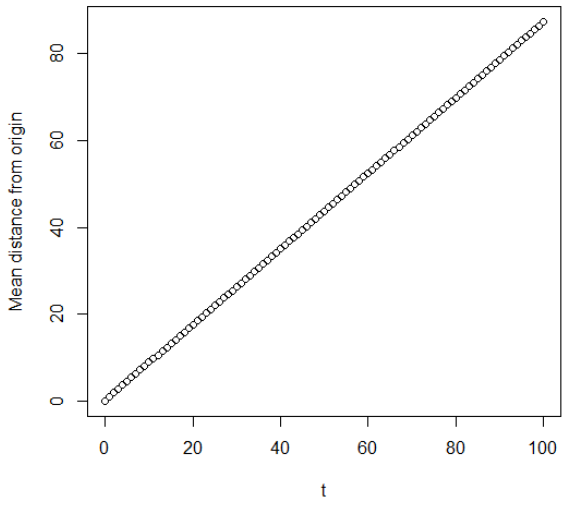

Figure 8.4: Plot of the mean Euclidean distance of individuals from the origin under diffusion with drift against timestep

From Figure 8.4 the mean distance is clearly very linear and has a gradient approximately equal to our MRL; this is consistent with the expected mean from equation 8.4 of a (equivalent to MRL) multiplied by time.

Controlling for this drift, i.e. taking the distance from the mean location for each timestep, would result in similar results to our simple diffusion.

This is entirely suitable for models where there is true advection (an exogenous flow whose bulk motion moves the particles); however, unlike particles, the drift for birds would be endogenous and would be determined by their homing instinct. A simple example of two birds (in the Republic of Niger) and their nesting grounds (in Turkey and Spain) shows why the traditional advection approach is not viable:

Figure 8.5: Locations of two birds (red) and their nesting grounds (blue)

Figure 8.5 shows that to reach their nesting grounds, each of the two birds is required to head in completely different directions and this would be the case for all birds. As such, each bird would have a drift or endogenous direction of it's own that will get increasingly different from those of other birds as they move up into Europe (initially all birds will be moving broadly north and so variation in direction will be low). Theoretical results will not be viable in these circumstances though controlling individually for each bird's drift would give results similar to simple diffusion.

To model this kind of movement, the wrapped Normal distribution will still be used, but with a mean calculated for each bird based on where it is at a particular time and where its nesting grounds are in relation to this.

8.2.4 Obstacle Avoidance

As discussed for our models in previous chapters, the homing instinct is only one of two main determinants of how birds migrate, the other being obstacle avoidance. In the more mathematical particle diffusion models, obstacles tend to be limits on the x or y plane and have an absorbing or reflecting property. This allows for analytical

solutions to be derived and reflects the behaviours of many physical and chemical processes. However, it is less realistic for avian migration where an avoidance strategy is often adopted. For many of the existing ornithological models (Erni et al., 2003; A. Vrugt et al., 2007), this is often dealt with using a 'zugknick' which is a shift in a birds endogenous direction either at a particular point in time or space. For very simple migration paths this can work, the papers in question (Erni et al., 2003), used birds migrating south from Scandinavia to Africa with a change in direction at a point in latitude and both papers use changes based on coastlines, but for many birds traveling from different wintering grounds to a variety of nesting grounds this is not possible.

Many other models, including some of the diffusion simulations, make use of perfectly circular obstacles in order to make use of the simple geometric properties this offers, but again this would not adequately describe the behaviour of migrating birds who must deal with irregularly shaped obstacles such as the Mediterranean or English channel.

Instead, a simple approach based on birds detecting multiple small obstacles will be used, such that the birds will tend to avoid water and therefore display obstacle avoidance behaviour. A number of points will be regularly distributed throughout the bodies of water in our geographic area with the points being used to approximately map the overall obstacle.

This can be illustrated by examining the case of a single bird:

Figure 8.6: Location of a bird (red triangle) and its nesting grounds (filled red triangle) and regular points indicating water (blue)

In Figure 8.6 if the bird was to take a direct route to its nesting grounds then this would entail crossing over a body of open water which the majority of bird species tend to avoid where possible (Dorst, 1962). The bird will instead react to this body of water and this combined with the variation in orientation will result in the much less direct route in Figure 8.7:

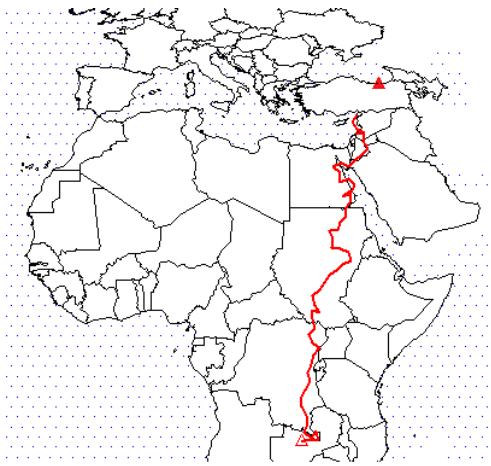

Figure 8.7: Location of a bird (red triangle) and its nesting grounds (filled red triangle) and regular points indicating water (blue) and the birds migratory path (red)

Focusing in on the north of the Red Sea, the behaviour governing obstacle avoidance can be examined.

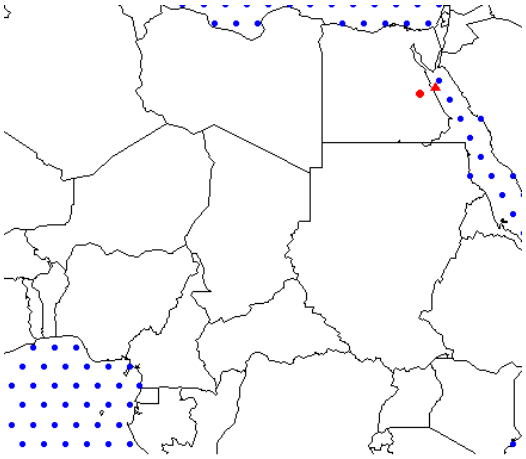

Figure 8.8: Position of bird at t=45 (red triangle) and at t=46 (red circle) and regular points indicating water (blue)

In Figure 8.8 the position of the bird at two consecutive time points is shown as well as the obstacle points. At timestep 45 (the red triangle), all obstacles within flight distance for the next time step will be found and their circular mean will be calculated. Assuming a set of angles $\alpha_1, \alpha_2 .... \alpha_n$ for the n obstacles within range then the mean ($\bar{\alpha}$) can be found:

$$\bar{\alpha} = tan^{-1}\left(\frac{1}{n}\sum_{j=1}^{n} sin\alpha_j, \frac{1}{n}\sum_{j=1}^{n} cos\alpha_j\right) \tag{8.20}$$

The arithmetic mean is not suitable for data from circular distributions, which is why Equation (8.20) is needed. This can be illustrated with a simple example; two data points of 0° and 360° which are effectively the same and so the true mean should again be 0° or 360° but if the arithmetic mean is used then the result would be 180°.

Taking the circular mean of the directions to all relevant obstacles would give the mean direction of obstacles. Adding 180° to this gives a direction that moves directly away from the mean obstacles. Combining this with the bird's endogenous direction

can give a direction of movement for the next timestep.

Considering the bird at t=45 in Figure 8.8, then all the angles that help determine its direction of movement for the next timestep can be plotted.

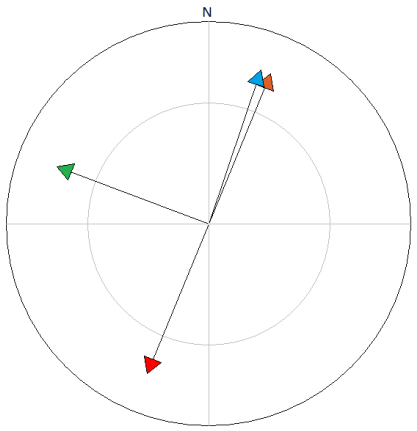

Figure 8.9: Directions measured for bird movement based on Figure 8.8 from t=45: direction to nesting grounds (blue), direction of obstacle within flight distance (orange), direction of obstacle avoidance (red) and direction of movement for next timestep t=46 (green)

In Figure 8.9 the blue arrow is the endogenous direction indicating the direct route between the bird and its nesting grounds. In this case, there is only one obstacle found within a day's migration distance and the direction between the bird and this point is shown by the orange arrow with the obstacle avoidance direction being directly opposite (the red arrow). The endogenous and obstacle avoiding directions are then averaged to find the direction of movement shown by the green arrow.

This should result in obstacle avoidance the majority of the time. For the model, the distance flown and the sample angle for each timestep will be stochastic and so there is the potential for a bird to move out into an area that birds would generally avoid but this would be an exception rather than the norm.

8.2.5 Continuous Space Model

Since there is no interaction between birds, then there is no need to model those that are not carrying CCHFV positive ticks. Therefore, the first parameter to be found for each bird will be the number of ticks it carries. This will be randomly sampled from a negative binomial distribution and any bird with zero ticks will be removed from the model. For each tick, a number of days of on-host attachment will be sampled.

The starting location for each bird will be randomly sampled from the wintering grounds for that species. For nesting grounds, i.e. finishing location, a country will be randomly selected for each bird with probabilities for a country based on its proportion of total bird numbers (birds for that country divided by total European population). A spatial point would then be randomly selected from that country.

For each timestep, the endogenous angle will be calculated for each bird. A migratory flight distance will be sampled for each bird for that timestep using the appropriate $v_{birdspecies}$ distribution; the obstacles within this range will be found and the direction of obstacle avoidance will be calculated. The circular mean of these two directions will then be found and used as the mean direction of movement for that bird in this timestep.

A sample direction from the wrapped Normal distribution with the calculated mean and a standard deviation based upon bird orientation studies, is taken for each individual bird and the movement in the x and y direction is found. A bird's current location will then be updated accordingly and each tick's days on-host attachment time will decrease by one. When the attachment time becomes zero, the current location of the relevant bird will be recorded. Once a bird reaches its nesting grounds, it will stop migrating and its location will no longer change.

Parameter estimates

The majority of parameter values were identical to those in Chapters 6 and 7. However, since migratory flight speed is now variable within timesteps and across individuals, then it will be described by a uniform distribution across the values used in these earlier chapters. The uniform distribution was chosen as the different parameter estimates are equally valid and so any value within the range should be equally likely.

Table 8.1: Estimates of parameter values for bird and tick behavioural factors in continuous space model

Parameter	Description	Value(s)	Reference
v_{Quail}	Distance (km) covered per day during spring migration by the common quail	U(150,160)	(Perennou, 2011)
$v_{Warbler}$	Distance (km) covered per day during spring migration by the willow warbler	U(62,114)	(Payevsky, 2013)
$v_{Wheatear}$	Distance (km) covered per day during spring migration by the northern wheatear	U(110,147)	(Payevsky, 2013)
v_{Pipit}	Distance (km) covered per day during spring migration by the tree pipit	U(57,106)	(Payevsky, 2013)
$v_{Bunting}$	Distance (km) covered per day during spring migration by the ortolan bunting	U(181,243)	(Payevsky, 2013)
a	Tick Length of Attachment	26	(Gale et al., 2011)
μ_{tick}	Mean number of ticks on migrating birds	0.049	(Gale et al., 2011)
ρ_{prev}	Prevalence of CCHFV in *Hyalomma* ticks	0.0001, 0.0007528	(Gale et al., 2011), (Lindeborg et al., 2012), (Palomar et al., 2013)
$\sigma_{orientation}$	Standard deviation of direction of bird migration	54°, 30°	(Batschelet, 1981), (Erni et al., 2003)

Number of runs

To determine the number of runs required for convergence, the model was run 150 times and the standard error of the number of ticks deposited in Spain (chosen as a country that consistently had ticks deposited and so non-zero data) was taken across increasing number of runs. There is no clear 'elbow' in the standard error (Figure 8.10) after a particular number of runs unlike in the previous chapters. There does seem to be a slightly steeper decrease up to about 50 runs so again this number was chosen.

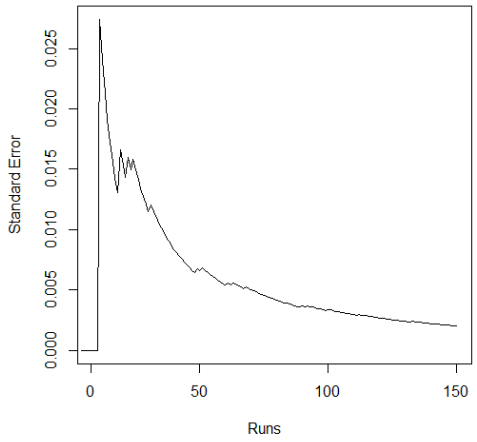

Figure 8.10: Standard error for number of ticks deposited in Spain under standard parameter values

8.2.6 Scenarios

To investigate the sensitivity of the model to its parameters four distinct scenarios were considered. These scenarios were identical to scenarios 1, 2, 3 and 5 used in Chapters 6 and 7. In all these scenarios the speed of migration is now a distribution across the range of estimated migratory speeds, hence why scenario 4 from the previous chapters (the increased migration speed) is not used.

The model was first run with the minimum of the defined parameters, for all tick related factors and for the orientation of bird migration, and the number of CCHFV infected ticks per country was calculated for each run. This is the baseline scenario

and was referred to as scenario one.

For scenario two, as before, everything was repeated with an increase in tick prevalence to the value calculated in Chapter 5 and used in the 2 previous models and the number of ticks was calculated to investigate the affect of an increase in CCHFV prevalence amongst ticks on the risk of the virus being imported into Europe via migrating birds.

The model being run with the higher population estimates from Table 4.5 was scenario three and the final scenario, still referred to as scenario five to maintain consistency with previous chapters, used a higher level of orientation amongst migrating birds.

8.3 Results

As in the individual agent cellular automata (Chapter 7), the results for this model will be restricted to integer values and, as in that chapter, to determine the best way of summarising these results the Normality of the data was inspected. A histogram of the results for the common quail for Spain under the baseline scenario was produced and with a fitted Normal curve (Figure 8.11). The Shapiro Wilk normality test was also carried out returning a p-value of $p < 0.001$.

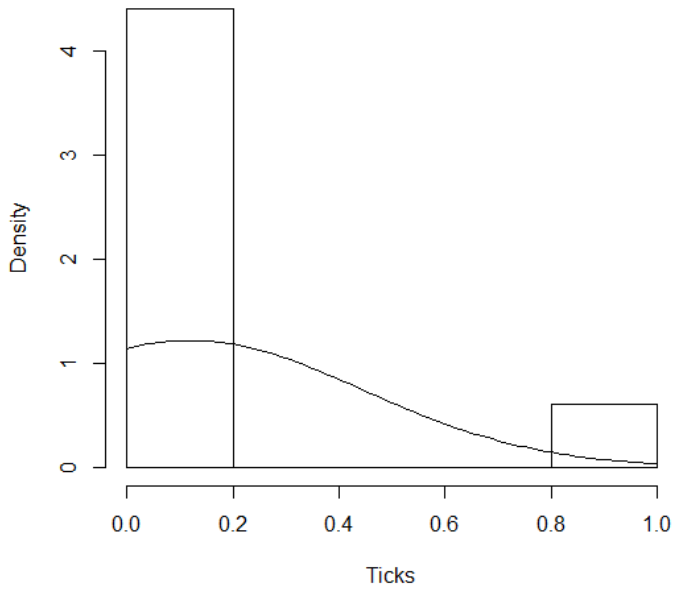

Figure 8.11: Histogram of level of tick incursion for Spain under standard parameters and average population numbers

In Figure 8.11, there is very little fit between the data and the proposed Normal distribution and, since the p-value from the Shapiro Wilk test is extremely small, then it is safer to assume results are not Normal. Therefore, results will summarised as in Chapter 7 with a full selection of summary statistics.

8.3.1 CCHFV Incursions to all of Europe

Examining the number of ticks being introduced into Europe, we can plot the expected numbers under each of our scenarios:

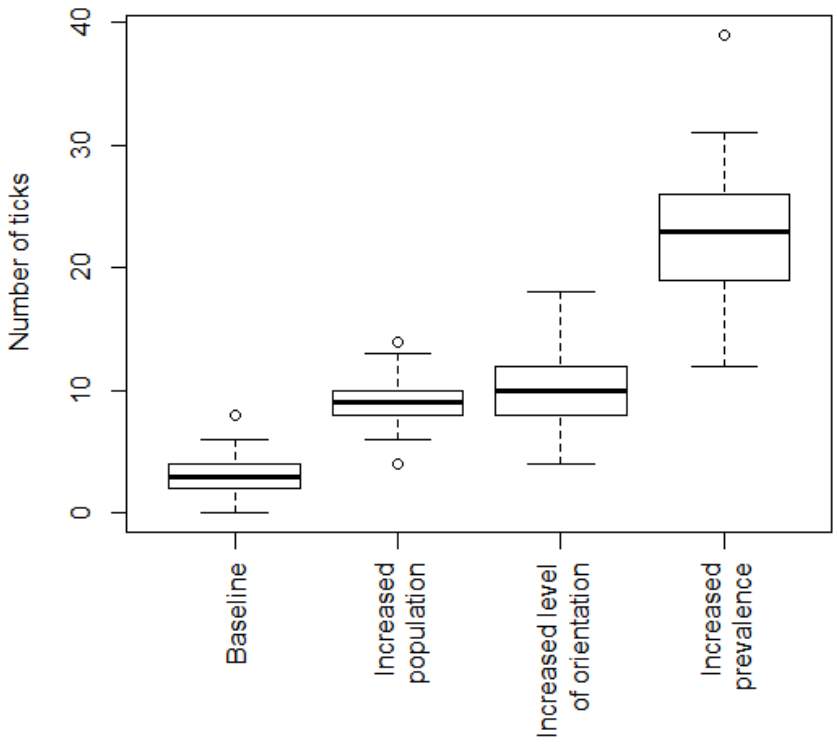

Figure 8.12: Boxplot of number of ticks introduced into Europe under each scenario

As was found in the previous models (chapters 4, 6 and 7), the level of sensitivity to changes in prevalence is greater than for any of the other parameters. From Figure 8.12 it can be seen that the level of variation in risk and the overall level of risk represented by the increased level of CCHFV prevalence is the highest for our scenarios. The baseline scenario, as expected, represents the lowest risk of CCHFV positive tick introduction and has the least variation in results. In contrast to the previous chapters, the difference between scenarios is much smaller, the ranges of number of CCHFV positive ticks for each scenario overlap with other scenarios much more than was seen previously.

Out of the remaining scenarios there does not seem to be much of a difference between the expected number of CCHFV positive ticks being introduced, though increasing the level of orientation of a migrant's flight results in more varied results than are produced under the higher population estimates.

Summary statistics of these results can be found in table 8.2, ordered by increasing mean, which illustrate the higher risk and level of variance present in scenario 2.

Table 8.2: Summary statistics of number of ticks introduced into Europe under each scenario

Scenario title (number)	Mean	Standard deviation	Median	Range
Baseline (1)	2.96	1.87	3	8
Increased Population (3)	9.14	2.35	9	10
Increased Orientation (5)	10	2.97	10	14
Increased Prevalence (2)	22.8	5.46	23	27

Table 8.2 supports the lack of a difference between scenarios 3 and 5, increased population and increased level of orientation, with their being only a small difference between their mean and median estimates.

To inspect whether the geographic distribution of risk changes between scenarios, plots can be produced (Figure 8.13). Once again, risk is not uniformly distributed across at risk countries and so as discussed in the previous chapter plotting as a proportion of total risk is not useful. To negate this issue then ordinal rankings of risk will be used.

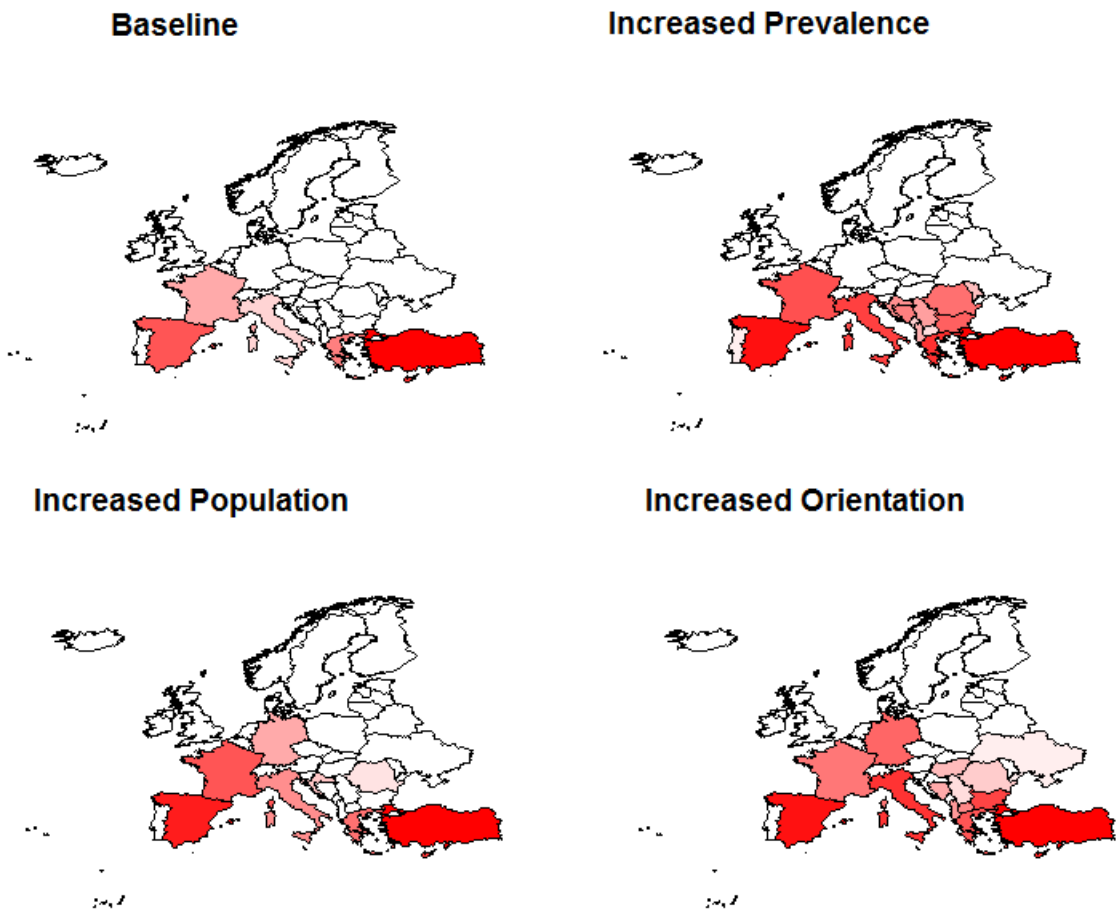

Figure 8.13: Geographic map of risk by rank; most at risk country (red) shading to no risk (white)

While there are some slight differences in Figure 8.13, in particular France and Germany under the increased prevalence scenario and baseline versus the increased population and increased orientation scenarios, generally, the geographic pattern of risk is fairly similar under each scenario as far as the more at-risk countries are concerned.

8.3.2 CCHFV Incursions by Individual Bird Species

Summary statistics for each individual bird species were derived to examine whether there is a species difference in the sensitivity to parameters. In Table 8.3 it can be

seen that the Ortolan Bunting represents the greatest risk under all of the scenarios. In addition, under all scenarios the Willow Warbler represents no risk and the Tree Pipit represents almost no risk.

The Ortolan Bunting is the only species of the five that represents a risk under every scenario and the Willow Warbler is the only species that has no risk of introducing CCHFV positive ticks despite any parameter changes. Of the three remaining species, the Common Quail and Northern Wheatear have means that are always greater than zero but the median values are zero for some scenarios; in particular, the baseline scenario. Increasing the population or prevalence results in medians that are greater than zero for both species. Both these scenarios result in an increased number of birds carrying CCHFV positive ticks, so an increase in the expected number introduced into Europe is not surprising. Increasing the orientation results in a higher median for the Common Quail only.

The final species is the Tree Pipit, which has a small risk of introducing infected ticks when migrating with an increased level of orientation. This is equivalent to increasing migration speed as better navigation results in a quicker migration though it is interesting that the same change did not yield a difference in results for the Willow Warbler.

Table 8.3: Summary statistics of number of ticks introduced into Europe under each scenario for each bird species

	Baseline (1)				Increased Orientation (5)			
Species	Mean	Standard deviation	Median	Range	Mean	Standard deviation	Median	Range
Ortolan Bunting	2.46	1.74	2	8	8.36	2.68	8	11
Tree Pipit	0	0	0	0	0.02	0.14	0	1
Common Quail	0.24	0.43	0	1	0.8	0.78	1	3
Willow Warbler	0	0	0	0	0	0	0	0
Northern Wheatear	0.26	0.44	0	1	0.82	0.98	0	3
	Increased Population (3)				Increased Prevalence (2)			
Species	Mean	Standard deviation	Median	Range	Mean	Standard deviation	Median	Range
Ortolan Bunting	7.72	2.41	7.50	10	18.64	4.87	19	28
Tree Pipit	0	0	0	0	0	0	0	0
Common Quail	0.68	0.77	1	3	2.50	1.52	2	6
Willow Warbler	0	0	0	0	0	0	0	0
Northern Wheatear	0.74	0.96	1	5	1.66	1.14	1	5

8.3.3 Countries

As in the previous chapters, it is worth looking at the most at-risk countries to identify areas of greater risk. The five most at-risk countries for CCHFV positive tick introduction are listed in Table 8.4 for each scenario.

Table 8.4: Mean number of ticks for top five most at risk countries for all bird species and for each scenario

Baseline (1)		Increased Population (3)		Increased Level of Orientation (5)		Increased Prevalence (2)	
Country	Mean	Country	Mean	Country	Mean	Country	Mean
Turkey	2.58	Turkey	8.22	Turkey	8.72	Turkey	18.86
Cyprus	0.16	Spain	0.36	Spain	0.42	Spain	1.7
Spain	0.12	Cyprus	0.2	Cyprus	0.26	Cyprus	0.8
Greece	0.06	France	0.14	Italy	0.16	Greece	0.42
France	0.02	Greece	0.1	Bulgaria	0.1	Italy	0.42

Examining Table 8.4, there are three countries that consistently occur although not always in the same rank; however, Turkey, the most at-risk country under every scenario, has a much greater level of introduced CCHFV positive ticks than any other country. Plotting the results in Table 8.4 yields Figure 8.14 which shows that there is a rapid decrease in risk after the most at-risk country.

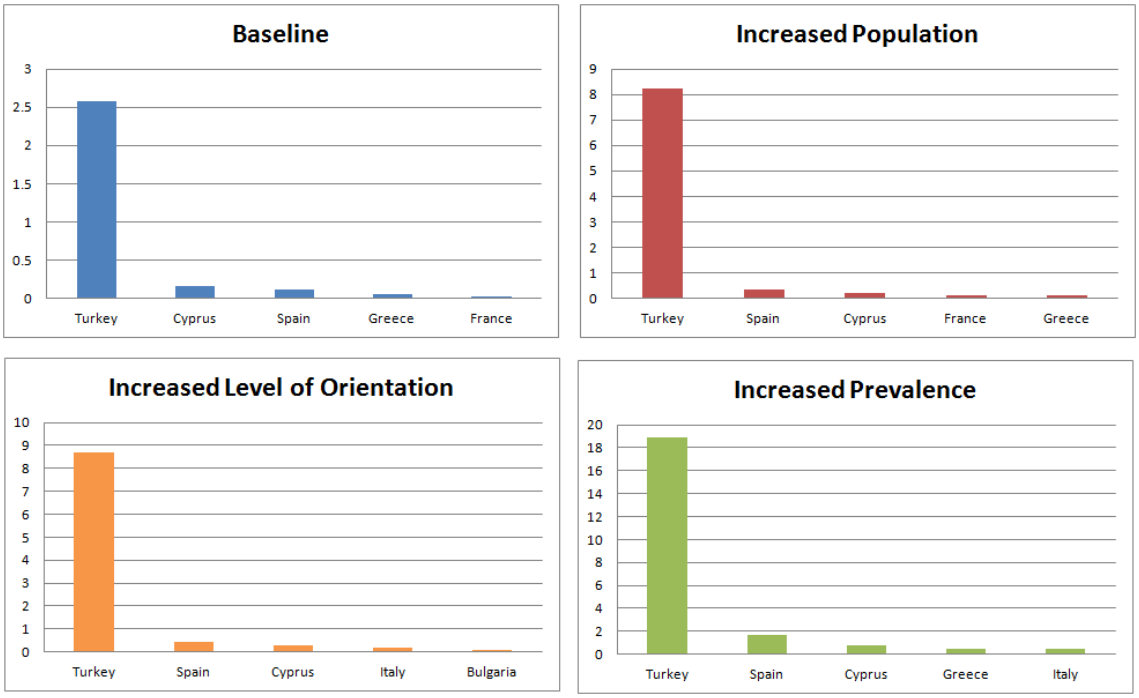

Figure 8.14: Histograms of level of tick incursion for top five most at risk countries for all bird species and for each scenario

The pattern or rate of reduction seems to be the same for each scenario and to further test whether this is the case, the proportion of risk for each country can be plotted (Figure 8.15). This is the proportion of each country out of the total risk for the top five countries. Examining the plot, it would seem that changing the parameters does not cause much of a change as far as the proportion of risk is concerned.

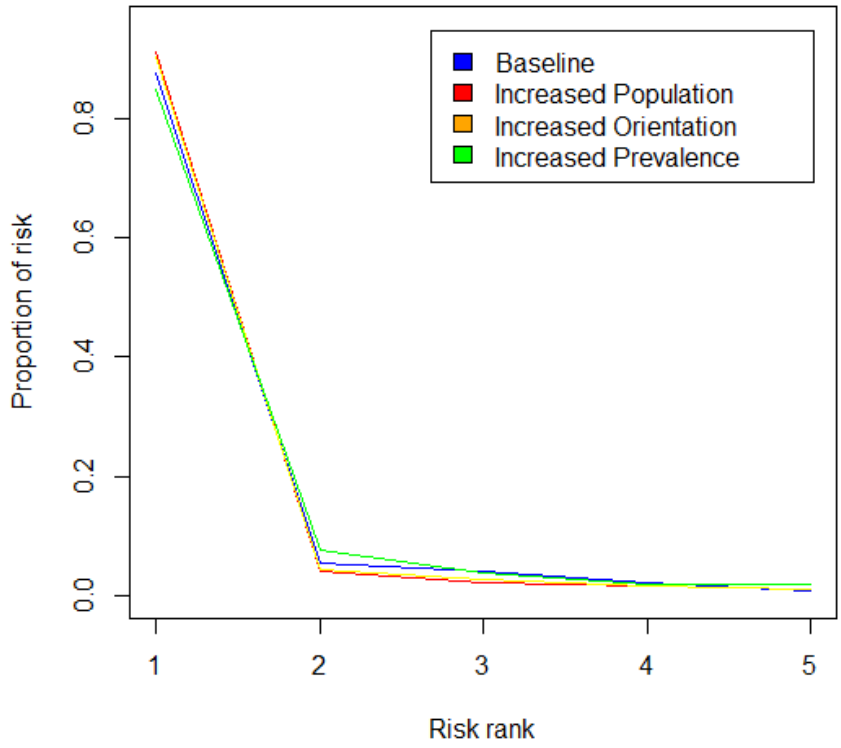

Figure 8.15: Proportions of level of tick incursion for top five most at risk countries for all bird species and for each scenario

8.4 Discussion

In this chapter, a continuous space model was developed to examine the level of risk represented by migrating bird species introducing CCHFV positive ticks into Europe. After being run for the five most populous and at-risk bird species, estimated numbers of CCHFV positive ticks were found for each country.

Compared to the previous two chapters, there are now only four scenarios, Figure 8.12 and Table 8.2 show that increasing the prevalence of CCHFV positive ticks has a greater impact than increasing the other parameters and also results in the greatest variation of results though not markedly so. As would be expected, the baseline scenario equates to the lowest level of risk and also the lowest variation and range of results. The remaining scenarios of increased bird populations and an increased level of orientation have a very similar effect on the number of CCHFV positive

introduced ticks.

There is much more overlap in results across scenarios, which is illustrated in Figure 8.12, meaning that it is difficult to say whether or not the scenarios are significantly distinct, though the inter quartile range for the increased prevalence does suggest that it is different from the remaining scenarios.

Comparing these results to those from the other three modelling approaches, it is clear that modelling birds as individual agents results in the estimated level of CCHFV positive tick incursions decreasing. The mean and median results are lower than the mean results for the aggregated cellular automata approach (Table 6.2) or minimum results for the geographic information system approach (Tables 4.8, 4.6 and 4.9). In comparison to those for the individual agent model (Table 7.3), there does not seem to be much of a difference. If the scenarios are taken as being distinct, then the order of the scenarios is, in effect, the same as for the aggregate cellular automata and so is in agreement with the sensitivities found in Chapter 4.

The maps of risk of tick introduction (Figure 7.4) show a pattern of risk that is fairly consistent between scenarios and is very much focused round countries that are major parts of the main migration routes for birds flying between Africa and Europe. The top at-risk countries for each scenario are, however, less consistent (Table 8.4) though the top place is always held by the same country. This country, Turkey, is consistently at the top in all the preceding chapters (Chapter 4, 6 and 7) and has a historical problem with CCHFV. Spain and Cyprus are also both repeatedly included where CCHFV has recently been detected in the former but is not believed to be present in the latter meaning that this continuous space modelling approach suggests a definite possibility of it being introduced.

Figure 8.13 shows a much tighter geographic distribution of risk in comparison to the maps for the aggregated cellular automata approach (Figure 6.4) or the individual agent based cellular automata (Figure 7.4) suggesting that freeing up the movement of migratory birds has reduced the speed of their northwards movements. Table 8.4 and Figures 8.14 and 8.15 show there is little variation in the geographic distribution

of risk between scenarios. Figures 8.14 and 8.15 shows that the most at-risk country, Turkey for all models, explains the vast majority of total risk even more so than the results in Chapters 6 and 7.

The risk represented by the different species in the continuous space model (Table 8.3) reveals that out of the five species modeled, the risk of CCHFV positive ticks being introduced into Europe can be attributed almost entirely to just one of them. While the Tree Pipit does introduce a single CCHFV positive tick under scenario five, it and the Willow Warbler can generally be regarded as posing no risk of introduction. The Common Quail and Northern Wheatear do carry a risk of introducing CCHFV positive ticks, but the vast majority (over 78%) are introduced into Europe by the Ortolan Bunting.

While there is a great deal of difference between the overall results for the two aggregated population approaches, the geographic information system and the aggregated cellular automata, there is less of a difference between the individual agent based cellular automata and this model. This could suggest that modelling birds as individuals is very important in terms of estimating risk but that the method of movement and settling behaviour can be simplified without a great difference being seen in the estimated risk.

However, when examining the geographic distribution of this risk within Europe, there is some difference between the two individual agent based models. In the continuous space model, a much greater proportion of the total number of CCHFV positive ticks are deposited in Turkey and a relatively small number are deposited elsewhere. For the individual agent based cellular automata, the majority of risk is still associated with Turkey but there is now a much greater risk associated with Spain as well. CCHFV has been found in Spain but the method of its introduction is less certain (Foley-Fisher et al., 2012).

Examining the two methods from a species approach also reveals a difference. In both cases, the Ortolan Bunting is the most important species in terms of introducing CCHFV positive ticks, but it plays slightly more of a role in the continuous space

model than in the individual agent based cellular automata.

The lack of certainty in the method of CCHFV introduction into Spain means that the much lower risk associated with Spain in this model could reflect a real world pattern, in which case, this more detailed modelling approach is required when properly evaluating the risk of CCHFV positive ticks being introduced into Europe. The large proportion of risk focused round the Ortolan Bunting and Turkey, and so the eastern migration route, means that surveillance and preventive methods could be restricted to key areas and species.

In summary, all of the modelling approaches suggest that surveillance for CCHFV positive ticks being introduced can be made more effective by focusing on key geographical areas. The more realistic the model, the more the results suggest a focus on the eastern migration route for surveillance; so a focus on Turkey. Focusing efforts on countries that lie along migratory routes is not a surprising result; what is of much greater interest is that all modelling approaches suggest that efforts should not focus on the most populous at-risk species, but that instead a combination of population and migratory speed should be used to select the species to be focused on. As the modelling approach became more realistic, the strength of this conclusion increased so any surveillance programs should take this into account.

Chapter 9

Discussion

The overall aim of this thesis was to identify useful sources of information and methods for examining the risk of introduction of exotic pathogens into Europe. This was addressed by firstly investigating available data sources including historical data and expert opinion data and the different methods used to analyse these data as well as some of the issues inherent to data of these types. These issues required the investigation of transformational techniques or variants of more common statistical tests and highlighted the dangers of applying the more well known tests to datasets without investigation. In relation to the search engine queries, it highlights the risks of poor sub-sampling techniques and why at the current level of data availability many of the results based on this source of data could well be spurious. From a detailed analysis of the expert opinion data, some key emerging pathogens were identified and, considering their characteristics, a potential biggest emerging risk was identified.

The second part of the thesis focused on this key risk, a particular method for the introduction of the identified pathogen, whereby geographical and host species data were used to predict the risk of it being introduced into Europe. Different models were investigated, ranging from the simplest but most efficient, up to the more complex but more computationally time consuming stochastic models. These models also progressed from the least realistic, with uniform spread of populations, through to the more realistic, with modelling of individuals through continuous space. Each of these models highlights issues that are worth considering for future work or for

the examination of current published work, and the variation in results emphasises the need to consider these issues closely.

9.1 Key Findings

In Chapters 2 and 3, expert opinion data on current and emerging threats collected at an EPIZONE conference were outlined and analysed. These data examined what the experts viewed as important in terms of a diseases impact i.e. economic, human health or animal health impact, which diseases were most likely to be introduced, spread and persist, both now and in the future, and what factors they thought might influence changes between now and then. In Chapter 2, a univariate approach for analysing these data was used to identify the diseases experts viewed as the most important emerging threats as well as factors that might contribute to these views. Delegate's scoring behaviours were examined separately for individual groups and it was found that the background or area of expertise of a delegate was not particularly important in determining their opinion. The sole exception seeming to be Foot-and-mouth disease, for which a delegate's background was found to be significant with those from an industry background viewing it as less of a threat. Many of these industry delegates worked in the development of vaccine and anti-viral development, so potentially viewed this disease as less of a threat due to a greater knowledge of its treatment. Stronger delegate bias was found in relation to a delegate's geographical region and whether a delegate has worked on a disease group. The diseases a delegate works on was also found to be significant in almost all cases and regional significance was found in many of the results. This suggested a possible bias whereby a delegate allowed their own work or outbreaks common in their area to affect their views on threats to all of Europe. However, it is also fair to assume that an expert would want to work on diseases that they genuinely believe to be important so for the first of these this may not be a true bias but just another reflection of a delegate's expert opinion.

An additional, though unsurprising, trend that ran through most of the results was that delegates were less certain about the future compared to the present, with one of

the key differences between the two being the increasing importance of the zoonotic arbovirus group, a group containing WNV, RVF and CCHFV. This was the most important conclusion of this chapter: that this particular disease group was viewed as a much more serious threat in the future, possibly at the expense of threats with a historical precedent such as Influenza and Foot-and-mouth disease. This means that a large number of delegates from diverse regions and backgrounds identified a small number of exotic pathogens as being the most important emerging threat.

In Chapter 3, a multivariate approach was used to examine all disease group scores simultaneously in order to take into account interactions and trade offs between them. Many of the same patterns were found as for the univariate approach, including evidence of regional bias and a delegate's bias towards the disease they worked on. Once again, there was less certainty relating to future risk and there was seen to be an increased future importance for the zoonotic arbovirus group. This reinforced the key messages of the previous chapter. Even when examining the views on all disease groups simultaneously, one group of diseases stood out as a threat and it was one member of this group of diseases that would form the basis for the work in the rest of the thesis.

The disease group identified in Chapters 2 and 3 contained three zoonotic arboviruses, namely West Nile Virus, Rift Valley Fever and Crimean-Congo Haemorrhagic Fever and, of these, the latter was selected for more detailed work, specifically the risk of introduction of the virus through ticks on migrating birds. As such, the second part of the thesis involved making use of available data on host populations and behaviours to model and estimate the risk of an exotic pathogen being introduced. CCHFV was selected as it has a viable vector present throughout Europe and, in Chapter 4, the lifecycle of the disease, the behaviour of the birds that could introduce it to new areas and the available parameter data were explored.

In the rest of Chapter 4, a model based upon one previously published (Gale et al., 2011) was developed in order to introduce a spatial distance element and analytical and simulated solutions were found. The analytical and simulated results were broadly similar to each other, but were quite different from those in the previously

published approach (Gale et al., 2011) on which the model was based. The bird species that represented the most risk was the Ortolan Bunting and it was found that the number of introduced CCHFV positive ticks is much more sensitive to a change in CCHFV prevalence than a change in any parameter relating to the birds themselves, and that a small change in prevalence will have a large effect on our main estimate of risk, that is, the number of CCHFV positive ticks introduced into Europe.

The effect of parameters relating to distance, migration speed and tick feeding time, showed more variation between different bird species. Relaxing the assumption of a uniform distribution of birds across Europe and introducing the spatial element to the model, where migratory distance became a factor, meant that for some species the larger part of their populations could not contribute to introducing CCHFV positive ticks, due to being unable to reach their nesting grounds within the required time. This means that, under this model, the two most populous of migratory bird species were in fact the least important, as their migratory flight times meant that a relatively small proportion of the species would reach Europe within the maximum on host attachment time for ticks. As such, even if ticks were to attach themselves to the host immediately before migration, there was still little chance of them being deposited in Europe. In contrast, some of the other bird species were found to introduce more CCHFV positive ticks. This means that any examination of a virus where time is a factor for either the vector being attached to the migratory bird, as in CCHFV, or for the infectious period of the migratory bird itself, such as WNV, must take account of the distance and length of time of migration.

Taking account of bird movement alone was not enough; the use of a uniformly distributed population in Europe as in Gale et al. (2011), even if distance of migration is explicitly taken into account, is still a seriously flawed assumption. Since it affects the number of birds at a particular geographic point, it will therefore affect the migratory distance for the birds; obviously, this could be in a negative or positive way, more birds could be assumed to have further to migrate thus lowering the potential risk, but in either case it does not result in an accurate assessment of risk.

The results for the most at-risk countries are consistent with what little real data

there are. However, results from this modelling approach must be viewed with some caution due to a couple of weaknesses with this approach. Firstly the possibility of ticks detaching from their host before reaching their breeding grounds but still detaching in another European country and so contributing to the risk of introduction. Secondly, the distributions of migratory distance make use of the Euclidean distance which is a straight line between the two points and this is not realistic given the routes used in migration. In Chapter 6, an aggregated cellular automata modelling approach with a more spatially explicit element was used to partially address these two issues. However, before doing this the level of parameter uncertainty was examined. The level of difference found in the analytical and simulated results means that tackling many of the parameter uncertainties, especially for the prevalence of CCHFV amongst larval and nymphal ticks on migratory birds, would be very important in estimating the risk represented by CCHFV and to examine some of the parameter values more fully, a Bayesian approach was applied in Chapter 5 to try and form a better estimate for the prevalence of CCHFV amongst ticks by combining expert opinion along with the limited data available. This resulted in an alternative higher estimate of CCHFV prevalence that was used as the higher risk scenario in the later chapters with the lower estimate being kept as the value derived from expert opinion.

In Chapter 6, a more spatially explicit cellular automata approach was used to address the issues identified with the model in Chapter 4. Despite a different modelling approach, it was still found that increasing the prevalence of CCHFV positive ticks, even by the amount calculated by the Bayesian approach in Chapter 5, still has far more impact than increasing any of the other parameters and that, generally, the pattern of parameter sensitivities seemed the same with or without the explicit spatial element and once again, the Ortolan Bunting represented by far the greatest level of risk. Two of the more populous species, the Tree Pipit and the Willow Warbler, represented almost no risk of introducing CCHFV in this modelling approach. In addition, the number of expected CCHFV positive tick incursions were much lower. Thus, modelling the actual movement of migratory bird populations had a noticeable affect on the estimated risk of CCHFV positive ticks being introduced into Europe. This approach, however, still did not truly represent bird migratory behaviour as it made the assumption that birds ceased migrating in the first viable

area in Europe and this is not what actually occurs where birds instead return yearly to the same breeding area. To more properly reflect this, a model was developed to include a form of random settling behaviour and this required an individual agent based model though it was still based round a cellular automata type model and so still had discrete spatial steps.

With all birds being modeled individually (Chapter 7), it was found again that increasing the prevalence of CCHFV positive ticks had the greatest impact on CCHFV incursions and compared to the previous model, the expected number of CCHFV positive ticks being introduced was much lower and there was a greater level of variation in the results. There was also less difference between scenarios when modeled from an individual agent perspective. This suggested that moving away from population models to an individual based model can reveal potentially important dynamics in migration behaviour.

As in the previous approaches, the Ortolan Bunting represented by far the greatest level of risk, and the species with the largest population, the Willow Warbler, represented no risk of tick introduction; however, migratory behaviour could still be more explicitly modeled: while birds no longer simply settled in the first available breeding space, the model still did not show the homing instinct of migratory birds. In Chapter 8, a continuous space model was developed and, in line with all modelling approaches, increasing the prevalence of CCHFV positive ticks had a greater impact than increasing any other parameters. This consistency of result under the simplest and more complex model suggest that there could be some real value in attempting to gather more data so as to properly estimate the true value of this parameter and so to be able to correctly estimate the risk of CCHFV being introduced to Europe.

However, under this more complex model, there was much more overlap in results between scenarios, meaning that it is difficult to say whether or not the risk of CCHFV introduction is that sensitive to change. The results from the different modelling approaches displayed an indirectly proportional relationship between the estimated level of CCHFV positive tick incursions and the complexity of the model. This would mean that previously published papers making use of very simple models may be

overestimating the level of risk of CCHFV introduction.

The maps of risk of tick introduction showed a pattern of risk that was fairly consistent between scenarios, and across the three different types of spatial models. As would be expected, these were very much focused round countries that are major parts of the main migration routes for birds flying between Africa and Europe. Turkey was consistently the most at-risk country; Italy, Spain and Cyprus also were repeatedly included in the largest results. CCHFV has recently been detected in Spain, but is not believed to be present in either of the other countries meaning that increased monitoring in these countries may be worthwhile.

As the complexity of the model increased, a much tighter geographic distribution of risk was seen, suggesting that freeing up the movement of migratory birds reduces the speed of their northwards movements which is reflected in the lower results for the more complex models as less birds reach Europe before any ticks they are carrying detach.

As the migration of birds became more realistically modeled, the importance of the more populous species as a proportion of the birds introducing CCHFV positive ticks decreases and the importance of the less populous but faster Ortolan Bunting increases. This might be one of the more important results, as any paper estimating the risk or surveillance effort would intuitively target the larger species under the assumption that sheer numbers will mean they pose the greatest threat and, even in this thesis, the most important species may have been missed as only the top five identified species were selected.

While there is a difference between the overall results for the two aggregated population approaches, the geographic information system and the aggregated cellular automata, there is less of a difference between the individual agent based cellular automata and this model. This could suggest that modelling birds as individuals is very important in terms of estimating risk but that the method of movement and settling behaviour can be simplified without a great difference being seen in the estimated risk.

However, in the continuous space model, a much greater proportion of the total number of CCHFV positive ticks are deposited in Turkey, and a relatively tiny number are deposited elsewhere. For the individual agent based cellular automata, the majority of risk is still associated with Turkey but there is now a more significant risk associated with Spain as well. CCHFV has been found in Spain, but the method of its introduction is less certain (Foley-Fisher et al., 2012), so whether this presence of the pathogen can be taken as support for the slightly less complex model is not clear. The lack of certainty in the method of CCHFV introduction into Spain means that the much lower risk associated with Spain in the final model could reflect a real world pattern, in which case, this more detailed modelling approach is required when properly evaluating the risk of CCHFV positive ticks being introduced into Europe. The large proportion of risk focused round the Ortolan Bunting and Turkey, and so the eastern migration route, means that surveillance and preventative methods could be restricted to key areas and species.

Overall, this thesis has found that a particular group of viruses are viewed as an important potentially emerging threat for Europe in coming years, and further work and investigation of the risk of introduction of these viruses should be considered. Any work that evaluates such a risk must take into account all factors, not just those that are disease related but also geographic factors, especially the distance across which these viruses might have to be carried; two of the three identified, WNV and CCHFV, can be introduced into a region by birds, but in both cases there is a time dependency. To properly model this risk, a spatially explicit model that correctly reflects bird migratory behaviour should be used. The approaches highlighted in this thesis show that, for CCHFV, there is a definite risk of introduction, but it may be smaller than that which has been estimated previously. The results also show that the bird species that should be focused on will not be those intuitively identified, and the migratory speed of birds is a key factor in identifying the species that represent the most risk of introducing CCHFV positive ticks into Europe.

9.1.1 Modelling approaches

As discussed in Chapter 1 the modelling approaches used in the latter half of the thesis move from less to more computationally complex and relax a number of biological assumptions as they do so. The initial modelling approach adapted the linear model from Gale et al. (2011) in order to introduce a spatial distance element and allow for non-uniform distribution of birds and resulted in three sets of results. The first made use of an assumed Normal distribution for the migration distance between wintering grounds in sub-Saharan Africa and Europe as well as a uniform distribution of bird populations to derive results analytically. As such it was extremely efficient since there was no time required for the running of simulations; but while it was an improvement from the model in Gale et al. (2011) it still had a number of very questionable assumptions. The second made use of the same assumptions to arrive at analytical solutions and as such had the same flaws, however it was very efficient to simulate due to the low level of model complexity and was a good initial baseline for the development of more complex models. The final approach relaxed the assumption of a uniform European population and used separate non-standard distributions of distance for each country and a separate uniform distribution of bird population for each country. Therefore it could be argued to be more biologically realistic without adding to the technical difficulty of coding since the approach was identical to the simulation before. Given the greater granularity the computation time was increased relative to the previous models but with a resultant increase in the detail of the final output.

The spatially explicit cellular automata model used in Chapter 6 allowed for ticks to be deposited anywhere along a birds migration route. This was based at the population level and so was still fairly efficient to run with the disadvantage that some biological assumptions were still in place, in particular the model included a form of automatic settling behaviour which while easy to simulate was not biologically realistic. In contrast a form of variable settling behaviour was added in Chapter 7 to an agent based cellular automata approach but to allow for this different settling behaviour each bird had to be modelled individually. This added an additional level of computational complexity to the model as each calculation had to be carried out for every bird in the data frame. Each of these approaches, despite increasing

the time required to produce an estimate, resulted in significantly different results suggesting that removing these flawed modelling assumptions could be important to correctly estimating the true risk of introduction. In the final chapter, a continuous space model was developed to allow for freer movement of migratory birds and for true homing behaviour, rather than an automatic or random settling behaviour, and this had the greatest computational complexity of all the modelling approaches with a commensurate increase in run time. As such it would have to be carefully considered whether this approach added enough value to be worth the additional time required, given that there was less difference in results between this continuous space approach and the much simpler cellular automata in Chapter 7 as opposed to the change between any of the other modelling approaches.

9.2 Further Work

The mathematical approaches used in the latter part of this thesis could be used to investigate other exotic pathogens. Additionally, it would be interesting to use another pathogen to see if a similar pattern of risk emerges, that is, do the more complex but possibly more realistic modelling approaches result in lower levels of risk? As for CCHFV, with every modelling approach emphasising the importance of the Ortolan Bunting, it would be worth going through each of modelling approaches with a much wider range of species, or potentially for the ease of modelling, doing a sub-selection of species based upon migratory flight speed rather than population size. In addition, while the majority of papers give the same estimate of tick attachment time, it would be worth investigating the effect of varying the on host attachment time of ticks. This will have a similar final effect as increasing migration speed, since it will increase the potential range of introduction, but given the scale of the parameter the models may be more sensitive to a change in attachment. There are also additional bird species and populations that migrate into Europe via Turkey from wintering grounds that are not in sub-Saharan Africa and there would be value in modeling the migration of these birds using similar approaches to investigate the threat of introduction of CCHFV, and other pathogens, from Asia.

With the apparent importance of the level of prevalence of CCHFV in larval and nymphal ticks it would be of great benefit to have a better estimate of the true value of this. This could be achieved by the gathering of additional data, or a similar modelling approach could be used, where this unknown parameter is estimated by using more well known ones. That is, a model of the tick lifecycle in sub-Saharan Africa could be developed and parameters on the spread of CCHFV between ticks could be used to investigate what kind of equilibrium values of prevalence there could be. The models developed here, and the results that were produced, can guide further work that will allow the key factors, with the greatest impact, to be investigated and, in this way, the models developed in this thesis could be updated to allow an even more accurate examination of the risk represented by CCHFV; with further applications to other pathogens that spread in similar ways and other geographic regions.

Appendix A

Individual Scores

Table A.1: Logistic Regression: Individual Score Confidence Intervals (Influenza)

Influenza Factor	Current CIs (Wald)	CIs (Profile Penalized)	Future (2020) CIs (Wald)	CIs (Penalized Profile)
Worked on Flu	(1.37, 5.30)	(1.39,5.32)	(1.97, 8.50)	(2.00, 8.61)
Region (East)	(0.31, 5.58)	(0.30, 5.25)	(0.42, 12.23)	(0.41, 11.85)
Region(Non)	(0.30, 4.96)	(0.29, 4.67)	(0.43, 11.98)	(0.42, 11.67)
Region(North)	(1.86, 12.71)	(1.94, 13.11)	(1.70, 19.20)	(1.86, 21.22)
Region(West)	(0.93, 5.78)	(0.97, 5.93)	(1.00, 10.41)	(1.09, 11.51)

Table A.2: Logistic Regression: Individual Score Confidence Intervals (BT & AHS)

BT & AHS Factor	Current CIs (Wald)	CIs (Profile Penalized)	Future (2020) CIs (Wald)	CIs (Penalized Profile)
Worked on BT & AHS	(2.35, 13.18)	(2.41, 13.62)	(1.99, 13.27)	(2.05, 13.88)

Table A.3: Logistic Regression: Individual Score Confidence Intervals (ASF)

ASF Factor	Current CIs (Wald)	Current CIs (Profile Penalized)	Future (2020) CIs (Wald)	Future (2020) CIs (Penalized Profile)
Worked on ASF	(2.74, 20.50)	(2.75, 20.30)	(2.25, 17.33)	(2.24, 17.02)

Table A.4: Logistic Regression: Individual Score Confidence Intervals (CSF)

CSF Factor	Current CIs (Wald)	Current CIs (Profile Penalized)	Future (2020) CIs (Wald)	Future (2020) CIs (Penalized Profile)
Worked on CSF	(0.95, 28.51)	(0.79, 31.08)	(2.83, 49.22)	(2.93, 54.18)
Background (Industry)	(0.58, 19.96)	(0.52, 20.58)	NA	NA
Background (Research)	(0.01, 4.04)	(0.00, 2.36)	NA	NA

Table A.5: Logistic Regression: Individual Score Confidence Intervals (Foot-and-mouth Disease)

FMD Factor	Current CIs (Wald)	Current CIs (Profile Penalized)	Future (2020) CIs (Wald)	Future (2020) CIs (Penalized Profile)
Worked on FMD	(0.89, 5.98)	(0.94, 6.15)	(1.00, 6.87)	(1.06, 7.09)
Background (EPIZONE)	(0.97, 6.42)	(1.02, 6.61)	(1.06, 7.53)	(1.12, 7.85)
Background (Research)	(1.37, 9.94)	(1.43, 10.23)	(1.54, 11.95)	(1.63, 12.44)

Table A.6: Logistic Regression: Individual Score Confidence Intervals (WNF, RVF, CCHFV)

WNF, RVF & CCHFV Factor	Current		Future (2020)	
	CIs (Wald)	CIs (Profile Penalized)	CIs (Wald)	CIs (Penalized Profile)
Worked on WNF, RVF & CCHFV	(2.01, 12.24)	(2.00, 12.20)	NA	NA
Region(East)	NA	NA	(0.31, 8.19)	(0.27, 7.15)
Region(Non)	NA	NA	(0.73, 11.47)	(0.71, 10.99)
Region(South)	NA	NA	(2.57, 19.72)	(2.69, 20.55)
Region(West)	NA	NA	(0.97, 6.24)	(1.01, 6.55)

Bibliography

(2008). Veterinary Sciences Core Team REPORT ON THE DISTRIBUTION OF BLUETONGUE INFECTION IN GREAT BRITAIN ON 15 MARCH 2008. Technical Report April, DEFRA.

A. Vrugt, J., van Belle, J., and Bouten, W. (2007). Pareto front analysis of flight time and energy use in long-distance bird migration. *Journal of Avian Biology*, 38(4):432–442.

Aitchison, B. Y. J. (1982a). Principal component analysis of compositional data. *Biometrika*, (1083):57–65.

Aitchison, J. (1982b). The Statistical Analysis of Compositional Data The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society*, 44(2):139–177.

Aitchison, J. (1999). *A Concise Guide to Compositional Data Analysis*. In: 2nd Compositional Data Analysis Workshop; Girona, Italy. Available: http://ima.udg.edu/Activitats/CoDaWork05/A concise guide to compositional data analysis.pdf.

Aitchison, J. and Egozcue, J. J. (2005). Compositional Data Analysis: Where are we and where should we be heading? *Mathematical Geology*, 37(7):1–30.

Aitchison, J. and Kay, J. W. (2003). POSSIBLE SOLUTIONS OF SOME ESSENTIAL ZERO PROBLEMS IN COMPOSITIONAL DATA ANALYSIS. In *Proceddings of CoDaWork '03, The 1st Compositional Data Analysis Workshop*, pages 1–6.

Albert, A. and Anderson, J. a. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Atkinson, B., Latham, J., Chamberlain, J., Logue, C., Donoghue, L. O., Osborne, J., Carson, G., Brooks, T., Carroll, M., and Jacobs, M. (2012). Sequencing and phylogenetic characterisation of a fatal Crimean Congo haemorrhagic fever case imported into the United Kingdom , October 2012. (October):5–8.

Batschelet, E. (1981). *Circular statistics in biology*. Academic Press Inc.

Bauchinger, U. and Klaassen, M. (2005). Longer days in spring than in autumn accelerate migration speed of passerine birds. *Journal of Avian Biology*, 36(1):3–5.

Baylis, M. and Githeko, A. (2006). Foresight Infectious Diseases : preparing for the future Contents :. Technical report, UK Government Foresight Project.

BBC (2014). Ebola Outbreak.

BBC (2015). New strain of 'deadly' bird flu.

Beauchemin, C., Samuel, J., and Tuszynksi, J. (2008). A Simple Cellular Automaton Model for Influenza A Viral Infections. *Journal of Theoretical Biology*, pages 1–12.

Beigel, J. H., Farrar, J., Han, A. M., Hayden, F. G., Hyer, R., de Jong, M. D., Lochindarat, S., Nguyen, T. K. T., Nguyen, T. H., Tran, T. H., Nicoll, A., Touch, S., and Yuen, K.-Y. (2005). Avian influenza A (H5N1) infection in humans. *The New England journal of medicine*, 353(13):1374–85.

Benyoussef, A., Boccara, N., and Chakib, H. (1999). Lattice three-species models of the spatial spread of rabies among foxes. *International Journal of Modern Physics*, 10(06).

BirdLife International (2012a). BirdLife International and NatureServe (2012) Bird species distribution maps of the world.

BirdLife International (2012b). BirdLife International and NatureServe (2012) Bird species distribution maps of the world.

Bivand, R. S., Pebesma, E. J., and Gomez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*.

Bliss, C. I. and Fisher, R. A. (1953). Fitting the Negative Binomial Distribution to Biological Data. *Biometrics*, 9(2):176–200.

Boender, G. J., Nodelijk, G., Hagenaars, T. J., Elbers, A. R. W., and de Jong, M. C. M. (2008). Local spread of classical swine fever upon virus introduction into The Netherlands: mapping of areas at high risk. *BMC veterinary research*, 4:9.

Bourn, J. (2002). The 2001 Outbreak of Foot and Mouth Disease. Technical Report June, National Audit Office.

Bruderer, B. (1997). The Study of Bird Migration by Radar. *Naturwissenschaften*, 84(2):45–54.

Burfield, I. and van Bommel, F. (2004). Birds in Europe: Population Estimates, Trends and Conservation Status. Technical report.

Caporale, G. (2009). SCIENTIFIC REPORT submitted to EFSA Scientific Review on Crimean-Congo Hemorrhagic Fever. Technical Report 178, Prepared by Istituto Zooprofilattico Sperimentale dell Abruzzo e del Molise Scientific Review on Crimean-Congo Hemorrhagic Fever.

Carrasco, L. R., Monteiro, D. M. S., Cook, A. J. C., and Moffitt, J. (2010). Economics of Robust Surveillance on Exotic Animal Diseases : the Case of Bluetongue. Agricultural & Applied Economics Association.

Claas, E. C., Osterhaus, a. D., van Beek, R., De Jong, J. C., Rimmelzwaan, G. F., Senne, D. a., Krauss, S., Shortridge, K. F., and Webster, R. G. (1998). Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet*, 351(9101):472–7.

Clarke, K. C. (2003). *Getting Started with Geographic Information Systems*. Prentice Hall.

Cleere, N., Kelly, D., and Pilcher, C. (2000). Results from a late autumn ringing project in Kuwait, 1995. *Ringing & Migration*, 20(2):186–190.

DEFRA (2014). Disease Control Strategy for African and Classical Swine Fever in Great Britain. Technical Report August, DEFRA.

DEFRA and AHVLA (2014). African Swine Fever in the Ukraine.

Deyde, V. M., Khristova, M. L., Rollin, P. E., Ksiazek, T. G., and Nichol, S. T. (2006). Crimean-Congo hemorrhagic fever virus genomics and global diversity. *Journal of virology*, 80(17):8834–42.

Dinh, P. N., Long, H. T., Tien, N. T. K., Hien, N. T., Mai, L. T. Q., Phong, L. H., Tuan, L. V., Van Tan, H., Nguyen, N. B., Van Tu, P., and Phuong, N. T. M. (2006). Risk factors for human infection with avian influenza A H5N1, Vietnam, 2004. *Emerging infectious diseases*, 12(12):1841–7.

Dorst, J. (1962). *Migration of Birds*. Heinemann.

ECDC (2014). autochthonous cases of chikungunya fever in France.

Edwards, A. L. (1948). Note on the "Correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3).

EFSA Panel on Animal and Welfare (AHAW) (2010). Scientific Opinion on the Role of Tick Vectors in the Epidemiology of Crimean-Congo Hemorrhagic Fever and African Swine Fever in Eurasia 1. *EFSA Journal*, 8(8):1–156.

Egozcue, J. J. and Barcel, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis 1. *Mathematical Geology*, 35(3):279–300.

Enøe, C., Georgiadis, M. P., and Johnson, W. O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine*, 45(1-2):61–81.

Erni, B., Liechti, F., and Bruderer, B. (2003). How does a first year passerine migrant find its way? Simulating migration mechanisms and behavioural adaptations. *The Nordic Society OIKOS*, 2(December 2002):333–340.

Estrada-Peña, a., Martínez Avilés, M., and Muñoz Reoyo, M. J. (2011). A population model to describe the distribution and seasonal dynamics of the tick Hyalomma

marginatum in the Mediterranean Basin. *Transboundary and emerging diseases*, 58(3):213–23.

Estrada-Peña, A., Sánchez, N., and Estrada-Sánchez, A. (2012). An assessment of the distribution and spread of the tick Hyalomma marginatum in the western Palearctic under different climate scenarios. *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)*, 12(9):758–68.

Fienberg, S. E. and Shmueli, G. (2005). Statistical issues and challenges associated with rapid detection of bio-terrorist attacks. *Statistics in medicine*, 24(4):513–29.

Foley-Fisher, M., Phipps, P., Medlock, J. M., Atkinson, P., Atkinson, B., Hewson, R., and Gale, P. (2012). Ticks on northward migrating birds in southern Spain during Spring, 2011. *Journal of vector ecology : journal of the Society for Vector Ecology*, 37(2):478–480.

for Animal Health, W. O. (2005). Old Classification of Diseases Notifiable to the OIE.

for Animal Health (OIE), W. O. (2012). World Animal Health Information Database (WAHID).

Formenty, P., Schnepf, G., Gonzalez-Martin, F., and Bi, Z. (2007). International Surveillance and control of Crimean-Congo Hemorrhagic Fever outbreaks. In *Crimean-Congo Hemorrhagic Fever*, pages 295–303.

Fraser, D. (2008). Animal welfare and the intensification of animal production. *The ethics of intensification*.

Gale, P., Brouwer, A., Ramnial, V., Kelly, L., Kosmider, R., Fooks, A. R., and Snary, E. L. (2010). Assessing the impact of climate change on vector-borne viruses in the EU through the elicitation of expert opinion. *Epidemiology and infection*, 138(2):214–25.

Gale, P., Stephenson, B., Brouwer, A., Martines, M., de la Torre, A., Bosch, J., Foley-Fisher, M., Bonilauri, P., Lindstrom, A., Ulrich, R., de Vos, C., Scremin, M., Liu, Z., Kelly, L., and Muoz, M. (2011). Impact of climate change on risk

of incursion of Crimean-Congo haemorrhagic fever virus in livestock in Europe through migratory birds. *Journal of Applied Microbiology*, 112(2):246–257.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1998). *Bayesian Data Analysis*.

Gibbens, J., Mansley, S., Thomas, G., Morris, H., Paton, D., and Al., E. (2000). Origins of the CSF outbreak. *Vetrinary Record*, 147.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4.

Goldenberg, A., Shmueli, G., and Caruana, R. (2003). Using grocery sales data for the detection of bio-terrorist attacks. *Statistics in medicine*, (412):0–36.

Gould, E. A. and Higgs, S. (2009). Impact of Climate Change and Other Factors on Emerging Arbovirus Diseases. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103(2):109–121.

Hasle, G. (2010). *Dispersal of ticks and tick-borne pathogens by birds Dynamics of birds transport of ticks to Norway*. PhD thesis.

Haydon, D. T., Kao, R. R., and Kitching, R. P. (2004). The UK foot-and-mouth disease outbreak the aftermath. *Nature*, 2(August).

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–19.

Hoogstraal, H. (1979). The epidemiology of tick-borne Crimean-Congo Hemorrhagic Fever in Asia, Europe and Africa. *Journal of Medical Entomology*, 15(4):307–417.

Hulth, A., Rydevik, G., and Linde, A. (2009). Web queries as a source for syndromic surveillance. *PloS one*, 4(2):e4378.

Izs-ve, G. C., Cidc, A. D., Vla, G., Sva, A. L., Lipowski, A., Iah, P. M., Vla, P. P., Vla, E. S., Fli, R. U., and Yin, H. (2006). Workpackage 7 . 4 : - Impact of environmental effects on the risk of the occurrence of epizootic diseases in Europe : Identification and prioritisation . Hazard Identification. Technical Report I.

Jameson, L. J., Morgan, P. J., Medlock, J. M., Watola, G., and Vaux, A. G. C. (2012). Importation of Hyalomma marginatum, vector of Crimean-Congo haemorrhagic fever virus, into the United Kingdom by migratory birds. *Ticks and tick-borne diseases*, 3(2):95–9.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, 2nd edition.

Kelly, L., Brouwer, A., Wilson, A., Gale, P., Snary, E., Ross, D., and de Vos, D. J. (2013). Epidemic threats to the European Union: expert views on six virus groups. *Transboundary and emerging diseases*, 60:360–9.

Kirby, J. S., Stattersfield, A. J., Butchart, S. H. M., Evans, M. I., Grimmett, R. F. a., Jones, V. R., O'Sullivan, J., Tucker, G. M., and Newton, I. (2008). Key conservation issues for migratory land- and waterbird species on the world's major flyways. *Bird Conservation International*, 18(S1).

Leguenno, B., Health, T. P., and Alfort, M. (1990). Crimean-Congo hemorrhagic fever in Senegal: temporal and spatial patterns. *Archives of Virology*, pages 323–340.

Lindeborg, M., Barboutis, C., Ehrenborg, C., Fransson, T., Jaenson, T. G. T., Lindgren, P.-E., Lundkvist, A., Nyström, F., Salaneck, E., Waldenström, J., and Olsen, B. (2012). Migratory birds, ticks, and crimean-congo hemorrhagic fever virus. *Emerging infectious diseases*, 18(12):2095–7.

Midilli, K., Gargili, A., Ergonul, O., Elevli, M., Ergin, S., Turan, N., Sengöz, G., Ozturk, R., and Bakar, M. (2009). The first clinical case due to AP92 like strain of Crimean-Congo Hemorrhagic Fever virus and a field survey. *BMC infectious diseases*, 9(1):90.

Mouritsen, H. (1998). Modeling migration: the clock and compass model can explain the distribution of ringing recoveris. *Animal Behaviour*, 56:899–907.

Naicker, P. R. (2011). The impact of climate change and other factors on zoonotic diseases. *Archives of Clinical Microbiology*, 2(2):2–7.

Navier-stokes, T. (1985). Two-Dimensional Cellular Automata. *Journal of Statistical Physics*, 38(March):211–249.

Newton, I. (2011). The ecology of bird migration patterns. Number Newton 2008, pages 1–3. BOU Proceedings - The Ecology and Conservation of Migratory Birds.

Palomar, A. M., Portillo, A., Santibáñez, P., Mazuelas, D., Arizaga, J., Crespo, A. n., Gutiérrez, O., Cuadrado, J. F., and Oteo, J. a. (2013). Crimean-Congo hemorrhagic fever virus in ticks from migratory birds, Morocco. *Emerging infectious diseases*, 19(2):260–3.

Papazoglou, C., Kreiser, K., Waliczky, Z., and Burfield, I. (2004). Birds in the European Union a status assessment. Technical report, Birdlife International.

Pawaiya, R. V. S., Dhama, K., Mahendran, M., and Tripathi, B. N. (2009). Swine flu and the current influenza A (H1N1) pandemic in humans : A review. *Indian Journal of Veterinary Pathology*, 33(1):1–17.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2006). Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications*, 264(1):1–10.

Payevsky, V. a. (2013). Speed of bird migratory movements as an adaptive behavior. *Biology Bulletin Reviews*, 3(3):219–231.

Pearson, K. (1897). Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proceedings of the Royal Society of London*, 60:489–498.

Peiso, O. O., Bronsvoort, B. M. D. C., Handel, I. G., and Volkova, V. V. (2011). A review of exotic animal disease in great britain and in scotland specifically between 1938 and 2007. *PloS one*, 6(7):e22066.

Pena-Rehbein, P. (2013). Use of a negative binomial distribution to describe the presence of Sphyrion laevigatum in Genypterus blacodes. *Revista Brasileira de Parasitol. Vet.*, 2961:602–604.

Pennycuick, C. J. (1982). The Ornithodolite: An Instrument for Collecting Large Samples of Bird Speed Measurements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 300(1098):61–73.

Pennycuick, C. J. (2001). Speeds and wingbeat frequencies of migrating birds compared with calculated benchmarks. *The Journal of experimental biology*, 204(Pt 19):3283–94.

Pennycuick, C. J., Å kesson, S., and Hedenström, A. (2013). Air speeds of migrating birds observed by ornithodolite and compared with predictions from flight theory Air speeds of migrating birds observed by ornithodolite and compared with predictions from flight theory. *Journal of The Royal Society*.

Perennou, C. (2011). European Union Management Plan 2009-2011, Common Quail, Coturnix coturnix. Technical report, Natura 2000, European Union.

Randolph, S. E. and Rodgers, D. (2007). Ecology of tick-borne disease and the role of climate. In *Crimean-Congo Hemorrhagic Fever*, pages 167–186.

Sambri, V., Capobianchi, M., Charrel, R., Fyodorova, M., Gaibani, P., Gould, E., Niedrig, M., Papa, a., Pierro, a., Rossini, G., Varani, S., Vocale, C., and Landini, M. P. (2013). West Nile virus in Europe: emergence, epidemiology, diagnosis, treatment, and prevention. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 19(8):699–704.

Sánchez-Vizcaíno, J. M., Mur, L., and Martínez-López, B. (2012). African swine fever: an epidemiological update. *Transboundary and emerging diseases*, 59 Suppl 1:27–35.

Sang, R., Lutomiah, J., Koka, H., Makio, A., Chepkorir, E., Ochieng, C., Yalwala, S., Mutisya, J., Musila, L., Richardson, J. H., Miller, B. R., and Schnabel, D. (2011). Crimean-Congo hemorrhagic fever virus in Hyalommid ticks, northeastern Kenya. *Emerging infectious diseases*, 17(8):1502–5.

Shaw, D. J., Grenfell, B. T., and Dobson, A. P. (1998). Patterns of macroparasite aggregation in wildlife host populations. *Parasitology*.

Shmueli, G. and Burkom, H. (2010). Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *Technometrics*, 52(1):39–51.

Sigmundsdottir, G., Gudnason, T., Ólafsson, O., Baldvinsdóttir, G. E., Atladottir, A., Löve, A., and Danon, L. (2010). Surveillance of influenza in Iceland during the 2009 pandemic. Technical Report April 2009, Euro Surveillance.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Thorup, K. and Rabøl, J. (2001). The orientation system and migration pattern of long-distance migrants : conflict between model predictions and observed patterns. *Journal of Avian Biology*, 2(1992):111–119.

Turell, M. (2007). Role of ticks in the transmission of Crimean-Congo Hemorrhagic Fever Virus. In *Crimean-Congo Hemorrhagic Fever*, pages 143–154.

Vose, D. (2008). *Risk Analysis: A Quantitative Guide*. Wiley, third edit edition.

White, S. H., del Rey, a. M., and Sánchez, G. R. (2007). Modeling epidemics using cellular automata. *Applied Mathematics and Computation*, 186(1):193–202.

Wiltschko, W. and Wiltschko, R. (1975). The Interaction of Stars and Magnetic Field in the Orientation System of Night Migrating Birds. *Zeitschrift fur Tierpsychologie*, 355:337–355.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining*. Morgan Kaufmann.

Xiao, X., Shao, S.-H., and Chou, K.-C. (2006). A probability cellular automaton model for hepatitis B viral infections. *Biochemical and biophysical research communications*, 342(2):605–10.

Yohannes, E., Biebach, H., Nikolaus, G., and Pearson, D. J. (2009). Migration speeds among eleven species of long-distance migrating passerines across Europe, the desert and eastern Africa. *Journal of Avian Biology*, 40(2):126–134.

Zaugg, S., Saporta, G., van Loon, E., Schmaljohann, H., and Liechti, F. (2008). Automatic identification of bird targets with radar via patterns produced by wing flapping. *Journal of the Royal Society, Interface / the Royal Society*, 5(26):1041–53.