Epidemiological investigations of surveillance strategies for zoonotic *Salmonella*

A dissertation presented
in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
at Massey University

Jacqueline Benschop
2009

Epidemiological investigations of surveillance strategies for zoonotic *Salmonella*

A dissertation presented

in partial fullfilment of the requirements

for the degree of Doctor of Philosophy

at Massey University

Jacqueline Benschop

Institute of Veterinary, Animal and Biomedical Sciences

Massey University

Palmerston North, New Zealand

2009

(Submitted 16 January 2009)

Institute of Veterinary, Animal and Biomedical Sciences

Massey University

Palmerston North, New Zealand

2009

Abstract

This thesis is concerned with the application of recently developed epidemiological and statistical tools to inform the optimisation of a national surveillance strategy of considerable importance to human health. The results of a series of epidemiological investigations of surveillance strategies for zoonotic *Salmonella* are presented. *Salmonella* are one of the most common and serious zoonotic foodborne pathogenic bacteria globally. These studies were motivated by the increasing focus on the cost-effectiveness of surveillance while maintaining consumer confidence in food supply. Although data from the Danish *Salmonella* surveillance and control programme has been used in these investigations, the techniques may be readily applied to other surveillance data of similar quality.

The first study describes the spatial epidemiological features of Danish *Salmonella* surveillance and control programme data from 1995 to 2004, using a novel method of spatially adaptive smoothing. The conditional probability of a farm being a case was consistently high in the the south-west of Sonderjylland on the Jutland peninsula, identifying this area for further investigation and targeted surveillance. The identification of clustering of case farms led into the next study, which closely examines one year of data, 2003, for patterns of spatial dependency. K-function analyses provided evidence for aggregation of *Salmonella* case farms over that of all farms at distances of up to six kilometres. Visual semivariogram analyses of random farm-level effects from a Bayesian logistic regression model (adjusted for herd size) of *Salmonella* seropositivity, revealed spatial dependency between pairs of farms up to a distance of four kilometres apart. The strength of the spatial dependency was positively associated with slaughter pig farm density. We describe how this might inform the surveillance programme by potentially targeting herds within a four kilometre radius of those with high levels of *Salmonella* infection.

In the third study, farm location details, routinely recorded surveillance information, and industry survey data from 1995 were combined to build a logistic seroprevalence model.

This identified wet-feeding and specific pathogen free herd health status as protective factors for *Salmonella* seropositivity, while purchasing feed was a risk factor. Once adjusting for these covariates, we identified pockets of unexplained risk for *Salmonella* seropositivity and found spatial dependency at distances of up to six km (95% CI: 2–35 km) between farms. A generalised linear spatial model was fitted to the Jutland data allowing formal estimation of the range of spatial correlation and a measure of the uncertainty about it. There was a large within-farm component to the variance, suggesting that gathering more farm level information would be advantageous if this approach was to be used to target surveillance strategy.

The fourth study again considers data from the whole study period, 1995 to 2004. A detailed temporal analysis of the data revealed there was no consistent seasonal pattern and correspondingly no benefit in targeting sampling to particular times of the year. Spatio-temporal analyses suggested a local epidemic of increased seroprevalence occured in west Jutland in late 2000. Lorelogram analyses showed a defined period of statistically significant temporal dependency, suggesting that there is little value in sampling more frequently than every 10 weeks on the average farm.

The final study uses findings from the preceding chapters to develop a zero-inflated binomial model which predicts which farms are most at risk of *Salmonella*, and then preferentially samples these high-risk farms. This type of modelling allows assessment of similarities and differences between factors that affect herd infection status (introduction) and those that affect the seroprevalence in infected herds (persistence and spread). The model suggested that many of the herds where *Salmonella* was not detected were infected but at a low prevalence. Using cost and sensitivity, we compared the results with those under the standard sampling scheme based on herd size, and the recently introduced risk-based approach. Model based results were less sensitive, but showed significant cost savings. Further model refinements, sampling schemes, and the methods to evaluate their performance are important areas for future work, and should continue to occur in direct consultation with Danish authorities.

Acknowledgements

Thank you to my three kind PhD supervisors who have provided guidance during this project. To Nigel French, who has mentored me and taught me about science and about myself. You have opened doors to many opportunities and been very influential in my career development. Thank you Nigel for encouraging me to publish and for always having a suggestion on how to move forward. To Mark Stevenson, who has mentored me and taught me a lot about epidemiology and the beauty there is in presenting data on a map. You have been very influential in my approach to teaching, learning, and thinking through a problem. Thank you Mark for your attention to detail and your tremendous ability to deliver complex information clearly. To Roger Morris, who invited me on board here at the EpiCentre and who made so many things possible for me, thank you.

Thank you to Martin Hazelton from the Statistics department for working alongside me on Chapter 3 of this thesis. Thanks to Simon Spencer for the same with Chapter 7. You both have shown me what can result from collaborating across disciplines and you have rekindled the interest I had in numbers.

Thanks to the people I have come to know at the Danish Meat Association. To Jan Dahl for provision of the data and advice on things Danish, I especially appreciated your hospitality while I was in Denmark. More recently Lis Alban has been my contact with the DMA. Jan and Lis, you both made it possible for me to work with data from the Danish *Salmonella* surveillance and control programme while I was in on the other side of the world. You provided thoughtful and frank comments on my work and have always been ready to consider new ideas. Also I thank Bodil Ydesen for providing additional risk factor and spatial data for this project.

For many friendly discussions and much helpful advice about modelling, spatial epidemiology, thesis writing, database management and the use of LaTeX, WinBUGS, and R, I thank my workmates and fellow students Caryl Lockhart, Daan Vink, Naomi Cogger,

Eve Pleydell, Thibaud Porphyre, Birgit Schauer, Solis Norton, Patricia Jaros, and Petra Müllner.

Thank you Cord Heuer. Working alongside you I have been able to develop interest in and experience with other projects while doing this PhD. Cord, you have always been ready to give support and advice and to take back the reins to free me up to do this work. You, Naomi Cogger, Mark Stevenson, Eric Neumann, and Deb Prattley, have taken up some of my share of teaching allowing me to complete my PhD. Thank you.

Thanks to Colleen Blair, Julie Dunlop, Simon Vershaffelt, Christine Cunningham, and Wendy Maharey for your administrative and computing support. Thanks to Ruth Upperton for proof-reading the final drafts of this thesis.

I thank my other fellow students for making the EpiCentre a diverse and fascinating place to work. Too often here I have had to say good bye as someone returns home to Zambia, Bhutan, Uruguay, or elsewhere overseas to take up a senior role in animal health in their home country. That is both a rewarding and a sad part of the job.

I leave the most important to the end now and thank those who are very dear to me. To my husband for making our home a warm and loving place to return to at the end of each working day, for taking care of me and our children, and for cherishing the differences between us. There is no way I could have all the blessings that I do, and have completed this work, without you, Tim. And also to our children, Ruth, Tess, Ben and Katrina. I am so proud of you and love you very much. This is just the beginning.

Nomenclature

ACF	Autocorrelation function
AFP	Acute Flaccid Paralysis
ANN	Artificial neural networks
ARIMA	Auto-regressive integrated moving average
BOSSS	Bovine Syndromic Surveillance System
BSE	Bovine spongioform encephalopathy
CCF	Cross-correlation function
CDC	Centers for Disease Control and Prevention
CHR	Central Husbandry Register
CI	confidence/credible interval
CJD	Creutzfeldt-Jakob Disease
DLM	Dynamic linear models
DMA	Danish Meat Association
DRB	Danish Meat Association risk-based sampling
DSSCP	Danish swine *Salmonella* surveillance and control programme
EARS	Early Aberration Reporting System
ED	Emergency department
ESR	Institute of Environmental Science and Research Limited
FMD	Foot and Mouth Disease
FoodNET	Foodborne Disease Surveillance Network
GAM	Generalised additive model
GIS	Geographic information system
GLM	Generalised linear model
HFRS	Haemorrhagic fever with renal syndrome
HIV	Human immunodeficiency virus

INAR	Integer-valued autoregressive models
MCMC	Markov Chain Monte Carlo
MRBA	Model derived risk-based sampling A
MRBB	Model derived risk-based sampling B
NNDSS	National Notifiable Disease Surveillance System
OD	Optical Density
OHS	Original herd size based sampling
OIE	World Organisation for Animal Health
RADAR	Rapid Analysis and Detection of Animal related Risk
RRV	Ross River virus
SARS	Severe Acute Respiratory Syndrome
SPF	specific pathogen free
SSI	Serological *Salmonella* Index
vCJD	variant Creutzfeldt-Jakob disease
WNV	West Nile Virus
ZIB	zero-inflated binomial

List of Publications

Benschop, J., Stevenson, M., Dahl, J., Morris R.S., French, N. (2009) Informing surveillance programmes by investigating spatial dependency of subclinical *Salmonella* infection. *Epidemiology and Infection* **137**:1348-1359

Benschop, J., Hazelton, M.L., Stevenson, M., Dahl, J., Morris, R.S., French, N. (2008) Descriptive spatial epidemiology of subclinical *Salmonella* infection in finisher pig herds: application of a novel method of spatially adaptive smoothing. *Veterinary Research* **39:02**

Benschop, J., Stevenson, M., Dahl, J., Morris R.S., French, N. (2008) Temporal and longitudinal analysis of Danish swine *Salmonella* control programme data: implications for surveillance. *Epidemiology and Infection* **136**:1511-1520

Benschop, J., Stevenson, M., Dahl, J., French, N. (2008) Towards incorporating spatial risk analysis for *Salmonella* seropositivity into the Danish swine surveillance programme. *Preventive Veterinary Medicine* **83**:347-359

Benschop, J., Hazelton, M., Stevenson, M., Dahl, J., Morris, R., French, N. (2007). 'Application of a novel method of spatially adaptive smoothing' in *Proceedings of the GisVet Conference*, University of Copenhagen, Denmark.

Benschop, J., Stevenson, M., Dahl, J., Morris, R., French, N. (2007). 'Using temporal and longitudinal analyses of accumulated data to inform sampling strategy' in *Proceedings from a Veterinary Epidemiology Seminar*, VetLearn Foundation, Palmerston North. ISSN: 1176-7979.

Benschop, J., Stevenson, M., Dahl, J., Morris, R., French, N. (2006). 'Can spatial analysis be used to target swine *Salmonella* surveillance for public health in Denmark?' in *Proceedings of GeoHealth Conference* 2006, Ministry of Health, Nelson, New Zealand. ISBN 0-478-30096-4.

Benschop, J., Stevenson, M., Dahl, J., Morris, R., French, N. (2006). 'second-order Spatial Effects: Danish Swine *Salmonella* Control Program' in *Proceedings of the 11th International Symposium on Veterinary Epidemiology and Economics*, Cairns, Australia.

Benschop, J., Stevenson, M., Dahl, J., Morris, R., French, N. (2006). 'Descriptive spatio-temporal epidemiology of sub-clinical *Salmonella* infection in Danish finisher pigs' in *Proceedings of the 11th International Symposium on Veterinary Epidemiology and Economics*, Cairns, Australia.

Benschop, J., Stevenson, M., Dahl, J., Morris, R., French, N. (2006). 'Spatial and risk factor analyses of *Salmonella* seropositivity in Danish pigs herds' in *Proceedings of the Epidemiology and Animal Health Management Branch, Food Safety, Animal Welfare and Biosecurity Branch of the New Zealand Veterinary Association Conference*, Vet Learn Foundation, Palmerston North, NZ. NZVA-FAVA Conference, Auckland, New Zealand. ISBN/ISSN: 1176-7979.

Contents

List of Figures

List of Tables

Preface

Family Story

‘Poor little shitter,’ her sister said. A tainted wonton
was what did it, that third night of the power cut.
Lines down everywhere, and us out in the sticks,

The power company put us in a priority queue
and sick of it all we hit town and queued instead
at the noodle-house - fuggy, crowded with pale

parents and fractious children also escaping
their darkened homes and wanting light, light
and warm food. The rain steamed off our backs

An ease came as we waited our turn, and children
played among tables. A kind of company, that’s what
we were, such as Chaucer or Boccaccio made a meal of,

only we made a meal of numbers on the menu,
our children too unsure to pronounce the names -
except Katrina, whose seven-year-old mouth

had eaten many wontons and called them out
now with conviction. On the wall, the Health
Inspection rating was unreadable behind yellowing

cellophane, a detail recalled as, hours later in the cold
bedroom, we stroked Katrina’s clammy forehead
while she writhed.… Poor little shitter, asleep now,

already a family story, her stained sheets churning
away in a washing machine that had at last,
when our hope was gone, our patience spent,

at the very last - then it was, as tiny red standby
lights throughout the house glowed into life,
the blessed machine shook, stirred itself and beeped.

Tim Upperton

from ‘A House on Fire’ Steele Roberts (2009)

CHAPTER 1

Introduction

1.1 Introduction

Veterinary medicine plays an essential role in protecting and promoting public health, especially in the prevention and control of zoonotic diseases. Zoonotic disease agents account for approximately 75% of emerging human pathogens and for over half of known human pathogens (Taylor et al. 2001). The recent and continuing spread of highly pathogenic avian influenza H5N1 (HPAI) and the emergence and spread of bovine spongiform encephalopathy (BSE) and severe acute respiratory syndrome (SARS) across many countries has caused concern internationally for human and animal health authorities.

New Zealand, by virtue of its relative geographical isolation and stringent biosecurity measures, has so far been spared from many of these disease threats but has, both in current times and historically, experienced significant zoonotic disease burdens (Crump et al. 2001). Past and present New Zealand significant foodborne zoonoses include tuberculosis, salmonellosis, and campylobacteriosis (Baker et al. 2007); direct zoonoses include hydatidosis, brucellosis, leptospirosis (Thornley et al. 2002), and *Salmonella brandenburg* and *Salmonella enterica* var Typhimurium DT 160 (Thornley et al. 2003). A response to these concerns in the past has been the application of direct government control programmes which have successfully led to provisional freedom from hydatidosis and brucellosis. However, in the present day, advances in control of zoonotic diseases are more likely to be industry lead, and more associated with risk management than eradication. The poultry industry risk management strategy for *Campylobacter*[1] and the meat

[1] `http://www.nzfsa.govt.nz/consumers/food-safety-topics/foodborne-illnesses/campylobacter/\strategy/index.htm`

industry input into work place protection from leptospirosis are cases in point (Keenan 2007).

Control programmes are built upon the back of sound surveillance strategies (Merianos 2007). This thesis uses the example of zoonotic *Salmonella* in Danish pig herds to investigate novel surveillance strategies that may be used in control programmes for zoonotic disease.

Non-typhoid salmonellosis is a serious zoonotic disease that mainly causes febrile gastroenteritis (Sanchez et al. 2002, Schlundt et al. 2004). Approximately 5% of those infected suffer severe sequellae such as polyarthritis, endocarditis, and, in rare cases, death. These sequellae are generally more common in the elderly or the very young (Fisker et al. 2003, Weinberger et al. 2004). In industrialised countries most cases of salmonellosis in humans are food-borne, and pork has been implicated as an important source (Baggesen & Wegener 1994, Wegener & Baggesen 1996). Clinically, asymptomatic finisher pigs and culled, older breeding stock may carry *Salmonella* that contaminates food product which is then capable of infecting humans (Borch et al. 1996, Swanenburg et al. 2001*a*).

In the early 1990s, it was estimated that approximately 15% of reported cases of salmonellosis were associated with the consumption of pork in Denmark, the Netherlands, and Germany (Mousing et al. 1997, Berends et al. 1998, Steinbach & Hartung 1999). However, up until 2007, the number of cases of salmonellosis in humans in Denmark attributable to pork consumption decreased by a factor of ten: from 1444 in 1993 to 101 in 2006 (Nielsen et al. 2001, Ministry of Family and Consumer Affairs 2007). In 2006, this represented 6.1% (95% CI: 4.6%–7.8%) of reported cases. The attribution of human salmonellosis cases to the major animal and food sources in Denmark is through a mathematical model developed by Hald et al. (2004). This model is based on a comparison of the number of human cases caused by different *Salmonella* sero- and phage types with the distribution of the *Salmonella* types isolated from the various animal-food sources. The decline in human cases from 1993 until 2007 has been attributed to the large-scale national control programme aimed at reducing the occurrence of *Salmonella* in pigs and focussing on interventions applied at the slaughterhouse. Data taken from ten years of this programme (1995 to 2004) forms the basis of the analyses presented in this thesis.

In this chapter I introduce the Danish swine *Salmonella* surveillance and control programme (DSSCP) and look at its performance compared with other European programmes.

This introduction also describes how this thesis came about and how it is structured.

1.2 The Danish Swine *Salmonella* Surveillance and Control Programme

The Danish swine *Salmonella* surveillance and control programme was set up in 1993 in response to an increase in the incidence of confirmed cases of human salmonellosis due to pork consumption (Baggesen & Wegener 1994), and a large, common source outbreak caused by *Salmonella infantis*, traced back to one slaughter plant and a small number of supplier pig herds (Wegener & Baggesen 1996). The objective of the DSSCP is to reduce the prevalence of *Salmonella* in pork to an acceptably low level so that domestically produced pork is no longer an important source of human infection.

The programme has components at all stages of pork production: breeding, multiplying and finishing herds, as well as controls on feed for pigs and at the slaughterhouse. This thesis considers data from the finishing herd component of the programme. This is based on the random testing of post-slaughter meat-juice samples from all finisher pig herds that have an annual kill of greater than 200 finishers (Mousing et al. 1997, Alban et al. 2002). The testing of meat-juice rather than blood facilitates both sample collection and carcass identification (Nielsen et al. 1998). All samples are analysed at the Danish Institute for Food and Veterinary Research using the Danish mix-ELISA which can detect O-antigens from at least 93% of all serovars that are known to be present in Danish pigs.

Sample results are used to categorise herds into one of three levels of a 'serological *Salmonella* finisher index' (Alban et al. 2002). The three levels are 'level 1' with an index of 1-39; 'level 2' with an index of 40-69; and 'level 3' with an index of 70 or more. Herds in levels 2 and 3 have requirements placed upon them. For example, producers must report their most recent weaner suppliers, pen faecal samples are collected for culture and typing, and there are penalty '*Salmonella* deductions' resulting in reduced payments. Furthermore, pigs from level 3 herds are slaughtered under special hygienic precautions. At the end of 2006, 2.5% and 0.9% of finisher herds were assigned to level 2 and 3, respectively.

Up until July 2005, the number of animals sampled at slaughter depended on herd size,

with 60, 75, or 100 pigs sampled per herd per year. Since that time a new level of 'serological *Salmonella* finisher index', level 0, has been created. Under this scheme herds that have had no positive samples in the previous three months are reduced to one sample per month (Ministry of Family and Consumer Affairs 2007). This change reduced the annual sample size gradually from approximately 570,000 meat-juice samples in 2004 to approximately 250,000 in 2006. This reduction in sampling has resulted in large cost savings for the Danish Meat Association (DMA) producers who bear the bulk of the cost of the programme.

The build-up to the introduction of risk-based surveillance was important in the genesis of this project. In 2002 a cooperative risk-based food safety research group between the EpiCentre and a number of European scientists was formed (the SaFoodChain group). This group included Jan Dahl and Lis Alban from Danske Slagterier, and one of the cooperative projects agreed to by members of this group was a joint study of the *Salmonella* surveillance data for Danish pigs, to be conducted by Massey University's EpiCentre and Danske Slagterier (now the Danish Meat Association). I was offered this data as the basis for my PhD, and in early 2004 Jan Dahl visited New Zealand and provided the data set and explanations about its interpretation. This was coincident with the interest both Danish authorities and the pig industry had in achieving the greatest reduction in *Salmonella* for their investment. They were already considering approaches such as the risk-based surveillance system that was introduced in 2005.

1.2.1 Results from the control programme

In mid-2008 a report on *Salmonella* in pigs and pork in the European Union (EU) was released (European Food Safety Authority 2008). This reported on an EU-wide baseline survey that was carried out in 2007 to determine the prevalence of pigs infected with *Salmonella* at the point of slaughter. Pigs were randomly selected from those slaughterhouses that together accounted for 80% of pigs slaughtered within each member state.

The findings for Denmark were that the adjusted pre-harvest prevalence, as measured by culture of ileo-caecal lymph nodes, was 7.7% (95% CI: 5.5%–10.7%). This was marginally lower than the average EU prevalence of 10.3% (95% CI: 9.2%–11.5%). These prevalence estimations accounted for clustering and the differences between the

complex survey design and simple random sampling. Given the short transport distances and holding times of pigs in Denmark, it would be reasonable to assume that this test reflects infection on the farm of origin.

The findings for Denmark in terms of post-harvest control, as measured by culture of carcass swabs, was 3.3% (95% CI: 1.3%–8.5%). This was substantially lower than the average EU prevalence of 8.3% (95% CI: 6.3%–11.0%). The prevalence of positive carcass swabs is a product of the risk of infection within a pig, the risk that the infection is released to the exterior of the carcass, and the risk of cross-contamination from other carcasses or the slaughterhouse environment. Since 2001, the movement of the focus of the *Salmonella* control programme towards increased slaughter-house interventions (such as hot water decontamination) is thought to be responsible for these results. It is important to remember that on-farm controls, such as penalties for level 2 and 3 herds, have been maintained and provide a powerful incentive for producers to perform well at the pre-harvest level.

1.3 The structure of this thesis

The aim of this thesis is to investigate techniques that have potential to be used to develop alternative surveillance strategies for zoonotic diseases using the example of *Salmonella* in the Danish pig population. Data has been sourced from the Danish swine *Salmonella* surveillance and control programme from 1995 to 2004. Detailed descriptions of these techniques comprise five of the eight chapters of this thesis. These five chapters are presented in the format of manuscripts for peer-reviewed publication. As a consequence of this style of thesis presentation, there is some repetition between chapters, especially in each of the introduction sections. Furthermore, the process of writing for publication and then subsequently responding to reviewers' and editors' requests requires a substantial amount of distilling of material and presentation of only the most pertinent results. This is evident in these five chapters, as I have maintained the published form of the manuscripts. The only alterations are some additional graphics and some extra text to provide linkage between the chapters and throughout the thesis as a whole.

The first of these manuscripts (Chapter 3) is an application of a novel method of spatially adaptive smoothing to describe the spatial epidemiology of subclinical *Salmonella* infec-

tion over the entire study period. The identification of clustering of cases leads into the second study, Chapter 4, an investigation of spatial dependency in the data from 2003. Random farm-level effects from a Bayesian seroprevalence model (adjusted for herd size) are used to explore spatial dependency. I describe how this might inform the surveillance programme by potentially targeting herds within a four kilometre radius of level 2 or 3 herds. In the third study (Chapter 5), data from a 1995 producer questionnaire gave additional farm-level covariate information (such as feed-type and feed source). This was used to build another seroprevalence model and investigate risk factors and first- and second-order spatial properties of the data to provide an informed framework for risk-based surveillance. Chapter 6 again considers the whole study period, 1995 to 2004, and investigates how temporal and longitudinal analytical techniques might be used to pinpoint sampling at certain times of the year, and how frequently herds should be sampled. The final study (Chapter 7) uses results from the previous chapters to develop a zero-inflated binomial model that predicts which farms are most at risk of *Salmonella* and uses this to inform a risk-based sampling approach.

The second chapter of the thesis is a literature review of methods for detecting and responding to changes within surveillance data for zoonotic disease, focussing primarily on temporal techniques. Temporal techniques are highly developed in public health surveillance systems but have been little researched within our group. The temporal focus of my literature review goes some way towards addressing that imbalance. Our group has some expertise in spatial and spatio-temporal methodologies and there have been a number of reviews of these techniques (see, for example Stevenson (2004), Lockhart (2008), and Porphyre (2008)). It is not my intention to provide another review of spatial and spatio-temporal techniques but to briefly describe these, and give a detailed account of the most recent developments.

The thesis concludes with a discussion of the study findings, identification of important lessons learnt, and plans for future work. Finally I present some future perspectives on surveillance for zoonotic disease.

CHAPTER 2

Literature review: Methods of detecting and responding to change within surveillance data

2.1 Introduction

Our lives are increasingly characterised by change and emergence. New products bombard us and our children. We are strongly encouraged to purchase Play Station 3 before Play Station 2 is out of its wrapper, and to buy the impossibly small iPod Nano. Mirroring this emergence of unprecedented gross consumption is the recent emergence or re-emergence of infectious diseases of veterinary public health importance (Vorou et al. 2007, Sargeant 2008). These include foodborne (*Escherichia coli* O157, Creutzfeldt-Jakob Disease) and occupational zoonoses (brucellosis, avian influenza), and those related to companion animals (monkey pox) and wildlife (rabies, West Nile Virus).

Alongside disease emergence and reemergence, there have been concerns about terrorism and bioterrorism since the September 2001 attacks on the World Trade Centre in the USA and the intentional release of anthrax in New York in October 2001 (Jernigan et al. 2001). The subsequent development of surveillance science associated with bioterrorism preparedness, and specifically early detection, has been rapid. Early detection is desirable because disease-causing agents have a prodromal phase that is relatively non-specific. If the disease is detected while patients are in this phase, they may be helped by specialised care. Furthermore, for contagious disease, early intervention, such as isolation or treatment to reduce shedding of infectious material, may slow down or stop disease progression.

Alongside the threat of emerging infections and the increased potential for bioterrorist attacks, there are increasingly limited resources for disease surveillance (Stärk et al. 2006). Coupled with increasing international trade, these are driving forces for improvements in disease surveillance and moves to optimise the resources used in disease surveillance. These include diseases of veterinary public health importance such as bovine spongiform encephalopathy (BSE) (Bohning & Greiner 2006), salmonellosis (Alban & Stärk 2005) and trichinellosis (Kapel 2005).

This literature review focusses on methods and tools for detecting changes within surveillance data. The focus is on zoonotic disease and techniques in time, space, and space-time. In situations such as epidemics, vaccination, or climate change, both human and animal health surveillance data present changes of magnitude and periodicity. For example, in June 2003, the Dutch National *Salmonella* Centre reported a significant excess isolation rate of *Salmonella enteritidis* from humans when compared with previous years (van Pelt et al. 2004). This is illustrated in Figure 2.1. The authors suggest that the increase in importation of contaminated eggs, as a result of the avian influenza outbreak, was the most probable reason for this excess. Although this example presents a real increase in cases it is important to be aware that there is much potential for artefactual changes in disease incidence e.g. through modifications of the case-definition, or the introduction of screening programmes.

Change does not always manifest in increased numbers of cases. For example, in 2001 there was a significant decrease in reports of human cryptosporidiosis in the north-west of the United Kingdom (UK), from 1382 cases in 2000 to 428 (Hunter et al. 2003). This was almost coincident with the introduction of control measures put in place around the 2001 outbreak of foot-and-mouth disease (FMD). During the 2001 outbreak of FMD in the UK, over six million ruminants were slaughtered, and there were strict and widespread bans on access to the countryside. The authors report that the decline in cases was most likely related to the outbreak of FMD. They conclude that these surveillance data support previous evidence that zoonotic transmission is a major route of infection for human cryptosporidiosis in this region of the UK.

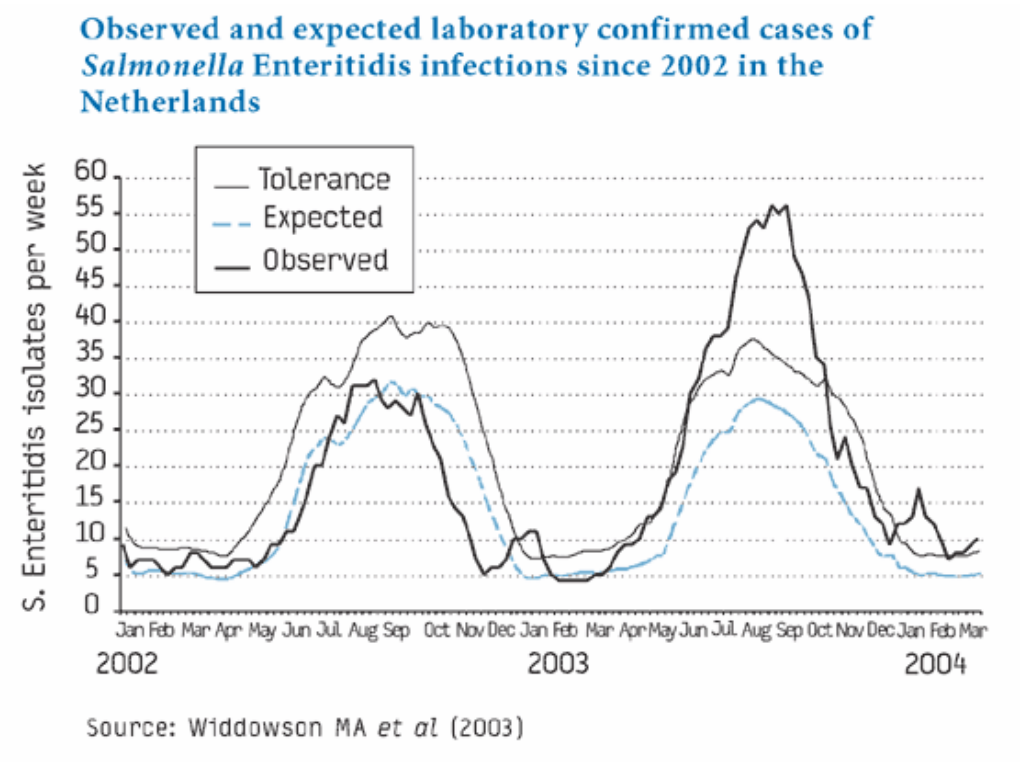


Figure 2.1: Observed, expected and tolerance levels for laboratory confirmed cases of *Salmonella enteriditis* in the Netherlands 2002 to 2004. Source: van Pelt et al. (2004)

The Belgian dioxin crisis in animal food in 1999 provides another illustration of change in surveillance data. The dioxin problem started at the end of January 1999, when contaminated feed was processed. However, official notification of the crisis did not occur until four months later at the end of May, when Belgian authorities ordered the withdrawal from sale of Belgian poultry and eggs. For the month of June 1999, the expected number of campylobacteriosis cases was 643, while the actual number of cases reported was 375 (Vellinga & Van Loock 2002). This reduction in cases was likely due to the withdrawal of chicken and its products from the market during the dioxin crisis. At the time of writing (December 2008), there is another European dioxin animal feed contamination concern, this time in Ireland (ProMED-mail 2008*a*).

Large amounts of data are collected by human and animal health surveillance systems. For example, the National Notifiable Disease Surveillance System (NNDSS) in the USA, run and maintained by the Centers for Disease Control and Prevention (CDC), has tracked 52 different diseases since January 1999. These are reported weekly at both the state and national level and include many zoonotic diseases e.g. brucellosis, salmonellosis, and

rabies. In New Zealand, the Notifiable Disease Surveillance system for humans currently covers about 50 diseases including leptospirosis, campylobacteriosis, and hydatidosis. Geo-referenced notifiable disease information is entered in real-time on a web-based application, EpiSurv, from laboratories, health providers, and regional medical officers of health. These data are analysed at the Institute of Environmental Science and Research (ESR) on behalf of the Ministry of Health.

The UK launched its strategy for enhancing veterinary surveillance in 2003, a major part of which is the Rapid Analysis and Detection of Animal Related Risk (RADAR) system. This is designed to bring together key surveillance information collected in other systems about animal diseases and conditions in a structured and consistent way. Types of data recorded by the system include agricultural holdings, land and livestock data from UK government databases, diagnostic data from veterinary laboratories, animal health data from private veterinarians and animal owners, and meteorological information.

Public health surveillance aims to detect disease outbreaks and clusters, identify changes or trends in health-related problems, and monitor the effectiveness of prevention and control programmes. Furthermore, the results of surveillance activities can assist in establishing public health programmes and priorities, understanding the natural history of disease, and stimulating analytical research.

The OIE (World Organisation for Animal Health) identifies similar aims for animal health surveillance: determining the occurrence or distribution of disease or infection, while also detecting exotic or emerging diseases as early as possible, monitoring disease trends, controlling endemic and exotic diseases, and providing data to support the risk analysis process, for animal health and/or public health purposes. There are two important additional aims of animal health surveillance that are not shared by human health surveillance: demonstrating the absence of disease or infection, and substantiating the rationale for sanitary measures.

To meet these aims, a surveillance system should have some key attributes that were first recorded by Thacker (Thacker et al. 1988). These include sensitivity, timeliness, flexibility, simplicity, and adequate positive predictive value. They continue to be important benchmarks of effective surveillance systems (German 2000, Zepeda & Salman 2003, Jajosky & Groseclose 2004, Babin et al. 2007, Buehler 2008).

Data from surveillance systems can be used prospectively or retrospectively, and it is

important to distinguish between the two. Retrospective and prospective analyses are used to respond to different types of health surveillance needs.

In a retrospective analysis, the data set is 'complete'. This type of analysis can fulfil most of the above aims and will be reviewed here. The retrospective analysis of surveillance data has been extensively developed throughout this thesis using data from the Danish swine *Salmonella* surveillance and control programme. Data that is retrospectively analysed is, by definition, not continually updated, so this type of analysis does not allow for real-time detection of outbreaks.

However, there are many situations when a repeated analysis of accumulating data over time is called for: the prospective case. For example, if we wish to detect an increased incidence of disease in real time. In this situation timeliness is important as the sooner the detection is made the sooner interventions can be put in place (Jajosky & Groseclose 2004). This is particularly important in the current global environment of disease emergence, climate change, increased trade, and the threat of bioterrorism. Prospective methods will also be reviewed here.

2.2 Surveillance

Before considering the methodology, it is important to consider different approaches to, and types of surveillance. The way in which surveillance data are collected has a large bearing on interpretation. For example, if surveillance data are collected only on a specific subset of the population we need to consider the representativeness of these data whilst interpreting them.

2.2.1 Active approaches to surveillance

An active approach to surveillance means that the onus is on the organisation conducting the surveillance to obtain the surveillance data from the providers, such as physicians, veterinarians and laboratories (Buehler 2008). This is considered to be the 'gold standard' of systems as it has the potential to provide comprehensive, accurate data on disease incidence that can be extrapolated to larger populations. This is providing that the information

that is supplied is an unbiased sample of the target population. Active surveillance is often the result of structured surveys with a particular question in mind, e.g. whether or not New Zealand is free of BSE. This approach is often costly and complex.

The Foodborne Disease Surveillance Network (FoodNET) is an active, laboratory based surveillance program for foodborne pathogens of humans in the United States of America (USA) (Scallan & Angulo 2007). Personnel routinely contact laboratories to ascertain confirmed cases of disease such as campylobacteriosis, cryptosporidiosis, salmonellosis and haemoltyic ureamic syndrome. Even though laboratories are audited twice yearly to ensure all cases are ascertained, there is still potential for under-ascertainment prior to and beyond laboratory confirmation. A study in the United Kingdom estimated that for every case of infectious intestinal disease detected by national laboratory surveillance, there were 136 in the community (Wheeler et al. 1999). A Canadian study reported a larger disparity, with each case of enteric illness reported to the province of Ontario reflecting an estimated number of cases in the community ranging from 105 to 1389 (Majowicz et al. 2005). The use of capture-recapture methods to remedy this incomplete counting of cases is developed by Hook & Regal (2004). Figure 2.2 shows the steps that must be followed for a case to be ascertained in a surveillance system.

2.2.2 Passive approaches to surveillance

These approaches are initiated by the provider of the data, and so relies on their willingness to report an infectious disease to public or veterinary health authorities (van Beneden et al. 2007).

In New Zealand, the national notifiable disease surveillance system (EpiSurv) is an example of a passive system. Human health professionals and laboratories are required to report notifiable disease that they suspect or diagnose to their local Medical Officer of Health (Population and Environmental Health Group 2008). Zoonotic diseases covered here include those caused by *Cryptosporidia* and *Salmonella spp*. Although there is a legal requirement to notify, this is truly a passive system, in that the onus to notify is on the data provider and not on the organisation conducting the surveillance.

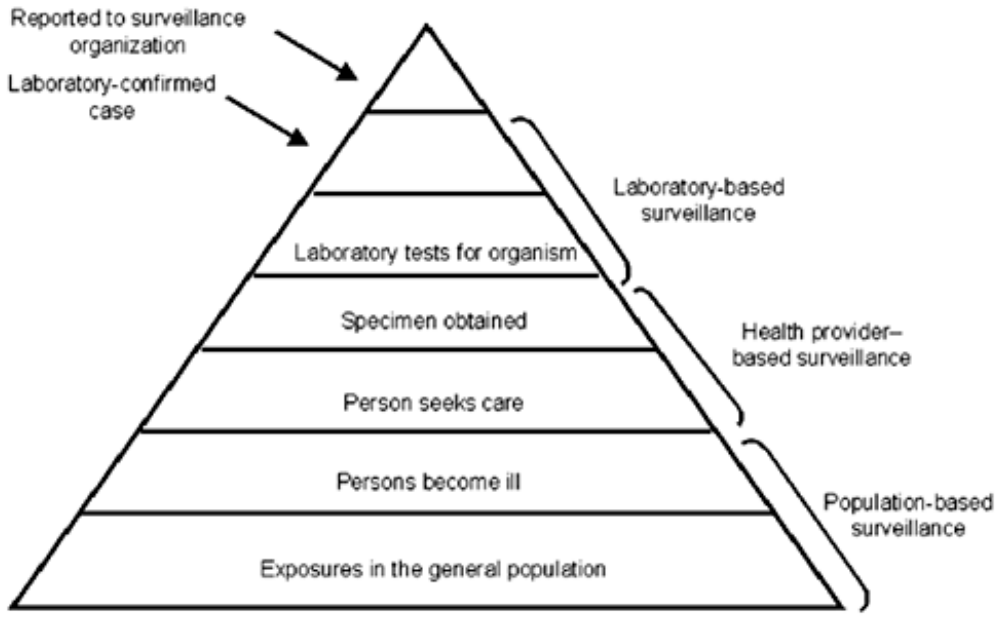


Figure 2.2: The cascade of surveillance steps that must occur for laboratory-confirmed cases to be ascertained through active surveillance. Source Centers for Disease Control and Prevention

.

Passive surveillance is less expensive and easier to operate than active systems, but typically underestimates the true incidence of many diseases. It is a highly specific form of surveillance but is neither timely nor sensitive. The issue of timeliness of public health surveillance systems for infectious diseases is addressed in a review by Jajosky & Groseclose (2004). Using NNDSS data from 1999 to 2001, it was reported that timeliness of reporting varied by disease, with salmonellosis, cryptosporidiosis, and *E.coli* O157:H7 infections having less than 40% of cases reported within the median incubation period for the disease.

The distinction between active and passive surveillance is not always clear and hybrids have been developed. For example, in New Zealand, surveillance of Acute Flaccid Paralysis (AFP) is carried out to fulfil the requirements of the World Health Organisation (WHO) for certification of polio eradication (Population and Environmental Health Group 2008). Every month specialists in paediatric practice are sent a reply-paid card from the New Zealand Paediatric Surveillance Unit. They are then prompted to fill in and return the card detailing if in the previous month they have seen any cases of AFP. Despite the follow up not all specialists respond, and the submission rate of stool sample testing does

not yet meet WHO's criteria.

2.2.3 Syndromic surveillance

If the goal is to detect change in surveillance data as early as possible, then the focus must move towards analysis of data collected before a definitive diagnosis is made. Examples of these data include supermarket sales of items like tissues, orange juice, or paracetamol, pharmacy sales, school and work absenteeism, daily GP office visits and laboratory test requests. These data will provide signals at different times of disease progression, e.g. in chronological order a person is likely to purchase an over-the-counter remedy before visiting the doctor (see Figure 2.3).

Developments in information technology have facilitated the capturing of these data and subsequent automation of surveillance systems (Mandl et al. 2004, Hauenstein et al. 2007, Lawson & Kleinman 2005). As these types of systems are in action before a diagnosis is made, the data need to be sorted into groups, or syndromes, before anomaly detection algorithms are applied. Health-care data sources use groupings like respiratory illness or gastro-intestinal illness, but other data sources such as work absenteeism records need different groupings. Lombardo et al. (2003) provides a widely used syndrome category set. This was developed by the architects of the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE).[1]

This type of prospective surveillance is both timely and sensitive. Buckeridge (2007) reviews 35 studies that used automated syndromic surveillance to detect disease outbreaks. The systems reviewed could detect large seasonally occurring outbreaks with sensitivity and timeliness that was comparable to, or better than, systems that relied on diagnostic data alone.

Rolka et al. (2007) report on the analytic issues surrounding syndromic surveillance data captured from multiple sources. These authors discuss creative analytic approaches that address issues around the use of secondary data such as over-the-counter (OTC), emergency department (ED), and laboratory test order data. They identify 'data lag', 'time alignment', and the 'unlinked data source' problem as key issues.

[1] `http://www.geis.ha.osd.mil`

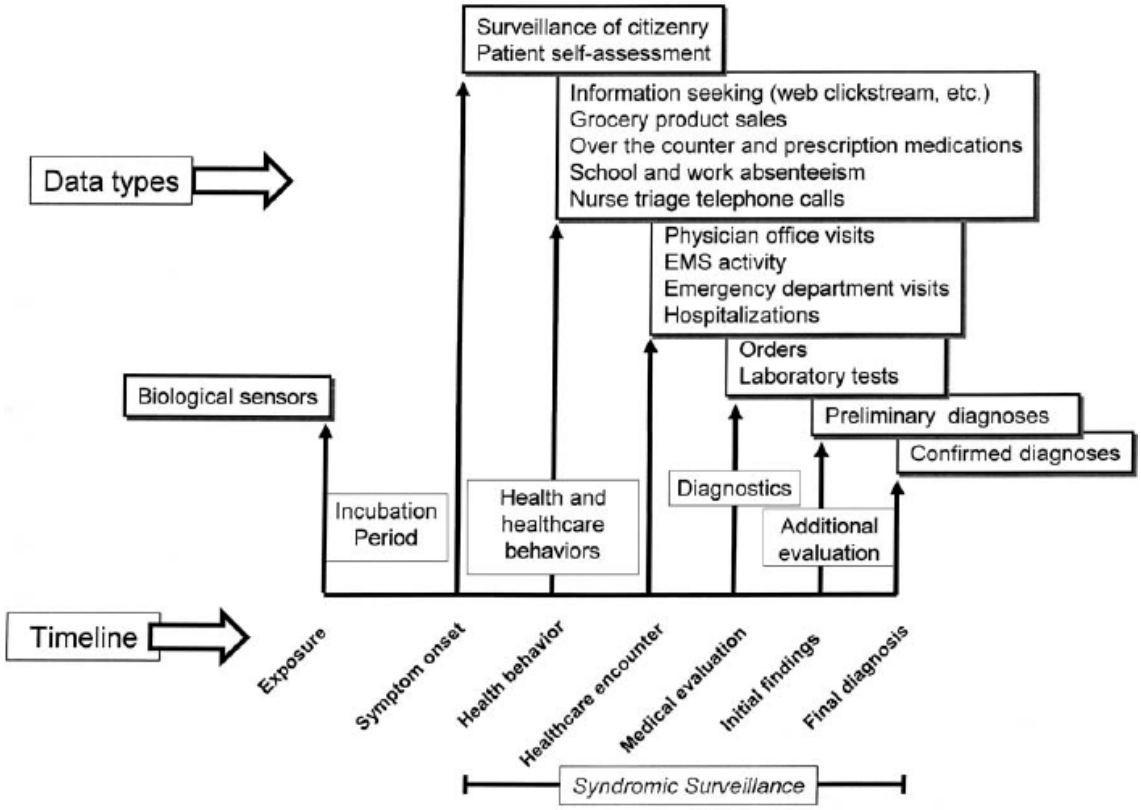


Figure 2.3: A progression of useful data sources for syndromic surveillance as related to the underlying infection and associated behaviours. Source Mandl et al. (2004).

Three information systems that capture syndromic animal health data from veterinarians and animal owners to detect emerging diseases have been reported by Vourc'h et al. (2006). The systems described are the Veterinary Practitioner Aided Disease Surveillance in New Zealand, the Rapid Syndrome Validation Project Animal in the United States, and 'Emergences' in France. It is envisaged that, if successfully incorporated, these systems will define the 'normal' clinical baseline for syndromes and rare diseases, allowing statistical confirmation that an atypical syndrome is emerging.

Highly sensitive and timely syndromic surveillance can suffer from lack of specificity and the occurrence of false alarms. These occur when an outbreak is detected when in fact there is none. False alarm rate is a key measure of the performance of syndromic surveillance systems.

Consider a time series of health events. For example the number of thermometers sold over the counter in Auckland, (X), per day, (t), as in Equation 2.1.

$$X = \{X(t); t = 1, 2, ...\} \tag{2.1}$$

Our interest is prospective in that we aim to detect a change in the series at an unknown time point, τ, as quickly and accurately as possible. The time when we inspect the series, say, daily, will be a predetermined decision point, s. We want to decide if the process is in-control or out-of-control. We call these $D(s)$ and $C(s)$ respectively. To make this decision, we use the accumulated observations $X_s = \{X(t); t \leq s\}$ to produce alarm sets $A(s)$, so that if $X_s \in A(s)$ then the process is out-of-control, (in state $C(s)$), and an alarm is triggered, invoking a response such as further investigation of cases, vaccination, or dissemination of information. Usually this is done by an alarm function $p(X_s)$, and a control limit $g(s)$, where the time of the alarm, t_A, is as follows:

$$t_A = min\{s; p(X_s) > g(s)\} \tag{2.2}$$

For health surveillance, a common usage is $D(s) = \{\tau > s\}$ and $C(s) = \{\tau \leq s\}$. The change to be detected in our scenario could well be a change in the mean level of X, the average number of thermometers sold in Auckland per day. Most of the literature on syndromic surveillance considers an abrupt or step form, where a parameter changes from one constant level to another. This might be caused by a sudden bioterrorist attack. However changes can also be linear, exponential or gradual. A gradual change is likely to be seen during a naturally occurring outbreak of infectious disease and these can be particularly problematic to detect.

Further detailed reports on syndromic surveillance are mentioned throughout this chapter and covered in detail in an extensive report by the Centers for Disease Control and Prevention (2004).

2.2.4 Sentinel surveillance

In sentinel surveillance, the health status of a population is periodically assessed. A sentinel is defined as a person or thing that watches or stands as if watching. In sentinel surveillance data is collected from only a subset of a larger population. Therefore, it is important to ensure the representativeness of those under surveillance. The sentinel

population may be a group that is at higher risk of developing the disease under surveillance. For example, measuring the prevalence of human immunodeficiency virus (HIV) infection in New Zealand amongst intravenous drug users (AIDS Epidemiology Group 2007).

Sentinel surveillance systems may also target an at-risk time period rather than an at-risk population. For example, the human influenza sentinel surveillance system (Population and Environmental Health Group 2008) operates from May to September each year in New Zealand. This gathers data on the incidence and distribution of influenza. In 2007, this was based on a network of 87 general practices that record the number of consultations for influenza-like illness each week by age group.

Sentinel surveillance systems may also target at-risk individuals due to their location. For example, to demonstrate freedom from arboviral disease, New Zealand regularly tests cattle from sentinel herds in areas where the local climate would favour establishment of *Culicoides spp*. Seventeen herds in 10 districts are tested annually for antibodies to bluetongue virus, epizootic haemorrhagic disease virus (serotype 2), Palyam (DAguilar) virus, Simbu viruses (Akabane and Douglas), and disease vectors in the genus *Culicoides*. Most of the herds are in northern coastal areas (see Figure 2.4) where the local climate is warm and wet, favoured habitat conditions for the *Culicoides* midges.

2.2.5 Risk-based surveillance

When a population perceived to be at a greater risk of a disease is preferentially sampled, this is called risk-based or targeted surveillance. The use of targeted surveillance is extensive in both the veterinary and human world as a tool to make best use of limited resources. A formative discussion paper on risk-based surveillance in veterinary public health from 2006 sought to develop a framework for practical implementation of risk-based surveillance (Stärk et al. 2006). This paper was the first to clearly identify the two key components of a risk-based surveillance system: (1) preferential surveillance for hazards that have serious consequences to human or animal health and trade; and (2) preferential sampling in sub-populations that have a higher risk of having a disease. Although it is formative in some of the conceptual issues around risk-based surveillance, it is somewhat limited in its discussion of the analytical challenges around processing data

from such systems. Since its publication there have been many examples of the application of risk-based surveillance in animal health (Chriel et al. 2005, Martin et al. 2007*a*,*b*) including in the field of food-borne zoonotic disease (Alban et al. 2008).

For example, surveillance for *Trichinella spp.* in Denmark is risk-based: only sows, boars, and outdoor-reared pigs are sampled (Alban et al. 2008). Outdoor-reared pigs present a higher risk of introduction because of the possibility of contact with wildlife. Sows and boars are at an increased risk as a result of their age which gives them a longer exposure time when compared with finisher pigs. Another example is in the Belgian swine *Salmonella* control programme, where only the 10% of pig herds with the highest *Salmonella* infection burden (denoted high-risk herds) participate. Identification of *Salmonella* high-risk pig herds is based on previous serological data (Bollaerts et al. 2008).

The principles of risk-based surveillance are central to the work presented in this thesis.

2.3 Temporal surveillance

Human and animal health surveillance data form a univariate time series when they consist of a single observation recorded at regular time intervals. These observations could be the number of cases of salmonellosis notified per month, the number of emergency department (ED) cases involving respiratory symptoms per week, or the number of thermometers sold per day. If observations are taken on two or more time series simultaneously then the series is bivariate or multivariate, respectively. For example, this would occur if in addition to recording the number of ED cases involving respiratory symptoms per week, we had simultaneously recorded the sex, age and address of each case.

When considering the detection of change in surveillance data, we need to ask the following questions: are there consistent patterns such as trend or seasonality? Is there an abrupt change in the observed incidence or prevalence of disease? And is our interest prospective or retrospective?

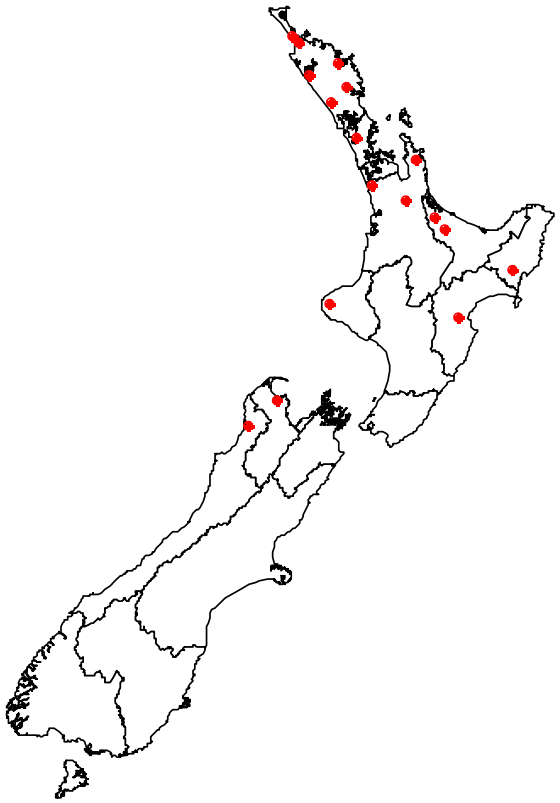


Figure 2.4: Map of New Zealand. The dots indicate sites of arbovirus sentinel cattle herds and *Culicoides spp* light traps. These herds are mainly in northern and coastal areas. Source: Ministry of Agriculture and Forestry 2008.

2.3.1 Time series methods

Time series methods have traditionally been used to describe trend, cyclicity, and autocorrelation in equally spaced data. There are many applications of time series analysis including those in finance, ecology, and increasingly public health monitoring and surveillance. There are two main purposes of time series analysis: to describe, explain or model the mechanism that gives rise to the data; and to forecast or predict the future values of

a time series based on its history and other related series or variables. In a univariate series, the interest is in how observations within a series are related to others in the same series. In a bivariate analysis, interest lies in identifying the relationship between the two series at the same time or with one leading by one or more lags (for example, between the residuals of two series of diseases or between the residuals of a disease and those of an effector variable such as weather).

Surveillance data often exhibit correlation at varying temporal scales. On a reducing scale, if we consider the example of the number of ED cases involving respiratory symptoms per day, this correlation could manifest as a long term trend of increasing incidence, annual winter peaks, day-of-the-week effect, to sequential autocorrelation. Statistical time series modelling is especially suited to accommodate these correlations. Correlations in themselves may be a nuisance e.g. we may wish to remove a seasonal pattern from the series to reveal the underlying structure. Conversely, we can exploit the correlation structure in surveillance data as it makes us reasonably confident of future behaviour.

Graphical exploration

The first approach for all data analysis is exploratory. For a time series analysis a simple plot of the data as a function of time is a crucial initial step towards understanding the data. For example, Figure 2.5 shows the monthly returns of deaths from bronchitis, emphysema, and asthma in the UK from 1974 to 1979 for males and females (adapted from Diggle (1990)). These types of plots are called time-series or run sequence plots (Chambers et al. 1983). This data set will be used throughout the time series methods part of this chapter to illustrate some of the methods discussed. It will be referred to subsequently as the respiratory deaths data.

Time series plots not only reveal the trend and seasonal variation (see Figure 2.5), but also missing values, outliers and abrupt changes in a series. This was illustrated in an eight year study of nosocomial infection in a Spanish hospital (Fernandez-Perez et al. 1998), when four distinct time periods were identified by simply graphing the incidence risk of infection as a function of time (Figure 2.6). These periods were: (1) a training period; (2) a period of minor incidence; (3) an abrupt change associated with a medical strike; and (4) a period of increased incidence.

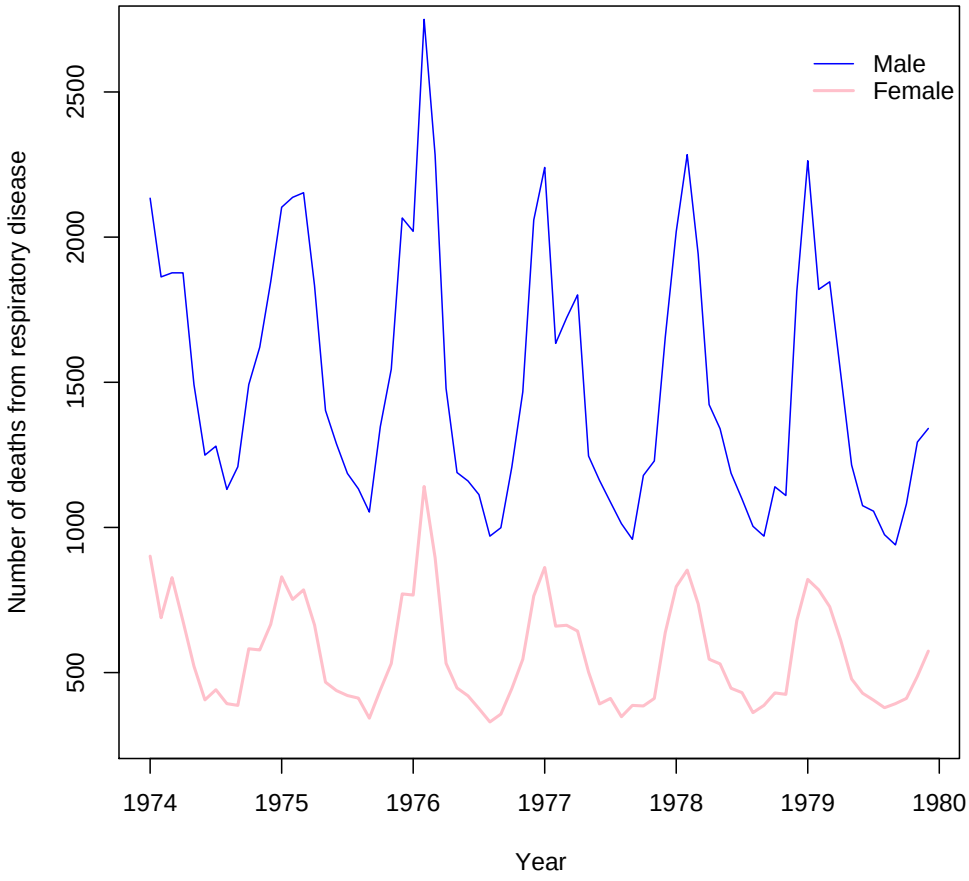


Figure 2.5: Monthly deaths from bronchitis, emphysema, and asthma in the UK, 1974-1979, stratified by sex. Source: Diggle (1990).

Overlying several years of data on the same plot allows a visual comparison to be made between years. Figure 2.7 illustrates this with weekly episodes of cases of *Salmonella typhimurium* infection in humans in Denmark. At the time of writing (December 2008), Denmark is experiencing the largest *Salmonella* outbreak since surveillance was initiated in 1980. The outbreak is caused by *Salmonella typhimurium* phage type U292 (Ethelberg et al. 2008*a*,*b*). At present, more than 1000 people have been infected since February 2008, and six have died. The source of infection has not been confirmed. The graph clearly shows the above average numbers for 2008 attributed to the outbreak of strain U292.

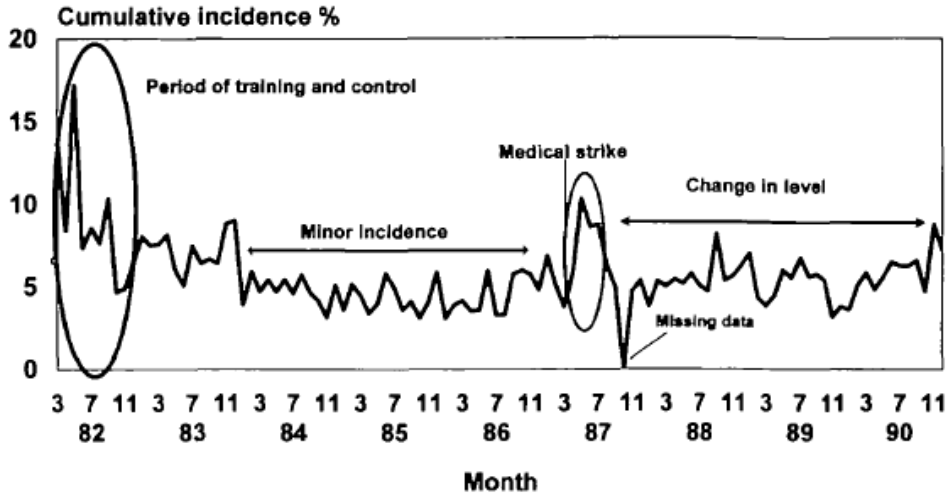


Figure 2.6: Monthly cumulative incidence of nosocomial infection in Hospital General de Guadalajara, Spain, between March 1982 and December 1990. Source: Fernandez-Perez et al. (1998).

Time series graphs also act as an aid for statistical analysis by showing the structure of the data and suggesting hypotheses for further investigation (Chatfield 2004). An example of this is the correlation structure in a time series plot of number of births per week in Bangladesh (Zeger et al. 2006) which suggests an autoregressive model (see forthcoming section on Box Jenkins methodolgy). Plotting two or more stationary series on the same axes may prompt an investigation of the association between them, as there was, for example, reported between weekly campylobacteriosis incidence and average weekly temperature in England and Wales (Louis et al. 2005).

Although it sounds straight-forward, plotting a time-series is not always so. Care must be taken with the choice of scales, where to place the intercept, and how the points are plotted. Furthermore, the data may show a lot of variation resulting in a plot that is difficult to interpret. This will be addressed in the next section.

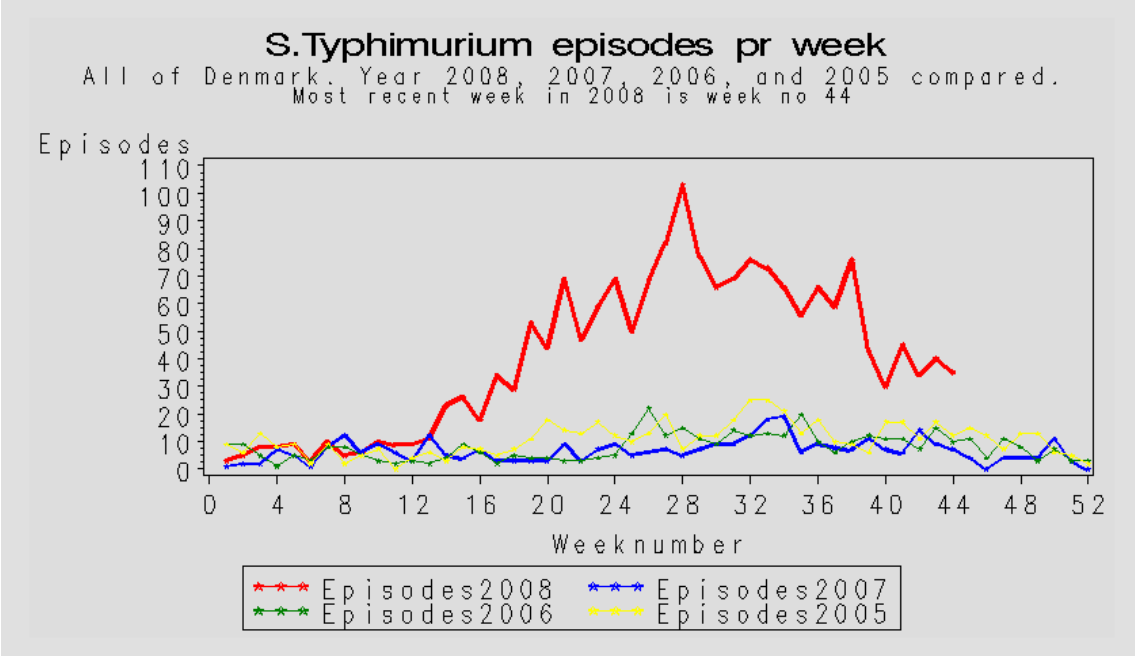


Figure 2.7: Reported *Salmonella typhimurium* cases per week in Denmark. The large sustained outbreak from week 12 in 2008 is largely due to phage type U292. Source: http://www.ssi.dk/graphics/html/Germ/germ/stm.htm

.

Temporal trend

Trend may be loosely defined as 'the long term change in the mean level of the time series' (Chatfield 2004). More formally it is the long term movement of the series, which is a systematic component that changes over time, and generally does not repeat itself within the time range of the available data. What is 'long term' is a subjective assessment, so when considering trend we must take into account the number of observations.

Methods used to detect change in surveillance data which show a trend will depend on whether the trend is of interest, or if it is of nuisance value only. Generally the further analysis of time series data requires that the trend be removed, so often the latter is the case. Trend removal is a major component of making a series stationary. In this context, stationary means the series has no systematic change in the mean or the variance, and periodic variations have been removed (Diggle 1990). It is an important prerequisite for further analysis of the data.

Temporal trend identification

Even though trend removal is required for further analysis, as epidemiologists our interest is also in the trends themselves in disease surveillance data.

As most surveillance time series data will contain considerable variation, the first step in the process of trend identification is likely to be smoothing. Smoothing generally involves local averaging of data so that the random components of individual observations are removed. Techniques include moving average, exponentially weighted moving average, and kernel and loess smoothing (Diggle 1990, Chatfield 2004). Figure 2.8 shows the unstratified respiratory deaths data in its raw state, and loess smoothed using two smoothing spans (0.2 and 0.5). This technique uses a locally-weighted polynomial regression, where the f value is the proportion of points in the plot which influence the smooth at each value (Cleveland 1979). Larger values of f provide a greater level of smoothing.

In three studies of human campylobacteriosis, different techniques for temporal trend identification were used. Baker et al. (2007) plotted the time series and used a chi-squared test for trend on New Zealand data from 1995 to 2003. As well as identifying an increase in incidence in the plot, the statistical test gave the probability of rejecting the null hypothesis of no trend, at $p < 0.01$. Time-series plots and regression analysis were performed to identify national and regional trends in Scottish cases of campylobacterosis from 1997 to 2001 (Miller et al. 2004). One region (Borders) had a significant trend ($p = 0.04$) where there was a reduction in cases over this time period. For the remainder of the regions and the country overall there was no significant trend. And in a spatio-temporal study, a simple smoothed trend line was fitted using a moving average filter to visualise Canadian cases of campylobacterosis from 1996 to 2004 (Green et al. 2006). No temporal trend was observed in these data. Although all these studies go beyond the identification of trend in their analysis, in each case they highlight the value of applying the straight-forward technique of plotting the time series as a preliminary step.

A seven-day moving average smoother was used to remove the effect of day-of-week and holidays on pharmacy over the counter (OTC) sales data (Centers for Disease Control and Prevention 2004). Another method used to detect trends in time series is based on cumulative sums which will be discussed in the forthcoming section on statistical process control in this chapter.

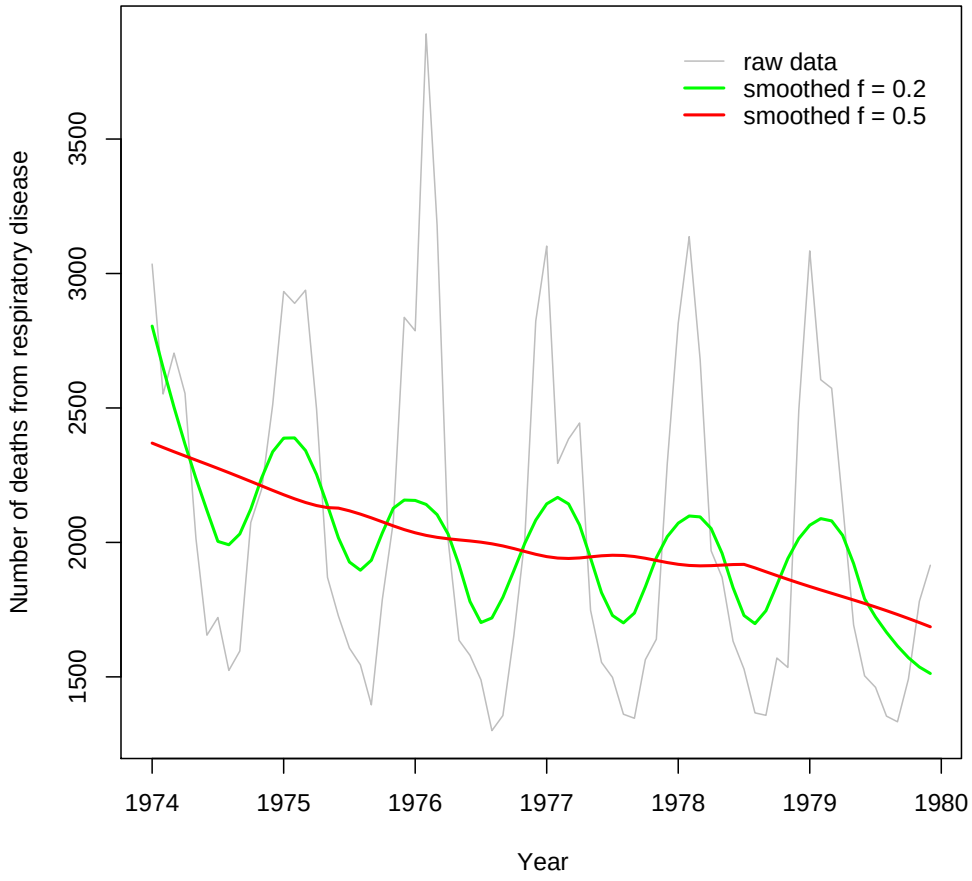


Figure 2.8: Loess smoothed monthly deaths from bronchitis, emphysema, and asthma in the UK, 1974-1979. The f value is the proportion of points in the plot which influence the smooth at each value. Source: Diggle (1990).

It is important to be aware that with any emerging pathogen such as *Campylobacter* the trend seen may largely reflect changes in laboratory or clinical practice, rather than real changes in disease incidence. Artefactual changes may also result from changes in the extent to which diagnosed disease is reported to public health authorities, and hence to the national notifiable disease surveillance system.

Temporal trend removal

There are two main approaches to trend removal: by curve/line fitting or by differencing (Diggle 1990, Chatfield 2004). In the first approach the raw series (shown as the grey line in Figure 2.8) could have the trend removed by fitting a straight line to the data

and examining the residuals from the fit. In the second approach, differencing, could be applied. This is a type of filtering whereby the raw series is converted into another series (the differenced series). The differenced series will be transformed as: $Y = Y - Y(lag)$. After differencing, the resulting series will be a vector of length $N - lag$ (where N is the length of the original series). Both options for trend removal are shown in Figure 2.9.

The different approaches to detrending are important for forecasting; fitting a line or curve to the series places global assumptions on the data which may poorly estimate the fit beyond the range of the period of interest (Diggle 1990). Most texts generally advise the use of differencing if forecasts are to be made (Diggle 1990, Chatfield 2004), and differencing is an integral part of autoregressive integrated moving average (ARIMA) modelling of time series (Box et al. 1994). More detailed discussion on the different approaches are well developed in the economic forecasting literature (Clements & Hendry 2001, Qi & Zhang 2008) but are beyond the scope of this review.

Seasonality

The seasonal distribution of infectious diseases has been well recognised and extensively studied (Grassly & Fraser 2006, Altizer et al. 2006). It is of interest in itself in providing clues to the possible aetiology of disease, but also as seasonal behaviour holds promise of predictability, there is great potential to exploit this for surveillance purposes. If it is possible to predict the occurrence of disease then the use of preventative measures and health-care resources can be targeted for those times, and we are provided with a baseline by which to compare future behaviour of the series.

Seasonal rhythms in human health have been observed to exist since at least 400 BC, when Hippocrates stated that *'Whoever wishes to investigate medicine properly should proceed thus: In the first place to consider the seasons of the year and what effect each of them produces'*, (cited by Hare (1975)).

Veterinary medicine is no exception: recent examples include the spread of bluetongue through northern Europe associated with a particularly hot summer (Enserink 2006), and the winter peak of rabies in Korea associated with the ethology of raccoon dogs in outbreak areas (Kim et al. 2006).

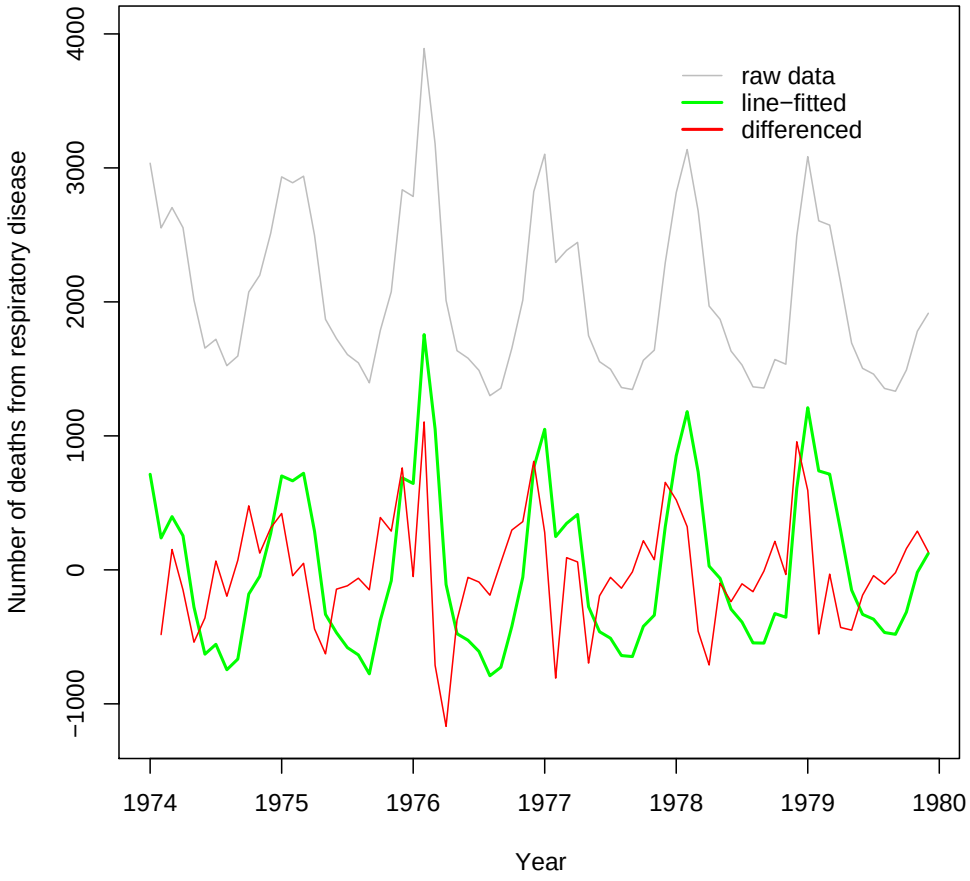


Figure 2.9: Monthly deaths from bronchitis, emphysema, and asthma in the UK, 1974-1979. The raw series (grey) has been detrended by straight line fitting (green) and by differencing (red). Source: Diggle (1990).

As well as the climatic reasons for the seasonal pattern of infectious disease, there are human-imposed annual drivers, sometimes called seasonal forcing. These include those associated with the management of livestock, such as a late spring peak of clinical mastitis in a spring calving dairy herd, and school terms, such as the September peak in childhood pertussis in the USA coinciding with the commencement of the school term (Shah et al. 2006).

Techniques to determine seasonality range from simple plotting of the time series, smoothing, and monthly box plots, to analytical techniques in both the time and frequency domains. Figure 2.10 is a monthly box plot of the respiratory deaths data. This shows a distinct summer trough and winter peak in respiratory deaths and that there is much less

variation in death numbers during the summer season.

Moineddin et al. (2003) proposed fitting an autoregressive model and determining the coefficient of determination, to measure the strength of the seasonal effect for asthma and atrial fibrillation hospitalisations. The coefficient of determination measures how well the next value can be predicted using the month as the only predictor. This technique was applied to influenza and pneumonia hospitalisations from 1988 to 2002 in a Canadian analysis (Crighton et al. 2004). This technique gave good results when used as part of a predictive model to forecast hospitalisations for 52 common discharge diagnoses from Ontario acute care hospitals (Upshur et al. 2005).

A common approach to handling a seasonal time series Y_t is to decompose its components of trend T_t, season S_t, and remainder R_t by a sequence of robust locally distance weighted regressions, where $Y_t = T_t + S_t + R_t$ (Cleveland et al. 1990). Decomposition using this technique identified a very similar seasonal pattern for the prevalence of *Salmonella* in pork and the incidence of salmonellosis in humans in Denmark (Hald & Andersen 2001). The late summer peak in pork prevalence appeared four weeks before the peak in human cases in this Danish study of data from 1995 to 2000. Figure 2.11 shows the decomposition of the lung deaths data clearly showing the components of season, trend, and residuals.

Further seasonality examples, including the use of generalised additive models will be demonstrated throughout this chapter.

Approaches to time series analysis

There are two overlapping streams within time series analysis: the time and frequency streams. Although mathematically these are equivalent and one can be derived from the other, this is of little practical use to us when applying these approaches to detecting change in surveillance data. The mathematical equivalence in the approaches does not equate to statistical equivalence, as the use of the underlying data in each approach highlights different aspects (Diggle 1990). In the time domain, the variation in the time series is described in terms of the way in which observations are related statistically with one another at different times. These methods focus on how a time series evolves from one time to the next, e.g. ARIMA models, cross correlation function, and hidden Markov

models.

In the frequency domain, inference is based on the spectral density function which describes how the variation in the time series may be accounted for by cyclic components at different frequencies. These methods measure the waves in a time series and include spectral analysis, Fourier models, and periodic regression.

Methods in the time domain

In the time domain, the variation in the time series is described in terms of the way in which observations are related statistically with one another at different times. Techniques here are centred around correlation. Inference is based on the autocorrelation of the series and autoregressive and/or moving-average models are fit to the data set after trend removal or seasonal adjustment (Box et al. 1994).

Autocorrelation refers to the correlation of a time series with its own past and future values. This is sometimes called 'serial correlation', which refers to the correlation between members of a series of numbers arranged in time. Alternative terms are 'lagged correlation' and 'persistence'. Three tools for assessing the autocorrelation of a time series are: (1) the time series plot, (2) the lagged scatterplot, and (3) the autocorrelation function. The lagged scatterplot, is a scatterplot of the time series against itself offset in time by one to n time periods. Figure 2.12 shows lagged scatterplots (up to 12 lags) of the log lung deaths data (previously detrended by fitting a straight line). The series is not stationary and still has a strong seasonal component, which can be seen by the strong autocorrelation at lags 6 and 12. Fitting month as a factor in the linear model and then re-running the lagged scatterplots resulted in no pattern at these lags, indicating that the autocorrelation at 6 and 12 months had been successfully accounted for by controlling for month.

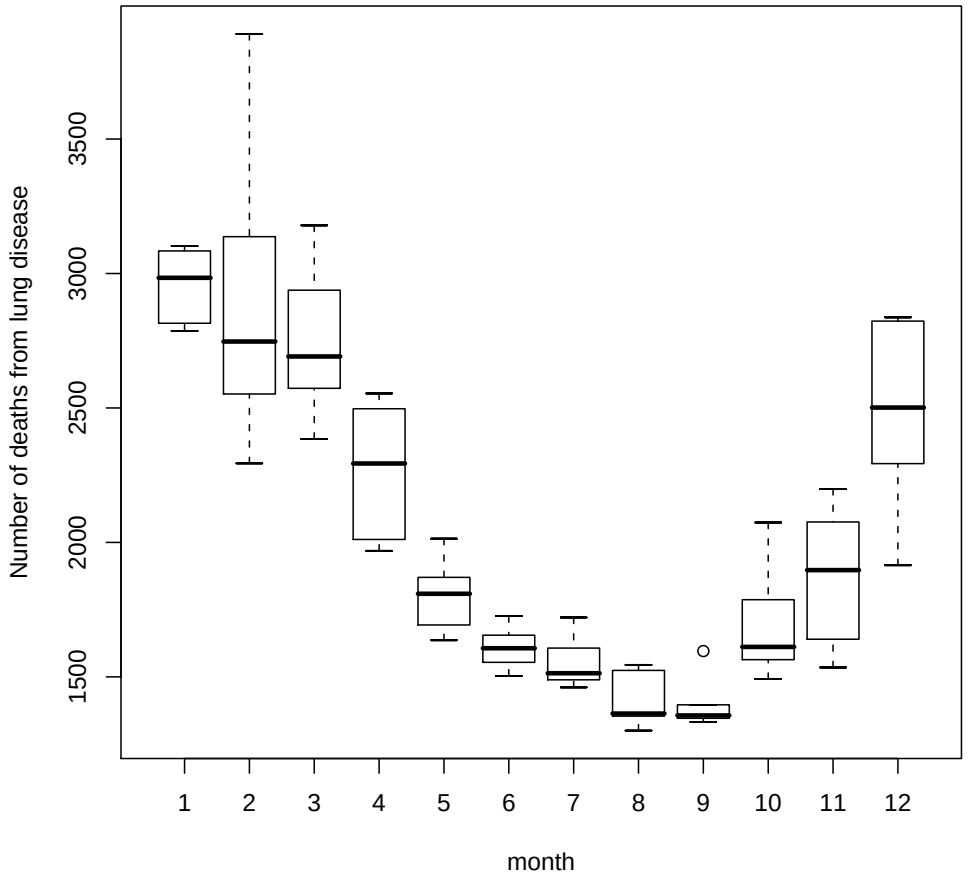


Figure 2.10: Box plot of monthly deaths from bronchitis, emphysema, and asthma in the UK, 1974-1979, showing a strong seasonal pattern with a winter peak and summer trough. Source: Diggle (1990).

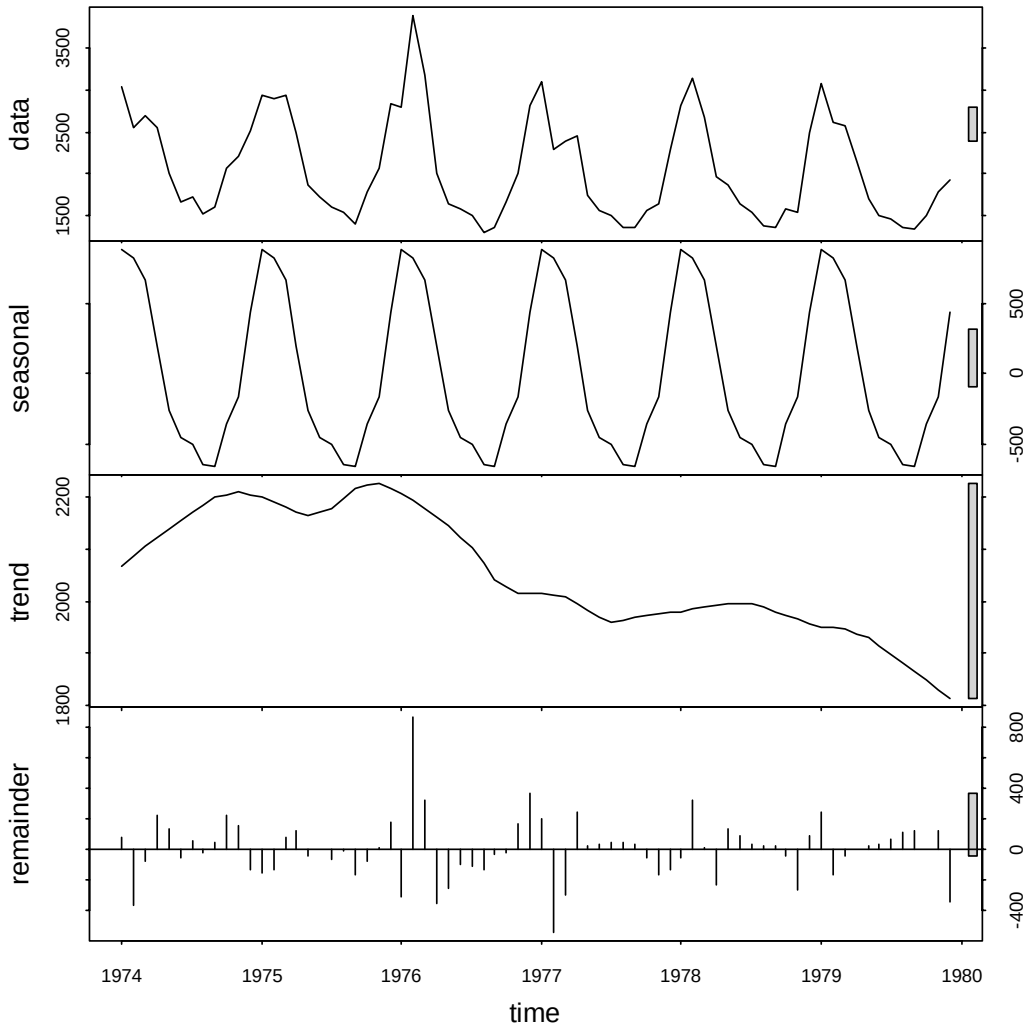


Figure 2.11: Decomposition of monthly deaths from bronchitis, emphysema, and asthma in the UK, 1974-1979 into seasonal, trend, and irregular component using the method of Cleveland et al. (1990). Source: Diggle (1990).

If lagged scatterplots reveal autocorrelation the next step is to confirm this by investigating the sample autocorrelation coefficients. These measure the correlation between observations at different times. The set of autocorrelation coefficients arranged as a function of separation in time is the sample autocorrelation function (ACF) (Diggle 1990, Chatfield 2004). Figure 2.13 (a) shows the ACF of the log respiratory deaths data previously detrended by fitting a straight line. There is a strong sinusoidal pattern in the ACF with peaks at lags 6 and 12 and no decay, indicative of a strongly seasonal pattern. Lag 0 always shows an autocorrelation of one by definition. The dashed lines are the 95%

confidence intervals. Figure 2.13 (b) shows the ACF of the log respiratory deaths data which has been made stationary by fitting month as a factor in the linear model. All of the autocorrelations fall within the 95% confidence limits and there is no apparent pattern. This is what is expected if the data are random.

If we identify autocorrelation, then lagged values of the same series can be added as covariates into a regression model. This technique is commonly used in the medical surveillance literature for infectious disease, as very often the number of disease cases in any given time period is strongly correlated to the levels of the preceding period. For example, Zhang et al. (2008) used a four-order autocorrelation of the number of salmonellosis cases in a study of the effect of temperature on salmonellosis in Adelaide, Australia. In a European study investigating the same association, a first-order autoregressive term was included in regression models (Kovats et al. 2004). In these examples, the inclusion of the lagged values of the salmonellosis case time series into the regression model ensures that the autocorrelation features of the data are adjusted for.

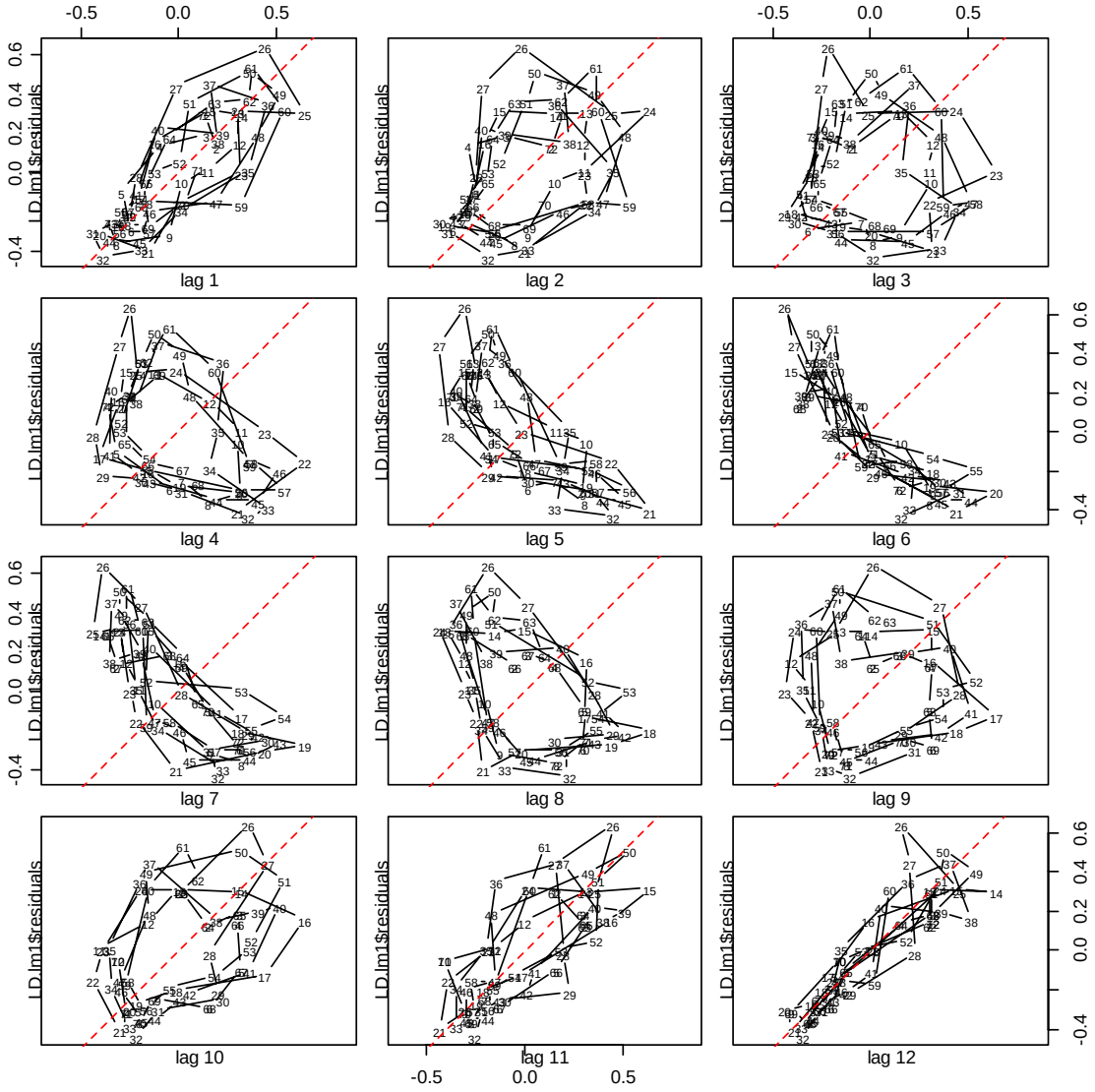
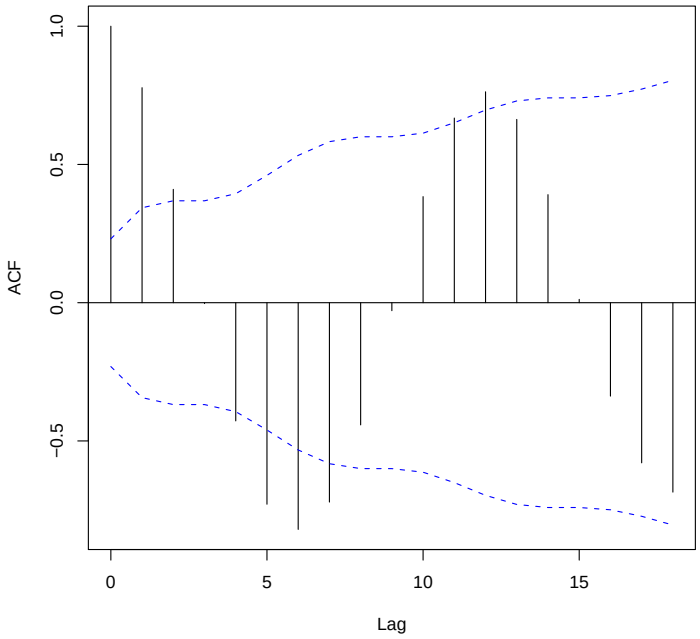
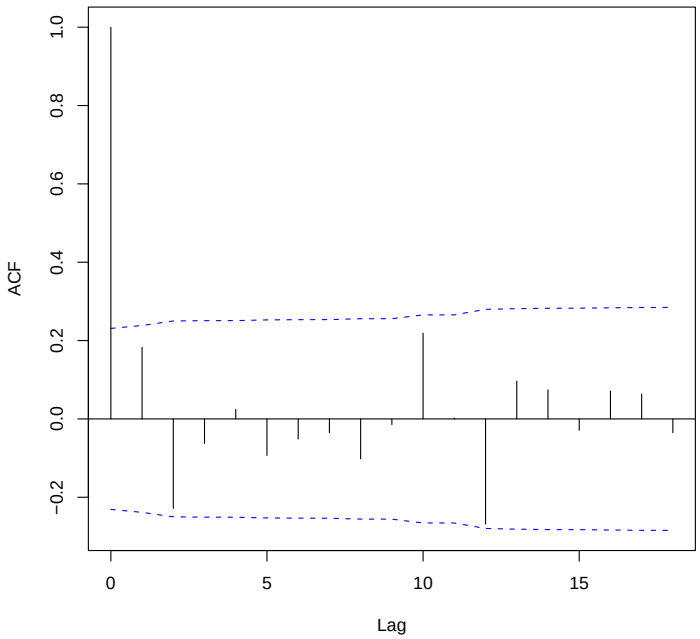


Figure 2.12: Lagged scatterplots of the detrended log respiratory deaths time series, showing strong negative correlation at lag 6 and strong positive correlation at lag 12 indicating a seasonal pattern. Source: Diggle (1990).



(a) ACF: straight line-fitted detrended



(b) ACF: month-fitted detrended

Figure 2.13: Autocorrelation function plot of the: (a) straight line-fitted detrended log respiratory deaths time series which still shows a seasonal pattern at 6 and 12 lags, and (b) month-fitted detrended respiratory deaths time series showing a random pattern to the ACF. Source: Diggle (1990)

Once an autoregressive term has been identified as being significant, different model families can be used to provide an optimal fit to the data. For example, three models were compared to determine risk factors for haemorrhagic fever with renal syndrome (HFRS) in Anhui Province, China, from 1983 to 1995 (Hu et al. 2006). The study of risk factors in disease surveillance may be used to make predictions in the face of changing environmental conditions. HFRS is a zoonosis caused by Hantaan type virus and these are a group of serious infectious diseases that have been endemic in many countries of the world. Rodents, mostly mice, are the reservoir of the disease and the source of infection. The first model was a standard linear regression time series model which assumes the expected value of $Y_{(t)}$ (the incidence of HFRS) has a linear form:

$$\hat{Y}_{(t)} = \phi Y_{(t-1)} + \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \tag{2.3}$$

The constant term is β_0, the autoregressive coefficient is $\phi Y_{(t-1)}$, the regression coefficients are $\beta_{1\ldots p}$, and ϵ is the error term. Explanatory variables used were the density of mice, crop production and water level difference in the Huai River. In the linear model, the distribution of Y is Gaussian.

The next model used was a generalised linear model (GLM) which extends Equation 2.3 to allow for the predictor variables to combine linearly to relate to the expected value of $Y_{(t)}$ (the incidence of HFRS) through a link function. The distribution of $Y_{(t)}$ in a GLM may be any of the exponential family distributions (e.g. Gaussian, Poisson or binomial) and the link function may be any monotonic differentiable function (like logarithm or logit). For time series regression modelling, there is a standardised methodology where the expected value of $Y_{(t)}$ is Poisson distributed with a mean of μ (Schwartz et al. 1996) and:

$$log(\mu) = \phi Y_{(t-1)} + \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \tag{2.4}$$

The next extension of Equation 2.4 was to a third model that replaced the linear function of the covariates with a smoothing function, Equation 2.5. This third model had the best goodness-of-fit (lowest deviance) and short-term predictive ability (lowest root mean squared error). This extension will be discussed further in the forthcoming section on generalised additive models (GAMs).

A model comparison was also made by Zhang et al. (2008) by performing modelling to quantify the relationship between climate variations and salmonellosis in Adelaide, Australia. Here standard Poisson regression, autoregressive adjusted Poisson regression, multiple linear regression, and SARIMA (seasonal autoregressive integrated moving average models) all found that the temperature occurring two weeks prior had the greatest significant association with the number of weekly salmonellosis cases. The SARIMA model had both the best forecasting ability and the best goodness-of-fit, suggesting it represents the most appropriate model for these data. SARIMA models have integrated functions controlling seasonal variation, autocorrelation, and long-term trend.

Papers which compare different modelling approaches, such as those by Zhang et al. (2008) and Hu et al. (2006), provide very useful additions to the literature. Benefits include providing accessible accounts of what can otherwise be complex modelling strategies, researchers using different approaches to verify the assumptions of their models by comparing outputs, and bringing techniques from other fields, such as finance, into the analysis of health surveillance data. Emerging complex areas for research such as answering questions around health and climate change are also driving these model comparisons (McMichael et al. 2006).

Introduction to Box-Jenkins Methods

Autoregressive moving average (ARMA) models analyse time-series as a function of its past values (autoregressive part) and past error (moving average part) (Box et al. 1994). As discussed previously, an autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. A moving average model is conceptually more difficult. To put it in the same context as an AR model, a MA model can be thought of as a linear regression of the current value of the series against the error of one or more prior values of the series. In other words the autoregressive model includes lagged terms on the time series itself, and the moving average model includes lagged terms on the noise or residuals. The order of the ARMA model is included in parentheses as ARMA (p, q), where p is the autoregressive order and q the moving-average order. These models are useful as they can approximate a large range of different behaviour in a time series using only these two parameters.

There are several possible reasons for fitting ARMA models to health surveillance data.

Modelling can contribute to understanding the physical system by revealing something about the process that builds persistence into the series. ARMA models can also be used to remove persistence from a time series so that the residual may be more suitable for studying the influence of climate and other outside environmental factors. ARMA models can also be used to predict behaviour of a time series from past values alone. Such predictions can be used as a baseline to evaluate possible importance of other variables to the system. The use of these models and their extensions on surveillance data has primarily been for the latter reason.

Fitting ARMA models is a three step process consisting of identification of the orders p and q, estimation of the parameters, and diagnostic checking. The ACF is used to identify the orders of the moving average part and the partial ACF, (PACF), the autoregressive part. Full description of these methods can be found in texts by Diggle (1990), Chatfield (2004) and Box et al. (1994).

ARMA models are suitable only for stationary time series. To allow for non-stationarity, the models can be extended in two main ways. Firstly, to account for the trend in the series, an integration term is added and the model is termed autoregressive integrated moving average (ARIMA). The integration term represents the order of differencing of the series which is identified by examining the ACF and estimating the variance of the series (Diggle 1990). The ARIMA is classified as an ARIMA (p, d, q) model, where *p* and *q* are as for ARMA models and *d* is the number of non-seasonal differences.

The second extension of the ARMA is to account for seasonality by seasonal differencing and the model can be generalised to a seasonal ARIMA (SARIMA) (Box et al. 1994). The seasonal ARIMA model is classified as an ARIMA $(p, d, q)x(P, D, Q)$ model, where P is the order of the seasonal autoregressive terms, D is the number of seasonal differences, and Q is the number of seasonal moving average terms.

Examples of uses of Box Jenkins methods

A review article by Allard (1998) covers the application of ARIMA modelling to the surveillance of infectious diseases, focussing on their use for forecasting and subsequent aberration detection. Examples of *Campylobacter* and measles notifications in Montreal are provided. An ARIMA model of order (1,0,3) was used to describe and predict hospital bed occupancy during the 2003 SARS outbreak in Singapore (Earnest et al. 2005). A sea-

sonal ARIMA of order (1,0,1) (1,1,1) was used to predict epidemics of Ross River virus (RRV) disease in Brisbane, Australia (Hu et al. 2004). Monthly precipitation was significantly associated with RRV transmission. However, there was no significant association between other climate variables (e.g., temperature, relative humidity, and high tides) and RRV transmission.

Helfenstein (1996) gives accessible coverage of Box Jenkins methodology to time series analysis of surveillance data. Its particular strength is in the step-by-step approach to application and the practical examples using data of interest to epidemiologists.

Box-Jenkins methods have also been applied to syndromic surveillance data. Reis & Mandl (2003) used an ARIMA (1,0,1) model of hospital emergency department visit rates for respiratory syndromes to forecast hospital utilisation needs.

ARIMA models are often used to predict an incidence rate that takes into account the serial correlation of the data. Inherent in their use is the underlying assumption that transformation of the data will lead to a stationary time series, for which a single underlying probability distribution can be assumed. This is not necessarily true, as human and animal health data may well present abrupt and wide changes of magnitude as well as irregular periodicity. Epidemics, environmental change, and vaccination are some of the reasons for these changes.

Notwithstanding the above, Trottier et al. (2006) analysed the stochastic dynamics of childhood infectious disease time series both before and after mass vaccination using Box Jenkins methodology. The authors found that time series of pertussis, mumps, measles, and rubella have about the same stochastic dependence in their consecutive data: generally the number of new cases in one period is given by the number of cases in the previous period and by periodically recurrent random shocks. Even though mass vaccination was expected to have a major impact on disease transmission dynamics (i.e., incidence, average age at infection, long-term periodicity, seasonal cycles), it did not clearly affect the stochastic dynamics.

Nevertheless the forecasting ability of ARMA models and their derivatives are heavily reliant on past values. If patterns of disease change abruptly, disrupting a previous history of stable trends, then aberrations should be readily detected using this methodology. Typically, effective fitting of Box-Jenkins models requires at least a moderately long series.

Chatfield (2004) recommends at least 50 observations. Although others would recommend at least 100 observations, these data are not always available (Hutwagner et al. 2005). The issue of detecting change in surveillance data with limited baseline data will be introduced in the forthcoming section on statistical process control.

Comparisons between Box Jenkins methods and other methods

Comparisons have been made between Box Jenkins methods and other methods for detecting change in health surveillance data. Tobias et al. (2001) and Zhang et al. (2008) compared Poisson regression models with ARIMA models for surveillance of daily emergency admissions and salmonellosis cases respectively. Tobias et al. (2001) found advantages and disadvantages with both approaches. On the one hand, the interpretation of the results from a Poisson regression model being more familiar for the epidemiologist in terms of relative risk estimates. On the other hand, regression models required estimation of a larger number of parameters to account for seasonality and trends when compared with the ARIMA model.

Nobre et al. (2001) used reported cases of malaria and hepatitis A from 1980 to 1995 to compare the forecasting performance of SARIMA and dynamic linear models (DLM). They report that no one method dominates over the other. DLM will be discussed in a later section of this chapter.

A comparison was made between ARIMA modelling of health surveillance data and integer-valued autoregressive modelling (INAR) (Cardinal et al. 1999). INAR methods have been used for rare event surveillance data such as meningococcal disease in Canada (Allard 1998, Le Strat 2005). Rare event surveillance data fit poorly with the standard approach of real-valued methods for analysing time series data. ARIMA models, which assume continuous outcomes, will be of limited value when outcome data are in the form of low-numbered counts. Only when the counts are large is the continuous approximation likely to be justified. INAR techniques have been applied to analysis and forecasting of the incidence of meningococcal disease in Quebec (Cardinal et al. 1999) where no more than six cases per 28-day period were recorded between 1986 and 1993. The INAR model provided a smaller relative forecast error than the ARIMA model in this example.

ARIMA methods have been used in combination with other tools to detect change in surveillance data. Williamson & Weatherby Hudson (1999) combined statistical process

control (SPC) with ARIMA time series modelling to detect aberrations in hepatitis A, meningococcal disease, typhus fever, and other infectious diseases. This is a two-stage modelling system that first provides a dynamic forecast of future expected disease reports from the ARIMA model, then uses SPC methods and control charts for comparison to the actual observed disease reports.

Goldenberg et al. (2002) combined Box Jenkins methods with wavelet analysis for detecting infectious disease outbreaks associated with bioterrorism. They report a two-stage prediction method suitable for non-stationary data that can be easily automated and yields accurate predictions. This method is used because the ARIMA type models alone do not perform well due to the changes in the behavior of the time series of OTC grocery and pharmacy data.

Other methods in the time domain

Use of the cross-correlation function

Our interest in the relationship between two time series occurs in two situations. In the first the series arise 'on an equal footing', and we are interested in the correlation between them. For example, the correlation between electrocorticographic signals that are recorded on a grid of many differently placed electrodes and used to localise seizure foci and to map brain functions (Zeger et al. 2006). The second is more attuned to our interest in surveillance and we ask: are the two series 'causally related'? Can we consider one series (e.g. ambient temperature), as an input to a linear system, while the other is an output (e.g. foodborne disease incidence). In surveillance our interest is in finding the properties of that linear system (Chatfield 2004). This is considering one time series (often lagged values of it) as an explanatory variable, and the other as an outcome, as in regression.

The relationship between two time series is called the cross-correlation function (CCF) (Diggle 1990). Cross-correlation functions are commonly used to explore the relationship between weather variables and disease outcomes. Three Australian studies illustrate this. A study of the association of short-term climate variation with Ross River virus (RRV) transmission was undertaken in Queensland, Australia (Tong & Hu 2002). Rainfall, temperature, relative humidity at a lag of 12 months, and high tide in the current month were found to be significantly associated with the monthly incidence of RRV. These were added

as explanatory variables to a Poisson regression model allowing quantification of their effect. In a model of salmonellosis transmission in Adelaide, Australia, Zhang et al. (2008) explored the lagged effects of climatic variables by cross-correlation analysis. They report a positive association between temperatures and salmonellosis, and a negative association between rainfall and salmonellosis. The lagged effects of climatic variables on *Campylobacter* infections from 1990 to mid-2005 in Adelaide and Brisbane were explored by cross-correlation analysis (Bi et al. 2008). The direction of the association was different in the different cities: negative in Adelaide, the temperate city, but positive in Brisbane, the sub-tropical city.

Bloom et al. (2007) report that the CCF is frequently used to identify lead/lag relationships in health surveillance time series but advise caution when using the cross-correlation function in the context of syndromic surveillance. In their example of clinical respiratory case counts, (lag), and aggregated sales of OTC cold and flu medicine from pharmacies, (lead), they demonstrate that the data must be treated to accentuate the influence of features of interest in a CCF analysis, otherwise the results may be misleading. Choosing a correct smoothing technique and period of interest is important, for example.

Other authors also warn against spurious correlations that can arise between time series when examined by the CCF (Diggle 1990, Chatfield 2004, Cryer & Chan 2008). Lagged correlations between time series can present misleading evidence of lagged relationships and dependence, especially if the individual time series are autocorrelated. The best protection against this is 'prewhitening' before estimating the CCF. Prewhitening in this context is intended to deal with the complicating effects of autocorrelation on the estimated CCF and its standard deviations. Detailed handling of this topic can be found in Diggle (1990) and Chatfield (2004).

Partially for the above reason, some authors choose to investigate the effect of individual lags of covariates by entering them into a regression model, *a priori*, rather than using the CCF to choose the most appropriate lag. For example, Hald & Andersen (2001) used stepwise selection to enter lagged values of meterological data and prevalence in pork into a model describing the number of human cases of *Salmonella typhimurium* in Denmark.

State-space approach to time series

State space models provide a cohesive framework in which any linear time series model

can be written. A state-space model of a time-series comprises a data generating process with a state that may change over time. For example, in health surveillance, data that may either be in an epidemic state or not. This state is often only indirectly observed, for example we may only observe 10% of individuals that are truly infected with polio, while 90% will be asymptomatic. In this review, two types of state-space models are discussed: dynamic linear models and hidden Markov models.

Dynamic Linear Models

The dynamic linear model (DLM) is a development of the state-space approach to the estimation and control of dynamic systems (West & Harrison 1997). Surveillance data can be modelled by means of a DLM, and forecasts based on prior knowledge and including former observations can be made. A comparison of the forecasting performance of DLM and SARIMA models was made using cases of hepatitis A and malaria in the USA from 1980 to 1995 (Nobre et al. 2001). Both gave comparable results but the DLM approach reportedly had some major advantages: (1) it is more appropriate for count data that may be from a rare disease or small areas; (2) the Bayesian nature of DLM allows inclusion of subjective information, such as expert opinion and historical data is not required; (3) it does not require a new cycle of identification and modelling when new data became available; (4) the assumption of stationarity is not a prerequisite; and (5) missing data are handled. Despite these advantages this technique has had little application in the health surveillance literature. This may be because the DLM forecasting approach requires the specification of several parameters that are not easily understood and the outputs require complex analysis.

Hidden Markov Models

A Markov property is exhibited when the state that a system is in, in the current time period, depends only on the state that the system was in, in the immediately preceding period. Hidden Markov models are a class of stochastic processes that are capable of modelling time-series data (Rabiner 1989, Allard 1998). A Markov model moves from state to state, e.g. epidemic to non-epidemic, according to a probability distribution of each state, called the transition probabilities. With each state visited a signal is emitted. Hidden Markov models move from state to state in the same way but emit a symbol, from a finite alphabet of the model, from each state visited, except silent states, according to the

probability distribution of the state, called the emission probabilities (Eddy 2004). Hidden Markov Models (HMMs) were developed in the early 1960s, and were initially used in the field of speech recognition. They are a convenient statistical tool for explaining serial dependency in data which assume that the observations form a noisy realisation of an underlying process that has a simple structure with Markovian dependence.

Le Strat & Carrat (1999) use a two-state hidden Markov model to correspond to either epidemic or non-epidemic states of two surveillance data sets. The two data sets were a French monthly series of influenza-like illness cases from 1985 to 1996, and a USA monthly series of poliomyelitis from 1970 and 1983. In the two-state model, a threshold can be computed directly from the non-epidemic state (and used as an early warning system, as in ARIMA models). For the series of French influenza-like illness cases they assumed the data were generated from a mixture of Gaussian distributions. For the series of USA poliomyelitis cases they used a mixture of Poisson distributions. In both cases the series were governed by an underlying Markov chain. In the models adjustment for trend was made by using a linear term, and adjustment for seasonality was made by using sine and cosine terms.

Rath et al. (2003) analysed the same French monthly influenza-like illness cases data set and showed that better detection accuracy can be achieved by modelling the data using a mixture of exponential and Gaussian distributions. By changing underlying distributions, the need to explicitly model trend and seasonal effects was removed. This removed these as potential causes of bias in the detection accuracy.

The use of an HMM to detect change in surveillance data is further extended in a paper that analyses hospital infection data (Cooper & Lipsitch 2004). Three classes of pathogens are investigated: methicillin-resistant *Staphylococcus aureus*; vancomycin-resistant *Enterococci*; and third generation cephalosporin-resistant Gram-negative rods. Three models were compared: a simple Poisson model; a standard hidden Markov model using a Poisson observation model; and a structured hidden Markov model (based on the susceptible--infectious–susceptible epidemic model (Isham 1993) assuming a mean intensive care unit stay of eight days). They conclude that structured hidden Markov models are a promising tool for analysing hospital infection count data for transmissible pathogens.

Further coverage of the application of hidden Markov models to surveillance data is pro-

vided by Madigan (2005).

Generalised additive models

A generalised additive model (GAM) is an extension of generalised linear models where the usual linear function of a covariate is replaced with a smoothing function e.g. natural cubic splines, locally weighted regression, penalised splines, or smoothing cubic splines (Hastie & Tibshirani 1990). These models are particularly useful in exploring the non-parametric relationship between disease incidence and climate variables such as seasonality, temperature, and humidity. These have had extensive use in health outcomes related to food safety (Hald & Andersen 2001, Kovats et al. 2004, 2005, Tam et al. 2006), and air pollution (Wilson et al. 2004, Touloumi et al. 2006). GAMs also have been used extensively in spatial and spatio-temporal epidemiology and surveillance (Kelsall & Diggle 1998, Vieira et al. 2008, Siqueira et al. 2008).

The example of the time series of HFRS in Anhui Province, China from 1983 to 1995 is used to illustrate this methodolgy (Hu et al. 2006). Equation 2.4 was extended in a third model allowing for smoothing non-linear functions of two of the three predictors (mice density and water level difference) as follows:

$$log(\mu) = \phi Y_{(t-1)} + s_0 + s_1 X_1 + \ldots + s_p X_p + \epsilon \tag{2.5}$$

where $s_0 ....s_p$ are natural cubic splines. GAMs provide a flexible, functional way of informing the relationship between exposure and outcome by fitting non-parametric functions. The GAM model had the best goodness-of-fit and short-term predictive ability.

Other climatic effects

The El Niño/Southern Oscillation climatic events are a natural phenomenon that occur every three to eight years and are reasonably predictable (Chen et al. 2004). The effect of these events on infectious disease has been reviewed by Kovats et al. (2003). The four-year super annual cycle in malarial incidence in Thailand has been attributed to El Niño/Southern Oscillation climatic events (Childs et al. 2006). A study investigating the relation between climate variability and daily admissions for diarrhoea in Peruvian children (Checkley et al. 2000) used GAMs. El Niño had an effect on hospital admissions greater than that explained by the regular seasonal variability in ambient temperature. Fu-

ture forecasts were made about potential disease risks for 2006-2007 based on the current El Niño's effect on vector abundance (Anyamba et al. 2006). To date, the forecasts made of Rift Valley fever in Kenya (ProMED-mail 2007) and malaria in India (Jelinek et al. 2007) have been proven correct indicating the benefits of this approach.

An arguably less predictable and non-natural phenomenon is that of global climate change, and there are a number of review articles on the potential effect of climate change on human health (McMichael et al. 2006). Global warming was cited as a reason for the increase in emerging viral diseases many of which are zoonotic (Kallio-Kokko et al. 2005, Ka-Wai Hui 2006). The ability to make long range forecasts of epidemics or epizootics is helpful for planning resource allocation for more intensive surveillance, prophylaxis, treatment, and warning. Climate change presents an emerging challenge for health surveillance research.

Methods in the frequency domain

As stated previously, in the time domain the variation in the time series is described in terms of the way in which observations at different times are related statistically to one another. These methods include ARIMA and hidden Markov models. This part of the literature review now considers the frequency domain which describes how the variation in the time series may be accounted for by cyclic components at different frequencies. Methods include harmonic (when the frequencies are predetermined) and spectral (when the frequencies are unknown), and in both cases the series must be first detrended. Spectral analysis is a modification of Fourier analysis which approximates a periodic signal (such as a consistent seasonal pattern) using a linear combination of sine and cosine waves (Chatfield 2004). When we consider regression in the frequency domain the inputs are periodic sine and cosine functions. Time series are represented as sinusoidal waves of different frequencies, amplitudes and phases.

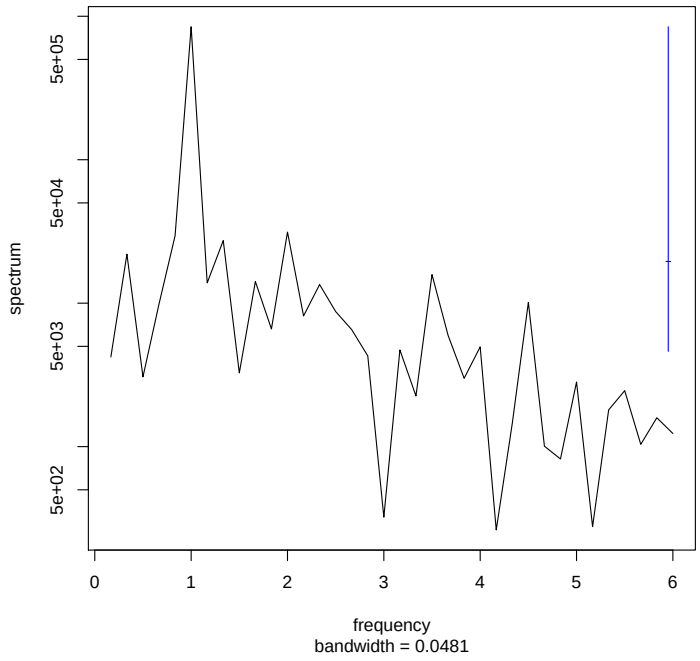
Although seemingly more complex than the time approach, the value of the frequency approach is that complex patterns can be described by only three data factors – the frequency, amplitude, and phase, whereas in the time domain they would take much more information to define accurately. In addition, the spectral analysis can identify frequencies that

are not predictable before the data are examined (Diggle 1990). Frequency domain analysis has been found to be especially useful in communications, engineering, geophysics, acoustics, and biomedical science.

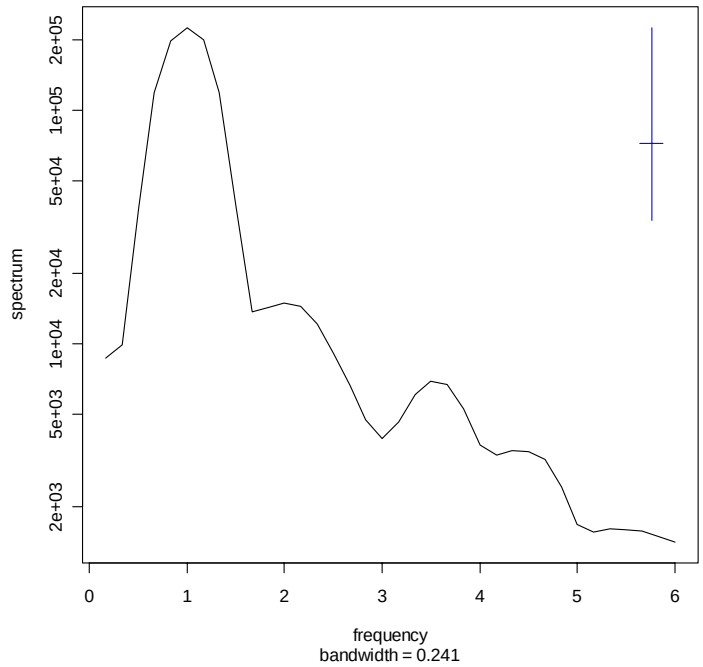
Figure 2.14(a) shows the raw periodogram of the respiratory deaths data. This was created by a mathematical procedure, termed a Fourier analysis. This analysis determines the collection of sine waves (differing in frequency and amplitude) that is necessary to make up the pattern under consideration (Chatfield 2004). The dominant peak is at a frequency of 6/72, giving a period of 72/6 = 12 months, this reveals the strong yearly pattern in these data. Figure 2.14(b) shows the periodogram Daniell (Daniell 1946) smoothed with spans of (3,5). This is a weighted moving average transformation used to smooth the periodogram to more clearly reveal the cyclicity. A more detailed description of Fourier analysis can be found in Diggle (1990) and Chatfield (2004).

Fourier models were used to describe the seasonal pattern of weekly cases of campylobacteriosis between 1997 and 2001 in Scotland (Miller et al. 2004). Strong seasonality was reported with an annual peak in late June to early July and successful predictions of both national and regional cases were made for 2002. Superimposed upon this seasonal pattern were irregular finer peaks and troughs which the authors termed 'bursts' of infection superseding the 95% prediction intervals. These bursts were both within and across regions of Scotland and thought to be due to previously unrecognised outbreaks.

There have been many studies on the effect of short-term temperature on infectious enteric diseases of humans (see above on CCF and GAMs). The specific question of interest here is 'is a change in ambient temperature in a time period associated with a change in disease reports x time periods later?' This short-term temporal association between climate and disease will very likely be confounded by trend and seasonal patterns other than those associated with temperature. The study design must adjust for these (Schwartz et al. 1996). Ways to make this adjustment include controlling for trend by adding indicator variables for each year of the series, and controlling for the seasonal patterns by adding Fourier terms (Kovats et al. 2004, Tam et al. 2006, Hashizume et al. 2008). This allows the assessment of any short-term effects of temperature on disease.



(a) Raw periodogram



(b) Daniell smoothed periodogram

Figure 2.14: (a) Raw and (b) smoothed periodograms of respiratory deaths showing a significant peak at a frequency of 6/72. The vertical bar on the right of the plot is the 95% confidence interval around the peak. The smaller horizontal bar bisecting the vertical bar is the smoothing bandwidth.

Kovats et al. (2004) found the greatest effect of temperature to be one week before the onset of *Salmonella* infections in ten European human populations. There were diminishing but positive effects for up to five weeks and they report a linear association between temperature and the number of cases of salmonellosis above a threshold of 6 °C. In the model of Tam et al. (2006) adjustment was additionally made for the delays in the effect of temperature on the number of reported cases by incorporating a six-week lagged temperature variable. These authors found a linear association between mean weekly temperature and the number of cases of human campylobacteriosis in England from 1989 - 1999, with a 1 °C rise corresponding to a 5% increase in the number of cases. Hashizume et al. (2008) found a strong association between hospital visits for rotavirus diarrhoea and temperature in Bangladesh. This was after adjustment for humidity, river level, public holidays, and seasonal and annual variations. The fact that ambient temperatures influence the incidence of these enteric diseases can facilitate targeting preventative action, as well as surveillance and resource allocation.

A study investigating the temporal association between climate and *Campylobacter* infection adjusted for confounding from seasonal factors other than temperature by matching on week (Kovats et al. 2005). This international study used 15 northern and southern hemisphere populations with most showing a peak of cases in spring, those with milder winters peaking earlier in the year.

The timing and intensity of seasonal peaks of six infectious enteric diseases was reported in a Massachusetts study using ten years of data from 1992–2001 (Naumova et al. 2007). *Campylobacter* and *Salmonella* closely followed the summer peak in ambient temperature (around the 24th of July) with a 2 – 14 day lag, while Giardia, Shigella, and Cryptosporidium infections peaked 40 days after the temperature peak. The difference in the lag phase between these two groups of diseases is suggestive of different routes of exposure. In this study it was the use of daily counts of cases (as opposed to aggregated weekly or monthly counts) that enabled recognition of the difference in the lag phase.

Periodic regression models were fitted to Danish data to investigate seasonality in different age groups with meningococcal disease (Jensen et al. 2003) and in the severity of non-typhoid *Salmonella* infections in humans (Gradel et al. 2007). The peak-to-trough ratios (PTR) were calculated to measure the magnitude of the seasonal variation. For meningococcal disease, the highest PTR were in the 5–9 year age group (4.9, 95% CI:

2.1–11.9) and the lowest were for children less than one year of age (1.4, 95% CI: 0.6–3.2), indicating that seasonality varies with age group. For non-typhoid *Salmonella* the seasonal pattern diminishes with increased severity of infection. This suggests that for severe *Salmonella* infections, endogenous factors play a more important role than exogenous factors. Also it may mean that patients with more severe non-typhoid *Salmonella* infections engage less in activities that increase the risk of acquiring infections in the warmer months (e.g. barbecuing or going on holiday).

The review paper from Zeger et al. (2006) provides accessible coverage of applications of both frequency and time-domain methodology to the time series analysis of health surveillance data.

2.3.2 Statistical process control

The essential difference between modelling data via time series methods (above) and using statistical process control methods is that time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend, or seasonal variation) that should be accounted for. Statistical process control can be thought of as a plotted time series with control limits applied.

In the 1920s, Walter Shewhart developed a number of business production analysis techniques, which were designed to detect changes in the quality of the output from continuous production processes (Shewhart 1931). Statistical process control (SPC), or quality control, is widely used in many industries to facilitate objective evaluation of business operations and production processes. SPC is used to monitor the level of a production trait, and to give a notification when the level changes beyond some predefined limit. The basic control charts are designed to monitor a process that is expected to be constant, although it allows for some random fluctuation.

The use of SPC has gone beyond industry to monitor any process, and identify when it changes to being 'out of control'. As early as 1942, Deming proposed the potential value of SPC for disease surveillance and rare events monitoring (Deming 1942). An overview of the use of SPC in health care and surveillance is given by Woodall (2006).

Consider the following as a time series of surveillance data $X = X(t) : t = 1, 2, ...$ Using this example, a standard Shewhart control chart would consider X as a continuously vary-

ing quality (e.g. number of thermometers sold in Auckland per day) with a mean of μX and a standard deviation of σX. Upper control limits would be set as UCL = μX + $K\sigma X$, the centre line at μX, and the lower control limits at LCL = μX - $K\sigma X$. Historically, $K = 3$ has become an accepted standard in industry.

It is important to remember that although statistical process control charts are among the most prevalent and valid methods for monitoring time series data, their use usually requires observations to be random variables when the process is in statistical control. Health surveillance data are not random variables and present problems that are not present in the case of industrial process control; health data often exhibit correlation, non-stationarity (in the mean and/or variance), and seasonality. However, these limitations may be substantially overcome by using one of two techniques (Stoumbos et al. 2000). Firstly, past-behaviour of the series can be corrected by inclusion of seasonal or historical adjustments. In other words, the original data is presented as a standard control chart but the control limits are adjusted for the autocorrelation in the series. A good example of this is the Early Aberration Reporting System (EARS).[2] This applies aberration detection algorithms to surveillance data and flags anomalies. Two methods are implemented: (1) a seasonally adjusted quality control statistic; and (2) a historical limits model that compares the current 4-week total to the mean of nine 4-week periods (using the previous, comparable and subsequent 4-week periods over the past three years). These methods can result in three different flags: (1) or (2), above, or (3), when both the models exceed the established thresholds. EARS uses Shewhart variants that use a moving sample average and sample standard deviation to standardise each observation.

The EARS system can be applied to daily, weekly and monthly data and allows for stratification of the data e.g. by geographic region and specified threshold limits. For rare diseases such as typhoid, the system can be set to flag every occurrence of a case. EARS is used in the national notifiable disease surveillance system (EpiSurv) in New Zealand. Figure 2.15 shows the system in use for flagging high numbers of campylobacteriosis cases. Flags consistently occur over the period December 2006 to February 2007 but not at Christmas/New Year time. As the historical mean also is reduced at Christmas/New Year time, this discrepancy is more likely a result of fewer notifications, due to people taking holidays, rather than fewer actual cases of disease.

[2] `http://www.bt.cdc.gov/surveillance/ears/`

Watkins et al. (2008) compared the use of three EARS cusum-based methods and a negative-binomial cusum for the retrospective detection of outbreaks of Ross River virus disease in Western Australia between 1991 and 2004. (See Equation 2.6 for an explanation of cusum). They found that the use of a negative binomial distribution accommodated the over-dispersion evident in disease notification data, and provided a lower rate of false alarms for a given sensitivity. However, these advantages were associated with decreased early timeliness performance when using the negative binomial cusum algorithm.

The second option for overcoming the problem of the lack of independence is to plot the residuals from a time series model on a standard control chart (Stoumbos et al. 2000). Williamson & Weatherby Hudson (1999) combine statistical process control with ARIMA time series modelling to detect aberrations in hepatitis A, meningococcal disease, typhus fever, and other infectious diseases.

One problem with this methodology is the need to have sufficient baseline data to produce a stable model. Generally to account for seasonality three years of data is considered the minimum (Diggle 1990). To overcome this problem methods have been developed that incorporate short seven-day baseline periods for threshold comparisons (Hutwagner et al. 2005). These thresholds were based on a cusum calculation and the baseline was varied to give different sensitivities. Cusum is a cumulative sum calculation as follows:

$$S_t = max(0, S_{t-1} + ((X(t) - (\mu X + K\sigma X))0/\sigma X)) \tag{2.6}$$

with a decision value of $S_t > 2$, where $X(t)$ is the count or percent e.g. number of thermometers sold in Auckland per day. The other parameters are described above, but here K is the detectable shift in the mean and not necessarily 3 as it is for a Shewhart chart.

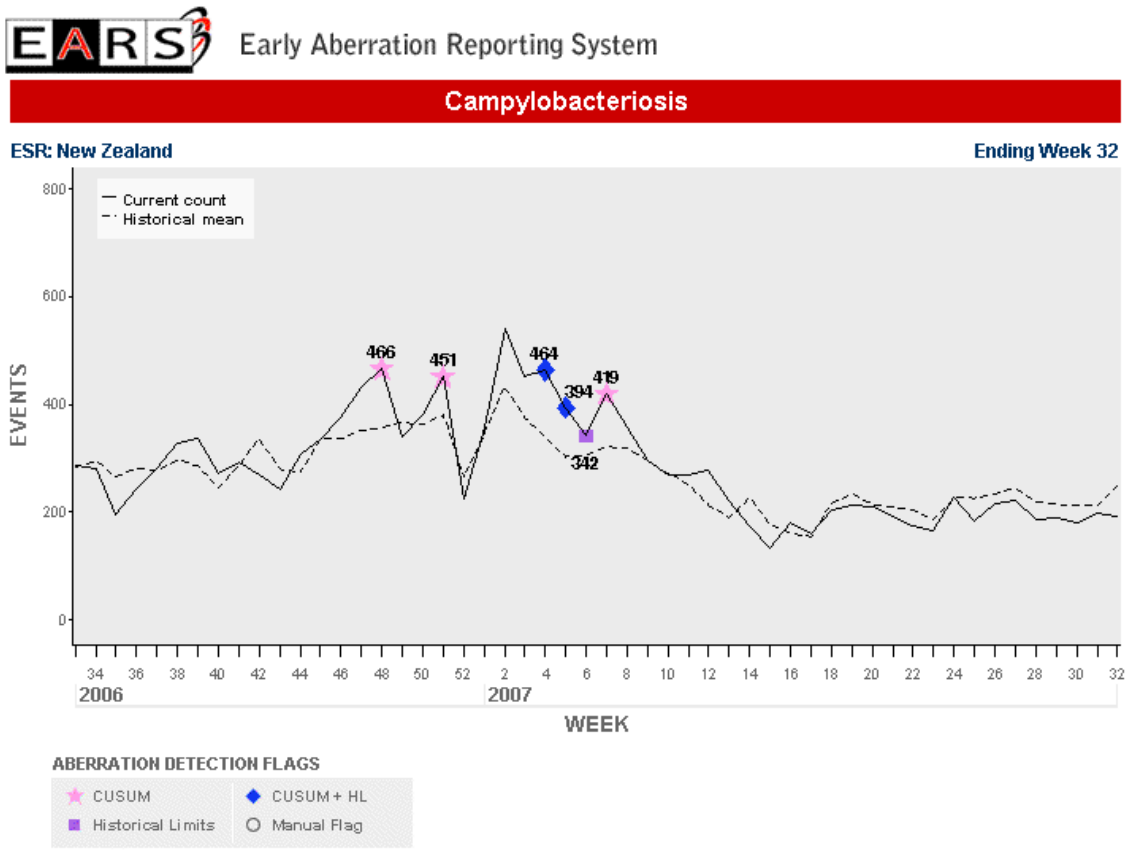


Figure 2.15: Use of EARS for Campylobacteriosis surveillance in New Zealand. Flags consistently occur over the period Dec. 2006 to Feb. 2007 but not at Christmas/New Year. Source: Institute of Environmental Science and Research Limited.

For example, the method with the least sensitivity used a baseline from the previous seven days in closest proximity to the current value. This is because if a flag which denotes an aberrant value is noted on a particular day, t, then the next day, $t+1$, it is less likely to produce a flag as the high count from the previous day is immediately incorporated into the new baseline. The methods were designed for enhanced bioterrorism surveillance to identify aberrations quickly e.g. within the first day or two of a special event such as the Olympic games. Since these methods are based only on current information, they would not be useful for identifying an infectious disease event that occurs gradually e.g. as in the start of the influenza season.

Generally changes will be of the step form, where a parameter changes from one constant level to another, but changes can also be linear, exponential or gradual; the latter

can be particularly problematic to detect. Exponentially weighted moving average charts and cusum are likely to perform better in this situation than traditional Shewhart charts (Stoumbos et al. 2000). A recent paper by Fricker et al. (2008) found that Shewhart-based methods, such as EARS, are not well suited for the syndromic surveillance problem in which outbreaks do not occur instantaneously and are transient.

Further information on the use of SPC methods for spatio-temporal surveillance is provided by Rogerson (2005).

2.3.3 Neural networks

Neural networks consist of a series of processing element (nodes) interconnected in a network that can capture and represent complex input/output relationships. This is achieved by training and exposing the network to known data sets. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform 'intelligent' tasks similar to those performed by the human brain. Artificial neural networks are powerful predictive tools that have multiple practical applications in imprecise systems (Sawyer 2000).

Neural networks have been used for financial time-series forecasting (Qi & Zhang 2008), and in the application of industrial process control (Yu & Xi 2009). Artificial neural networks (ANN) have been successfully applied on various areas of medicine, such as diagnostic systems, biomedical analysis, image analysis, and drug development (Patel & Goyal 2007). This technique has been applied to health surveillance data to detect excess deaths from cholera in Brazil (Penna 2004). In this paper the author compared the use of a recurrent neural network with a negative binomial regression model as a predictive technique for the cholera time series. Estimates from both models showed good agreement, indicating the adequacy of using ANNs for health-related time series. Although the neural network had good predictive ability, it was less sensitive than negative binomial regression in the detection of abnormal values. The author reported that the main advantage was the lower level of statistical knowledge required to implement it. However this is a novel method and researchers generally have little familiarity with the process, compared to other statistical methods. Furthermore, a long time series is often required to train ANNs (De Gooijer & Hyndman 2006).

2.3.4 The Temporal Scan statistic

Two Canadian studies have used the temporal scan statistic (Kulldorff 1997), implemented in SaTScan[3], to retrospectively identify temporal clusters of *Salmonella* infection in animals in Ontario (Zhang et al. 2005), and of both animals and humans in Alberta (Guerin et al. 2005*a*,*b*). In the latter paper, the temporal scan statistic was used to test the null hypothesis that isolates of *Salmonella* serovars were randomly distributed with respect to time. They used the Bernoulli model, where cases are contrasted with a control group. This statistic scans each possible predetermined time window and compares the proportion of cases and controls within the window with the proportion of cases and controls outside the window of time.

2.4 Spatio-temporal Surveillance

Even though the detection of change in surveillance data over time is important it should not be considered in isolation from changes in space. Recent and accelerating technological advances mean that we have both increasing availability of data that contains spatial information, and improved ways of visualising spatial data. This is mainly due to advances in computing and ready access to commercially available geographical information system (GIS) software. Open source GIS software has made this technology more accessible, no doubt playing some role in the advances that have been made in recent years. Spatial techniques in epidemiology are advancing rapidly (Martin 2004, Lawson & Kleinman 2005, Pfeiffer et al. 2008, Bivand et al. 2008, Lawson 2009), allowing us to describe and explore disease in space and space-time, and increasingly to model and predict disease risk.

This part of the literature review is purposively less well developed than that on temporal surveillance as it is not my intention to repeat a review of spatial and spatio-temporal techniques as outlined in Chapter 1. Here I will review techniques that have been applied to zoonotic disease surveillance, giving the most detail to those most frequently used and to the most recent developments to date.

[3] `http://www.satscan.org/`

Analogies can be made between aspects of data in one dimension (time) and in two dimensions (space). For example, a temporal trend over a number of years may be thought of as similar to a first-order or broad spatial pattern. The serial autocorrelation between measures in a time series is similar to second-order effect or spatial correlation. A spatial analog of regular measurements in time e.g. daily counts of ED visits, may be that of the spatial correlation between regularly shaped polygons e.g. the division of France into regular hexagons to investigate BSE (Abrial et al. 2005) or the division of Vietnam into squares to investigate avian influenza (Lockhart 2008). A spatial analog of irregular measurements in time e.g. the sequential, but not every day, nor regularly spaced, daily counts of clinic visits, may be that of the spatial correlation between irregular shaped polygons e.g. disease counts in areas defined by political boundaries (Sanchez et al. 2005). There is no obvious spatial analogy for the seasonality of a time series.

It is natural to investigate the spatial distribution of cases to gauge whether or not they occur closely in space as well as in time. The value of this spatial information is twofold. Firstly, all cases are localised at some spatial scale — if our surveillance is done at a broad spatial scale, (e.g. aggregated regional counts of cases), then even large and sharp increases in relatively small aggregated counts (e.g. neighbourhoods) may be hidden because they are localised and of insufficient size to be detected. Spatial surveillance of defined regions will increase the probability of detecting these events (Lawson & Kleinman 2005). Secondly, if cases are spatially aggregated, then an appropriate response, (such as quarantine, vaccination or dissemination of public health messages), can be targeted to that area. However, it is important to consider that with the distribution of food over large areas, cases of foodborne disease will not necessarily occur close together in space. This point was well illustrated by an outbreak of *E. coli* O157 that was linked to consumption of fresh, bagged, baby spinach produced in three counties on the central California coast (Jay et al. 2007). Due to the dissemination of product, illness was widespread throughout 26 states of America and Canada resulting in 205 cases and three deaths. This points to the need for integration of another rapidly developing tool, that of social network analysis, into surveillance methodology (Christley et al. 2005, Kao et al. 2007).

However, when cases are confined to one area this may point to a local source of infection. A current outbreak under investigation illustrates this point. The on-going large outbreak of *Salmonella typhimurium* U292 infection in Denmark (Ethelberg et al. 2008*a*,*b*) appears

to be fully confined to Denmark; no cases have been reported from other countries including neighbouring Scandinavian countries and Germany. This could point to a local source of foodborne infection that is not consumed beyond Denmark.

Not only might transportation of food carry pathogens from one location to another, but the movements of animals and people themselves must be considered when observing disease surveillance data. For example, movements of people by air travel have been implicated in the spread of SARS globally (Hufnagel et al. 2004), and influenza and pneumonia in the USA (Paul et al. 2008).

Just as we have temporal trends we also have spatial trends, often called first-order spatial effects. This describes the mean value of a process in space. Second-order spatial effects are analogous to temporal autocorrelation in a time series. Second-order effects result from the spatial correlation or dependency of a process. They are local effects.

Both first- and second-order effects can lead to clusters of disease even though they are driven by different underlying mechanisms, the first-order effect being from a variation in the intensity of events while the second-order effect being due to dependence between events (Diggle 2003). Identification of these clusters is potentially a powerful tool for surveillance. We may be able to identify specific areas in which to target sampling or to study more intensely. Clusters may give clues as to underlying aetiologies such as vector abundance or locally acting risk factors, or may be an indication of spatial dependency that might lead us to consider contagion.

One of the earliest examples of disease mapping was the point map produced by John Snow in 1854 of the addresses of cholera victims related to the location of water supplies (Snow 1854). This provides a classic example of surveillance of cases, mapping their location and then using the map to come up with a putative source. Figure 2.16 provides a re-analysis of this data by Bivand et al. (2008). This map is composed of a number of layers of spatial data: the roads in Soho around Broad Street; the superimposed heat colours indicating the distance from the Broad Street pump in the centre of the image; representation of pump locations; and the bubbles, with size proportional to number of mortalities, representing the street distances from each mortality dwelling to the nearest pump, the grey representing Broad Street and the pink other pumps.

Whether the legendary removal of the pump handle actually happened or not (McLeod 2000), it is an important example of an action taken that stemmed from surveillance. An

example from September 2006, a full 152 years after Broad Street, was the advice of the USA Food and Drug Administration to avoid eating fresh spinach or fresh spinach-containing products that had been grown in three California counties (Centers for Disease Control and Prevention 2006). Although the technology had advanced (involving rapid diagnosis of *E. coli* O157:H7 infection, culture, PFGE analysis, and reporting to the CDC), the timely intervention based on surveillance data of this geographically spread epidemic is not dissimilar to that seen 152 years before at Broad Street in London.

2.4.1 Spatial variation in risk

Point maps have limitations, including lack of confidentiality, and if there are multiple events that happen at the same location they can be difficult to distinguish from each other. Smoothing of points can be applied to overcome these limitations. The spatial risk function, estimated by kernel density methods, was developed for case control data for rare diseases by Bithell (1990). Lawson & Williams (1993) and Kelsall & Diggle (1995) further improved the estimation.

A smoothed map surface that accounts for the population at risk as well as the cases themselves will allow us to identify areas of the greatest risk of disease. These can be targeted as areas in which to do more focussed studies to identify risk factors or as areas in which to enhance surveillance. A kernel density estimate of the relative risk of canine faecal contamination in Naples highlighted two areas at greatest risk close to the eastern and northern border of the city (Biggeri et al. 2006*a*). This formed part of a larger study investigating the risk of zoonotic infection from this contamination using a number of spatial techniques. The study reported that the larger risk areas were identified at the city border where wild dogs mixed with domestic dogs and human or urban barriers were less present (Biggeri et al. 2006*b*). Kernel smoothing techniques have also been used in the assessment of spatial risk of BSE in the UK epidemic (Stevenson et al. 2000) and in West Nile virus (WNV) surveillance to identify clusters of dead crows (Johnson et al. 2006). These clusters of WNV were then used as part of the information to predict human cases.

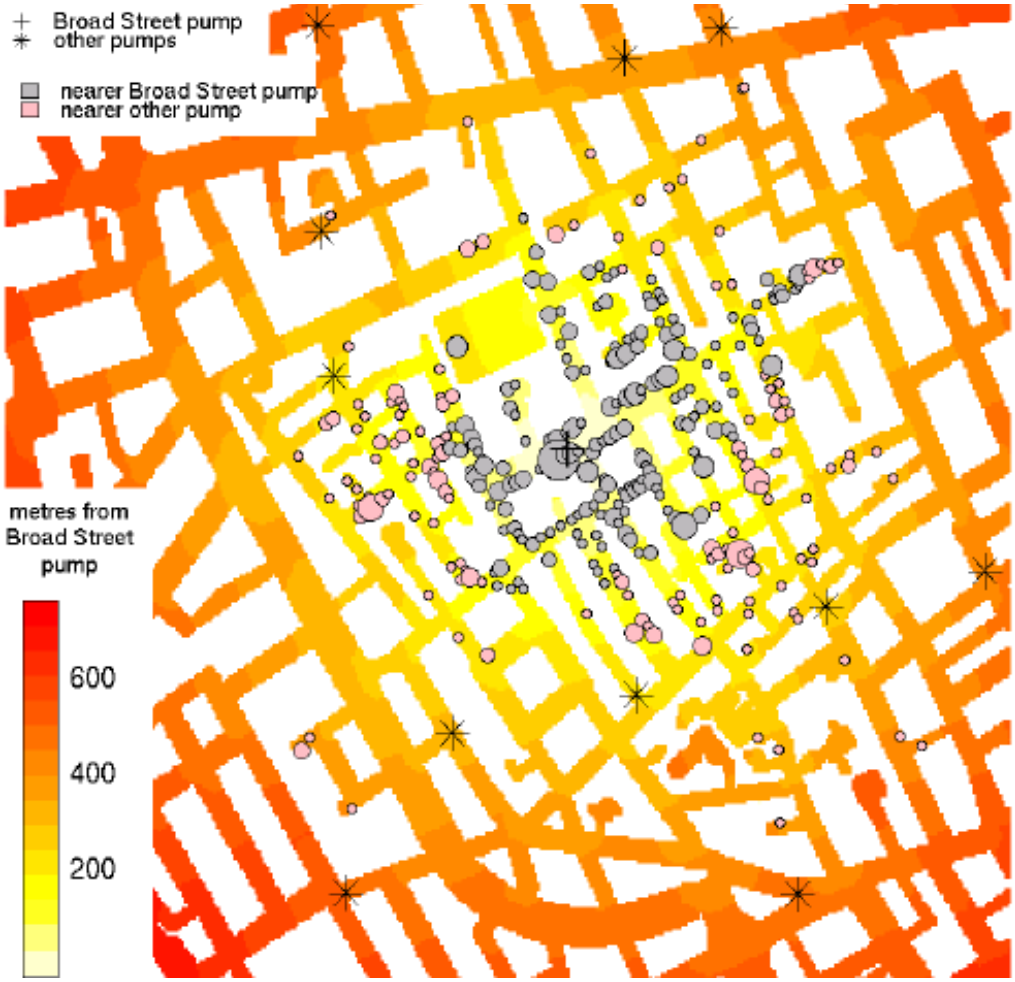


Figure 2.16: Point map of Broad Street cholera cases, showing location of pumps, London, 1854. The bubbles, with size proportional to number of mortalities, represent the street distances from each mortality dwelling to the nearest pump, the grey representing Broad street and the pink other pumps. Source: http://www.bias-project.org.uk/ASDARcourse/.

Two key issues in the use of kernel density methods are those of bandwidth selection and dealing with edge effects. The bandwidth, or smoothing parameter, requires careful selection: if it is too large, the resultant map is over-smoothed, potentially masking areas of increased risk, and if the bandwidth is too small, too much detail may be seen, resulting in over-interpretation of the pattern. Pfeiffer et al. (2008) provides a detailed discussion of issues related to bandwidth selection. Edge effects arise due to non-existent or incomplete data that occurs near the edges of a study area. These edges may be natural boundaries such as the coast or man-made boundaries such as administrative borders. As the kernel smoothing technique borrows strength from neighbours, distortions can result when these neighbours are absent. These are called edge effects. More details on edge effects can be

found in Lawson et al. (1999) and Zheng et al. (2004).

There is much ongoing work in this field: a novel approach to the estimation of spatial variation of relative risk by the use of local polynomial regression is under development (Fernando 2008). Planned work includes applying veterinary surveillance data to the development of appropriate space-time kernels for the estimation of time-varying relative risk.

2.4.2 Spatial and spatio-temporal clustering

Spatial clustering occurs when cases occur more closely together in space than would be expected with a random sample from the population at risk (Diggle 2003). The assessment of disease clustering was largely born out of the need for public health authorities to respond to public concerns about putative sources of environmental contamination, such as waste dumps, incinerators, and steel foundries (Lawson & Williams 1994).

Clustering tests are commonly defined as either being focussed or general (Tango 1999). The following definitions I use are from Tango (1999); others can be found in Besag & Newell (1991), Diggle (2003), and Wartenberg (2001). A focussed test of clustering would be used in the steel foundry example, whereby the location of the cluster is identified a priori and the likelihood of the location truly being a cluster centre is determined. Compare this with general tests which determine whether or not clustering occurs over the study region. This latter type of test can be further divided into two groups: those that examine the tendency to cluster, and those that find the location of the cluster/s.

Clustering of disease can occur for a variety of reasons, such as the aggregation of risk factors in a specific area or environmental factors that affect vector abundance. A cluster of cases of disease that are close in both space and time is highly suggestive of an infectious process. Key questions are: whether the cluster is statistically significant, or if it has occurred by chance, or if it simply reflects the underlying spatial distribution of the population at risk.

Ward & Carpenter (2000*a*,*b*) and Carpenter (2001) provide an accessible introduction to this topic from the veterinary perspective. More general accounts are provided by Waller et al. (2006), Pfeiffer et al. (2008), Wakefield et al. (2001), and Lawson (2006).

The space and space-time *K*-Function

The spatial K-function of a spatial point pattern is defined as the expected number of further points within a distance r of an arbitrary point, divided by the overall density of the points (Ripley 1976). For a clustered pattern, each point or event, e.g. a case of disease, is likely to be closely surrounded by other cases, so for a given small distance r the K-function will be relatively large. On the other hand, if cases are randomly spaced each case is more likely to not be surrounded by other cases, so for a given small distance r the K-function will be smaller. This is a general, rather than focussed test for clustering that examines the tendency to cluster.

An important assumption of the K-function is that there are no first-order effects in the spatial pattern. The inhomogeneous K-function (Baddeley et al. 2000) allows for the non-uniform intensity of the spatial locations for hypothesis testing for aggregation over and above that of the population at risk. An approach to this issue that enjoys common usage is to use the observed-difference K-function. This is a measure of the difference between the K-function for cases compared with that of controls or of the population at risk. The null hypothesis tested is that there is no extra aggregation of cases over that of the population (corresponding to the cases being a random sample from the population).

Fenton et al. (2008) used the observed difference K-function to provide evidence for spatial clustering of *Salmonella* serovars in UK dairy herds. Serovars Agama and Dublin showed evidence of spatial clustering at distances up to 30 kilometres. This suggests either a contagious process or the presence of spatial localised factors which increase the risk of infection such as contaminated feed or other animal reservoirs (e.g. birds, rodents, and badgers).

Broman et al. (2006) used the observed-difference K-function to investigate clustering of ocular chlamydia in households in a Tanzanian village population. The authors detected clustering of households with high loading of ocular chlamydia among children, at distances up to two kilometres. The observed-difference K-function was calculated between households with a high loading of chlamydia and those with a low load. As the analysis did not examine direct transmission, it was conceded that possibly households with heavy loads of ocular chlamydia simply share the same risk factors for infection. These risk factors that might affect clustering of infected households include annual seasonal variations, and proximity to water.

The observed-difference K-function was used to investigate early life residence with regard to subsequent risk of breast cancer in western New York women (Han et al. 2004). Cases were women, aged from 35 to 79 with incident, primary, pathologically confirmed breast cancer diagnosed during the period from 1996 to 2001. Controls were frequency matched to cases on age, race, and county of current residence; controls under 65 years of age were randomly selected from a New York State Department of Motor Vehicles list and those 65 years and over were chosen from a Health Care Finance Administration list. All cases and controls used in the study provided lifetime residential histories. The authors report that their analysis of breast cancer clustering in space provided evidence of geographic clustering of pre-menopausal, but not post-menopausal, breast cancer cases at the time of birth and menarche. They conclude that there is a possible influence of environmental risk factors on breast cancer at these times in a woman's life.

One way to improve surveillance is to develop methods of identifying temporal and geographic clusters of events that may merit additional evaluation, rather than rely on merely temporal or spatial methods alone. The space-time K-function $K(s,t)$ is an extension of the spatial K-function that was first proposed by Diggle et al. (1995). It compares the observed spatio-temporal pattern with that with the same temporal and spatial properties as the original data, but with no space-time interaction. This test gives a measure of both the nature and scale of the space–time interaction.

Examples include the use of $K(s,t)$ to support a role for infectious aetiologies for glioma in the Netherlands (Houben et al. 2005) and for some childhood cancers in Great Britain (McNally et al. 2006). The space-time K-function was used to describe the risk of infection with FMD attributable to spatiotemporal interaction in two counties in the UK during the 2001 epidemic (Wilesmith et al. 2003). This identified the extent of contagiousness in space and time which could be used to inform policy for pre-emptive culling distances. French et al. (1999) used $K(s,t)$ in an investigation of sheep scab outbreaks in Great Britain between 1973 and 1992. They report that a large proportion of the cases within 12 km and five months of each other can be attributed to space-time clustering, supporting the highly contagious nature of this infection. Sanchez et al. (2005) used $K(s,t)$ in an investigation of the 1999 outbreaks of infectious bursal disease (IBD) in Denmark. These authors found that cases of IBD were more likely to occur during a short period of time and over relatively short distances, indicating that local factors facilitated the spread of

the virus.

The spatial and space-time scan statistic

A commonly used test that can find the location of spatial clusters is the spatial scan statistic (Kulldorff 1997). The first application of this was in finding a circular cluster in case control data of breast cancer in the north-east USA (Kulldorff et al. 1997). Scan statistics are used to detect and evaluate clusters in a temporal, spatial, or space-time setting (see preceding section on the use of the temporal scan statistic). This is done by gradually scanning a window across time and/or space, while noting the number of observed and expected observations inside the window at each location. In the SaTScan software[4], the scanning window is either an interval (in time), a circle or an ellipse (in space) or a cylinder with a circular or elliptic base (in space-time). Multiple different window sizes are used. The window with the maximum likelihood is the most likely cluster and a p value is assigned to this cluster.

There have been many developments in the spatial scan statistic including the ability to find an elliptical cluster in case control data (Kulldorff et al. 2006), an application for ordinal data (Jung et al. 2006) and for survival data with adjustment for covariates (Huang et al. 2007), and one which accounts for the movement of people between home and work with a putative workplace exposure (Duczmal & Buckeridge 2006).

As well as Kulldorf's own elliptical cluster other workers have developed techniques for detecting irregularly shaped clusters (Duczmal & Assunçao 2004, Tango & Takahashi 2005, Assunçao et al. 2006).

The spatial scan statistic has been used in many investigations of zoonotic disease, including locating clusters of dead crows to predict human cases of WNV in New York State (Johnson et al. 2006), BSE cases in the UK (Stevenson et al. 2000) and Japan (Kadohira et al. 2008), and tuberculosis in cattle in Argentina (Perez et al. 2002). Green et al. (2006) used the spatial scan statistic to identify statistically significant ($p < 0.05$) high and low rate clusters of *Campylobacter* incidence in Manitoba. This study used a diverse set of spatial techniques including spatial scan statistic, spatial smoothing, and Poisson regression with a range of socio-demographic and landscape factors. They found a pronounced geographic variation of *Campylobacter* incidence associated with agricultural

[4] `http://www.satscan.org/`

animal density.

A number of papers have compared different spatial cluster detection techniques as follows. A full description of other mentioned cluster detection techniques can be found in Pfeiffer et al. (2008).

1. Wheeler (2007) uses the K-function, Cuzick and Edward's method, and the kernel intensity function to test for significant global clustering. This author also uses the kernel intensity function and Kulldorff's spatial scan statistic in SaTScan to test for significant local clusters. He finds consideration of the potential shape of clusters in the study area to be an important issue.

2. Waller et al. (2006) compared Tango's index of clustering and the spatial scan statistic using data from 1981 of severe cardiac birth defects in California. They report the dependence between the statistical power of tests of disease clustering and the strength, type, and location of suspected disease clusters. Consideration of the spatial distribution of the population at risk is also required for interpreting power comparisons between the different methods.

3. Song & Kulldorff (2003) compared eight test statistics for their power to detect disease clusters. They used simulated clusters based on the 1990 female population in the northwestern USA and conclude the power varies greatly for different test statistics. They recommend using the spatial scan statistic locally, and Tango's MEET for a general evaluation of clustering, if the the size and scale of the cluster is not known.

It is recommended that more than one method is applied to the data and results are compared. Also test comparisons such as these reported raise the issue of biological as opposed to statistical significance. Wartenberg (2001) questions the stringent application of statistical tests on decision criteria and instead advises considering the public health significance.

The spatial scan statistic implemented through SaTScan has enjoyed wide application and will continue to do so. However, two issues make using the method and interpreting its results complex. SaTScan does not provide cartographic support to view the identified clusters, nor a visual interface to explore cluster characteristics. Furthermore, it

is difficult to determine an optimal setting for SaTScan scaling parameters e.g. the default maximum-size setting of 50% may not produce usable, informative results, because the reported primary cluster often occupies a large proportion of the study area scanning window (Haining 2003). To address these issues, Chen et al. (2008) propose a novel geovisual analytics approach that combines the strength of advanced visualisation methods with the analytical capabilities of the spatial scan statistic. Geovisual analytics goes beyond traditional map output, allowing users to interactively explore visual representations of geographic information, use their own ability to process patterns and outliers from a visual scene, link these patterns and outliers to existing knowledge bases, and arrive at an appropriate course of action given the visual input (Keim et al. 2006). Chen et al. (2008) apply their methods to cervical cancer mortality data for the United States between 2000 and 2004 and conclude that their proposed geovisual analytics approach complements traditional statistical methods in cluster identification, enhancing the interpretation of identified clusters.

The space-time scan statistics were developed for retrospective data analysis of brain cancer in New Mexico (Kulldorff et al. 1998). This type of geographical disease surveillance tests if the disease is randomly distributed over space and time for a predefined geographical region during a predetermined time period. Kulldorff (2001) further developed the space-time scan statistic for prospective disease surveillance to detect active geographic clusters of disease. The statistic adjusts for the many possible time lengths and geographic sizes of the space-time clusters, and for multiple testing.

The space-time scan statistic has been utilised for research and surveillance of many zoonotic diseases. Use of the statistic to evaluate the potential for using reported clinical equine cases of WNV as an estimate of risk of human infection has been reported in Saskatchewan using data from a 2003 outbreak (Corrigan et al. 2006). Inopportunely, most clusters of human cases were not preceded by horse case clusters in the same areas. Ward (2002) investigated clustering of cases of canine leptospirosis in veterinary teaching hospitals in the USA using the space-time scan statistic. Results of this study suggest that cases were diagnosed at a higher rate than expected between 1993 and 1998 at hospitals located in the midwest of the United States. Possible explanations for the clustering detected include a shift in the serovars causing disease, climatic factors, and referral, diagnostic, and reporting biases. Recuenco et al. (2007) identified statistically

significant clusters of raccoon rabies in specific areas of New York from 1997 to 2003. They recommend that cluster areas identified should be considered for raccoon rabies control interventions such as use of the oral rabies vaccine. Furthermore these authors advise that public education on raccoon rabies exposures and the need for increasing pet vaccination activities should be prioritised in areas where clusters were identified.

The prospective space-time scan statistic was further developed to account for naturally occurring temporal and spatial trends (Kleinman et al. 2005) and for use when no information is available on the population at risk (Kulldorff et al. 2005). The prospective space-time scan was applied to Chicago's 2002 shigellosis surveillance data to assist in the detection and tracking of human shigellosis investigations (Jones et al. 2006). The authors suggest that this methodology could help prioritise the assignment and investigation of cases, particularly when an agency's resources are stressed by other events, such as outbreaks. They propose that other reportable endemic diseases, particularly those that, like shigellosis, are easily transmitted via close personal contact, could be monitored using this techniques. These diseases include hepatitis A virus infection and influenza.

The prospective space-time scan statistic has recently been incorporated into an online geographical information system (EpiScanGIS) for the detection of clusters of meningococcal disease in Germany (Reinhardt et al. 2008). An additional layer of information, the DNA-sequence typing of the bacteria, allows for detailed surveillance important for the monitoring of clonal spread and for the assessment of vaccine coverage. The combination of automation and typing allows the system to identify clusters of disease caused by a single type in close to real time. This results in an up-to-the-minute assessment of the disease burden that can trigger preventative actions such as public health campaigns.

Takahashi et al. (2008) proposed a flexible shaped space-time scan statistic and applied it to respiratory syndromic surveillance data in Massachusetts. These authors also tested the power of the flexible and the cylindrical scan to detect outbreaks. They report that for large and narrow clusters, as you might see on a peninsula or river bank, the flexible scan statistic would perform better than the cylindrical one. However they comment, that due to it being computationally demanding, the flexible scan is slower to run, so it may be more appropriate to use cylindrical scan for early detection, as timeliness is the key objective in this situation. However, when monitoring an occurring outbreak, geographical accuracy becomes the key objective, so once the outbreak has spread to a larger area, using the

flexible scan statistic is an appropriate next step to monitor spread.

2.4.3 Other spatio-temporal surveillance techniques

Point process methodology for surveillance

Currently in the UK, gastrointestinal disease surveillance is principally pathogen-specific based on isolations from routine faecal samples submitted by primary care. The Infectious Intestinal Disease study found that for every 136 cases that occurred in the community, only one is reported to national surveillance (Wheeler et al. 1999). This loss of epidemiological information limits the ability to detect outbreaks within local communities and reduces the opportunities for intervention. Therefore, surveillance based on pathogen isolations is highly specific but lacks sensitivity and speed. The aim of the Ascertainment and Enhancement of Gastrointestinal Infection Surveillance and Statistics project was to develop a surveillance system with enhanced sensitivity and speed to provide more opportunities for intervention and prevention in the local population (Diggle et al. 2004). In principle, this type of surveillance could also be applied to other types of public health needs.

The statistical objective of the analysis is to estimate the 'normal' pattern of spatial and temporal variation in the incidence of cases, and to identify quickly any anomalous variations from this normal pattern. This is addressed by decomposing the space-time intensity of incident cases into three separate terms (Diggle et al. 2005): firstly, the temporal variation in the mean number of incident cases per day, which is modelled parametrically through a combination of day-of-week and time-of-year effects; secondly, the overall spatial variation, modelled non-parametrically as a smoothly varying surface; and thirdly, the residual space-time variation, modelled as a spatiotemporal stochastic process.

Bayesian hierarchical modelling for surveillance

Knorr-Held & Richardson (2003) analyse space-time surveillance data on meningococcal disease using an hierarchical formulation, where latent parameters capture temporal, seasonal, and spatial trends in disease incidence. Spencer et al. (2008) modify this approach for the surveillance of cases of campylobacteriosis. The aim is to detect spatially localised point source outbreaks in campylobacteriosis notification data. Spencer et al. (2008) use 'outbreak indicators', spatio-temporal parameters that change from zero to one during a

period of increased incidence, to do this. The posterior distribution of these indicator variables consists of a probability that an outbreak is occurring at each point in space and time. This model has been applied to data in the Manawatu region of New Zealand and will be used in a real-time study of campylobacteriosis in Canterbury in 2009. In the proposed study, all notified cases will be typed by both pulse-field gel electrophoresis and multi-locus sequence typing, facilitating source attribution.

Joint disease modelling for surveillance

The issue of under-reporting of surveillance data was raised at the beginning of this chapter. Under-reporting has many components that may vary across space, for example, the proportion of sick patients for whom faecal samples are analysed, and the reporting behaviour of clinicians (Wheeler et al. 1999) and laboratories.[5] This can make a detailed spatial analysis of the variation in disease incidence difficult, as areas where disease incidence appears high may simply have higher rates of reporting.

One method of overcoming this problem is to develop a joint model of more than one disease under surveillance, that estimates and adjusts for under-reporting (Held et al. 2006). The incidence of the four zoonotic diseases campylobacteriosis, yersiniosis, and infections with two serovars of *Salmonella*, *Salmonella enterica* serotypes Enteritidis and Typhimurium, were jointly modelled. The model adopted is the so-called shared component model of Held et al. (2005) that assumes that the underlying risk surface for each disease can be partitioned into components that may be shared with the other diseases and that the geographical variation in under-reporting should be similar for the diseases considered. In this study, the risk pattern resulting for *S. typhimurium* and yersiniosis, representing infections hypothesised to be associated with raw or undercooked pork, show a clear regional pattern likely associated with raw food consumption habits. A geographically even distribution of risk, as was seen for *S. enteritidis*, could be explained by poor hygiene in food preparation. Unlike raw food consumption habits, lapses in hygiene are much less likely to show a clear regional pattern.

[5] http://www.svepm.org.uk/posters/2008/Spatial%20analysis%20of%20Salmonella.pdf

2.5 Conclusions

This review has focussed on methods and tools for detecting changes within zoonotic disease surveillance data. It is appropriate now to return to the key attributes of surveillance that were first recorded by Thacker et al. (1988). These include sensitivity, timeliness, flexibility, simplicity, and positive predictive value, and they continue to be important benchmarks of effective surveillance (Babin et al. 2007, Buehler 2008). Do the methods and tools presented in this review reach this benchmark? As computing power and statistics advance, there is a tendency for model complexity to increase. With increasing complexity comes less flexibility and simplicity. Furthermore, our surveillance data sets are becoming larger. It would be easy to lose sight of these key attributes as our ability to become more sophisticated increases. I propose three guidelines:

1. Strive to keep the analyses as simple as possible. Complexity should only be added if there are good grounds for doing so. Simplicity facilitates 'buy in' and trust from all those using the system, increasing the likelihood that the results of analyses will be carefully evaluated, and acted on in the event of a detected anomaly.

2. Improve sensitivity by using multiple techniques. No one methodology will suit all data. If similar results are found when using more than one analytical technique then this adds support to study findings.

3. Do not forget the human element. While we have many useful surveillance tools, they do not replace the human element of case investigations. Tracking down animal owners or patients and communicating with them, or with the administrators of at-risk settings such as day care centres, cannot be done without skilled investigators. Similarly, any tools we use are dependent on the accuracy, timeliness, and completeness of case report data. It would be too easy to not see the wood for the trees and forget about the importance of having well trained and remunerated staff working to ensure quality data is collected.

CHAPTER 3

Descriptive spatial epidemiology of subclinical *Salmonella* infection in Danish finisher pig herds: application of a novel method of spatially adaptive smoothing

Benschop, J., Hazelton, M.L., Stevenson, M., Dahl, J., Morris, R.S., French, N. (2008) Descriptive spatial epidemiology of subclinical *Salmonella* infection in finisher pig herds: application of a novel method of spatially adaptive smoothing. *Veterinary Research* **39:02**

3.1 Abstract

We describe the spatial epidemiological features of the 6.8 million meat-juice serological tests that were conducted between 1995 and 2004 as part of the Danish swine *Salmonella* surveillance and control programme. We investigated pig and farm density using edge-corrected kernel estimations. Pigs were aggregated at the county level to assess county-level risk, and then we investigated farm-level risk by giving farms a case or non-case label using a cut-off of 40% of pigs positive. Conditional probability surfaces, correcting for the underlying population at risk, were produced for each year of the study period using a novel kernel estimator with a spatially adaptive smoothing bandwidth. This approach improves on previous methods by allowing focussed estimation of risk in areas of high population density while maintaining stable estimates in regions where the data are sparse. Two spatial trends in the conditional probability of a farm being a case were evident: (1)

over the whole country, with the highest risk in the west compared to the east; and (2) on the Jutland peninsula, with the highest risk in the north and south. At the farm-level, a consistent area of risk was the south-west of Jutland. Case farms tended to aggregate indicating spatial dependency in the data. We found no association between pig or farm density and *Salmonella* risk. We generated hypotheses for this spatial pattern of risk and we conclude that this spatial pattern should be considered in the development of surveillance strategies and as a basis for further, more detailed analyses of the data.

3.2 Introduction

Exploratory data analysis is the cornerstone of sound epidemiological investigation (Tukey 1977). In exploratory spatial data analysis we aim to describe spatial variation in disease without any explicit attempt to represent this variation in terms of a probability model. Suitable tools include point maps, kernel smoothing, and mapping of relative risk and conditional probability surfaces. These allow us to generate hypotheses that give insight as to why disease is abundant in some areas but not in others.

Our motivation was to explore data from the Danish Swine *Salmonella* Surveillance and Control Programme (DSSCP) established by the Danish Ministry of Food, Agriculture and Fisheries in 1993 (Mousing et al. 1997). The DSSCP was set up in response to a general increase in the incidence of confirmed cases of human salmonellosis due to pork consumption (Baggesen & Wegener 1994) and a large, common source outbreak caused by *Salmonella infantis*, traced back to one slaughter plant and a small number of supplier pig herds (Wegener & Baggesen 1996). The objective of the DSSCP is to reduce the prevalence of *Salmonella* to an acceptably low level so that domestically produced pork is no longer an important source of human salmonellosis. At the time of writing this, the objective has largely been achieved: the number of cases of salmonellosis in humans attributable to pork consumption has reduced from 1,444 in 1993 to 142 in 2004 (Nielsen et al. 2001, Ministry of Family and Consumer Affairs 2005). At present, the focus of the DSSCP is increasingly on the efficiency of *Salmonella* surveillance, with both industry and authorities wanting to achieve the greatest reduction in the prevalence of *Salmonella* for their money.

In the light of this focus on efficiency, we used a novel kernel estimator with a spa-

tially adaptive smoothing bandwidth to produce estimates of the conditional probability of *Salmonella* risk across Denmark. Identified areas with a high probability of risk could be targeted for increased surveillance, while those with a low probability could be less frequently surveyed. This risk-based approach to surveillance would complement the recent strategy initiated in July 2005, where herds with a negative history of *Salmonella* are sampled less frequently (Ministry of Family and Consumer Affairs 2006).

Earlier spatial analyses of data from the DSSCP have described the geographical distribution of seroprevalence (Mousing et al. 1997), and fitted county of origin as a fixed effect in regression models that aimed to quantify the effect of factors that influence *Salmonella* seroprevalence (Carstensen & Christensen 1998). These studies have only considered data from 1995 from the DSSCP while the analyses presented here consider data from 1995 to 2004. The identification of spatial patterns in the data will inform further, more detailed analyses as well as highlight areas with abundant *Salmonella* infection for targeted surveillance.

3.3 Materials and methods

3.3.1 The *Salmonella* Surveillance and Control Programme

The DSSCP is based on the random testing of post-slaughter meat-juice samples from all finisher pig-herds that have an annual kill of greater than 200 finishers. The number of animals sampled at slaughter depends on herd size, with 60, 75, or 100 pigs sampled per herd per year (Alban et al. 2002). The testing of meat-juice rather than blood facilitates both sample collection and carcass identification (Nielsen et al. 1998). All samples are analysed at the Danish Institute for Food and Veterinary Research using the Danish mix-ELISA (Nielsen et al. 1995). This test can detect O-antigens from at least 93% of all serovars that are known to be present in Danish pigs (Mousing et al. 1997).

The sample results are used to categorise swine herds into one of three levels of a serological *Salmonella* finisher index (Alban et al. 2002). The index for each herd was calculated using a three step process: (1) totalling the monthly number of positive samples (optical density greater than 20) for the last three months; (2) weighting the months totals by 3:1:1 so the most recent month carries the most weight; and (3) adding the weighted totals

which are then divided by five. The three levels are level 1 with an index of 1-39; level 2 with an index of 40-69; and level 3 with an index of 70 or more. Herds in levels 2 and 3 have requirements placed upon them, e.g. pen faecal samples must be collected from the herd, producers must report their most recent weaner suppliers, and there are penalty *Salmonella* deductions resulting in reduced payments to these producers. Furthermore, pigs from level 3 herds are subject to special slaughter conditions. At the end of 2004, 3.5% of finisher herds were assigned to level 2 and 1.1% to level 3.

There have been a number of changes to the DSSCP since its inception. These include a change in the sampling strategy in August 2001, which was introduced to give more precise estimates for seroprevalence in smaller herds (Alban et al. 2002).

3.3.2 The data

Two extracts of data were obtained from the central database of the DSSCP from 1st January 1995 until 31st December 2004 (inclusive). The first data extract provided farm level information and included a unique identifier, as well as the identification of the county and commune in which the farm was located. There are 275 communes within 14 larger counties in Denmark. The unique identifier is attached to the physical locality, and not the owner, thus ensuring correct identification of the farm over the ten-year study period. Due to the near complete coverage of sampling in the DSSCP (all herds producing more than 100 finishers per annum prior to 1st August 2001, all producing more than 200 at or after 1st August 2001 to date), our data set was effectively a census comprising 99% of the population of Danish finisher swine herds.

Details of location of the farm house were provided for all farms that were registered with the Danish Central Husbandry Registry in March 2004. Amongst the 22,344 farms sampled between 1st January 1995 and 31st December 2004 which had individual pig and farm level data, 14,319 had recorded easting and northing coordinates. The remaining 8025 farms had coordinates randomly generated within their respective communes. This was an appropriate means for dealing with this type of missing data for two reasons. The size of those communes that produce pigs are small (range 20 to 58 km^2) relative to the entire land area of Denmark (43,000 km^2), and inferences drawn from this study were made at a broad national level rather than at the small commune level. Furthermore, the

need to randomly generate coordinates was time dependent: in 1995 this was necessary for 5940 of 16,095 farms (37%); in 1998, 4313 of 15,790 farms (27%); in 2001, 1204 of 11,977 farms (10%) and in 2004, only 173 of 9813 farms (2%).

The second data extract provided information relating to the 6.8 million individual carcasses that were tested (out of a total kill of approximately 200 million), and included the date of sampling, the unique farm identifier, and the result of the Danish mix-ELISA. For this analysis, a result of greater than 20 optical density (OD%) was classified as positive: this is the cut-off for positivity that has been used by the DSSCP since 1st August 2001 (Alban et al. 2002). The individual test sensitivity and specificity of the Danish mix-ELISA at this cut-off using meat-juice has been estimated at 60% and 100% respectively (Enoe et al. 2003).

3.3.3 Statistical analyses

Summary statistics

We provided summary statistics at the county level for the ten year study period to determine a broad pattern of the variation in spatial risk. For each of the 14 counties (see Figure 3.1) the number of farms and the proportion of samples yielding positive serology results were determined. The unit of analysis was the individual animal and confidence intervals for the proportion of positive serology results were calculated, taking into account the presence of clustering at the farm level (Dargatz & Hill 1996). The proportion of positive serology results for a county was expressed as the county-level incidence risk (IR). For each county the OD% values of positive serology results were aggregated and the lower, middle and upper quartiles calculated.

Spatial analyses

The land area of Denmark is comprised of a large peninsula (Jutland), two main islands (Fyn and Zealand), and 441 smaller islands (Figure 3.1). Southern Jutland forms a border with northern Germany. Since the start of the DSSCP over 85% of slaughter pigs have originated from the Jutland peninsula and Fyn. To visualise the broad scale variability in farm density we calculated edge-corrected Gaussian kernel estimations (Diggle 1985) of

the intensity function of the farm locations for each year of the study period. A fixed bandwidth was chosen for each year, determined using the Gaussian optimal method (Bowman & Azzalini 1997).

To visualise broad-scale variability in sampled pig density we produced edge-corrected Gaussian kernel estimations of the intensity function of the count of sampled pigs, by weighting the point locations of farms by the count of sampled pigs (Baddeley & Turner 2005). Since we did not have access to details of the number of pigs on each farm, we used the number of sampled pigs as a proxy for the farm-level pig population.

An important concept when describing spatial variation in risk is to adjust for the underlying population structure. The spatial relative risk function, estimated by kernel density methods, was developed for case control data for rare diseases by Bithell (1990). Lawson & Williams (1993) and Kelsall & Diggle (1995) further improved the estimation. In our analysis, we depart from the classic case control design since we have the unusual situation of what is effectively a census of pig farms. That is, we were not constrained by a strategy that required sampling of cases and controls. This means we can directly estimate the intensity of disease and use the following function, $p(x)$, to estimate the conditional probability of a farm being a case at location x:

$$p(x) = \frac{\lambda_1(x)}{\lambda_0(x) + \lambda_1(x)} \tag{3.1}$$

Equation 3.1 represents the probability of a farm being a case, conditional on its location. For each year of the study period farms were defined as cases if the proportion of sampled pigs that were positive was greater than or equal to 0.40. If otherwise farms were defined as non-cases. We chose this cut-off since it is the cut-off between levels 1 and 2 of the serological *Salmonella* finisher index. The function $\lambda_1(x)$ is the intensity of cases at an arbitrary point x, and $\lambda_0(x)$ is the intensity of non-cases at an arbitrary point x.

We obtained plug-in estimators of λ_1 and λ_0 using kernel smoothing:

$$\hat{\lambda}_0(x; h) = \sum_{i=1}^{n_0} K_h(x - x_i) \tag{3.2}$$

$$\hat{\lambda}_1(x; h) = \sum_{i=n_0+1}^{n} K_h(x - x_i) \tag{3.3}$$

where $x_1, ..., x_{n_0}$ are the locations of the non-case farms, and $x_{n_0+1}, ..., x_n$ are the locations of the case farms. Also, $K_h(x) = h^{-2}K(h^{-1}x)$ with $K()$ being a radially symmetric kernel function, and h a smoothing parameter called the bandwidth. The performance of the kernel density estimator depends on the choice of h. Fixed bandwidth kernel density estimators, as in Equations 3.2 and 3.3, have a major limitation when there is large spatial variation in the density of the population (as there is in the distribution of farms in Denmark). This results in an over-smoothing in areas of high density that could mask clusters of disease, and under-smoothing in areas of low density which could result in artificial spikes of disease risk.

We addressed this limitation by estimating conditional probability surfaces of a farm being a case for each year of the study period, using a method for bandwidth selection that implements spatially adaptive smoothing. Specifically, we used sample point dependent bandwidths in Equations 3.2 and 3.3, so that for the ith term in summation the bandwidth becomes $h_i = h(x_i)$. This bandwidth can be conveniently decomposed as $h_i = h\gamma_i$ where γ_i is a local bandwidth factor (constrained to have a geometric mean of one over the samples of cases and non-cases) and h a global smoothing multiplier common to both $\hat{\lambda}_0$ and $\hat{\lambda}_1$.

We selected the local bandwidth factors and global multiplier using a novel generalisation of the methodology of Hazelton (2007), extended from the case of one dimensional x values to two dimensional geographical locations. Specifically, we computed the local bandwidth factors by $\gamma_i = \delta_0/\sqrt{\tilde{\lambda}_0(x_i)}$ and $\gamma_i = \delta_1/\sqrt{\tilde{\lambda}_1(x_i)}$ for non-cases and cases respectively, where $\tilde{\lambda}_0$ and $\tilde{\lambda}_1$ are pilot estimates of the respective intensity functions (constrained to be equal at coastal boundaries), and δ_0 and δ_1 are constants chosen to fix the geometric means of the bandwidth factors as described above. As a consequence, the local bandwidth factors are selected so as to apply the appropriate degree of smoothing in a relative sense (i.e. that which minimises mean square error), so that regions with a high intensity, λ, of farms have smaller bandwidths than regions with low intensity. This means that data rich areas of fine detail (e.g. Viborg and Nordjylland) receive (far) less smoothing than geographical regions where the data are sparse (e.g. on Zealand). Meanwhile, the global smoothing multiplier is adjusted to optimise the conditional probability surface as a whole (in terms of minimising an estimate of mean squared error, integrated over the geographical region), rather than the individual intensity functions *per se*.

Another important consequence of this methodology is that the estimated conditional probability surfaces are consistent at the boundary without any additional correction for bias introduced by edge effects. This boundary bias in a kernel comprises contributions from both the numerator (intensity of case farms at an arbitrary point) and denominator (intensity of all farms at an arbitrary point) of the conditional probability function. However, the use of the global smoothing multiplier leads to the cancellation of the leading bias terms from both numerator and denominator, resulting in a great reduction in the magnitude of the bias for the conditional probability function itself.

Inspection of maps of slaughter pig farm densities (Figure 3.2) and the conditional probability surfaces (Figure 3.4) provided a visual guide as to possible association between farm density and the risk of disease.

3.4 Results

3.4.1 Summary statistics

The results from the descriptive analysis for counties are shown in Table 3.1. There was a large variation in the ten year county-level IR. The counties at the extreme north (Nordjylland) and south (Sonderjylland) of Jutland had the highest IR of 10%, with very little uncertainty surrounding these estimates. The next three counties with 9% IR were Fyn, Arhus, and Viborg. The counties with the highest IR also demonstrated higher OD% values for their lower, middle, and upper quartiles of positive serology results. Counties on the islands of Zealand and Bornholm had 4% or less IR, with the exception of Vestsjalland which had 6%.

3.4.2 Spatial analysis

Edge-corrected Gaussian kernel estimations of the intensity functions of the farm locations and of the count of sampled pigs for each year of the study period were mapped. Slaughter pig farm density (Figure 3.2) showed large variability across the country (from zero farms per square kilometer in Copenhagen to 0.6 in western Viborg in 1995). For presentation purposes, Bornholm is excluded from the figures. Farm density on Bornholm

was similar to that on Fyn. The solid and dashed contour lines delineate the upper fifth and twenty-fifth percentiles of farm density respectively. There was a reasonably uniform decrease in farm density over the ten year study period with the relative densities remaining very similar. For example, even though farm density in western Viborg reduced from 0.6 to 0.4 farms per square kilometre over the ten year period, it remained the region with the highest farm density. Both farm (Figure 3.2) and sampled pig density (Figure 3.3) was the highest in northern Nordjylland, western Viborg, central Arhus, eastern Vejle and south-eastern Sonderjylland (specifically on the island of Als). Throughout the whole study period, this farm and pig density pattern was reasonably consistent.

Figure 3.4 presents density plots of the conditional probability of a farm being a case for four years of the study period. The solid contour lines delineate the upper fifth percentile of *Salmonella* risk, clearly showing areas where the conditional probability of a farm being a case is high. The dashed contour lines delineate the upper twenty-fifth percentile of risk. The variable bandwidths ranged between 4.4 and 62.5 km. The geometric means of the bandwidths were between 8.6 and 12.6 km. These maps show two spatial trends that are consistent throughout the whole study period: (1) over the whole country there is increased conditional probability of a farm being a case on Jutland and Fyn when compared to Zealand and Bornholm; and (2) on the Jutland peninsula high conditional probabilities are in the south (especially the south-west) and to a lesser extent in the north. In addition to the consistent spatial trends in the conditional probability of a farm being a case over the ten years, there is a further trend that becomes apparent from 1998 onwards. This is a shift south in the northern polarity of risk on Jutland, and by 2004 there were pockets of central Jutland that had become areas of increased risk.

There was also a temporal pattern, with highest risk of *Salmonella* at the beginning and end of the study period. Six percent (905 of 16,095) and 4% (399 of 9813) of farms were cases in 1995 and 2004 respectively. Three percent of farms in both 1998 (404 of 15,790) and 2001 (303 of 11,977) were cases. Visual appraisal of the map series suggests no association between areas of high farm (Figure 3.2) or sampled pig density (Figure 3.3) and areas with increased risk of *Salmonella* (Figure 3.4).

In addition to the broad spatial trends, all maps showed evidence of aggregation of case farms visible as dark spots of varying diameter on the probability surfaces in Figure 3.4. This was most obvious in 1995 when the number of cases was the highest ($n = 909$).

Table 3.1: *Salmonella* seropositivity in Danish finisher pigs, 1995-2004. Descriptive results stratified by county. Data originate from the Danish swine *Salmonella* surveillance and control programme.

County Name	Number of			Percent positive	Percent positive		
	Farms tested	Results[1]	Positive results[2]	(95% CI)[3]	25th	50th	75th
Copenhagen	8	1719	57	3 (2–5)	24	32	40
Frederiksborg	238	40,254	1501	4 (3–4)	24	32	49
Roskilde	255	68,322	2638	4 (3–5)	24	33	54
Vestjalland	1577	406,690	23,257	6 (5–6)	26	38	69
Storstrom	1176	335,530	12,459	4 (4–4)	24	33	53
Bornholm	468	149,040	4994	3 (3–4)	24	31	46
Fyn	1874	644,171	58,723	9 (8–9)	27	41	75
Sonderjylland	1830	659,878	64,759	10 (9–10)	28	44	80
Ribe	1163	303,373	24,636	8 (8–9)	28	44	79
Vejle	1724	545,750	39,164	7 (7–7)	26	48	69
Ringkobing	2797	946,542	74,245	8 (8–8)	26	38	70
Arhus	2527	754,141	67,774	9 (8–9)	25	36	64
Viborg	3278	919,622	79,121	9 (8–9)	26	38	69
Nordjylland	3429	993,813	103,438	10 (10–11)	26	40	76
Total	**22,344**	**6,768,845**	**556,766**	**8 (8–8)**			

[1] Individual pig serology results.
[2] An OD% of greater than 20.
[3] Adjusted for farm-level clustering.
[4] Quartile range of all positive results.

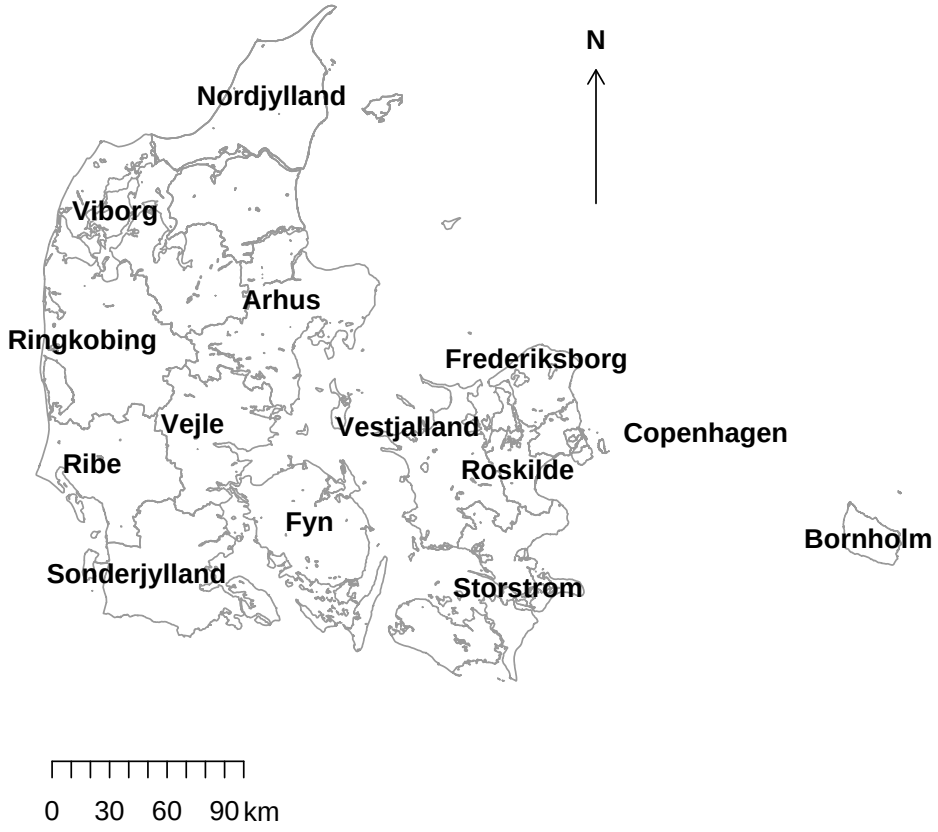


Figure 3.1: Map of Denmark showing the location of counties. The central island group containing Vestjalland, Roskilde, Storstrom, Copenhagen, and Fredriksborg is Zealand. The largest land mass containing Nordjylland, Viborg, Arhus, Ringkobing, Vejle, Ribe, and Sonderjylland is the Jutland peninsula. Southern Sonderjylland forms a border with northern Germany.

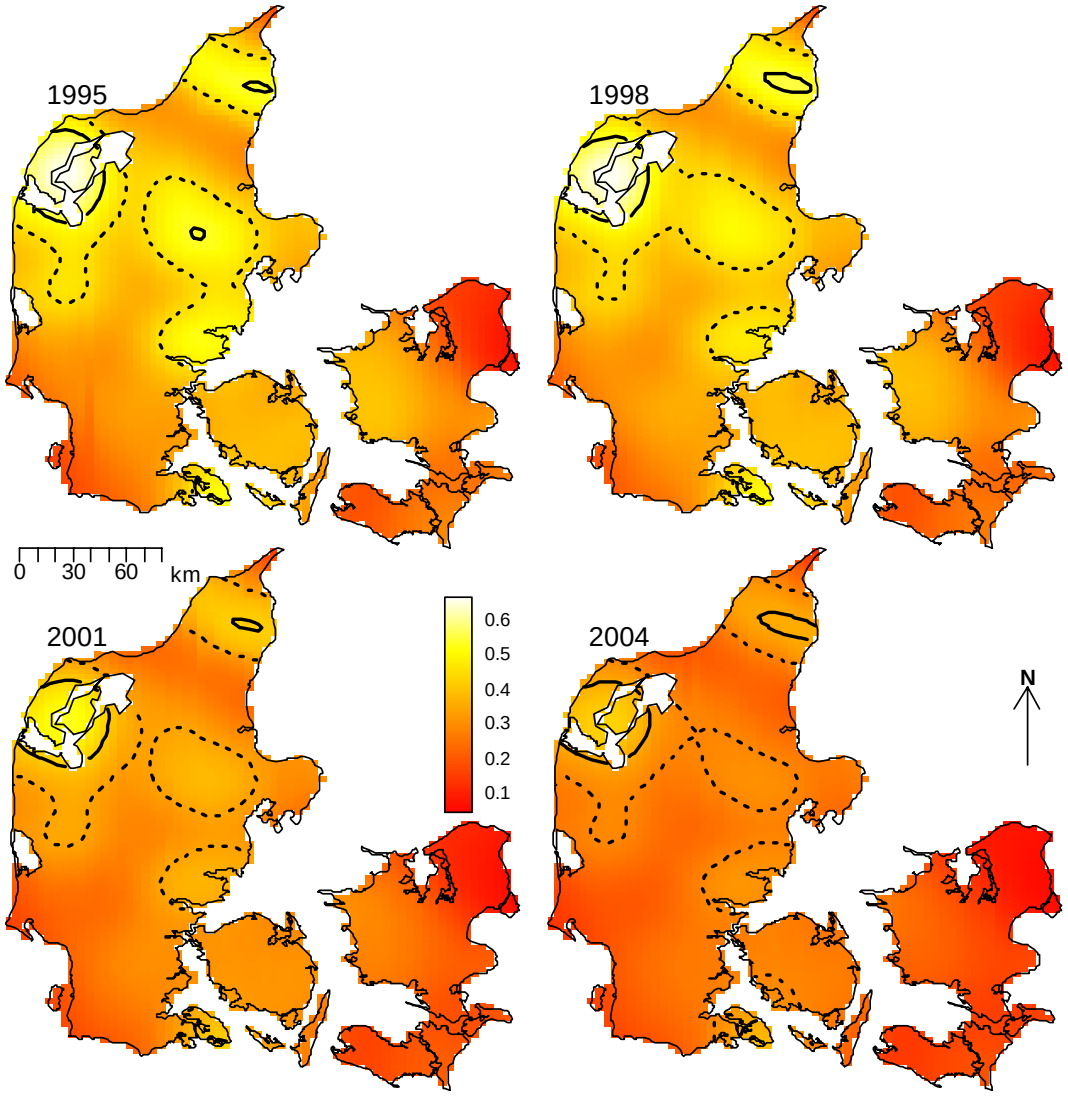


Figure 3.2: Edge-corrected kernel smoothed maps showing *Salmonella* meat-juice tested slaughter herd densities across Denmark in 1995, 1998, 2001, and 2004. Units are farms per square kilometre. The solid and dashed contour lines delineate the upper fifth and upper twenty-fifth percentiles of herd densities respectively. A fixed bandwidth of between 15.8 km (1995) and 17.2 km (2004) was used. Data originate from the Danish swine *Salmonella* surveillance and control programme.

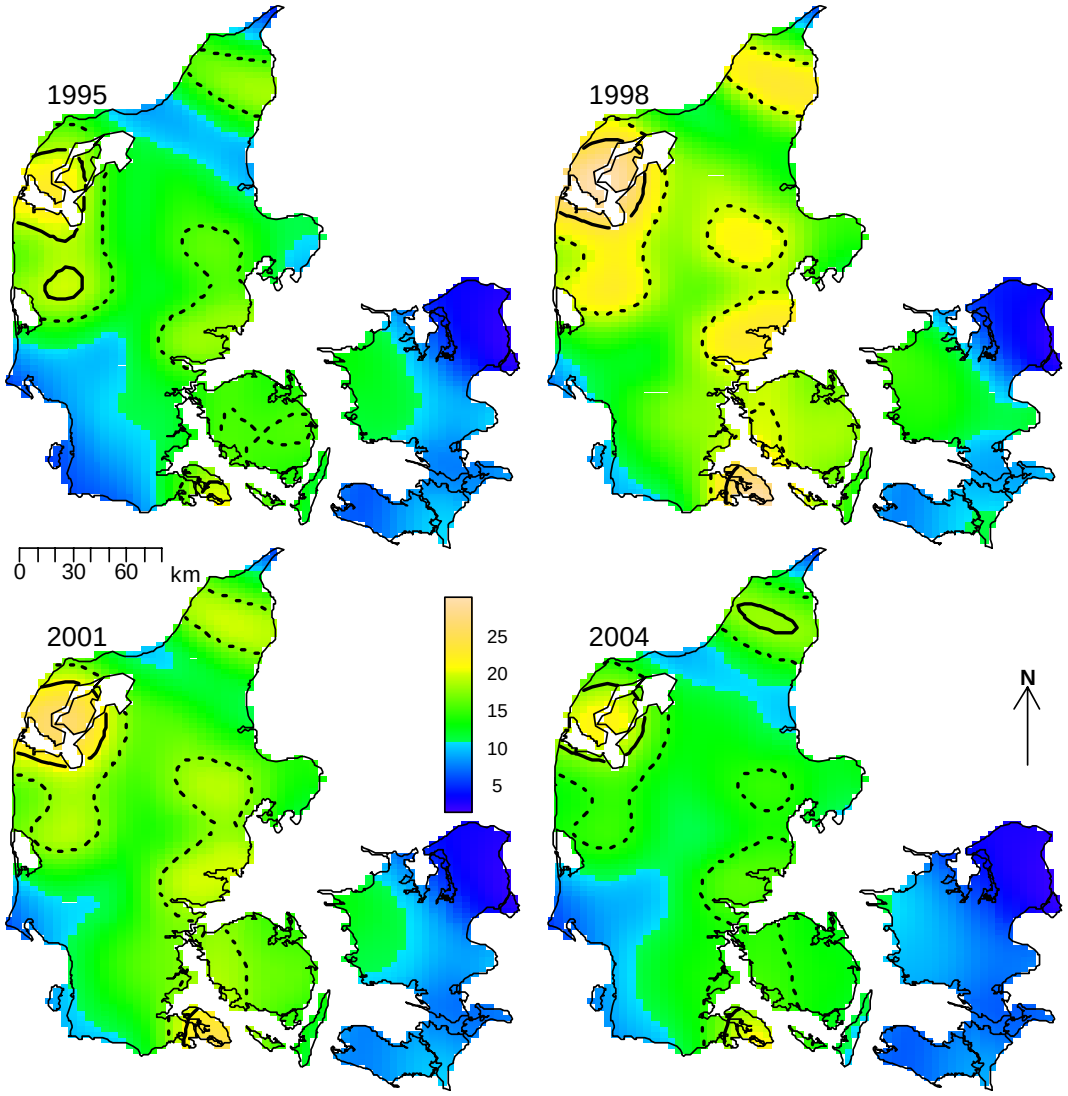


Figure 3.3: Edge-corrected kernel smoothed maps showing sampled *Salmonella* meat-juice tested slaughter pig densities across Denmark in 1995, 1998, 2001, and 2004. Units are sampled pigs per square kilometre. The solid and dashed contour lines delineate the upper fifth and upper twenty-fifth percentiles of herd densities respectively. A fixed bandwidth of between 15.8 km (1995) and 17.2 km (2004) was used. Data originate from the Danish swine *Salmonella* surveillance and control programme.

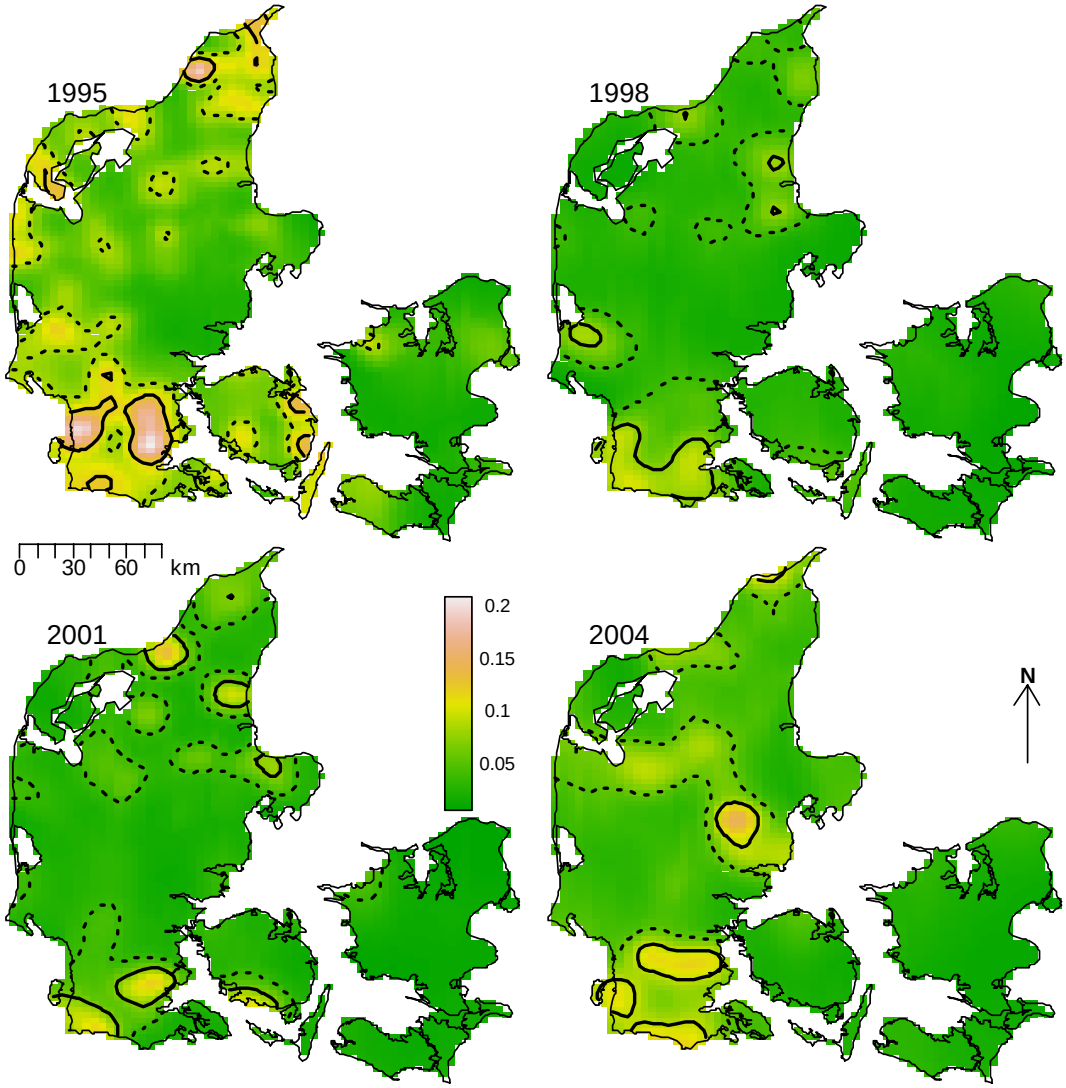


Figure 3.4: *Salmonella* seropositivity in Danish finisher pigs in 1995, 1998, 2001 and 2004. Density plots show the conditional probability of a farm being a case. The solid contour lines delineate the upper fifth and twenty-fifth percentiles of *Salmonella* risk respectively. Data originate from the Danish swine *Salmonella* surveillance and control programme.

3.5 Discussion

This study has used a novel spatially adaptive smoothing technique to identify spatial patterns of slaughter pig *Salmonella* risk throughout Denmark. We found large variation in both county- and farm-level risk throughout Denmark, with increased risk in the south and north of Jutland. The spatio-temporal analysis carries this further with visual confirmation of this pattern over the whole ten year study period. It has become apparent from this study that farms are most at risk in south-west Jutland and that those in central Jutland experienced increased risk from 1998 onwards.

We found that the geographical area within which a farm is located influences the risk for *Salmonella* and although farm location is not something that can be altered, it is useful to be aware of the high risk areas. More intensive sampling and gathering of epidemiological data in these areas may provide clues to underlying aetiologies. Qualitative assessment of the risk surfaces showed that case farms tend to aggregate, especially in 1995 when they were more numerous. This suggests that the underlying process is naturally clustered pointing to the need to explicitly model this structure in later work. When we visually assessed the relationship between farm and pig density and the farm-level risk of being a *Salmonella* case, we saw no association between them.

The kernel estimation of spatial relative risk is a useful tool in epidemiology which has been previously constrained by instability in estimates in areas where data are sparse (Clark & Lawson 2004). Our application of a method for bandwidth selection which implements spatially adaptive smoothing overcomes this constraint and has the added advantage of self-correcting for edge effects. When compared with estimates computed using a fixed bandwidth, we were able to demonstrate appreciable improvements in performance with correction of over-smoothing in areas of low density, and under-smoothing in areas of high density. We believe this technique is a valuable addition to the repertoire available to spatial epidemiologists.

Our findings at the county-level are in agreement with earlier work on data from the DSSCP (Carstensen & Christensen 1998, Mousing et al. 1997). Counties in the north and south of Jutland (Nordjylland and Sonderjylland) experienced increased *Salmonella* risk. Counties with the highest animal-level IR also demonstrated higher OD% values for the lower, middle and upper quartiles of their positive serology results. One possible

explanation for this is that differences in the serological response depend on the serovars involved and the time since infection (Nielsen et al. 1995). It is also very likely that the infection pressure within a herd contributes to an individual pig's level of seropositivity. The pattern of a low positive summary for counties in Zealand (with the exception of Vestjalland) and Bornholm (Table 3.1) may be a result of either low herd-level infection pressure or infection with serovars resulting in low test sensitivity, or both.

Caution must be used when interpreting the results from aggregated data since it may be subject to ecological bias (Biggeri et al. 1999). If a county has a high IR, this will not necessarily mean that all farms within that county are so affected; this becomes apparent when we investigate the farm level data. Notwithstanding our use of different measures of risk for county (proportion of samples positive) and farm, (case or non-case with cut-off of 40% of samples positive), we can identify that it is only parts of the most-affected counties that have an increased conditional probability of being a case. Sonderjylland experienced a high county-level IR at 10%, but it is only the south-west of the county which consistently had the highest conditional probability of a farm being a case.

Our findings must be considered in the light of changes that have occurred in both the DSSCP sampling strategy and the Danish swine population during the study period. A small part of the reason for the decrease in farm density throughout the study period was due to a change in the DSSCP that occurred in August 2001 (Alban et al. 2002). The cut-off for herds to be eligible for sampling was raised from 100 to 200 pigs killed per year resulting in fewer herds being sampled with those that remained eligible being larger. This potentially results in selection bias, the direction and magnitude of which is unknown. However, we are confident it is likely to be of little consequence as over the study period herd sizes were increasing. In fact Danish pig herds doubled in size: in 1995 the number of finishers produced per herd was 1757 compared to 3043 in 2004 (Anonymous 2006).

However, the reduction in the total number of suppliers producing pigs for slaughter was the more important reason for this decrease in farm density over the study period. For example, the number producing more than 200 pigs annually was approximately 13,200 in 1994 reducing to 7800 in 2004 (Anonymous 2006).

Another major change implemented in the DSSCP in August 2001 resulted in more samples being taken from smaller herds and less samples from larger herds. Again, this poten-

tially results in bias of unknown direction and magnitude since the relationship between herd size and *Salmonella* risk is not straightforward (we address this more fully later in the discussion). Overall, both DSSCP changes instigated in August 2001 led to a 13% reduction in the total number of samples taken (Alban et al. 2002). These changes in the sampling strategy confound our results for sampled pig density, allowing only empirical comparisons to be drawn.

The spatial patterns of *Salmonella* risk reported here allow us to generate a number of hypotheses. The pattern was possibly due to risk factors acting on a broad spatial scale such as a common infected source of pigs, contaminated feed, or regional biosecurity practices. A Danish study of the latter identified southern Jutland as an area where all surveyed sites hired commercial transport to the abattoir (Boklund et al. 2004). Transportation can be a biosecurity risk if, for example, transport personnel are allowed onto the farm (Baum et al. 1998).

The south of Jutland has been identified as a high risk area for post-weaning multisystemic wasting syndrome from 2001 to 2003 (Vigre et al. 2005) and porcine reproduction and respiratory syndrome virus. There are published reports of the links between these diseases and salmonellosis in pigs (Belœil et al. 2004, Ha et al. 2005, Murakami et al. 2006, Wills et al. 2000). Concurrent infections may act systemically to lower a pig's overall resistance to *Salmonella*, or cause disturbances in the normal gut flora, leading to increased susceptibility for *Salmonella* colonisation.

Denmark enjoys relative geographic isolation when compared with other northern European countries: Fyn, Zealand and Bornholm are true islands, and only the south of Sonderjylland is connected to the European mainland. The boundary between south Jutland and northern Germany may play an as yet undetermined role in the consistent risk of seropositivity seen in the south-west of Sonderjylland. Although there is no movement of pigs across the German border into Denmark, there is movement of pig feed, since farmers in the south of Jutland purchase feed across the border when the price is lower. We speculate that this could be a potential source of *Salmonella* for herds in south Jutland, but further studies would need to be done to elucidate this.

Other risk factors acting on this scale could include regional variations in management practices such as home-mixing of feed in grain-producing areas, with corresponding less use of pelleted feed. It is generally agreed that the feeding of pelleted feed is a risk for

salmonellosis on pig farms (Leontides et al. 2003, Mikkelsen et al. 2004).

The two counties with the highest county level IR (Nordjylland and Sonderjylland) have large pig populations and generally larger farms. The association between herd size and *Salmonella* risk is complex: some studies report a positive association (Carstensen & Christensen 1998, Mousing et al. 1997) others a negative (van der Wolf et al. 2001) or no association (Lo Fo Wong et al. 2004*b*, Stege et al. 2001). Larger herds may have facilities that, in themselves, are sparing on *Salmonella* risk, e.g. feeding of fermented liquid feed (Belœil et al. 2004, van der Wolf et al. 2001) and improved biosecurity (Funk et al. 2001, Lo Fo Wong et al. 2004*b*).

Since this is a descriptive study, we set out to generate the above hypotheses rather than specifically address the factors that may explain the patterns observed. By identifying these spatial patterns, we have highlighted the south-west of Sonderjylland on the Jutland peninsula as an area for further investigation and targeted surveillance. Furthermore, our exploration of the data has identified that case farms tend to aggregate. A method for investigating this issue further would be to model the data, accounting for both the broad spatial trend and the local spatial correlation structure. We explore the latter in the next chapter.

CHAPTER 4

Informing surveillance programmes by investigating spatial dependency of subclinical *Salmonella* infection

Benschop, J., Stevenson, M., Dahl, J., Morris R.S., French, N. (2009) Informing surveillance programmes by investigating spatial dependency of subclinical *Salmonella* infection. *Epidemiology and Infection* **137**:1348-1359

4.1 Abstract

The aim of this paper is to investigate local spatial dependency with regard to *Salmonella* seropositivity in data from the Danish swine *Salmonella* surveillance and control programme and its application in informing surveillance strategies. We applied inhomogeneous and observed-difference K-function estimation, and geo-statistical modelling to data from the Danish swine *Salmonella* surveillance and control programme. Slaughter-pig farm density showed large variation both at the country-wide and at the local level in Denmark (median 0.23, range 0.02-0.47 farms per square kilometre). The spatial distribution of pig farms followed a random inhomogeneous Poisson process but was not aggregated. We found evidence for aggregation of *Salmonella* case farms over that of all farms at distances of up to six kilometres and semivariogram analyses of *Salmonella* seropositivity revealed spatial dependency between pairs of farms up to four kilometres apart. The strength of the spatial dependency was positively associated with slaughter pig farm density. We propose sampling more intensively those farms within a four kilometre

radius of farms that have been identified with a high *Salmonella* status, and reduced sampling of farms that are within this radius of '*Salmonella*-free' farms. Our approach has the potential to optimise sampling strategies while maintaining consumer confidence in food safety, and also to be used in other zoonotic disease surveillance systems.

4.2 Introduction

The value of geo-referenced data in veterinary surveillance of both endemic and exotic diseases is immense. Recent examples in the literature show that these data have been used not only to identify areas with excess disease (Haine et al. 2004, Sanchez et al. 2005) and target areas for further studies (Graham et al. 2005), but also to produce hypotheses about means of disease introduction (Vigre et al. 2005), identify the likely site of incursion of an exotic disease (Stevenson et al. 2005) and for predictive modelling of alternative control strategies (Yoon et al. 2006).

The geo-referenced locations of livestock farms can be considered a spatial point process (Diggle 2003). An underlying assumption in the analysis of these processes is that of stationarity or spatial homogeneity, i.e. the intensity of the process does not depend on the location in space (Diggle 2003, Banerjee et al. 2004). A point pattern representing the location of livestock farms will typically not meet this assumption - farms will likely be distant from large urban centres and often will be located near areas that meet their needs for specific inputs e.g. feed supply and market access. Furthermore, in developed countries, legislation now dictates the location of intensive production units due to their effects on the environment, such as emissions of ammonia and phosphorus and requirements to spread slurry.

Statistically, spatial point patterns can be partitioned into first- and second-order properties that capture their global and local behaviours respectively (Banerjee et al. 2004). If the pattern shows a global trend (i.e. is non-stationary or inhomogeneous) then it exhibits a first-order effect. A second-order effect is due to spatial dependency and results from the spatial correlation structure in the data; these are small-scale or local effects. Somewhat ambiguously, both first- and second-order effects produce point patterns that exhibit local concentrations of points and it can be difficult to clearly identify one from the other (Diggle et al. 2007).

When looking specifically at slaughter-pig production in intensive farming areas, there are concentrated areas of pig production within which the distances between farms can be very small. Denmark, as the world's largest exporter of pig meat, provides a good example of intensive pig farming. The first aim of this paper is to capture the spatial distribution of these farms with regard to first- and second-order effects, using farm location data from the Danish Central Husbandry Register in 2003. Our second aim is to investigate the second-order spatial properties by marking the locations with disease status and with a random farm effect value from a generalised linear mixed model. We then determine the implications for surveillance. This methodology could be used on suitable data from any national disease control programme. We used data from the Danish swine *Salmonella* surveillance and control programme from 2003. Many other countries that intensively farm pigs look to the Danish control programme as a model e.g. the Zoonoses Action Plan in the United Kingdom (Armstrong 2003), Ireland (Casey et al. 2004) and the German QS system (Blaha 2004). The Danish programme was developed in 1993 in response to an increase in the incidence of salmonellosis in humans attributable to consumption of pork (Alban et al. 2002, Mousing et al. 1997) and is based around the random testing of meat-juice samples from slaughtered pigs. All herds that produce greater than 200 finishers per year are tested and then categorised into one of three levels of a 'serological *Salmonella* index' for intervention strategies (Alban et al. 2002). An in-depth review of the programme is given by Christensen (2003).

In Denmark the number of human Salmonellosis cases due to pork consumption has substantially reduced from 1444 in 1993 to 164 in 2004 (Nielsen et al. 2001). This reduction in the number of human cases provides some indication that the interventions that have been applied have been effective but raises questions about where to go to next in terms of resource allocation within the programme (Alban & Stärk 2005). There have been a number of recent stochastic models, both Danish (Alban & Stärk 2005) and from elsewhere (van der Gaag et al. 2004*a*, Miller et al. 2005) which have addressed the question, and results were variable. The North American model found higher cost-benefit ratios for improvements in the post-slaughter phase (Miller et al. 2005), while both the Danish (Alban & Stärk 2005) and the Dutch (van der Gaag et al. 2004*a*) identified both pre- and post-slaughter interventions as efficient.

In terms of pre-slaughter interventions, little consideration has been given to small-scale

spatial risk factors. Work on the Danish programme has described a strong first-order spatial effect with a higher prevalence of farm-level seropositivity in the north and south of Jutland, and in the west of the country compared with the east (Mousing et al. 1997, Carstensen & Christensen 1998, Benschop et al. 2008*a*). Our recent work (Benschop et al. 2008*a*) has identified that case farms tend to spatially aggregate, but we are not aware of any work specifically investigating the second-order properties of the data. Both increased pig density within a region (Fedorka-Cray et al. 2000) and small distances to other pig farms (Berends et al. 1996, Langvad et al. 2003) have been identified as risk factors for *Salmonella* infections. The bacteria are long-lived in the environment (Winfield & Groisman 2003) and contaminated faecal matter can act as a reservoir (Gray & Fedorka-Cray 2001), so processes acting locally, such as sharing contaminated agricultural machinery or poor biosecurity between farms, make the small-scale spatial structure worthwhile investigating. This has the potential to inform models that may lead to improved resource allocation in the Danish and other similar programmes.

4.3 Materials and methods

4.3.1 The data set

Two extracts of data from 1st January 2003 until 31st December 2003 were obtained from the Danish swine *Salmonella* surveillance and control programme (DSSCP) (Alban et al. 2002, Mousing et al. 1997). These extracts comprised pig and farm level data. We chose data from 2003 for analysis since this was the period with the highest proportion of geo-referenced farms (96.2%).

Data were managed using a relational database (Microsoft Access 2002 for Windows; Microsoft Corporation, Washington, USA) and spreadsheet software (Microsoft Excel 2002 for Windows; Microsoft Corporation). Statistical analyses were performed using the R statistical package version 2.2.0 (R Development Core Team 2007) and WinBUGS version 1.4.1 (Imperial College and MRC, UK). R contributed packages spatstat (Baddeley & Turner 2005), geoR (Ribeiro Jr. & Diggle 2001), splancs (Rowlingson & Diggle 1993) and sm (Bowman & Azzalini 1997) were also used.

4.3.2 Pig-level data

There were 578,268 individual finisher pig meat-juice results. Each included the date of sampling, the central husbandry register number identifying the farm of origin, and the result of the Danish-mix ELISA. A result of greater than 20 OD% was classified as positive. This is the cut-off for positivity that has been used by the DSSCP since 1st August 2001 (Alban et al. 2002).

4.3.3 Farm-level data

Of the 10,571 farms for which individual pig results were available 10,166 also had easting and northing coordinates of the farm house. This represented 96.2% of the contributing farms. The 405 farms without coordinate information were excluded from the analyses. Each farm had its central husbandry register number which included a number indicating within which of the 15 Danish counties the farm was located. Because they contributed very few farms, the two counties that constituted the county of Copenhagen were merged.

4.3.4 Spatial analysis

To investigate the spatial distribution of slaughter pig farms we used three techniques: kernel estimation, nearest neighbour distance, and the inhomogeneous K function.

We calculated kernel density estimates (Diggle 1985) of farm locations to visualise the broad scale variability in farm density. Spatially adaptive smoothing was implemented by weighting the global bandwidth at each data point with weights derived from a pilot estimate (Marshall & Hazelton 2008). Regions that are data rich (e.g. Jutland), therefore, receive less smoothing so as to preserve fine detail, whereas regions where the data are sparse (e.g. Zealand) receive more smoothing. A linear boundary kernel with a Gaussian base, was used to reduce boundary bias, and a global smoothing bandwidth of 17 km was chosen using the normal optimal method (Bowman & Azzalini 1997).

For each county, we calculated the distance from every farm location to its nearest neighbour.

We estimated a non-stationary analogue of the standard *K* function, the inhomogeneous *K*-function (Baddeley et al. 2000), to investigate for evidence of local aggregations of pig farms after allowing for their non-uniform density. The *K*-function is defined as the expected number of further points within a distance r of an arbitrary point, divided by the overall density of the points (Ripley 1976).

$$K(r) = \frac{N(r)}{\lambda} \tag{4.1}$$

In Equation 4.1 $K(r)$ is the standard *K*-function, $N(r)$ is the expected number of neighbouring farms within a distance r of an arbitrary farm and λ is the farm density. Inhomogeneous *K*-function analysis was performed using five large approximately square areas that included 82% of the sampled farms (Figure 4.1). Square areas were chosen to avoid the instability that may be associated with unusual window geometry (Ripley 1988). Analysis of the whole of the country was prevented by computational and geographical constraints. To reduce the instability due to edge effects, Ripley's isotropic corrections were implemented (Ripley 1988). One hundred simulated realisations of an inhomogeneous Poisson process were generated and the inhomogeneous *K*-functions of these were calculated to produce an envelope around the observed data. This provides a way of testing if the observed pattern of farms is aggregated even after allowing for its non-uniform density. The practical value of the inhomogeneous *K*-function over the standard *K*-function is that the former permits a more global measure of aggregation as it allows for spatial inhomogeneity of the pattern (a varying λ).

To investigate if there were spatial aggregations of case farms over that of all farms, the observed-difference *K*-function was calculated. A farm was defined as a case if it had a proportion of positive pigs greater than or equal to 0.4. We chose this cut-off as it is the cut-off between levels 1 and 2 of the serological *Salmonella* finisher index. If herds are in level 2 or 3, there are requirements placed upon them e.g. pen faecal samples must be collected from the herd, and there are penalty '*Salmonella* deductions' reducing payments to these producers. Approximately 3% of herds were in levels 2 or 3 during 2003.

For each county, separate *K*-functions at distances r were calculated for both case farms, $K_{case}(r)$, and for all farms, $K_{pop}(r)$, and the observed difference function $D(r)$ was calculated as follows:

$$D(r) = K_{case}(r) - K_{pop}(r) \tag{4.2}$$

The null hypothesis was of no extra aggregation of cases over that of the population corresponding to the cases being a random sample from the population. This permits the use of randomisation tests which do not require the underlying point process to be stationary (Diggle et al. 2007). Upper and lower permutation envelopes were produced by 99 random re-labellings of the cases and population. Values of the observed difference function were calculated for each permutation to investigate if there was any significant deviation of the observed difference function from zero (Chetwynd & Diggle 1998).

Our second approach to determine if there were any second-order effects was to investigate the hypothesis that geographically close farms were more similar than those geographically distant. The relationship between the outcome response (the proportion of pigs positive per farm) and the effect of herd size and farm was examined by fitting a generalised linear mixed model as follows:

$$\log(p_{ij}/1 - p_{ij}) = \beta_0 + \beta_1 x_{ij} + U_i \tag{4.3}$$

In Equation 4.3 the logit of the observed probability of the *j*th pig from the *i*th farm being seropositive, p_{ij}, was estimated as a function of a binary variable representing large herd size category, and a random effect term, U_i, which was normally distributed with a mean of zero and variance σ^2.

The model was applied to all farms in Denmark that had easting and northing coordinates supplied and were producing pigs for slaughter in 2003. The model was sequentially run on all Danish pig producing counties as computational constraints prevented modelling all farms at once.

Model parameters were estimated using a Bayesian approach, implemented in WinBUGS version 1.4.1. Markov chain Monte Carlo (MCMC) methods were applied to the observed data to simulate values from the joint conditional distributions of the unknown quantities. We chose relatively non-informed prior and hyperprior distributions for all model parameters: for the fixed-effects we chose Normal(0, 0.000001) and for σ^2 (the variance of the farm random-effect term), we chose Gamma(0.1, 0.001). Three chains were run and

convergence was judged to have occurred on the basis of visual inspection of time series and Gelman-Rubin plots (Toft et al. 2007). The length of the chain was determined by running sufficient iterations to ensure the Monte Carlo standard errors for each parameter were less than 5% of the posterior standard deviation. A total of 30,000 iterations were run with a burn in of 5000 iterations.

The farm level random effects from the model were plotted onto county map outlines in an initial investigation into the presence or otherwise of second-order spatial effects. Then omni-directional binned semivariograms were plotted. These illustrate the difference between pairs of data points (farm level random effects) within a given spatial lag (the distance between pairs of farms) (Isaaks & Srivastava 1989). If there was spatial dependency between farms we would expect an upward trend in the variogram. Conversely, little or no spatial auto-correlation would produce an essentially flat variogram. Directional semivariograms at angle sizes of 0, 45, 90, and 135 degrees (tolerance of ± 22.5 degrees) were plotted to investigate if the spatial structure was anisotropic.

The significance of the spatial auto-correlation was determined by permuting the data values on the spatial locations to produce simulation envelopes. As permuted data should not exhibit spatial dependency any points lying outside these simulation envelopes indicate significant spatial auto-correlation. The magnitude of the spatial auto-correlation was determined by calculating the ratio of nugget to total semivariance. The nugget semivariance is the point at which an extrapolated fitted line would cross the vertical axis. A nugget to total semivariance ratio of less than 25% indicated strong spatial dependence, between 25 and 75% indicated moderate spatial dependence, and greater than 75% indicated weak spatial dependence (Cambardella et al. 1994).

As we were interested in small-scale spatial dependency for both K function and semivariogram analysis, the maximum distance investigated was ten kilometres.

4.4 Results

There were 10,166 farms sampled in 2003 in the Danish program with coordinate information. Figure 4.2 is the edge-corrected kernel smoothed map of the farm density. Smoothed farm density was normally distributed with a mean of 0.20 and a standard

deviation of 0.09 farms per square kilometre. The range of smoothed densities varied through-out the country from zero in Copenhagen to 0.47 per square kilometre in Viborg.

Figure 4.1 shows the location of counties and the five areas used in the investigation of inhomogeneous K-function estimation. Table 4.1 gives the area, number of farms, and farm density for each of the five areas selected for inhomogeneous K-function analysis. In total, the areas encompassed 8286 of the 10,166 farms sampled for 2003. Over all five areas there was a wide range of farm densities from a median of 0.30 (range 0.03–0.38) farms per square kilometre in North Jutland to 0.14 (range 0.01–0.30) in Zealand.

The inhomogeneous K-function analysis of all large square areas showed that the observed pattern of farms was not aggregated. Figure 4.3 shows the inhomogeneous K-function for the square area in the north of Jutland, all estimates lie within the simulation envelopes that are produced under the null hypothesis of no aggregation.

The median nearest-neighbour distance was 0.77 km (IQR: 0.69; range: 0.01–11.56 km). Figure 4.4 is a boxplot of these distances for four of the Danish counties.

Using the cut-off of greater than or equal to 40% meat-juice ELISA positive pigs in a herd produced 272 case farms. The case incidence risk was 3%. Figure 4.5 shows the observed-difference K-function between case and population farms for the counties of Nordjylland, Arhus, Ringkobing, and Sonderjylland. Nordjylland, Ringkobing, and Arhus show evidence of local spatial aggregation of case farms over that of all farms. The extent of the aggregation was 1 km for Nordjylland and 4 km for Arhus. For Ringkobing it was statistically significant at 6 km with points beyond the simulation envelope. Together, these three counties held 40% of the Danish pig population in 2003. The results for the remaining counties were similar to that of Sonderjylland, showing no evidence for local spatial aggregation of case farms over that of all farms.

When the farm level random effects were plotted by their coordinates, there were no apparent aggregations of similar sized random effects. This pattern was seen in all counties. However, semivariograms for most large pig-producing counties showed evidence of spatial dependency with an upward trend in the variogram at up to four kilometres distance. Although most counties had all points lying within the simulation envelopes, the four main pig producing counties, Nordjylland, Viborg, Arhus and Sonderjylland had points below the envelopes indicating significant spatial auto-correlation from two to four kilometres (Figure 4.6). Together, these four counties held 50% of the Danish pig population

in 2003. The nugget to total semivariance ratios of these four counties was approximately 70%, indicating moderate spatial auto-correlation. The strength of the dependency was proportional to slaughter pig density with the exception of Fyn. Variograms for the remaining pig producing counties are shown in Figure 4.7.

Table 4.2 shows the farm-level prevalence unadjusted for herd size, proportion of farms in the large herd size category, odds ratios for large herd size, and the variance of the random effects with 95% Bayesian credibility intervals for each county. The unadjusted farm-level prevalence was highest at approximately 5% in the north of Jutland (Nordjylland and Arhus) and lowest, at approximately 1%, in the east of Denmark (Bornholm and Roskilde). All counties in Jutland and Fyn had 43% or more farms in the large herd size category. Odds ratios for Nordjylland, Fyn, Ribe, Vejle, and Viborg were significant, suggesting that pigs in these counties were at more risk of being seropositive if herd size was large (greater than 2000 finishers produced annually) than if it was medium (between 200 and 2000 finishers produced annually). The variance of the random effects was greatest in Sonderjylland, indicating that farms in this county showed the most variation in farm level prevalence of *Salmonella*.

Table 4.1: Area, number of farms, number of case farms, and farm density for the five approximately square regions used in K-function analysis, Denmark, 2003. Data originate from the Danish swine *Salmonella* surveillance and control programme.

Location	Area (sq. km.)	Number of farms	Farm density (farms per sq. km.)	
			Median[1]	Range[1]
North Jutland	2645	832	0.30	0.03–0.38
Central Jutland	15,960	4232	0.23	0.06–0.39
South Jutland	8712	1623	0.17	0.02–0.32
Fyn	3248	738	0.23	0.05–0.36
Zealand	5625	859	0.14	0.01–0.30

[1] Calculated using the spatstat library (Baddeley & Turner 2005) in R.

Table 4.2: Unadjusted farm-level *Salmonella* seroprevalence, odds ratios for large herd size, proportion of farms in the large herd size category, and variance of the farm level random effects for Danish pig producing counties in 2003. Data originate from the Danish swine *Salmonella* surveillance and control programme.

County	Prevalence % (95% CI[1])	Odds Ratios (95% CI[1])	Proportion farms in large size category[2]	Variance random effects (95% CI[1])
Nordjylland	5.9[3] (5.5–6.3)	1.15[4] (1.01–1.32)	0.4	1.3 (1.2–1.5)
Bornholm	1.0 (0.7–1.3)	0.92 (0.48–1.77)	0.18	2.3 (1.7–3.3)
Sonderjylland	3.3 (2.8–3.8)	1.04 (0.80–1.36)	0.5	3.7 (3.2–4.2)
Fyn	3.6 (3.3–4.0)	1.25 (1.03–1.52)	0.46	1.6 (1.4–1.8)
Viborg	3.5 (3.2–3.9)	1.27 (1.07–1.52)	0.44	2.0 (1.8–2.3)
Storstrom	1.4 (1.2–1.7)	1.33 (0.95–1.85)	0.32	2.0 (1.6–2.5)
Ribe	2.0 (1.6–2.5)	1.66 (1.14–2.41)	0.43	2.8 (2.3–3.5)
Ringkobing	3.5 (3.2–3.9)	0.92 (0.77–1.10)	0.48	2.0 (1.8–2.2)
Arhus	4.9 (4.5–5.4)	1.09 (0.91–1.30)	0.48	1.7 (1.5–1.9)
Roskilde	1.1 (0.7–1.7)	1.65 (0.75–3.75)	0.42	1.7 (1.0–3.2)
Vejle	2.9 (2.5–3.3)	1.34 (1.05–1.70)	0.48	2.2 (1.9–2.5)
Vestsjalland	1.8 (1.5–2.1)	1.01 (0.76–1.34)	0.3	1.8 (1.4–2.1)
Frederiksborg	1.5 (0.8–2.3)	0.87 (0.34–2.26)	0.26	1.6 (0.8–3.2)

[1] 95% Bayesian credible intervals.
[2] had over 60 pigs tested in 2003 (equates to an annual slaughter of greater than 2000 finishers).
[3] unadjusted farm-level prevalence.
[4] *Interpretation*: In Nordjylland the odds of a pig being seropositive was increased by a factor of 1.15 (95% Bayesian credible interval 1.01-1.32) if the pig was from a large (greater than 2000 finishers produced annually) herd than if it was from a medium (between 200 and 2000 finishers produced annually) herd.

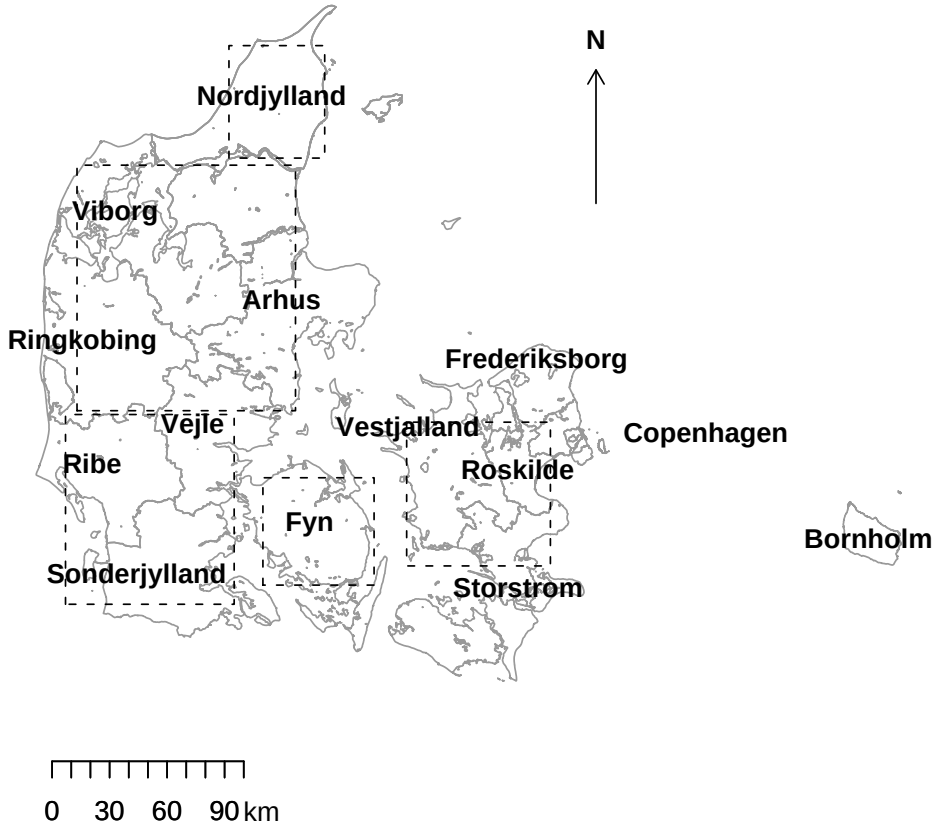


Figure 4.1: Map of Denmark showing location of counties and of the five areas used in the investigation of inhomogeneous K-function estimation.

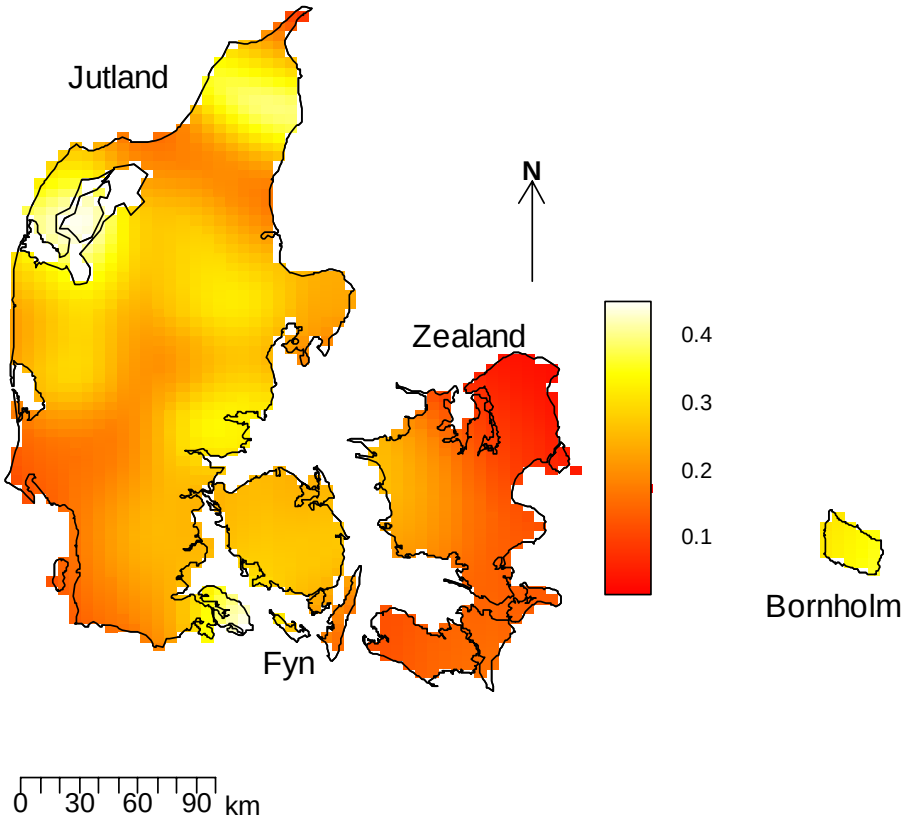


Figure 4.2: Kernel smoothed map showing variation in *Salmonella* meat-juice tested slaughter herd densities across Denmark in 2003. Herds that produced less than 200 pigs for slaughter annually were not tested. Units are farms per square kilometre. Data originate from the Danish swine *Salmonella* surveillance and control programme.

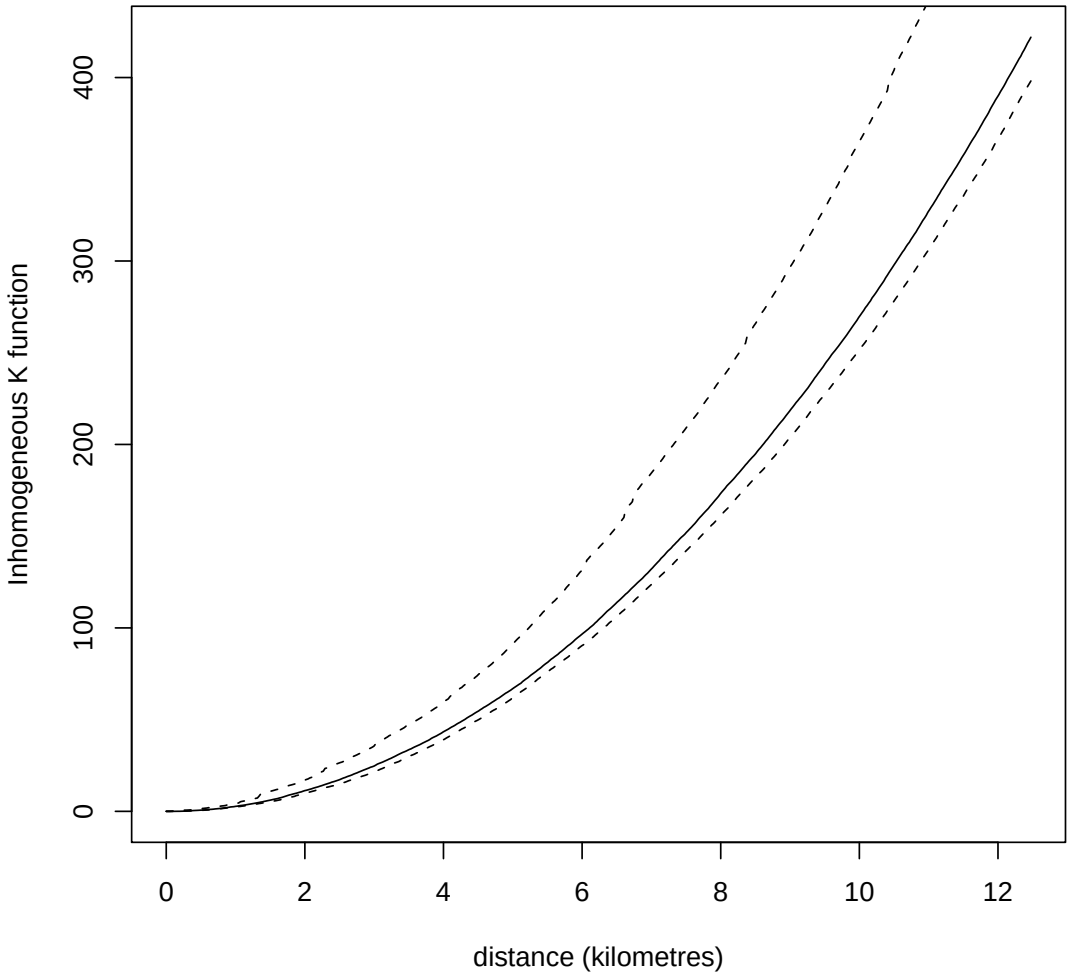


Figure 4.3: Inhomogeneous K-function for farms in north Jutland, 2003 (solid line). The dashed lines represent the inhomogeneous K-function of 100 simulated realisations of a inhomogeneous Poisson process. All points fall within the simulation envelope showing that the observed pattern of farms was not aggregated. Data originate from the Danish swine *Salmonella* surveillance and control programme.

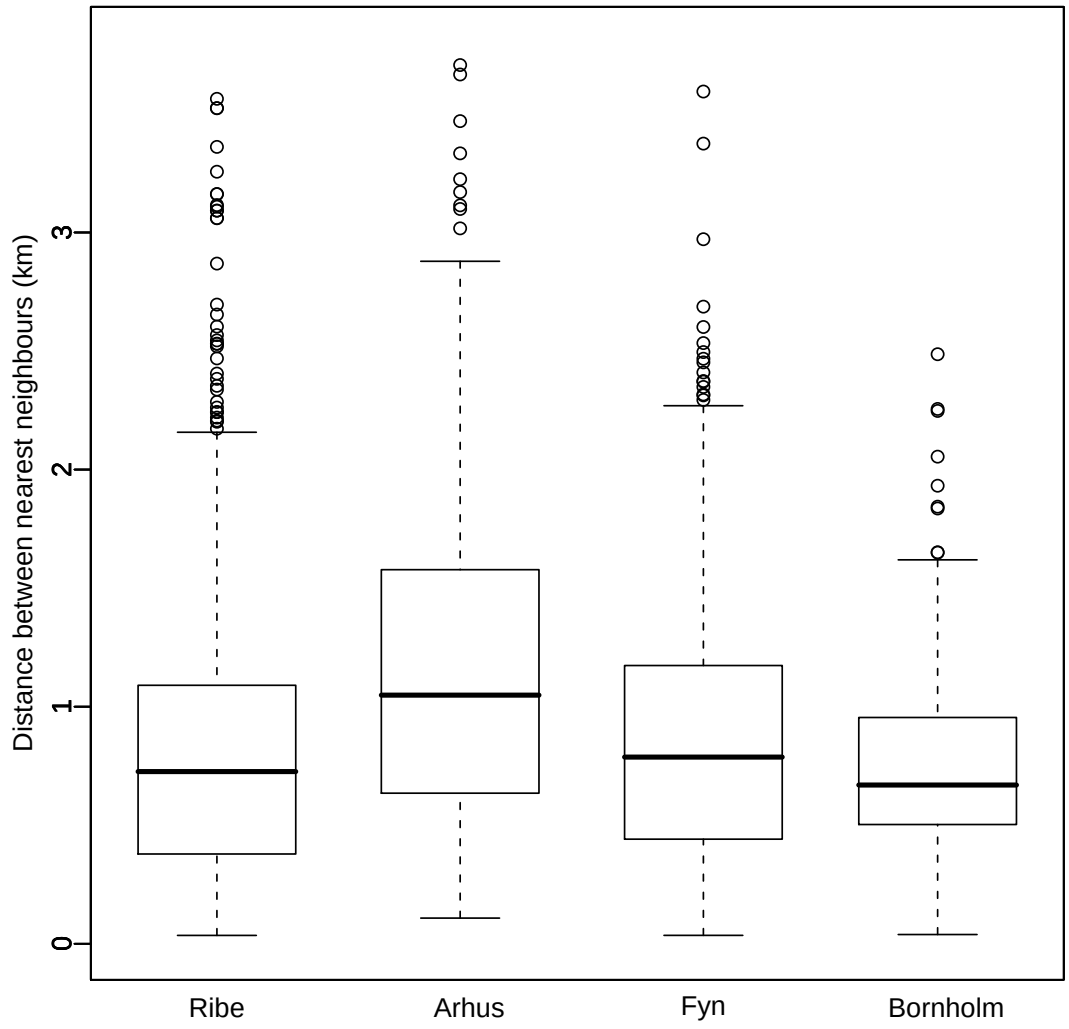


Figure 4.4: Boxplot of nearest neighbour distances for four Danish counties, 2003. Data originate from the Danish swine *Salmonella* surveillance and control programme.

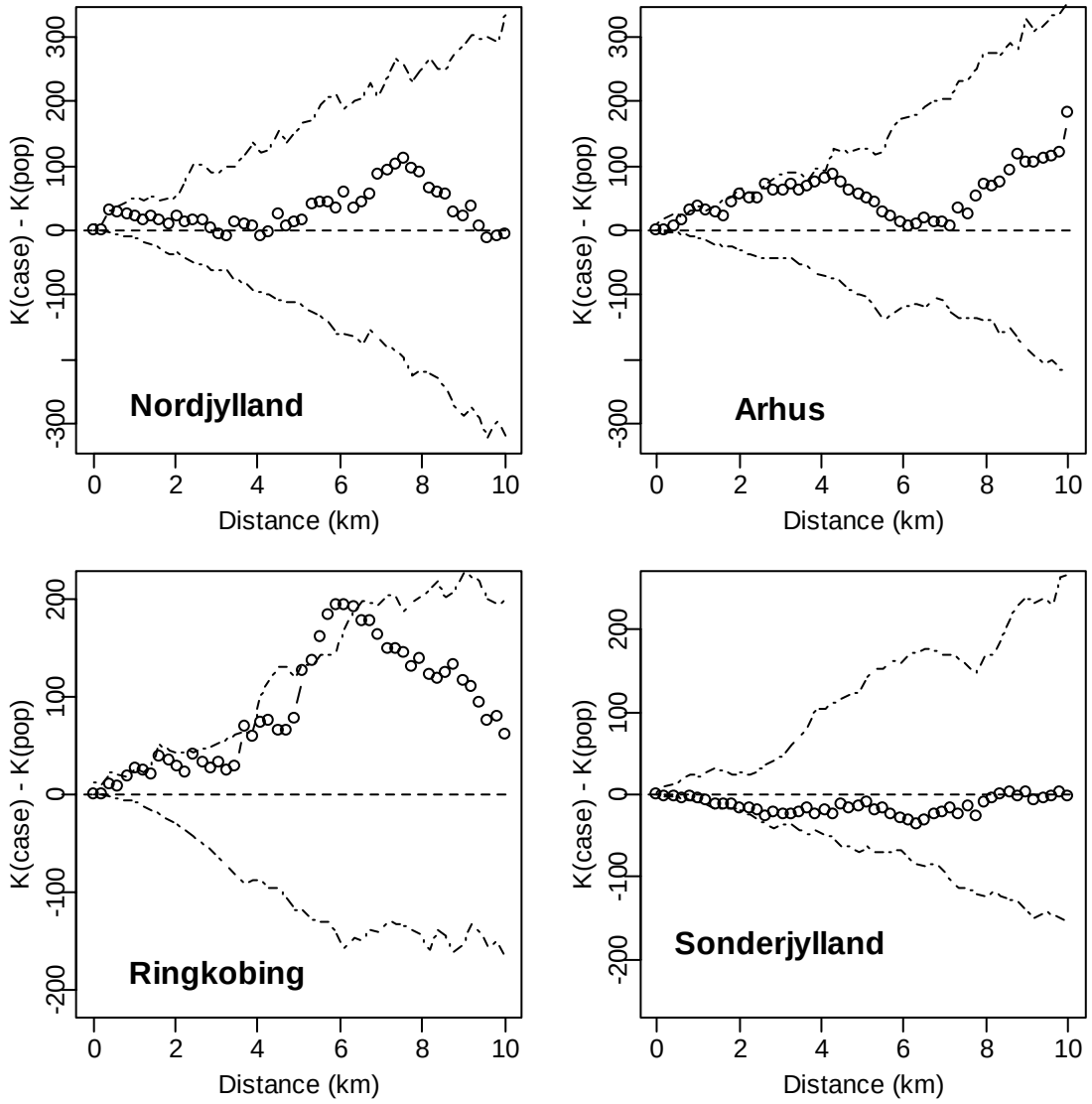


Figure 4.5: Observed difference K-function between case and population farms for Nordjylland, Arhus, Ringkobing and Sonderjylland. The circles represent the difference between the two K-functions and the dot-dashed lines the simulation envelope based on 99 random re-labellings of the cases and population. A farm was defined as a case if in 2003 the proportion of positive results was greater than 40%. Data originate from the Danish swine *Salmonella* surveillance and control programme.

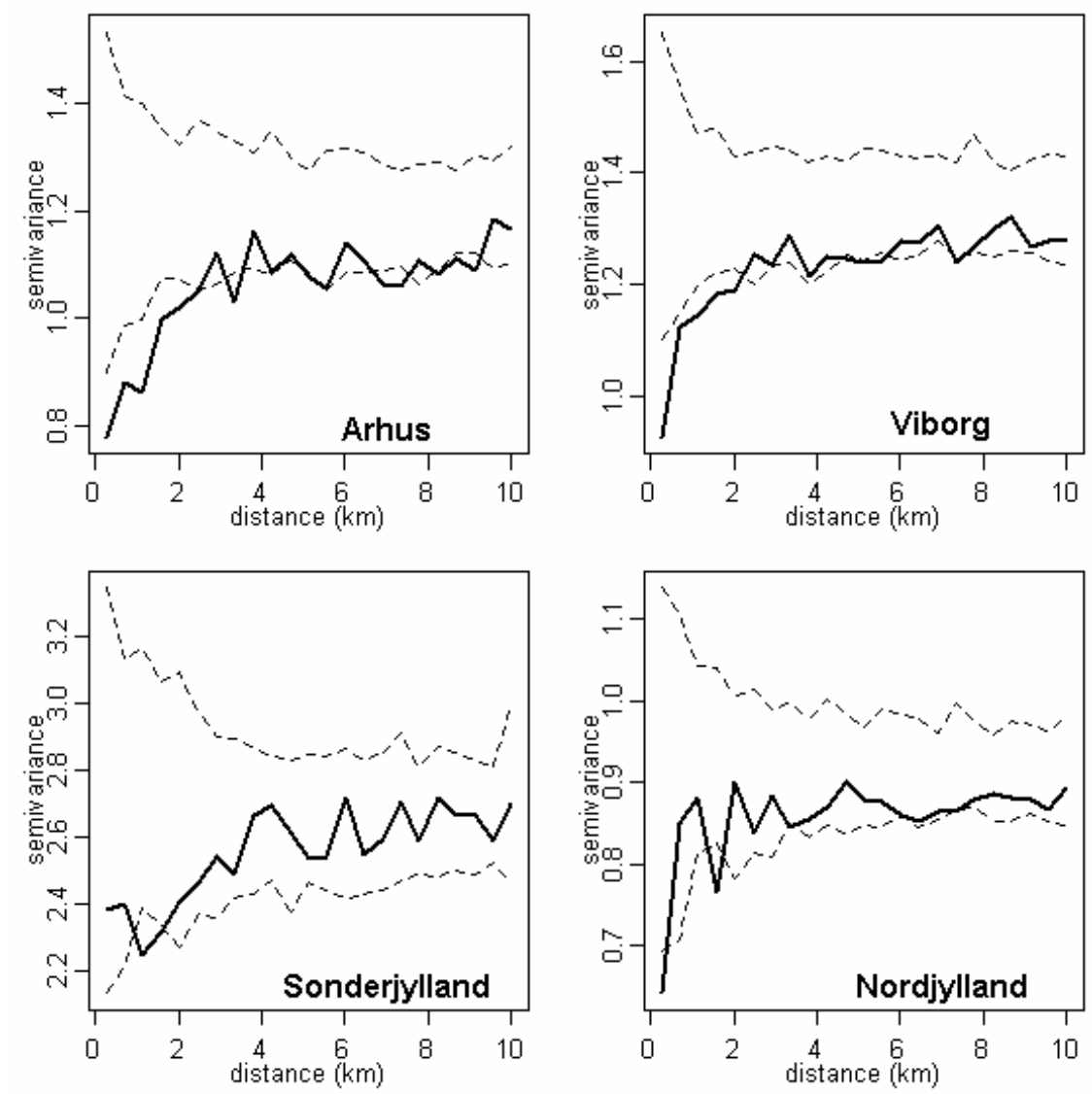


Figure 4.6: Spatial semivariograms fitted to the herd-size adjusted farm level random effects for the counties of Arhus, Viborg, Sonderjylland and Nordjylland. The solid line represents the semivariance and the dot-dash lines the simulation envelopes obtained by permutation of the data on the spatial locations. Data originate from the Danish swine *Salmonella* surveillance and control programme.

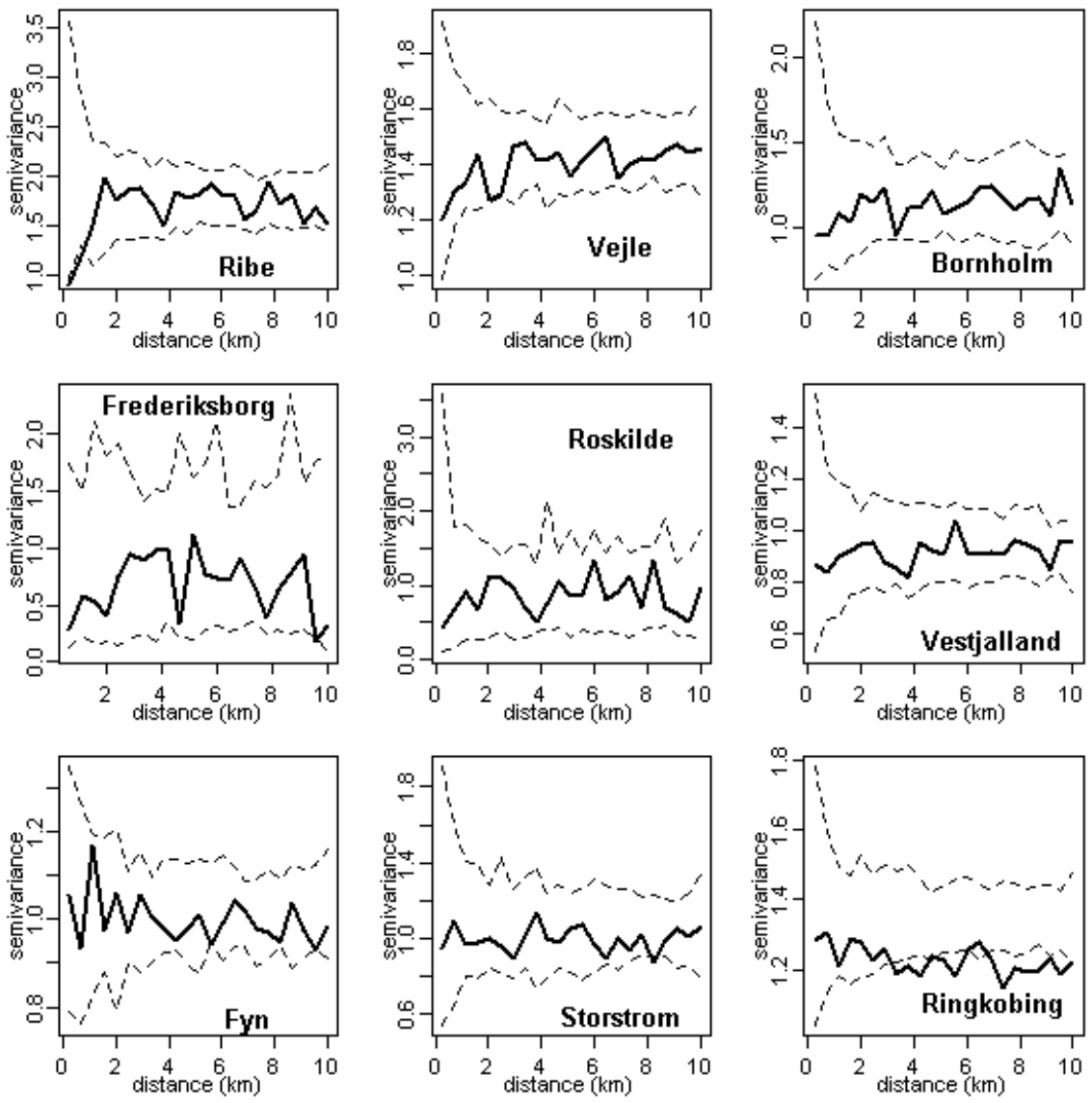


Figure 4.7: Spatial semivariograms fitted to the herd-size adjusted farm level random effects for remaining pig producing counties. The solid line represents the semivariance and the dot-dash lines the simulation envelopes obtained by permutation of the data on the spatial locations. Data originate from the Danish swine *Salmonella* surveillance and control programme.

4.5 Discussion

Slaughter-pig farm density showed large variation both at the country-wide and at the local level in Denmark in 2003. The areas of highest farm density are in Viborg and Nordjylland on the Jutland peninsula (0.47 farms per square kilometre); the lowest are on the island of Zealand. The distribution pattern of farms followed a random inhomogeneous Poisson process, and although farms had near neighbours they did not spatially aggregate. With regard to *Salmonella* seropositivity, we found consistent evidence for spatial dependency at distances of approximately four kilometres. The strength of the spatial dependency varied throughout the country being proportional to farm density. Our findings were in concordance with those that have reported short distance between farms (Berends et al. 1996), neighbouring an infected farm (Langvad et al. 2003) and pig density (Fedorka-Cray et al. 2000) as potential risk factors for *Salmonella* infection in pigs.

This local spatial dependency adds to the current knowledge of the epidemiology of subclinical *Salmonella* in Danish slaughter pig farms and can inform future strategies aimed at optimising the control program. For example, more intensive sampling of farms within a four kilometre radius of identified problem farms, such as those in level 2 or 3, on the Jutland peninsula is likely to capture more positive results, leading to interventions that may result in enhanced food safety. Likewise, we propose the concept of reduced sampling of farms that are near neighbours of '*Salmonella*-free' farms. '*Salmonella*-free' refers to farms enrolled in the 'risk based' scheme which has been running since July 2005. This scheme requires one sample per month to be taken from herds with a *Salmonella* index level of nil and a minimum ten negative meat-juice samples in the last six months. To date, 50% of herds meet these criteria. Our study has identified that when spatial dependency is present, such as on Jutland, there are farms that provide essentially redundant information that could potentially be eliminated from the surveillance programme. Spatial sampling optimization for groundwater monitoring has been achieved using the variogram (Ling et al. 2003, Cameron & Hunter 2002), and we propose it may be used to optimise sampling in the DSSCP. If spatial dependency is present in other disease programmes, both within and beyond Denmark, then these strategies could be applied to these programmes. For example, evidence for spatial dependency has been found between bulk milk tank titres for *Salmonella* Typhimurium in Texas dairy herds (Graham et al. 2005) and between cattle

herds in Denmark with regard to *Salmonella* Dublin infection (Ersbøll & Nielsen 2008).

Local farm density is a well recognised risk factor when investigating epidemics of animal disease (Gibbens et al. 2001, Mintiens et al. 2003, Le Menach et al. 2005, Sellers 2006). The density of neighbouring herds was associated with so-called 'neighbourhood infections' during the 1994 classical swine fever epidemic in Belgium (Mintiens et al. 2003), and 'local' spread accounted for 79% of means of spread in the first five months of the 2001 foot-and-mouth disease epidemic in Great Britain (Gibbens et al. 2001). High farm density implies that the distance between farms is short; in these examples a neighbourhood was an area of one kilometre radius around an infected farm and local meant within three kilometres of an infected place.

Although we are less familiar with farm density investigation in relation to a sub-clinical endemic infection such as *Salmonella* in Danish finisher pig herds, there are compelling reasons to investigate it. If *Salmonella* is not already present, or if a novel serovar is in circulation, then pig herds are at risk from its introduction through many routes, the two main ones being the introduction of infected pigs and contaminated feed (Lo Fo Wong et al. 2002). The latter is thought to be of minor importance as there are stringent controls on animal feed in Denmark; in 2005 the prevalence of *Salmonella* in animal feed was low. There is much support for the theory that the introduction of infected pigs is a likely source of *Salmonella* for Danish pig farms (Berends et al. 1996, Baggesen et al. 1996*b*, Stärk et al. 2002, Lo Fo Wong et al. 2004*a*). It is common farming practice to purchase stock from a geographically close supplier and this could lead to small-scale spatial dependency in the data. Denser farming areas probably offer more choice of supplier. Lo Fo Wong et al. (2004*a*) reported that the odds of seropositivity increased significantly if greater than three suppliers were used.

The other 'external' sources of *Salmonella* such as visitors (Funk et al. 2001), vermin (Steinbach & Kroell 1999, Fedorka-Cray et al. 2000), and sharing of contaminated equipment (Langvad et al. 2003) can also be farm density dependent. Rodents and flies have been found to carry *Salmonella* (Letellier et al. 1999, Barber et al. 2002), and the small distances between many of the Danish pig farms are well within the range of the brown rat (Endepols et al. 2003). In addition, airborne spread is possible at least over short experimental distances (Proux et al. 2001, Oliveira et al. 2006). Our findings of spatial dependency between farms with regard to *Salmonella* seropositivity, and aggregation of

Salmonella case farms over that of all farms at distances of up to four kilometres could be due to these locally-acting processes or the contagious nature of the disease. Temporal studies would help elucidate this.

The inhomogeneous K-function is a relatively new technique. It has been used to highlight significant differences in the spatial aggregation of vacuoles in mice brains infected with different transmissible spongiform encephalopathies (Webster et al. 2006). The use of the inhomogeneous K-function to summarise the spatial pattern of farms seems sensible. It allows for the spatial variation in intensity of the underlying point pattern which is likely to occur in animal production systems and is clearly seen in pig farm density in Denmark. By allowing for the non-uniform intensity of the spatial locations of farms, it permits hypothesis testing for aggregation. Our results support the hypothesis that the farm distribution pattern follows a random inhomogeneous Poisson process with no aggregation beyond that.

Even though our data set was effectively a census of Danish finisher swine herds in 2003, there was potential for selection, misclassification, and confounder bias in our study. Selection bias may have occurred when we excluded 405 of 10,571 (4%) of farms because coordinate information was unavailable. As our database was drawn from herds registered in March 2004 the 10,571 farms with available coordinate information were still in production then and were likely to be different from the 405 that no longer were. However this is likely to be of little importance as this group of farms represents only 4% of the total.

Further selection bias may have occurred in selecting the five large areas for the inhomogeneous K-function analysis. These were approximately square and excluded some areas of pig farms (notably Bornholm) and restricted the sites for consideration to those on large land masses. Nonetheless we believe the coverage of farms within the five areas was suitably representative of all pig farms tested in 2003; 82% were included and the case incidence risk (3%) was the same as that for all farms.

The use of the farm house locations over that of the actual polygonal boundaries of the farm may potentially lead to an over-estimation of the distance between farms (misclassification bias). This would be of much significance in extensive sheep or beef cattle farming systems were farm sizes are large. However it is likely to be of little consequence in intensive production systems such as the Danish pig farms we are investigating here.

The adjustment for herd-size in the geostatistical model was made as a number of earlier Danish studies (Baggesen et al. 1996*b*, Carstensen & Christensen 1998) and a recent Canadian one (Farzan et al. 2006) have reported large herd size as a risk factor for increased seropositivity in slaughter pig herds. However, later studies in Denmark (Stege et al. 2001) and Europe-wide (Lo Fo Wong et al. 2004*a*) showed no association, and a Dutch study (van der Wolf et al. 2001) showed that large herd size was protective. In this study, the effect of herd size was not investigated *per se*, but adjusting for herd size was undertaken in the context of its effect on spatial dependency. The odds ratios reported suggest increased risk in some counties as herd size increases. The reason for this may be that there are local practices, such as more movements of pigs between farms or higher within-farm pig density, which make large herd size more of a risk for increased *Salmonella* seroprevalence in these counties. Nevertheless, these results must be interpreted with caution as the effect of herd size is likely confounded by other covariates, such as feeding and biosecurity practices that we have no information on. The next chapter incorporates more covariates for a 1995 subset of the data to address this issue.

Distance can be defined in different ways; Euclidean, time of travelling, or in terms of social networks (Haining 2003). Ideally, all three definitions should be considered in the spatial epidemiological investigations and we should not constrain 'locality' to only imply spatial proximity. This study focussed on Euclidean distance between farm houses, but future studies in relation to social networks would appear to be a logical next step. This could be particularly helpful in tracing the dissemination of infected pigs.

We have outlined an approach to combine geo-referenced farm location information and routinely collected control programme data using techniques from spatial point pattern and geostatistical analysis. This has extended the current knowledge of the epidemiology of sub-clinical *Salmonella* in Danish slaughter pig farms. Furthermore we have demonstrated how our approach has the potential to optimise sampling strategies while maintaining consumer confidence in food safety. These techniques could be readily applied to data from other programmes in different countries.

4.6 Acknowledgement

Thanks to Adrian Baddeley from the University of Western Australia for his assistance with the inhomogeneous K-function analyses. And to Jonathan Marshall from Massey University for his assistance with the spatially adaptive smoothing.

CHAPTER 5

Towards incorporating spatial risk analysis for *Salmonella* seropositivity into the Danish swine *Salmonella* surveillance and control programme

Benschop, J., Stevenson, M., Dahl, J., French, N. (2008) Towards incorporating spatial risk analysis for *Salmonella* seropositivity into the Danish swine surveillance programme. *Preventive Veterinary Medicine* **83**:347-359

5.1 Abstract

An increased incidence of pork-related human salmonellosis in Denmark led to the development of a national control programme for *Salmonella* in Danish swine herds in 1993. The aim of the programme has been met and now the issue of cost-effectiveness is receiving greater attention. An appropriate way to address this is to bring a risk-based focus to the programme. We describe a practical approach to risk-based surveillance through spatial risk assessment using serological and questionnaire data from 2280 herds in 1995. A mixed effects logistic regression model was fitted and both first- and second-order spatial properties of the random effects were investigated. We identified wet-feeding (OR: 0.64; 95% CI: 0.54-0.75) and SPF health status (OR: 0.65; 95% CI: 0.52-0.81) as protective factors for *Salmonella* seropositivity. Purchasing feed (OR: 1.81; 95% CI: 1.61-2.04) was a risk factor. The west of the study area generally, and the north of Jutland in particular, experienced the greatest disease risk after controlling for the covariates. There was some evidence for spatial dependency between farms at distances of 6 km (95% CI: 2-35 km)

on the Jutland peninsula. We conclude that when farm location details are analysed in conjunction with routinely recorded surveillance information (such as that collected by the Danish swine *Salmonella* surveillance and control programme) and targeted industry surveys (such as those conducted by slaughterhouse co-operatives), our knowledge of the behaviour of disease in animal populations is enhanced providing a more informed framework for designing efficient, risk-based surveillance strategies.

5.2 Introduction

The Danish swine *Salmonella* surveillance and control programme was initiated in 1993 by the Danish Ministry of Food, Agriculture and Fisheries in response to increasing numbers of human cases of salmonellosis attributable to pork consumption (Alban et al. 2002, Mousing et al. 1997). This on-going programme is based on the serological surveillance of all herds that produce more than 200 pigs per annum and their subsequent assignment into one of three levels of a Serological *Salmonella* Index (SSI). SSI levels 1 to 3 represent low, medium and high levels of *Salmonella* in the herds, respectively, with level 2 and 3 herds paying penalties and undergoing on-farm investigations. The programme's objective is to lower the prevalence of *Salmonella* so that domestically produced pork is no longer an important source of salmonellosis in humans (Mousing et al. 1997). Since 2001 the prevalence of *Salmonella* in Danish pork (monitored at the slaughterhouse) has reduced from 1.5% to 1% of carcass swab samples taken. The number of cases of salmonellosis in humans in Denmark attributable to pork consumption decreased from 1444 in 1993 to 142 in 2004 (Nielsen et al. 2001, Ministry of Family and Consumer Affairs 2005). An in-depth discussion of the programme is provided by Hald et al. (2005).

Since 2004, the cost-effectiveness of *Salmonella* surveillance has received greater attention, with both industry and regulatory authorities wanting to achieve the greatest reduction in *Salmonella* for their financial investment (Ministry of Family and Consumer Affairs 2005). A simulation study using data from 2001 to 2002 of the Danish programme found that the number of samples taken from low prevalence herds could be reduced without jeopardising food safety (Enoe et al. 2003). This led to the development of a risk-based approach to *Salmonella* surveillance implemented in mid-2005 (Ministry of Family and Consumer Affairs 2005). Herds with no positive samples from the previous

three months testing are sampled once per month instead of the previous random sampling based on herd size. For those herds under the risk-based scheme this represents a four-fold reduction in the number of samples taken as the prior average number of samples per herd per month was 4.3 (Enoe et al. 2003).

A recent discussion paper on risk-based surveillance in veterinary medicine and veterinary public health states that although the risk-based concept is generally accepted, its practical basis is undeveloped (Stärk et al. 2006). In this paper we describe a practical approach to risk-based surveillance in the use of spatial risk assessment to direct surveillance activities. We use Danish *Salmonella* surveillance data from 1st October 1995 to 31st December 1995. A previous statistical analysis of a subset of these surveillance data reported a pig-level seroprevalence of 4.3% and found that the risk of seropositivity increased with increasing herd size, with dry- versus wet-feeding, and with purchased versus home-mixed feed (Dahl 1997). We refine and extend this analysis to investigate unaccounted for variation in *Salmonella* risk and propose that our findings have the potential to inform surveillance strategy. Many European studies have investigated risk factors such as herd size, feed type, and hygiene, for salmonellosis in pig herds (Dahl 1997, van der Wolf et al. 2001, Belœil et al. 2004, Nollet et al. 2004, Lo Fo Wong et al. 2004*a*). We and others have reported on the variation in spatial patterns of seroprevalence of *Salmonella* in Denmark (Mousing et al. 1997, Carstensen & Christensen 1998, Benschop et al. 2008*a*). However, disease risk is affected by factors which themselves exhibit spatial variation, such as vector or feed abundance. Thus, spatial variation is generally of interest if it persists after accounting for known risk factors.

Our objectives are to target limited investigative resources using a risk-based approach; firstly, at farms with risk factors associated with the presence of disease, and secondly, at farms in those areas where there is an excess of *Salmonella* risk beyond that explained by those identified risk factors.

5.3 Materials and methods

5.3.1 Data description and handling

The data comprised 45,103 individual pig meat-juice serology results obtained from 1st October 1995 until 31st December 1995 from 3784 herds in the Danish swine *Salmonella* surveillance and control programme (Mousing et al. 1997). These data were linked, through a unique herd identifier, to responses from a processor co-operatives questionnaire collected during the same time period. The questionnaire was mailed to all suppliers (approximately 10,000) of two of the three processor co-operatives active at the time. Its intent was to gather information on suppliers for an industry database. Those herds in the west of Jutland that supplied the third slaughterhouse co-operative (approximately 6000) were excluded (Figure 5.1).

Questionnaire responses provided details of herd demography and details of herd management, including the type of feed offered, specific pathogen free health status, floor type, feed source, and whether or not pigs had regular access to straw throughout the growing period. The original three questionnaire categories for feed type (dry, liquid, or both) and feed supply (purchased, home-mixed, or both) were collapsed into a two-level response as the both category for each represented less than 2% of farms. The original seven questionnaire categories for floor type were combined into a two-level response as there was little variability among the seven categories. The remaining categorical variables (health status and access to straw) were left in their original format and herd size was analysed as a continuous variable. Table 5.1 shows the number of herds in each level of the categorical variables.

5.3.2 Risk factor analysis

A complete case analysis comprising a subset of 2280 of the 3784 farms was performed. This subset of 2280 farms contributed 37,825 of the 45,103 serology results (83%).

A binary response variable at the pig level was created from the serology data. If a meat-juice serology result had an optical density percentage of greater than 20 the sample was deemed positive; this is the cut-off currently used in the Danish swine *Salmonella* surveillance and control programme (Alban et al. 2002).

The continuous variable herd size was checked to see if it was linear in its log odds (Hosmer & Lemeshow 1989). Polynomials of herd size and biologically plausible two-way interaction terms between the main-effect variables were considered for inclusion.

We developed a logistic regression model using a Bayesian approach to identify factors associated with a meat-juice sample being *Salmonella* positive.

$$\log(p_{ij}/1 - p_{ij}) = \beta_0 + \sum_{k=1}^{6} \beta_k x_{ik} + U_i \tag{5.1}$$

In Equation 5.1 the logit of the observed probability of the *j*th pig from the *i*th farm being seropositive, p_{ij}, was modelled as a function of k farm-level explanatory variables and a random effect term, U_i, which was normally distributed with a mean of zero and variance σ^2.

Markov chain Monte Carlo methods were applied to the data to simulate values from the joint conditional distributions of the unknown quantities using WinBUGS version 1.4.1 (Gilks et al. 1994). Initially, we stipulated data augmentation priors for the intercept term and for the following covariates: feed type, feed supply, and health status (Table 5.2). These were based on subjective information about the likelihood ascribed to various combinations of covariate values (Congdon 2001). For example, we assumed the proportion of positive pigs in a typical herd to be most likely 0.10 with 95% certainty that it would be less than 0.25. This information was then expressed as a conjugate prior beta density (Congdon 2001). In this context, the term typical means a herd with mean herd size and all categorical covariates set to the reference category: dry feed type; home-mixed feed supply; conventional health status, with access to straw, and slatted flooring. Relatively non-informative normally distributed priors centred at zero and with a variance of 1 were used for the following covariates: herd size, access to straw and slatted flooring. A non-informative gamma distribution (shape parameter of 0.1 and inverse scale parameter of 0.001) was used for the variance of the random farm effect term.

Sensitivity to the covariate priors was evaluated by re-running the models with non-informative normally distributed priors centred at zero and with a large variance (10,000). Three chains were run and convergence was judged to have occurred on the basis of visual inspection of time series and Gelman-Rubin plots (Toft et al. 2007). The length of the chain was determined by running sufficient iterations to ensure the Monte Carlo standard

errors for each parameter were less than 5% of the posterior standard deviation. A total of 40,000 iterations were run with a burn in of 5000 iterations.

5.3.3 Spatial analysis

We investigated both first- and second-order spatial patterns in the data by evaluating if farms with similar values of random effects tended to be closer together in space. The term first-order relates to the large-scale trend in the pattern, the variation in the mean value of a process in space. Second-order patterns are local or small-scale effects that result from the spatial correlation of the process.

Of the 2280 complete-case farms 1820 (80%) had recorded easting and northing coordinates. The remaining 460 farms had easting and northing coordinates randomly drawn from within the boundaries of their respective communes. Communes were the smallest spatial area available to the authors and in 1995 there were 255 pig-producing communes ranging in area from 20 to 58 km^2. This was an appropriate means for dealing with this type of missing data for two reasons: the size of those communes that produce pigs are small relative to the entire land area of Denmark (43,000 km^2); and we use these within-commune randomly generated coordinates to make inferences only at the broad (first-order) spatial scale.

An edge-corrected Gaussian kernel estimate of the intensity function of the random farm effects was produced to visualise the first-order spatial pattern. Kernel estimation is a mathematical function that can be applied to point data (such as farm locations) to smooth it (Bowman & Azzalini 1997). It is particularly useful when the farm density is so high it is not possible to obtain a visual impression of the point pattern. The estimate was produced by weighting the point locations of farms by the random farm effects (Baddeley & Turner 2005). The initial choice of bandwidth of 21 km was made using the normal optimal method (Bowman & Azzalini 1997). However, for the purpose of investigating the spatial variation the pattern appeared over-smoothed, so a bandwidth of 7 km was iteratively selected. To test the sensitivity of the result to farms given randomly selected coordinates, two kernel estimation surfaces were produced: (1) for all 2280 complete-case farms and (2) for those 1820 farms with recorded coordinate information. We tested the null hypothesis that the proportion of pigs positive was the same for farms both with

(n = 460) and without (n = 1820) missing coordinate information (Newcombe 1998).

The second-order spatial pattern was evaluated by plotting semivariograms of the random farm effects produced for the subset of 1820 farms that had recorded easting and northing coordinates (Isaaks & Srivastava 1989). Semivariograms were produced for the four geographically distinct regions of Denmark: the Jutland peninsula, and the islands of Funen, Zealand, and Bornholm (Figure 5.1). The semivariogram plots the semivariance as a function of the distance between pairs of farms (Isaaks & Srivastava 1989). If farms with more similar random effects were closer together in space than those with less similar random effects, we would expect that the semivariance would increase as a function of distance to reach an asymptote. This indicates the range of influence, the distance at which random farm effects are no longer correlated. To estimate this distance, we visually appraised the semivariograms and produced envelopes for each variograms based on 999 Monte Carlo permutations of the data. Here, the random farm effects were randomly allocated to each farm location and, as permuted data should not exhibit spatial dependency, any points lying outside these simulation envelopes indicated significant spatial auto-correlation. Directional semivariograms at angles of 0, 45, 90, and 135° (with a tolerance of $\pm 22.5^{\circ}$) were plotted to determine if the spatial distribution of random farm effects varied with direction.

The within-farm or locally erratic component of the total variance was estimated from the semivariograms by comparing the nugget variance to the total variance (Cambardella et al. 1994). The nugget variance is the point at which an extrapolated fitted line would cross the vertical axis and is a measure of purely random variation (Oliver & Kharyat 1999). If the proportion of nugget to total variance is high, there is evidence for a strong within-farm component to the variance.

Having identified spatial correlation in the variogram from the Jutland peninsula, we further analysed data from that region. We fitted a generalised linear spatial model to the Jutland data by extending the model shown in Equation 5.1 with the addition of a term $S(yi)$ (Diggle et al. 2002*b*). $S(yi)$ is a zero mean normal process with a variance ν and a Matérn correlation function $\rho(d, \Phi, \kappa)$ (Matérn 1960). Here Φ is the range parameter, which controls the rate at which correlation in random farm effects approach zero with increasing distance d, and κ controls the smoothness of the decrease in auto-correlation as a function of distance. This model allows formal estimation of the range of spatial

correlation and a measure of the uncertainty about it.

The computer programme R (R Development Core Team 2007) was used for data handling and management. Spatial analyses were performed in R using the contributed packages spatstat (Baddeley & Turner 2005) and geoRglm (Christensen & Ribeiro Jr. 2002).

5.4 Results

The response rate to the mailed questionnaire was approximately 37%. For complete case data the median herd size was 400 pen placers for slaughter pigs (IQR: 250–700), the median number of pigs sampled per herd was 48 (IQR: 35–63), and the proportion of pigs positive was 10.1%. Table 5.1 shows the distribution of categorical variables for complete case farms stratified by positivity.

The variable herd size was linear in its log odds and no interaction terms were included in the model. The results from the complete case analysis (n = 2280) are shown in Table 5.3. The practice of purchasing feed increased the odds of a pig being *Salmonella* positive by a factor of 1.81 (95% CI 1.61–2.04). SPF herd status and the practice of wet-feeding decreased the odds of a pig being *Salmonella* positive by a factor of 0.65 (95% CI 0.52-0.81) and 0.64 (95% CI 0.54-0.75) respectively. There was no difference in the monitored parameters when the covariate priors were varied.

The edge-corrected kernel estimate of the intensity function of the random farm effects is shown in Figure 5.2. Throughout the whole study area there were more positive random effects associated with farms on the Jutland peninsula and on Funen (in the west), compared with those on Zealand and Bornholm (in the east). The upper fifth (solid line) and twenty-fifth (dashed line) percentiles have been superimposed on the map to highlight areas with the most positivity. The largest area was in the north of Jutland and smaller pockets were found in centre and south east of the peninsula, and in the east of Funen. There was no visible difference between the first-order pattern between the two kernel estimation surfaces (including and excluding farms with randomly generated coordinates). The proportion of pigs positive was 9.5% for farms with (n = 460) missing coordinates and 10.2% for farms with recorded coordinates (n = 1820). The 95% confidence interval for the difference was 0.2-1.1% and the associated p-value was 0.01. This indicates that the null hypothesis of no difference between these two groups could be rejected.

Semivariograms of the random farm effects are shown in Figure 5.3. For all regions, all points are within the simulation envelopes indicating no significant spatial dependency, and the plots for Funen, Zealand and Bornholm are essentially flat. However, the plot for the Jutland peninsula clearly shows an upward trend in the variogram indicating some spatial dependency in the data for this region at distances to approximately 4-8 km. The within-farm component of the variance accounted for approximately 95% (on the islands) to 50% (on Jutland) of the total variance. The generalised linear spatial model fitted to Jutland data estimated the range parameter, Φ, at 6 km with a 95% credible interval (CI) of 2-35 km.

Table 5.1: Categorical variables, number of herds and missing cases, and pig-level seropositivity for 3784 Danish finisher pig herds from 1st October to 31st December 1995. Data originate from the Danish swine *Salmonella* surveillance and control programme.

Variable	Level	Number of herds (%)	Number pigs positive[1,2] (%)	Number pigs negative[2] (%)
Feed type	dry	2250 (59)	3106 (11)	25,253 (89)
	wet or both	448 (12)	693 (7)	8773 (93)
	missing	1086 (29)		
Feed supply	purchased	1339 (35)	1807 (14)	10,444 (86)
	home mixed or both	2175 (57)	1992 (8)	23,582 (92)
	missing	270 (7)		
Herd health status	conventional	2336 (62)	2432 (10)	21,281 (90)
	SPF (with *Mycoplasma*)	788 (21)	1160 (11)	9794 (89)
	SPF	239 (6)	207 (7)	2951(93)
	missing	421 (11)		
Straw available	yes	2988 (79)	2554 (10)	22,680 (90)
	no	796 (21)	1245 (10)	11,346 (90)
	missing	NA	NA	NA
Floor type	solid	1906 (50)	1070 (9)	10,398 (91)
	slatted or mixed	1763 (47)	2729 (10)	23,628 (90)
	missing	115 (3)		

[1] Danish-mix ELISA serology result with optical density percent of greater than 20.
[2] Pig data is from complete case data set only (2280 farms).

Table 5.2: Informed priors used for fixed effects in modelling factors associated with *Salmonella* seropositivity

Variable	Prior distribution
Intercept	beta(3.44, 22.99)
Feed type	beta(2.95, 26.96)
Feed supply	beta(5.04, 23.90)
Health status: SPF	beta(2.88, 36.70)
Health status: SPF (with *Mycoplasma*)	beta(2.45, 23.78)

Table 5.3: Factors associated with *Salmonella* seropositivity in 37,825 meat-juice ELISA results, taken from 2280 Danish finisher pig herds (complete-cases), from 1st October to 31st December 1995. Data originate from the Danish swine *Salmonella* surveillance and control programme.

Variable	Level	Posterior Mean	Posterior SD	MC error	OR(95% CI)
Herd size[1]	continuous	0.03	0.01	<0.01	1.04 (1.02–1.05)
Feed type	wet or mixed	-0.45	0.08	<0.01	0.64 (0.54–0.75)[2]
	dry	reference			
Feed supply	purchased	0.59	0.06	<0.01	1.81 (1.61–2.04)
	home mixed or both	reference			
Health status	SPF	-0.43	0.11	<0.01	0.65 (0.52–0.81)
	conventional	reference			

Model Statistics: Intercept, -2.89; DIC, 9797.91.
SD: Standard deviation; CI: Bayesian credible interval; MC error: Monte Carlo standard error of the posterior mean; OR: odds ratio
[1] Number of pen places for finishers (rescaled by subtracting the mean, then dividing by 100).
[2] *Interpretation*: Once adjusted for herd size, feed supply, and health status, a pig on a farm using wet-feeding had 0.64 times the odds of being *Salmonella* positive compared with a pig on a dry-feeding farm (95% CI: 0.54-0.75).

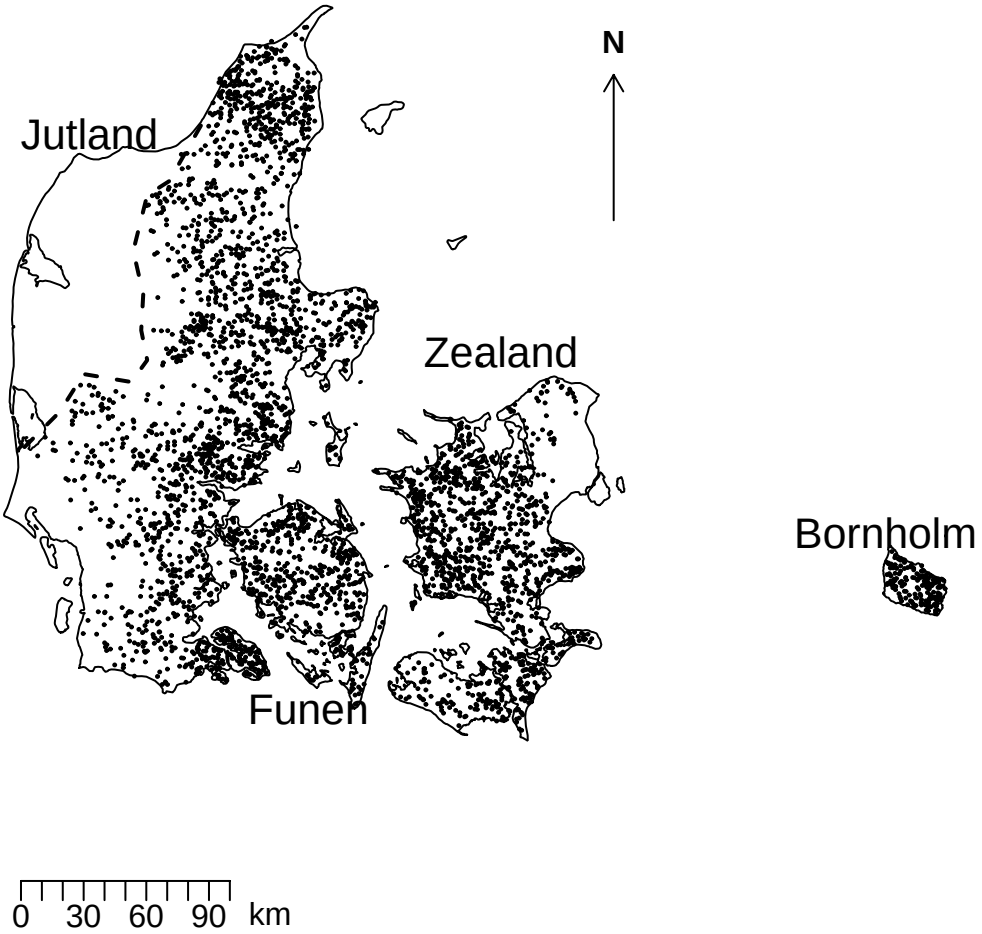


Figure 5.1: Map of Denmark showing Jutland peninsula and main islands. Study herd locations are shown as points. Herds in the area to the west of the dashed line were not surveyed.

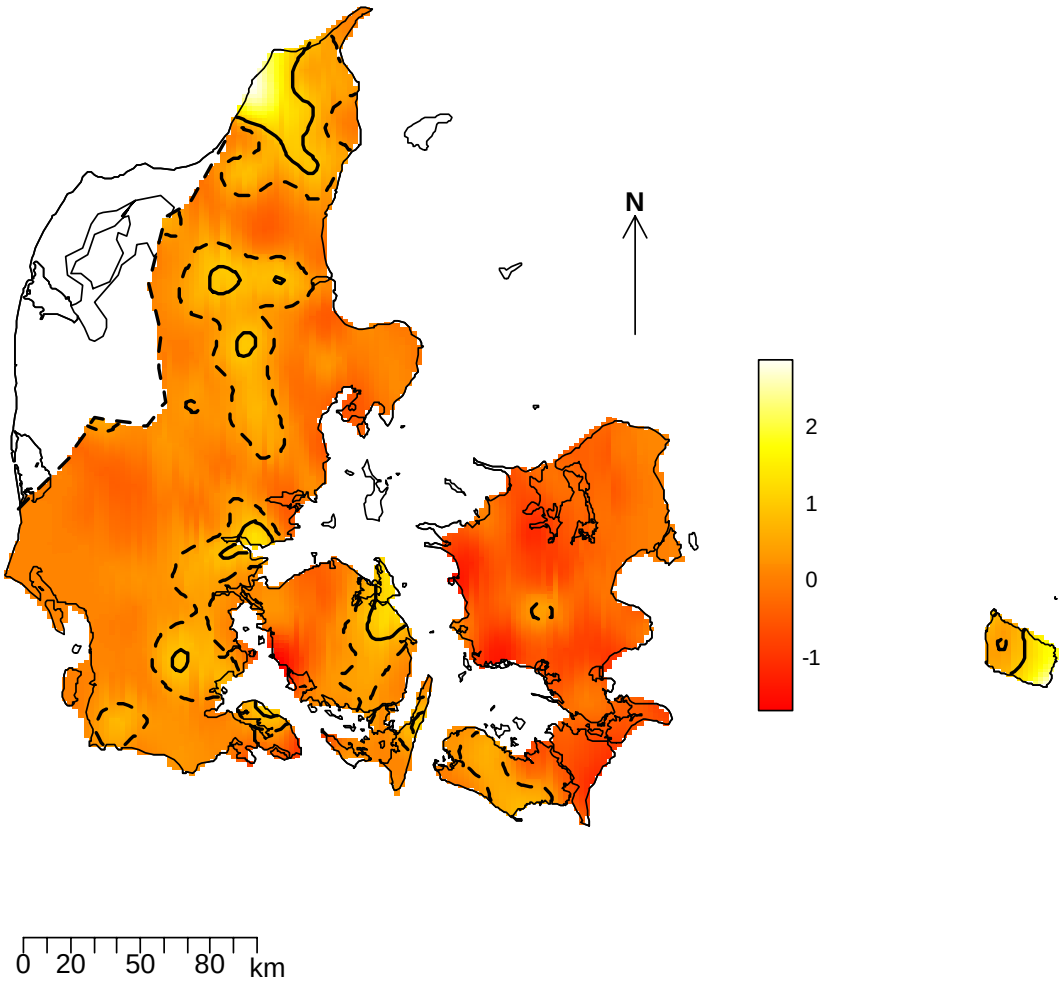


Figure 5.2: Edge-corrected kernel estimate of the intensity function of the random farm effects with the upper fifth percentile (solid line) and twenty-fifth percentiles (dashed line) superimposed. Herds in areas in the west of Jutland were not surveyed. Data originate from the Danish swine *Salmonella* surveillance and control programme.

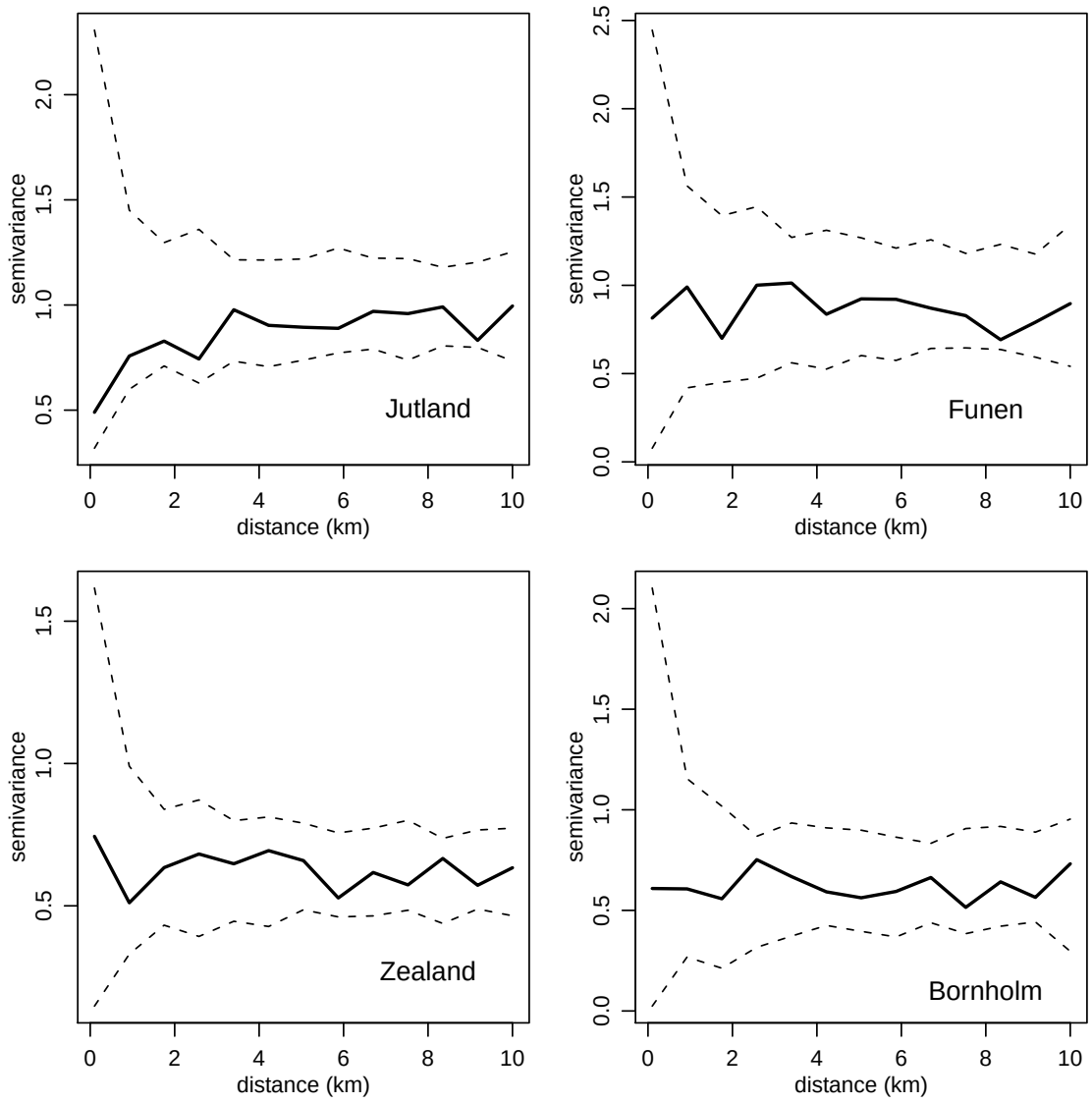


Figure 5.3: Spatial semivariograms fitted to random farm effects from farms in Jutland, Funen, Zealand and Bornholm. The dotted lines represent the Monte Carlo simulation envelopes. Data originate from the Danish swine *Salmonella* surveillance and control programme.

5.5 Discussion

Our analyses show that the use of purchased feed was positively associated with *Salmonella* seropositivity in this population. SPF herd status and wet-feeding were negatively associated with *Salmonella* seropositivity. The random farm effects showed a strong first-order spatial pattern with identified areas of unexplained risk for *Salmonella* seropositivity in the north of Jutland, and in the west of the study area. Variogram analyses of the farm random effects showed evidence of local spatial dependency between farms on the Jutland peninsula but not between those on the islands. However, throughout the whole study area there was a strong within-farm component to the variance, indicating unexplained variation beyond that accounted for by the risk factors and/or spatial dependency.

A previous analysis of a subset of these data identified similar risks for *Salmonella* seropositivity to those reported here (Dahl 1997). Many other studies have identified a positive association between the practice of dry-feeding (van der Wolf et al. 2001, Belœil et al. 2004, Farzan et al. 2006) and *Salmonella* risk.

It is important to explain further our finding with regard to the practice of purchasing feed being positively associated with *Salmonella* seropositivity. In Denmark pig feed has been heat treated and monitored for *Salmonella* since 1993 (Christensen 2003). The prevalence of *Salmonella* in feed has been low (Ministry of Family and Consumer Affairs 2006) and the main serovar found in Danish pigs (*Salmonella typhimurium*) has not been found in Danish produced feed (Baggesen et al. 1996*a*). However, purchased feed is almost exclusively produced from finely ground meal which is sold as pellets, and there is much evidence that it is these physical characteristics of the feed that constitute the *Salmonella* risk (Bach Knudsen 2001, Lo Fo Wong et al. 2004*a*, Mikkelsen et al. 2004, Hedemann et al. 2005) rather than the fact that feed is purchased *per se*.

We have used industry-derived questionnaire results and routinely collected national control programme data to quantify farm-level risk factors for disease and we propose that the methodology reported here might be used to further refine the conduct of surveillance for salmonellosis in the Danish pig industry. For example, farms with management practices associated with a higher than expected prevalence of *Salmonella* (such as dry-feeding) might be preferentially sampled over those with management practices associated with a lower prevalence of disease (such as wet-feeding). However, the operation of such a sam-

pling strategy will require more data than is currently routinely collected. An analysis of the potential benefits gained through such targeting would need to be weighed up against the costs of both implementing the sampling strategy, and gathering the necessary data.

The first-order spatial pattern evident in the random farm effects might also inform sampling. Data on the location of farms is routinely collected so the operational cost of this sampling strategy would be entirely in the implementation. The positive-sign random effects evident in the north of Jutland, and in pockets on the peninsula and in Funen, indicate disease risk that was present after controlling for the farm-level covariates included in the model. This is suggestive of unmeasured influences at either the farm-level, such as biosecurity and hygiene (Belœil et al. 2004), and/or at a regional level, such as a common contaminated pig supply or the effect of a farm adviser. Identifying the reasons for this unaccounted for level of disease via investigative effort targeted specifically at the region of interest should allow for more efficient use of surveillance resources.

Our finding of spatial dependency with regard to *Salmonella* seropositivity at distances of 6 km (95% CI: 2–35 km) between farms on the Jutland peninsula has the potential to inform surveillance. We have demonstrated that the random effects associated with each study farm on Jutland are correlated to an extent determined by the distance between them. This may allow further targeting of investigative effort, for example, preferentially sampling of farms within a 6 km radius of an identified problem farm. Nevertheless, determining what constitutes a problem herd is not straightforward: herds categorised at SSI level 3 (representing high levels of *Salmonella* risk), are an obvious choice. However, due to the dynamic nature of the current system herds may only be in this category for as little as one month. Capturing this changing problem herd status will be challenging.

The strong within-farm component to the variance also has important ramifications for surveillance. This raises issues for our targeted surveillance proposal as there is variation unexplained by the risk factors and spatial dependency. Investigating what determines this variation over and above that we have already accounted for is an important area for further study. Possible other determinants of risk include both those associated with the introduction of *Salmonella*, such as the number of weaner suppliers, and those associated with the spread of *Salmonella* through a herd, such as pen construction (Lo Fo Wong et al. 2004*a*).

Our use of incomplete data from both industry and control programme sources has po-

tentially introduced bias into our results. In the first instance we were aware of selection bias as we had questionnaire information from suppliers of only two of the three Danish processor co-operatives active at the time. The area excluded from the study (the west of Jutland), includes a large number of pig producers. This reduced our effective eligible population from approximately 16,000 to 10,000 farms. Secondly, the response rate to the mailed questionnaire was 37%. Poor response rates to questionnaires might introduce bias if there are systematic differences between those that respond and those that do not.

Other sources of bias included those related to misclassification when we collapsed the number of categories for floor type, feed type, and feed source for the purposes of the analysis. Because this was non-differential, it would only bias our results towards the null (Sackett 1979).

Further selection bias occurred during our investigation into second-order effects. Because the exact distance between pairs of farms up to 10 km apart was required, we only included farms with exact coordinate information available. The 1820 complete-case farms for which we had this information were still in production in 2004 and were likely to be different from the 460 farms that were no longer in production. We found that the population of pigs from farms no longer in production were significantly different from those that were: they were from smaller herds and had a lower proportion of seropositivity.

We acknowledge that our use of only complete-cases of data is a limitation of this study as in most situations this will introduce bias (Donders et al. 2006). Even though producers were unaware that their questionnaire responses were to be linked to *Salmonella* serology in their herds, we have evidence from our analysis that the non-response bias was differential. Generally the population of pigs from farms with missing covariate information had a higher proportion of seropositivity than those where the covariate information was complete. However, there were exceptions, e.g. pigs from farms with missing health status information showed the opposite effect. It is, therefore, difficult to speculate as to the direction and the magnitude of the bias introduced by complete case analysis.

While we do not present the results in this chapter, we advise that we have performed sensitivity analyses by comparing our results using imputed data (all 3784 farms) with those from complete-case analysis and found no significant difference between the two. More information about this can be seen in Appendix A. A detailed explanation of the handling

of these missing data with an extension into more sophisticated imputation techniques is beyond the scope of this study but has potential for future work.

The use of targeted surveillance is extensive in the veterinary world as a tool to make best use of limited resources and there are many examples of on-going systems, such as fallen stock for bovine spongiform encephalopathy in Europe (Giovannini et al. 2005), previous infection history for tuberculosis in cattle in Australia (Radunz 2006) and importation history for brucellosis in cattle in the United Kingdom (Stack & Perrett 2005). A recent discussion paper on risk-based surveillance in veterinary medicine and veterinary public health states that although the risk-based concept is generally accepted, its practical basis is undeveloped (Stärk et al. 2006). We demonstrate a practical basis for risk-based surveillance by combining and analysing information from different sources (in this case a national disease surveillance programme and industry-derived survey data), and including farm location information to enhance our knowledge of the behaviour of disease in animal populations, and provide a more informed framework for designing efficient surveillance strategies.

5.6 Conclusion

In this population, the use of purchased feed for pigs was positively associated with *Salmonella* seropositivity, whilst SPF herd status and wet-feeding were negatively associated. Once adjusting for these covariates we identified pockets of unexplained risk for *Salmonella* seropositivity, the largest being in the north of Jutland. We found spatial dependency with regard to *Salmonella* seropositivity at distances of 6 km between farms. We propose that there is potential to exploit these spatial and risk factor findings for targeting surveillance. However, there was much unexplained non-spatial variation between farms and investigating what determines both the pockets of risk and the farm-level variation is an important area for further study. We conclude that by combining farm location details, routinely recorded surveillance information, and industry surveys, we have put the concept of risk-based surveillance into practice and further identified another valuable use for geo-referenced data in veterinary epidemiology.

5.7 Acknowledgements

Thanks to Wes Johnson (University of California, Irvine, USA) for Bayesian and general statistical advice. And to Paulo Ribeiro (Universidade Federal do Parana, Brazil) for geostatistical advice.

CHAPTER 6

Temporal and longitudinal analysis of Danish swine *Salmonella* surveillance and control programme data: implications for surveillance

Benschop, J., Stevenson, M., Dahl, J., Morris R.S., French, N. (2008) Temporal and longitudinal analysis of Danish swine *Salmonella* control programme data: implications for surveillance. *Epidemiology and Infection* **136**:1511-1520

6.1 Abstract

The control programme for *Salmonella* infection in Danish swine has reduced the number of human cases attributable to pork consumption and the focus is now on cost-effectiveness. We applied time-series and longitudinal analyses to data collected between January 1995 and May 2005 to identify if there were predictable periods of risk that could inform sampling strategy; to investigate the potential for forecasting for early aberration detection; and to explore temporal redundancy within the sampling strategy. There was no evidence of seasonality hence no justification to change to targeted sampling at high-risk periods. The forecast of seropositivity made using an ARIMA (0, 1, 2) model had a root-mean-squared percentage error criterion of 8.4%, indicating that accurate forecasts are possible. The lorelogram identified temporal redundancy at up to 10 weeks, suggesting little value in sampling more frequently than this on the average farm. These findings have practical applications for both farm-level sampling strategy and national-level aberration detection which potentially could result in a more cost-effective surveillance strategy.

6.2 Introduction

In industrialised countries, most cases of human salmonellosis are foodborne, and pork has been implicated as an important source (Nielsen & Wegener 1997, Hald et al. 2004). In Denmark, the estimated number of cases attributed to domestically produced pork has decreased substantially from 22 cases per 100,000 head of population in 1993 to 2.6 cases per 100,000 in 2004 (Ministry of Family and Consumer Affairs 2005). This decrease has been largely attributed to the Danish swine *Salmonella* surveillance and control programme (DSSCP) which was established by the Danish Ministry of Food, Agriculture and Fisheries in 1993.

Finisher pigs are known to carry *Salmonella* and it is these that can contaminate the food product, which is then capable of infecting humans (Botteldoorn et al. 2003). Carriage occurs in the gut, lymph nodes, and tonsils. Bacteria may directly or indirectly contaminate the carcass during evisceration and other processes that occur in the slaughterhouse such as scalding or cutting (Pearce et al. 2004). In addition, there is accumulating evidence to suggest that pigs may become infected during transport to slaughter, or while waiting in lairage (Rostagno et al. 2003). However, in Denmark 95% of pigs have less than 3 hours of transport and lairage [1] (Alban & Stärk 2005), so the chance of infection occurring during these processes is likely to be minimal. Therefore, the predominant source of *Salmonella* that contaminates the carcasses and presents a risk to human health will be the farm of origin. What is unknown about the farm-level risk of infection is whether or not it is constant over time or fluctuates on a seasonal basis. Although this issue has been addressed by other authors (Carstensen & Christensen 1998, Christensen & Rudemo 1998, Hald & Andersen 2001) the follow-up period for each of these studies was short.

In this paper our first aim was to apply time-series methods to data collected by the DSSCP over a ten-year period and to describe the key temporal features of trend, cyclicity, and autocorrelation. Subsequent to the successful reduction in pork-attributed human cases, attention is now focused on the cost-effectiveness of *Salmonella* surveillance (Anonymous 2006) and an understanding of the temporal pattern of *Salmonella* seroprevalence in pigs can partially address this need. If identified predictable periods of high risk are found then a risk-based approach would involve sampling more frequently

[1] http://www.danskeslagterier.dk

at these times and less frequently at periods of less risk (Stärk et al. 2006).

Our second aim was to evaluate the predictability of the seroprevalence time series with a view to using the data for forecasting. By comparing forecasts with the data that was actually observed there is the possibility of being able to readily detect aberrant behaviour. The identification of an aberration in national herd seroprevalence could be used as a screening test to alert authorities to emerging problems.

Our third aim is to address the question of temporal redundancy within the sampling strategy. In this study, temporal redundancy encompasses the situation where farms are sampled so often that the dependency between these repeated samples could deem some testing unnecessary. We use the lorelogram, a technique related to that previously used in animal tracking (Salvatori et al. 1999), environmental radiometry (Dowdall et al. 2003), and ground-water monitoring (Cameron & Hunter 2002), to suggest an optimised sampling strategy for the DSSCP.

6.3 Materials and methods

6.3.1 The Danish swine *Salmonella* surveillance and control programme

The ongoing DSSCP has been based on the testing of meat-juice samples from finisher pig herds since 1995. The current sampling strategy has been in place since August 2001 and includes all herds with an annual kill of $>$200 slaughter pigs (representing 99% of all finisher herds in Denmark). The number of animals sampled at slaughter depends on herd size, with 60, 75, or 100 pigs sampled per herd per year. All samples are analysed at the Danish Institute for Food and Veterinary Research using the Danish mix-ELISA. This test can detect O-antigens from at least 93% of all serovars known to be present in Danish pigs (Mousing et al. 1997). On the basis of testing, herds receive a monthly serological *Salmonella* index which is based on a weighted average of the results from the previous 3 months. The levels of index are low (index 1–39); medium (index 40-69); and high (index $\geq$70) (Alban et al. 2002). In addition, testing is carried out in feed mills, multiplying and breeding herds, slaughter plants, and on fresh pork.

There have been a number of changes to the DSSCP since its inception. These include penalising farmers supplying pigs with an unacceptable level of serological *Salmonella*

index, lowering the positive cut-off for the Danish mix-ELISA, and introducing a fourth level of index (nil) where low risk herds are sampled less frequently (Anonymous 2006). The latter has been in place since July 2005.

6.3.2 The data

Two extracts of data were obtained from the central database of the DSSCP. These comprised the results of meat-juice testing conducted from 1st January 1995 to 31st May 2005 inclusive. The first data extract provided for each farm a unique identifier and details of farm location. The second data extract provided details of the 6,992,082 carcasses tested over the 10-year study period. Details included the unique farm identifier, the date of sampling, and the result of the Danish-mix ELISA. For these analyses an ELISA optical density percentage (OD%) >20 was classified as positive. This is equivalent to an adjusted OD% of >10: the cut-off for positivity that has been used by the DSSCP since 1st August 2001 (Alban et al. 2002).

For the period 8th May 2002 until 30th September 2004 inclusive, herd-size data were extracted from the Danish Central Husbandry Register and health status data were extracted from the Danish Specific Pathogen Free (SPF) Company.

6.3.3 Statistical analysis

Time-series analysis

We used time-series methods to describe the components of trend, cyclicity, and autocorrelation in the data. We reasoned that the exploration of trends in the data would facilitate the evaluation of the control programme, and that identification of seasonality or other temporal autocorrelation has the potential to inform surveillance.

To describe the trend, the weekly incidence risk (IR) of seropositivity was presented as a time-series graph. The presence of trend was formally tested using a bootstrapped Spearman test (Henderson 2005). The positive results were then stratified into three levels: low (10–20 adjusted OD%); medium (21–50 adjusted OD%); and high positive (>50 adjusted OD%) and the weekly IR of the three strata were plotted. For all time-series plots loess

smoothing splines were applied to the raw time series to emphasise the major features while reducing distraction from random variation (Bowman & Azzalini 1997).

The time series was stratified by region to investigate spatial variation in temporal patterns. Each regional series was further stratified by level of positivity. Because there were very few farms in the Copenhagen region (n=8), they were aggregated with farms from a larger adjacent region, Frederiksborg. A regional map of Denmark is shown in Chapter 3 (Figure 3.1).

The data were rendered stationary to provide a degree of replication within the series facilitating further statistical analysis (Diggle 1990). We first took the log of the proportion of samples positive to stabilise the variance and then detrended the series by fitting *a priori* a second-order polynomial to stabilise the mean. These transformations were performed on the overall time series and then for each region individually, allowing a more detailed examination of spatio-temporal variation. The model residuals were then examined as a near-stationary time series.

To identify the presence of seasonal effects, we first examined grouped box plots and seasonal sub-series plots. Secondly, we tested the statistical significance of calendar month by fitting an autoregressive linear model and calculating the coefficient of determination ($R^2_{autoreg}$) using the method proposed by Moineddin et al. (2003). Thirdly, to identify cyclicity including that produced by seasonality, a periodogram was plotted (Box et al. 1994). This involved fitting sinusoidal waves with a discrete set of frequencies (Fourier frequencies) to the data. Using this technique, the data were transformed to reveal cyclical behaviour. The practical value of the periodogram is that it can identify frequencies which are not always predictable before data are examined.

Temporal autocorrelation in the aggregated weekly data was identified using lagged scatterplots, autocorrelation (ACF) plots, and partial autocorrelation (PACF) plots. An autoregressive moving-average (ARMA) process was fitted to the data with examination of ACF and PACF plots used to identify the starting orders of the process (Box et al. 1994). Model parameters were estimated by maximum-likelihood methods using the Kalman filter. Overall model fit was assessed using Akaike Information Criteria (Diggle 1990). Residuals were examined in three ways: plotting as a time series; checking the autocorrelation function; and applying the LjungBox test for independence (Ljung & Box 1978). Model performance was estimated by calculating the amount of the original series vari-

ance that was explained by the model (Lopez-Lozano et al. 2000). This is analogous to the coefficient of determination (R^2) of a linear model.

Predictive modelling

For forecasting two subsets of the series were taken: (1) from weeks 1–487 to construct the model; and (2) from week 488 onwards for validation. We applied log transformation and differencing of the raw series to achieve stationarity. For forecasting we did not apply a polynomial fit since it places global assumptions on the data which may poorly estimate the fit beyond the range of the period of interest (Diggle 1990).

Identification of the order of differencing was by examination of ACF plots and estimating the variance of the series. An autoregressive integrated moving-average (ARIMA) process was fitted using the methods and diagnostics described previously. The predictive ability of the model was evaluated by two methods: (1) plotting the forecast and its 90% tolerance intervals and comparing it with the observed data; and (2) calculating the root-mean squared percentage error criterion (De Gooijer & Hyndman 2006).

Farm-level investigation

For this part of the study we used data from 8th May 2002 to 30th September 2004. This was for two reasons: (1) for this 127-week period we had additional farm-level variables for 86% of farms allowing a stratified analysis to be conducted, and (2) computational constraints.

Repeated measures on the same farm would be expected to be correlated, and determining the correlation structure has the potential to inform sampling strategy. For every week that a particular farm was sampled, the outcome was defined as positive if at least one result for the week was positive, and negative if all results for the week were negative. There were 11,754 farms contributing 697,877 farm weeks which were approximately uniformly distributed throughout the period of interest.

For stratification, a subset comprising 86% of the original 11,754 farms for which farm-level variables were available was extracted. This comprised 10,064 farms and 609,537 farm-weeks. These farms were stratified into two herd-size levels: small ($<$700 slaughter pigs, n=5605) and large ($>$700 slaughter pigs, n=4459). A second stratification was made

by health status: conventional (n=7028) and SPF (n=3036). We defined conventional herds as those not in the Danish SPF system. The proportion of positive farm-weeks for the period was expressed as the IR.

The serial correlation of these repeated binary outcomes was explored using the lorelogram (Heagerty & Zeger 1998). This is the mean log odds ratio between observations at each weekly time lag. It is a counterpart to the variogram approach for continuous longitudinal responses (Diggle et al. 2002*a*). Both allow for the irregular spacing between sampling times on different farms and approximate an average measure of temporal dependency of the repeated measures on farms. We produced lorelograms for all farms, and then for the different strata, to identify if herd size and health status influenced the temporal dependency structure.

6.4 Results

Time-series analysis

Figure 6.1 presents a loess smoothed plot of all positive results over the study period, stratified by levels of positivity. With regard to all positive results, the first three years of the control programme had the highest IR (10-16%). This was followed by a decline in IR from mid-1997. From 1999 to 2002 there was a period of relative stability with the percentage positive between 4% and 8%. In mid-2003 there was a large rise in positivity from 5% to 10% over one month. The IR remained at about 10% until the end of the study period. The Spearman test gave a value of -0.28 (p<0.001), confirming the overall decreasing trend. The variance in the series was higher with higher IRs: the variance from 1995 to 1997 was 7.0, from 1998 to 2002 was 1.4, and from 2003 onwards was 3.9. In addition, in the first three year period (1995–1997) there is an apparent seasonal pattern to the data with visible peaks in February and November.

These trends were broadly followed for all three strata, although in the first three years of the DSSCP the series for the high positive strata (>50 adjusted OD%), was consistently lower than the other strata and, in the period from late 1996 to late 1997, did not show the peak shown by the other strata. The heterogeneity seen in the time series for high positive strata was accentuated when these were stratified by region. There was considerable vari-

ation across the country (Figure 6.2a–d). However, even in the regions in Zealand (Figure 6.2d) and on Bornholm (Figure 6.2c, solid line), where the level and variation in positivity was generally lower, the overall trend showed an initial fall in positivity until 1998, then a plateau until a rise in 2003. Figure 6.2a (regions in the north of Jutland) shows that the time series for Arhus has a different pattern compared with the other regions in the area.

When the series was stratified by region alone there were time periods when it appeared as if all regions were simultaneously experiencing a rise in the percentage of pigs positive (see Figure 6.3). This was most noticeable in early 1996, early 1997, and late 2003. Moreover, there were clear regional differences in IR: regions in the north and south of Jutland, and in the west of the country had higher IRs compared with the east.

Figure 6.4 presents the model residuals for the series once detrended by taking the log and fitting a second-order polynomial. The results for Denmark and for the main pig-producing counties are broadly similar. All show peaks of positive residuals in early 1996, early 1997, and late 2003, as in the raw data. However, a prominent peak of positive residuals in late 2000 only occurred in Nordjylland, Viborg, Ringkobing, and Ribe (circles in Figure 6.4b, c, e, and f). These four counties are contiguous in the north-west of Jutland. This peak was less evident in Arhus, Vejle, and Fyn and absent in Sonderjylland and on Zealand and Bornholm.

There was large variability between years and no indication of cyclicity in the grouped box and seasonal sub-series plots (Figure 6.5). The $R^2_{autoreg}$ for the autoregressive linear model was $<1\%$ indicating that calendar month was a very poor predictor of the series. The periodograms of the complete time series (n=544 weeks) showed no clear seasonality. There was a minor and non-significant peak at a frequency of three: this indicates a weak 181-week cycle (Figure 6.6).

Lagged scatter-plots indicated serial dependence up until at least 16 weeks. The auto-correlation plot showed strong autocorrelation but no indication of a regular pattern as would be expected if seasonality were a factor. Examination of the ACF and PACF plots indicated that an autoregressive (AR) model, with order >1, may be appropriate. Model fitting resulted in a best fit of AR(3) to the stationary series. The residuals were randomly distributed in time. They were not significantly autocorrelated when tested by the ACF plot and the Ljung-Box test. The amount of the stationary series variance that was explained by fitting the model was 69%.

Predictive modelling

For forecasting, an ARIMA (0, 1, 2) model gave the best fit. The residuals were not significantly autocorrelated and the amount of the logged series variance that was explained by fitting the model was 86%.

When the forecasts were superimposed on the logged observed data, the observed value lay well within the 90% tolerance limits of the forecast (Figure 6.7). The root-mean-squared percentage error criterion for the one-week-ahead forecast was 8.4% indicating the model had a reasonable forecasting ability.

Farm-level investigation

The farm-week IR (the proportion of positive farmweeks) for all 11,754 farms was 10.6% (95% CI: 10.4-10.9). For the subset of 10,064 farms, there were statistically significant differences between the strata: small and large farm IRs were 9.3% (95% CI: 9.0-9.7) and 12.2% (95% CI: 11.8–12.6) respectively. Conventional and SPF farm IRs were 11.3% (95% CI: 10.9–11.6) and 10.1% (95% CI: 9.6-10.5) respectively.

The lorelograms for the stratified data are shown in Figure 6.8. There were two points of interest that apply to these plots, and to the equivalent plot for all 11,754 farms. Firstly, the log odds ratios decrease rapidly up to a lag of 10 weeks, beyond which the decrease is more gradual. At lag 1 the odds ratio is approximately seven indicating that positive farm-weeks tend to follow each other. These first 10 weeks show statistically significant temporal dependency, but beyond that there is still a strong downward trend in the lorelogram with the mean stabilising at about 66 weeks. Secondly, the mean remains at levels well above zero, suggesting that positive correlation remained at large time separations. This occurs as a result of heterogeneity between farms: some farms having relatively frequent positive weeks and some having very few or none.

There were also subtle differences in the plots for different strata. The two strata with the lowest IR (small and SPF herds) had mean odds ratios of eight at lag one, and stabilised at week 56 with a mean odds ratio of 2.5. The two strata with the highest IR (large and conventional herds) had mean odds ratios of 6.5 at lag one, and stabilised at week 76 with a mean odds ratio of 2.

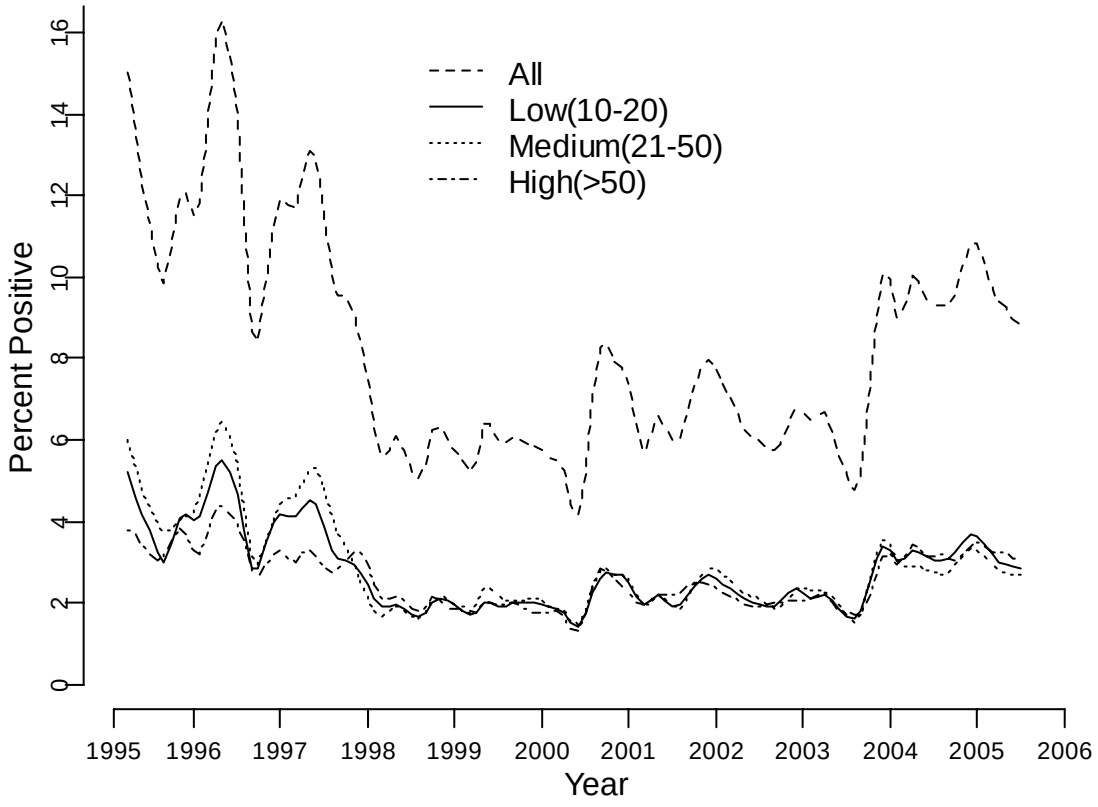


Figure 6.1: Loess smoothed plot of the percentage of pigs positive for the Danish mix-ELISA stratified by level of positivity. Data originate from the Danish swine *Salmonella* surveillance and control programme.

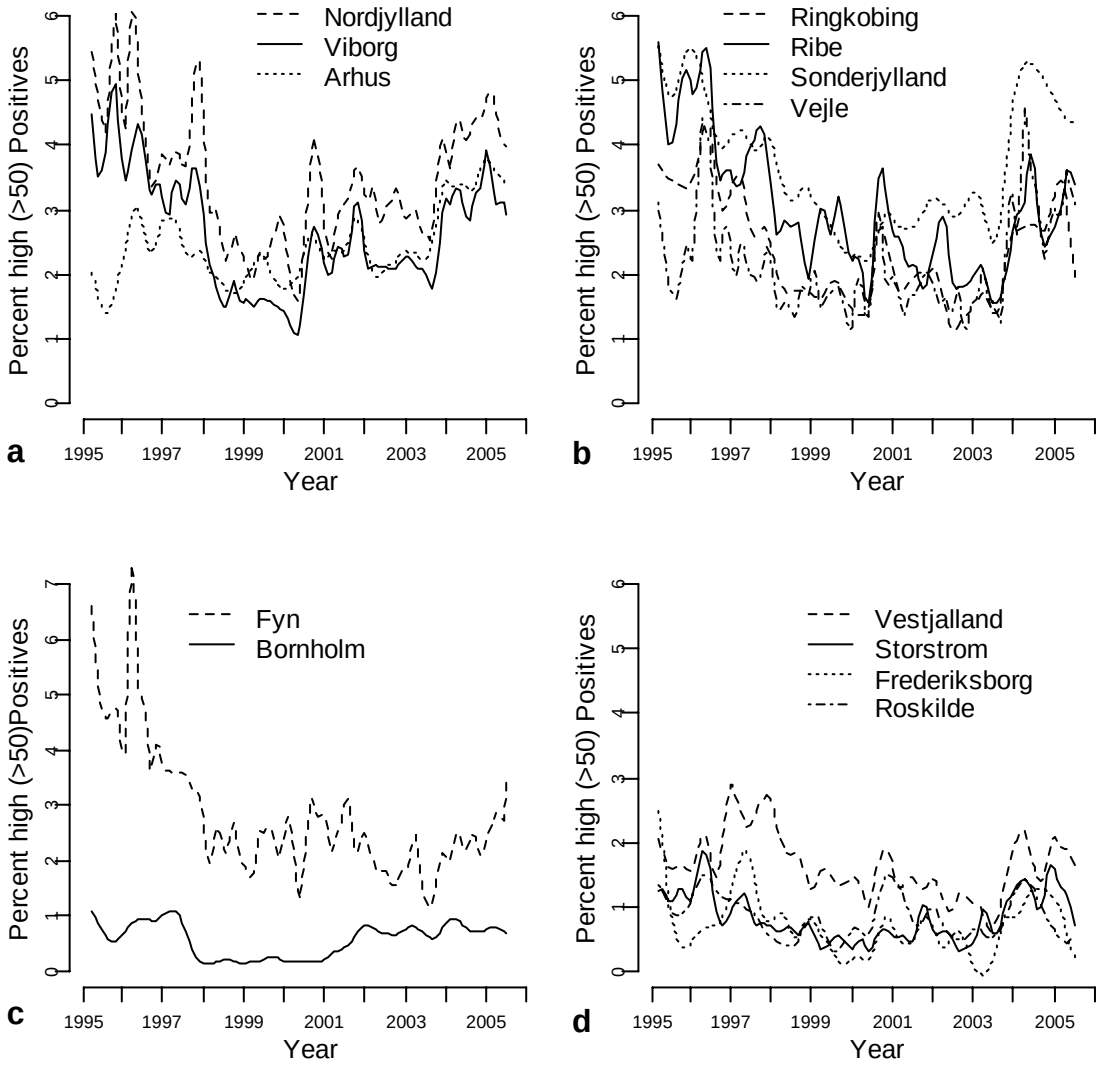


Figure 6.2: Loess smoothed plots of the percentage of pigs in the high positive strata for the Danish mix-ELISA stratified by region. Data originate from the Danish swine *Salmonella* surveillance and control programme.

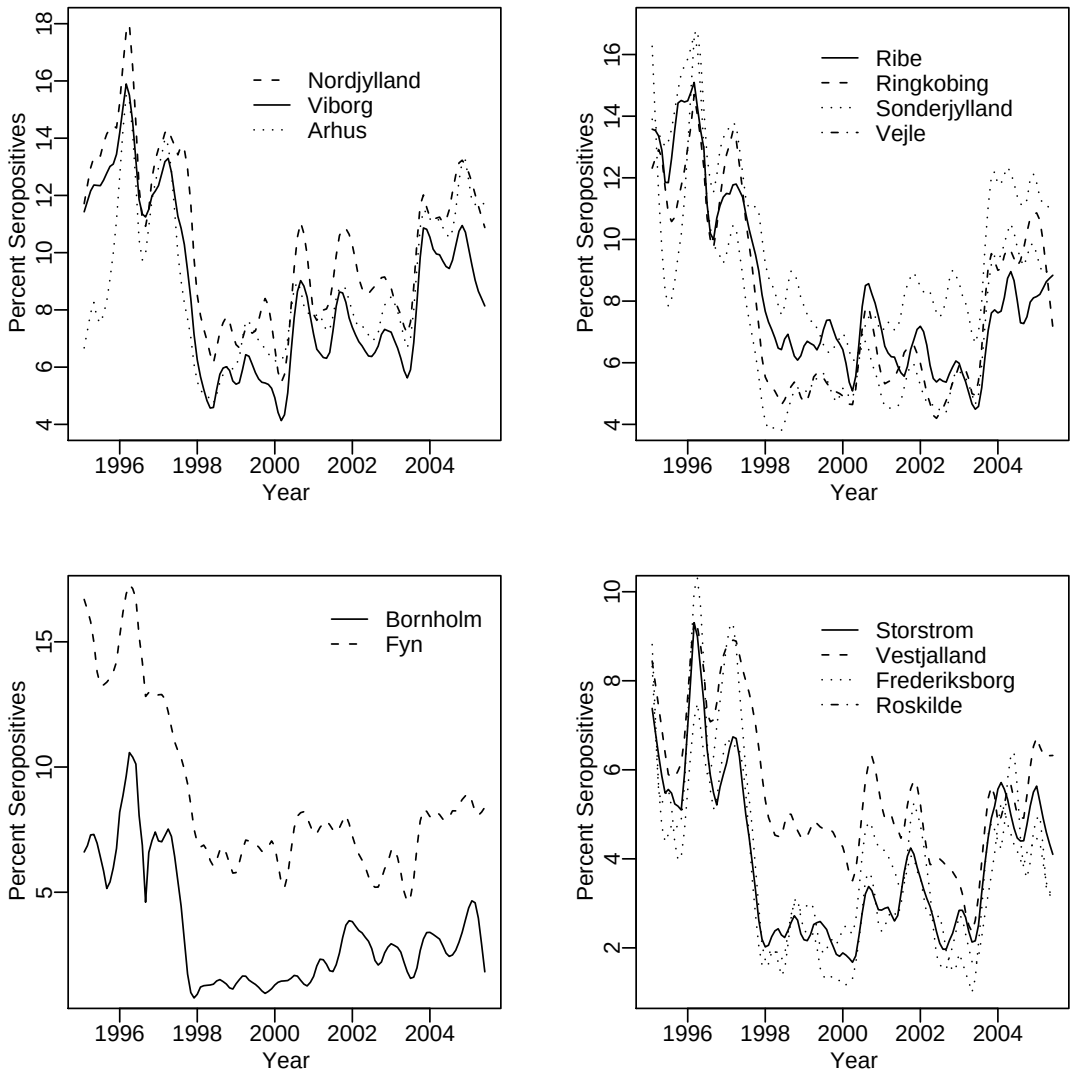


Figure 6.3: Loess smoothed plots of the percentage of pigs positive for the Danish mix-ELISA stratified by region. In 1996, 1997 and 2003 all regions experienced simultaneous rises in the percent of seropositive pigs. Data originate from the Danish swine *Salmonella* surveillance and control programme.

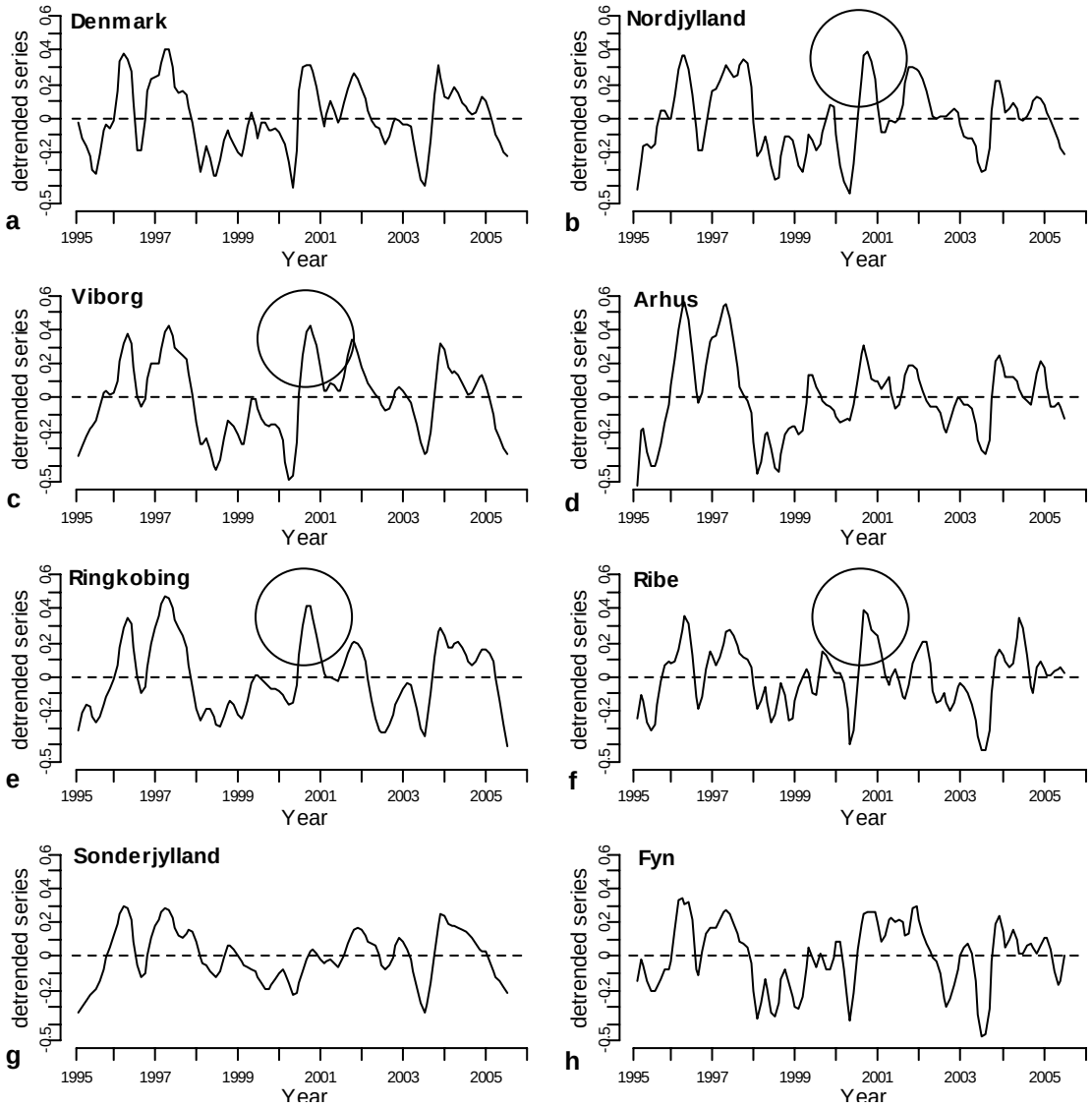


Figure 6.4: Plot of stationary time series for (a) Denmark, (b) Nordjylland, (c) Viborg, (d) Arhus, (e) Ringkobing, (f) Ribe, (g) Sonderjylland, and (h) Fyn respectively. The raw series has been logged to stabilise the variance and then detrended with a second-order polynomial. The dashed horizontal line is at the median percentage positive (0). The circles highlight the peak of positive residuals seen in counties in the north-west regions of Jutland in late 2000. Data originate from the Danish swine *Salmonella* surveillance and control programme.

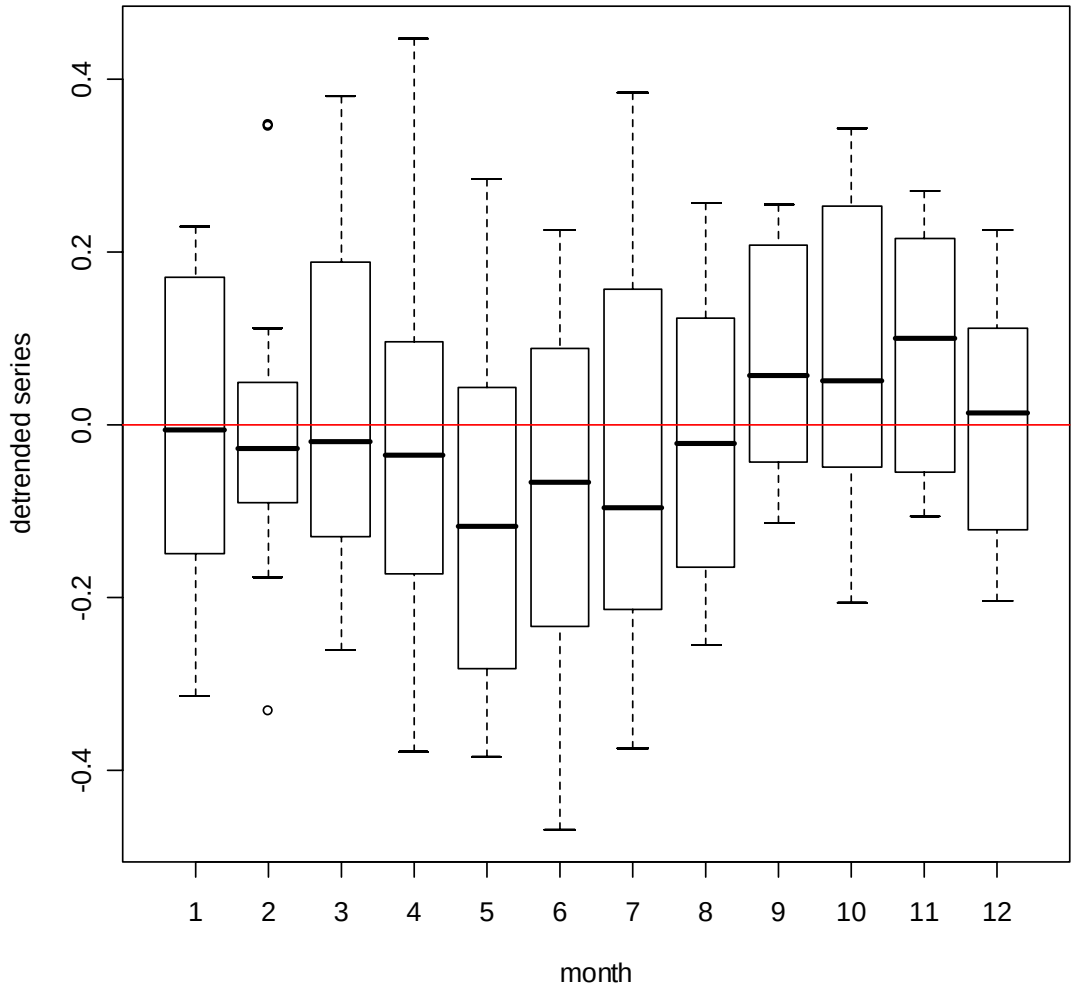
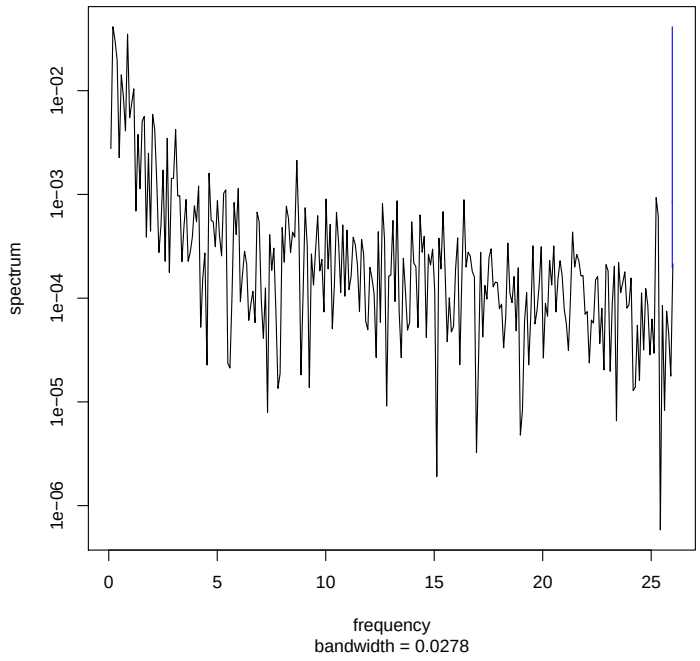
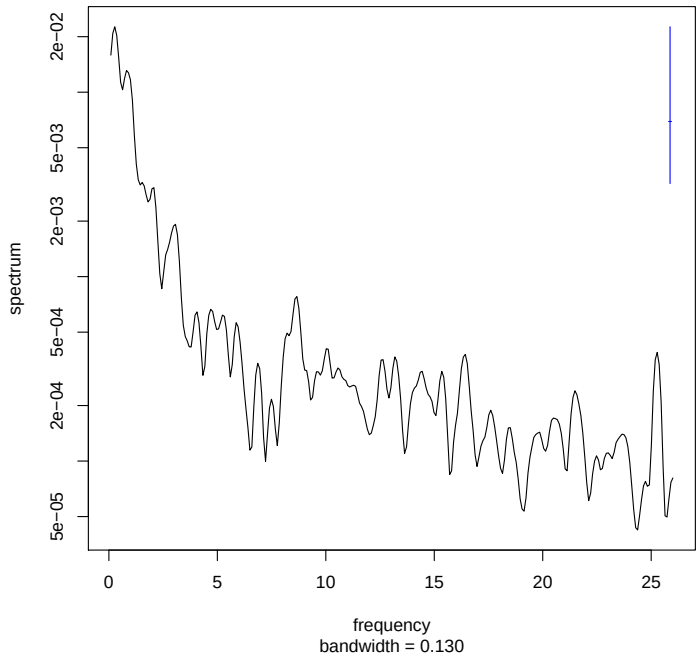


Figure 6.5: Monthplot of the time series residuals. The raw series has been logged to stabilise the variance and then detrended with a second-order polynomial. Data originate from the Danish swine *Salmonella* surveillance and control programme.



(a) Raw periodogram



(b) Smoothed periodogram

Figure 6.6: (a) Raw and (b) Daniell smoothed periodograms of the weekly time series residuals, Denmark, 1995–2004. The vertical bar on the right hand side of the plot indicates the 95% confidence interval illustrating that there is no clear peak that would indicate significant cyclicity.

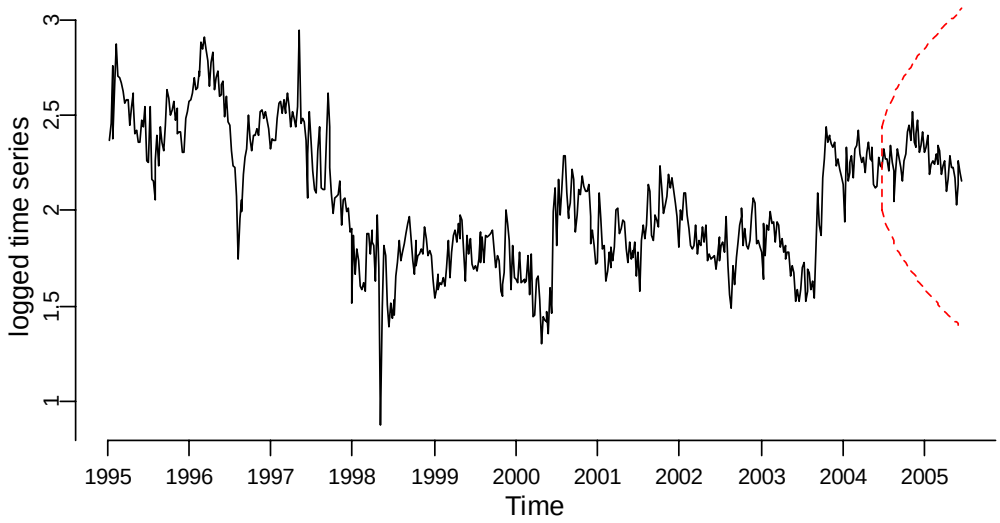


Figure 6.7: Plot of 90% tolerance limits of the forecasts made using a ARIMA (0, 1, 2) superimposed (dashed line) on the logged observed data. Data originate from the Danish swine *Salmonella* surveillance and control programme.

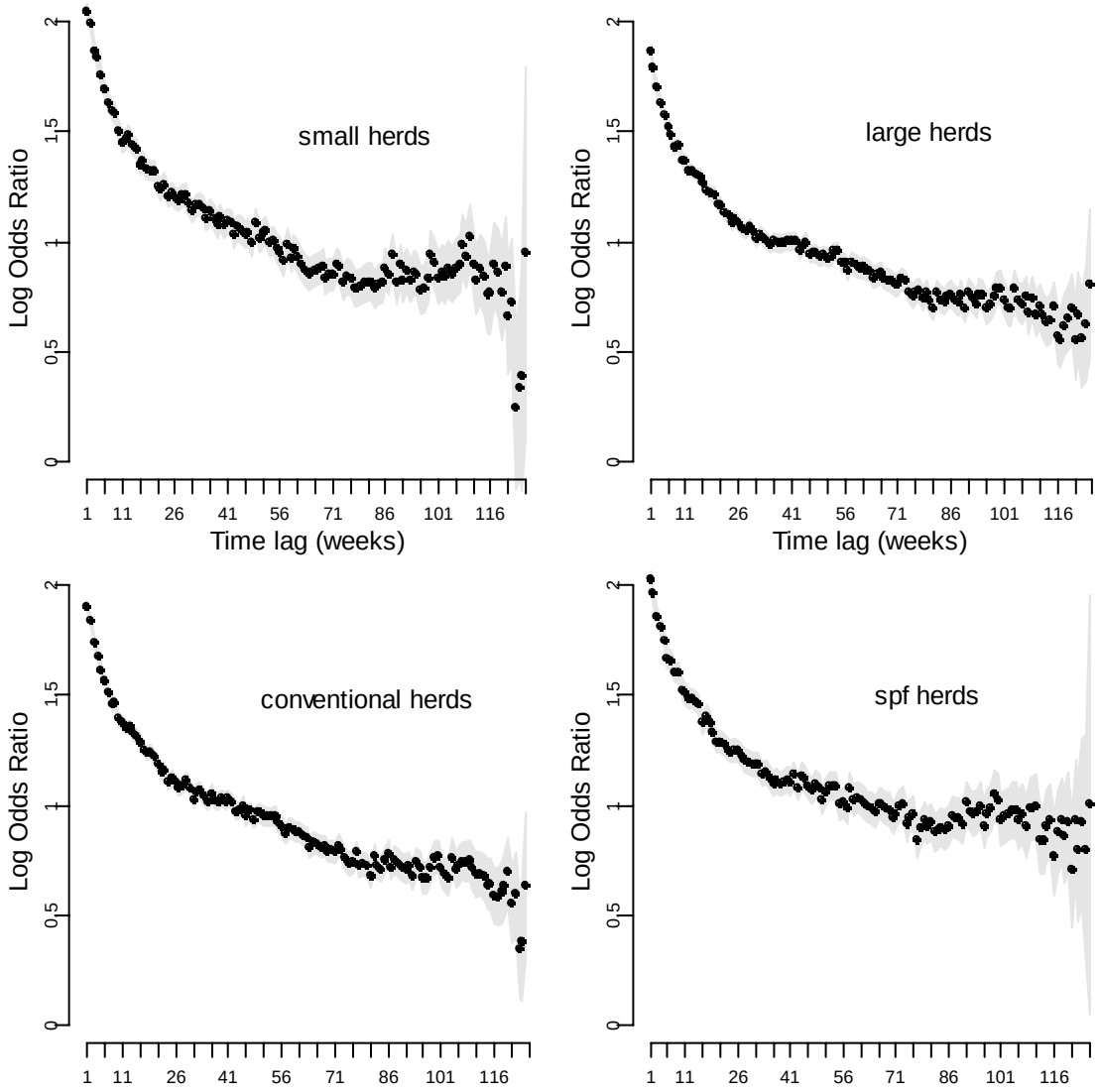


Figure 6.8: Lorelograms (estimated mean log odds ratio as a function of lag time) for stratified data from May 2002 until September 2004. Grey shading is the 95% confidence intervals around the log odds ratios. Data originate from the Danish swine *Salmonella* surveillance and control programme.

6.5 Discussion

This study has used accumulated data from the first 10 years of the DSSCP to investigate key temporal epidemiological features, explore the potential for forecasting, and address the question of temporal redundancy. Our focus has been on the application of both established and novel techniques to respond to the need for optimisation in the DSSCP.

The raw series showed a significant decreasing trend which was complex. The dramatic reduction in IR over the first $3\frac{1}{2}$ years was predominant and probably due to the effect of the control programme (Hald & Andersen 2001). From mid-1998 there was relative stability until a sudden rise in September 2003 when the IR rose from 6% to 10%. This rise was so sudden that a laboratory error was suspected and it was subsequently found that a problem in August 2002 had resulted in underreporting of positive results (Bak et al. 2007). This increase in seroprevalence may have been a factor in the small rise in pork-derived human salmonellosis cases in 2003, compared with 2002. However, 2004 saw a continuation in the long-term trend of a reduction in human cases that had been evident since 1996 (Ministry of Family and Consumer Affairs 2005). We decided to stratify by level of positivity because differences exist in the pigs' serological response depending on the serovars involved, time since infection, and infection pressure within a herd (Nielsen et al. 1995). Most serovars other than *Salmonella typhimurium* give only a moderate serological response with the Danish Mix-ELISA (Lo Fo Wong et al. 2003). Therefore, our finding of a different trend for the high positive strata (>50 adjusted OD%), in the first three years of the DSSCP, may indicate that at that time there was relatively greater contribution from serovars other than Typhimurium (Figure 6.1). This could explain why the time series for the high positive strata in Arhus shows a different pattern compared with the other regions in the north of Jutland (Figure 6.2a). The pattern of a low, almost stationary percentage of pigs in the high positive strata seen in the eastern counties (Figure 6.2d) may be a result of either infection with serovars with low test sensitivity, or low herd-level infection pressures, or both.

Regional stratification identified times (early 1996 and early 1997) when the whole country was experiencing an increased IR. This may point to a seasonal pattern or to laboratory anomalies. The regional differences in IR are in agreement with earlier work (Carstensen & Christensen 1998, Mousing et al. 1997) and may be associated with large

herd size: regions with high IR have the largest pig populations and larger pig farms. However, the relationship between herd size and *Salmonella* seroprevalence is complex, with some studies identifying a positive association between herd size and seroprevalence (Carstensen & Christensen 1998) and others identifying a negative association (van der Wolf et al. 2001).

The aggregations of positive residuals in the northwest regions of Jutland in late 2000 suggest a local epidemic. This could be due to the dissemination of contaminated feed or infected pigs throughout the area, possibly in combination with a regional practice that would favour rapid spread such as sharing of seasonal workers or machinery.

We found no evidence of a seasonal pattern in seroprevalence in our analysis of these 10 years of data. This is in agreement with a European-wide longitudinal study of seroprevalence in finisher pigs from October 1996 and May 1999 (Lo Fo Wong et al. 2004*b*). Similarly, temporal studies of passive laboratory-based surveillance data from 1991 to 2001 in Ontario (Zhang et al. 2005) and Alberta, Canada (Guerin et al. 2005*b*) found no seasonal pattern in *Salmonella* isolation from production animals, including pigs. Pigs are likely to be infected with *Salmonella* from a variety of sources including herd mates, introduced animals, contaminated feed, rodents, and the environment. None of these need be seasonally distributed.

Previous work on the DSSCP reported a seasonal pattern of seroprevalence throughout the period 1995-1997 (Carstensen & Christensen 1998, Christensen & Rudemo 1998, Hald & Andersen 2001). This was characterised by a summer trough and a late winter-early autumn peak. We also found visible peaks in February and November from 1995 to 1997 which may be due to seasonal effects, laboratory error, or simply random variation. It would appear that, if inspected over a long enough period, the distribution of *Salmonella* seroprevalence does not follow a consistent seasonal pattern. Thus, we conclude that there is no apparent benefit in targeting sampling to particular times of the year. The seasonal pattern in human cases is much clearer, with consistent reports of a late summer-early autumn peak (Hald & Andersen 2001, Guerin et al. 2005*a*).

Our finding of a minor, non-significant peak in the periodogram analysis equating to a cycle of 181 weeks (Figure 6.6) probably reflects the residual peaks that are visible in Figure 6.4 in late 1996, 2000 and late 2003. This apparent cyclicity is probably artefactual and due to laboratory problems in late 2003 (Bak et al. 2007) and in late 1996

(P. Willeberg, personal communication, 2005). The peak in 2000 was due to a regional excess of *Salmonella* seropositivity in the north-west of Jutland.

Fitting the AR process to the data provided a temporal summary for the logged polynomialdetrended series. A second-order polynomial was fitted *a priori*, allowing us to visualise and explain the complex trend in the national time series in three stages : (1) the dramatic effect of a successful control program (1995-1998); (2) a period of relative stability (1999-2002); and (3) a period of increasing national seroprevalence coincident with the laboratory problem in September 2003. We found that the current value of the percentage of pigs positive for the Danish-mix ELISA in any given week was dependent on the three preceding values, with most weight on the immediately preceding value. Our findings are identical to the relative weighting of 3:1:1 which is currently used to calculate the index for identifying herds for intervention in the DSSCP (Alban et al. 2002).

Forecasting is inherently challenging as it involves making assumptions based on past behaviour in the face of random variation. We used an ARIMA model that analyses the series as a function of its past values (AR), its trend (I) and its abrupt changes in the near past (MA) (Box et al. 1994). As it does not require a stationary series at the outset, the potential problem of imposing global assumptions on data that may poorly estimate the fit beyond the range of the period of interest are overcome (Diggle 1990). ARIMA models have traditionally been applied in the financial sector but are increasingly being used in medicine; recently as a tool for anomaly detection public health surveillance (Le Strat 2005). Forecasting nationally one-week-ahead, as done here, would provide a baseline to which observed data could be compared. Observations outside a preset alarm threshold could signal an investigation to ascertain if the problem was geographically widespread or clustered within particular regions. Forecasts themselves could be done on a regional basis allowing for regional differences in seroprevalence. Within this context we envisage real-time calibration of the model, as additional data fed into the model should improve its predictive ability.

The lorelogram has been used to explore serial correlation of repeated binary outcomes for constipation treatment efficacy (Choi et al. 2005) and schizophrenia symptoms (Heagerty & Zeger 1998). Although we are not aware of the previous use of the lorelogram for optimising a sampling strategy, the temporal variogram (a counterpart to the lorelogram for a continuous response variable), has been used for this purpose (Salvatori et

al. 1999, Dowdall et al. 2003, Cameron & Hunter 2002). Our finding of a ten-week period of statistically significant temporal dependency indicates that at the farm level there is a strong temporal memory up to (and to a lesser extent beyond) that lag. Extending this idea further would suggest that there is little value in sampling more frequently than every 10 weeks on the average farm. However, this theoretical approach needs to be balanced against the practicalities and advantages of the current continuous sampling and intervention strategy. Nonetheless, sampling at this reduced frequency (every 10 weeks) might be investigated for those herds enrolled in the risk based scheme which has been running since July 2005 (Anonymous 2006). This scheme requires one sample per month to be taken from herds with a *Salmonella* index level of nil and a minimum of 10 negative meat-juice samples in the last three months. To date 50% of herds meet these criteria.

It appears that there is a stronger episodic effect (i.e. faster decay in the lorelogram before reaching relative stability) in small and SPF herds when compared with large and conventionally managed herds. The more persistent temporal dependency in the large and conventional herds indicates that these herds could benefit most from a reduced frequency of sampling. However, the results for the different strata must be interpreted with caution as there is potential for bias here. Our choice of binary outcome (i.e. a farm-week is positive if it has at least one result positive) would predispose large herds to having more positive farm-weeks than small herds purely by chance as they are sampled more frequently. The log odds ratios between observations at very long lag periods should also be interpreted with caution as there are few pairs contributing to this lag.

Due to the near complete coverage of sampling in the DSSCP (all herds producing $>$100 finishers per annum prior to 1st August 2001, all producing $>$200 pigs at or after 1st August 2001) our dataset is effectively a census of the population of Danish finisher herds. This has enabled inferences to be made about the *Salmonella* status of all Danish finisher herds over the period January 1995 to May 2005. However, the sampling scheme used prior to August 2001 resulted in pigs from large herds being proportionally over-represented. The results from this early period may be subject to this ascertainment bias.

We have applied time series and longitudinal analytical methods to identify patterns in both national and farm-level data for sub-clinical salmonellosis in Danish swine. These findings have direct and practical applications for both farm-level sampling strategies and national-level aberration detection, which potentially could result in a more cost-effective

surveillance strategy.

6.6 Acknowledgements

Thanks to Bodil Ydesen (Danish Meat Association) for accessing the herd size and health status data.

CHAPTER 7

Bayesian zero-inflated predictive modelling of herd-level prevalence for risk-based surveillance

7.1 Abstract

The national control programme for *Salmonella* in Danish swine herds in 1993 has led to a large decrease in pork-associated human cases of salmonellosis. The pork industry is increasingly focussed on the cost-effectiveness of surveillance while maintaining consumer confidence in the pork food supply. Using national control programme data from 2003 and 2004, we developed a zero-inflated binomial model to predict which farms were most at risk of *Salmonella* and then preferentially sampled these high-risk farms. This type of modelling allows assessment of similarities and differences between factors that affect herd infection status (introduction) and those that affect the seroprevalence in infected herds (persistence and spread). The model suggested that many of the herds where *Salmonella* was not detected were infected but at a low prevalence. Using cost and sensitivity, we compared the results to those under the standard sampling scheme, based on herd size, and the recently introduced risk-based approach. Model based results were less sensitive but show significant cost savings. Further model refinements, sampling schemes, and the methods to evaluate their performance are important areas for future work and should continue to occur in direct consultation with Danish authorities.

7.2 Introduction

New challenges for animal health surveillance for zoonotic disease in the 21st century are manifold and include those brought about by increased trade, limited resources, consumer awareness, and disease emergence (Woolhouse & Gowtage-Sequeria 2005, Hodges & Kimball 2005, Fevré et al. 2006, Vorou et al. 2007). This paper is focussed on the additional challenge of developing exit or reduction strategies for surveillance systems for diseases that in the past represented an important risk, when today the risk to consumers is substantially reduced (Willeberg 2006). Bovine spongiform encephalopathy (BSE) in the United Kingdom and *Salmonella* in Danish pork are examples of diseases that meet these criteria.

When compared with the approximately 184,000 (Bradley et al. 2006) cases of BSE diagnosed in United Kingdom (UK) cattle born before the reinforced feed ban (on 1st August 1996), there has been a very small and diminishing number of cases (144 to date) born after the ban.[1] Although it is clear that the UK BSE epidemic has neared its end, tests still continue on all slaughtered cattle aged over 30 months and all fallen stock aged over 24 months (Hueston & Bryant 2005). At the time of writing (December 2008), there are plans to increase the age limit to 48 months for healthy slaughter stock.[2]

The Danish swine *Salmonella* surveillance and control programme (DSSCP) was instigated in 1993 by the Danish Ministry of Food, Agriculture and Fisheries in response to a human epidemic of salmonellosis during the summer of that year. This was traced to *Salmonella infantis* in pork and involved some 550 recorded cases. The programme's objective is to lower the prevalence of *Salmonella* so that domestically produced pork is no longer an important source of salmonellosis in humans (Mousing et al. 1997). Since 2001, the prevalence of *Salmonella* in Danish pork (monitored at the slaughterhouse) has reduced from 1.5% to 1% of carcass swab samples (Ministry of Family and Consumer Affairs 2006). The estimated number of cases of salmonellosis in humans in Denmark attributable to pork consumption decreased from 1444 in 1993 to 215 in 2005 (Nielsen et al. 2001, Ministry of Family and Consumer Affairs 2006).

At the time of writing (December 2008), there is a large and sustained outbreak of human salmonellosis due to *Salmonella enterica* serotype Typhimurium phage type U292

[1] http://www.defra.gov.uk/animalh/bse/controls-eradication/feedban-bornafterban.html
[2] http://www.food.gov.uk/news/newsarchive/2008/dec/bse

occurring in Denmark. Interestingly, no cases have been observed in countries outside Denmark despite the fact that 85% of Danish pork production is exported. In the face of this epidemic, the current climate in Denmark is probably not conducive for proposing a surveillance reduction strategy. Such strategies require a delicate balance between satisfying producer and industry concerns about cost-effective testing and maintaining consumer confidence in food supply. *Salmonella* and BSE were the food risks most dreaded in a UK survey of food risk perception undertaken in 1999 (Kirk et al. 2002) and a recent survey of consumers identified meat as the food item in which confidence had decreased the most (Verbeke et al. 2007). It makes sense that any strategy involving a reduction in testing should demonstrate an equal or greater sensitivity as the existing one, regardless of the potential efficiency gains.

The means to evaluate the sensitivity of a surveillance programme and subsequently compare alternatives has been recently explored in the veterinary epidemiological literature (Audigé et al. 2001, Cannon 2002, Martin et al. 2007*a*). In this context, a surveillance programme is considered as a diagnostic system which aims to correctly identify the presence or absence of an unwanted agent. By quantifying the characteristics of the diagnostic system (such as its specificity and sensitivity), a surveillance programme can be formally evaluated. For example, Audigé et al. (2001) defined surveillance sensitivity as the probability of declaring an area infected, given that infection exists, for the evaluation of surveillance for porcine reproductive and respiratory syndrome (PRRS) in Switzerland.

Quantification of the sensitivity of a surveillance system allows one to compare alternative surveillance strategies. For example, the comparison of the sensitivity of the currently targeted surveillance system for classical swine fever (CSF) in Denmark with that of a simulated non-targeted system identified that the current system was twice as sensitive compared with the simulated, non-targeted system (Martin et al. 2007*a*). In another Danish example, the sensitivity of the current surveillance programme for infectious bovine rhinotracheitis (IBR) was compared with three other surveillance scenarios targeting specific geographical areas and risk periods (Chriel et al. 2005).

To date, techniques involving scenario tree methodology have been used for proof of disease freedom for exotic, non-zoonotic, and clinically severe animal infections such as PRRS, CSF and IBR. In this paper we apply zero-inflated binomial modelling to the endemic, zoonotic, and sub-clinical infection of Danish finisher pigs with *Salmonella*

spp. Proof of disease freedom is not the end-point here. The issue is rather to maintain the status of domestically produced pork as a minor source of salmonellosis in humans.

We propose that it is possible to both maintain consumer confidence in food supply, and fulfil industry requirements for a surveillance reduction strategy with a targeted approach, whereby populations with higher risk of infection are preferentially sampled. Our objective is firstly to develop a model that predicts which farms are most at risk of *Salmonella*. Secondly, we preferentially sample the high-risk farms and compare our results to those under: (1) the standard sampling scheme, based on herd size, and (2), the recently introduced risk-based approach (Ministry of Family and Consumer Affairs 2006). In this way, we are able to evaluate the impact of alternative sampling strategies on overall system performance.

7.3 Materials and methods

7.3.1 Data sources

Data were obtained from three sources. Firstly, every pig herd is required to register with the Danish Central Husbandry Register. This provided a unique identifier (the CHR number), details of farm location, herd size, and the number of sows in the herd.

The second source of data was from the central database of the DSSCP. We used the results from 9735 farms in 2003 (n = 578,260 individual samples) for initial model building. The DSSCP database also provided results from the 8151 farms sampled at least 10 times in 2004 that were also sampled in 2003. This comparison was required to investigate the performance of our different sampling schemes. Details retrieved from the DSSCP database included the CHR number, the date of sampling, and the result of the Danish-mix ELISA (DME). This test measures antibodies in meat-juice to determine the previous exposure of finisher pigs to *Salmonella spp.* and can detect O-antigens from at least 93% of all serovars known to be present in Danish pigs (Mousing et al. 1997). The principal advantages of serological methods for *Salmonella* detection is the ability to assay a large number of samples rapidly at relatively low cost and high sensitivity when compared to bacteriology (€2 per sample).

For these analyses, an ELISA optical density percentage (OD%) greater than 20 is classified as positive. This is equivalent to an adjusted OD% of greater than 10: the cut-off for positivity that has been used by the DSSCP since 1st August 2001 (Alban et al. 2002). All samples included in this study were analysed at the Danish Institute for Food and Veterinary Research using the DME. On the basis of testing, herds receive a monthly 'serological *Salmonella* index' which is based on a weighted average of the results from the previous three months. The levels of index are low level or no antibodies (index 0–39, level 1); medium (index 40–69, level 2); and high (index 70 or greater, level 3) (Alban et al. 2002). Herds in the medium and high index have reduced payments for finisher pigs sent to slaughter and must collect pen-faecal samples to determine the subtype and distribution of *Salmonella* in the herd.

The third source of data was the Danish Specific Pathogen Free (SPF) Company which provided health status details associated with each farm.

We chose to analyse data from 2003 and 2004 as we had access to additional farm-level details such as herd size, health status, and the number of sows on the farm for those respective years.

7.3.2 Sampling schemes

Four sampling schemes were used or developed:

1. **Original herd size based sampling (OHS)**

   This sampling strategy was in place from August 2001 until July 2005. Using this approach, the eligible population comprised all herds with an annual kill greater than 200 slaughter pigs (representing 99% of all finisher herds in Denmark). The number of samples taken depended solely on herd size: the aim was to take 60, 75, or 100 samples annually from herds with an estimated annual kill of 200–2000 (small), 2001–5000 (medium), and greater than 5000 (large) slaughter pigs respectively (Alban et al. 2002). For the purposes of this paper, we have used this sampling scheme to represent the bench-mark to which we compare the alternative sampling strategies. Figure 7.1 shows the distribution of pigs that were actually sampled per herd for 2003 and 2004.

2. **Danish Meat Association risk-based sampling (DRB)**

   In July 2005, the surveillance system became performance-based which reduced the annual sample size by approximately one-third. For herds that had no positive meat-juice samples over the previous three months, the sample size was reduced to one sample per month (Enoe et al. 2003, Ministry of Family and Consumer Affairs 2006). If a herd then had one or more positive samples, the strategy reverts to one based on herd-size (OHS). We apply a modified version of these sampling criteria to herds in 2004 based on their performance in 2003. Our modification is that we have extended the time period over which herds are assessed to determine their prevalence to be the whole year, rather than the previous three months.

3. **Model derived risk-based sampling A (MRBA)**

   We developed a targeted surveillance sampling strategy based on our previous risk-factor, spatial, and temporal analyses of the DSSCP data (Benschop et al. 2008*a*,*b*,*c*). All herds with a predicted median within-herd seroprevalence at or below a model determined cut-off in 2003 were identified as low risk and were placed on the DRB scheme. This prediction was based on the farm's covariate pattern and random farm effect. All other herds (above the predicted within-herd seroprevalence threshold) were left on the current sampling scheme for 2004 based on herd size (OHS).

4. **Model derived risk-based sampling B (MRBB)**

   As in MRBA above all herds with a predicted median within-herd seroprevalence at or below a model determined cut-off in 2003 were identified as low risk and were placed on the DRB scheme. The remaining herds were then assigned to two different sampling schemes depending on their predicted seroprevalence in 2003: (1) those with a predicted seroprevalence that was ≤ 0.25 or ≥ 0.55 were left on the current sampling scheme based on herd size; (2) those with a predicted seroprevalence of between 0.25 and 0.55 were more intensively sampled to provide 95% confidence that we were within 0.05 of the true value of the predicted seroprevalence. This range was chosen as these herds were near the cut-off for level 2 *Salmonella* status (0.40).

7.3.3 Model development for the sampling schemes

The frequency histogram of the herd-level prevalence based on the actual test results from the OHS sampling strategy for 2003 and 2004 (Figure 7.2) showed a large amount of variation with a predominance of test-negative herds. These test-negative herds can come from two types of disease-negative herds: (1) those that are truly uninfected and therefore every sample is negative, and (2); those that are, in fact, infected but provide insufficient samples to detect the presence of infection. This led us to propose a zero-inflated binomial (ZIB) approach to model herd level *Salmonella* prevalence as it reflected our understanding of what is happening on the farm. The ZIB model has two herd level outcomes, the probability of infection and — conditional on infection being present — an estimate of herd-level seroprevalence. This type of modelling can provide an added advantage over logistic regression: an ability to assess the extent of the similarities and differences between factors affecting herd infection status (invasion) and those affecting the seroprevalence in infected herds (persistence and spread).

Variables that might explain both the presence of infection and herd level prevalence included herd size, farm location, the number of sows present, and herd health status. Herd size was the actual number of slaughter pigs produced for the year; this was centred by subtracting the mean and dividing by 1000. Health status was a three level categorical variable: conventional, specific pathogen free (SPF), and SPF with *Mycoplasma*. The presence of sows was expressed as a three level ordinal variable: farms with no sows, farms with less than or equal to 125 sows (some), and farms with over 125 (many).

Logistic regression model was used for initial model building. Bivariate analyses found all covariates significant at the $p \leq 0.25$, level and using data from 2003 we built a multivariable model using the statistical software R (version 2.5.1) (R Development Core Team 2007). All putative risk factors were significant at $p \leq 0.05$. The continuous variable herd size was checked to see if it was linear in its log odds (Hosmer & Lemeshow 1989). Polynomials of herd size and biologically plausible two-way interaction terms between the main-effect variables were considered for inclusion.

Once satisfied with the model structure, we developed a logistic model within a Bayesian framework using WinBUGS version 1.4.1 (Gilks et al. 1994). Initially, we stipulated informed priors for the intercept term, and covariates relating to location, health status,

and the number of sows present on farm. We based these on published literature supplying subjective information about the likelihood ascribed to various combinations of covariate values (Congdon 2001). For example, from earlier work on other data from the Danish *Salmonella* surveillance and control programme we believed that it would be a protective factor for a herd to have SPF health status (Benschop et al. 2008*b*). Moreover, residing in the district of Sonderjylland in the south of Jutland would be a risk factor for herd-level seropositivity (Benschop et al. 2008*a*). Based on available literature, an increased number of sows on farms was considered a risk factor for *Salmonella* in finishers (Hautekiet et al. 2008).

Priors for the Bayesian logistic regression model were expressed in terms of a conjugate beta density (Congdon 2001). We used a non-informed, normally distributed prior centred at zero and with a variance of 1 for the effect of herd size, given information about the effect of this variable on seropositivity was not certain or conflicting. Three chains were run and convergence was judged to have occurred on the basis of visual inspection of time series plots and Gelman-Rubin plots (Toft et al. 2007). The length of the chain was determined by running sufficient iterations to ensure the Monte Carlo standard errors for each parameter were less than 5% of the posterior standard deviation. A total of 40,000 iterations were run with a burn in of 4000 iterations.

The logistic regression model was extended to a zero-inflated binomial model and specified as follows:

$$Cases_i \sim Binomial(pop_i, p_i) \tag{7.1}$$

Here, the number of cases from the *i*th herd is binomially distributed as a function of the number of trials (tests for *Salmonella* antibodies in meat-juice) pop_i, and the probability of of a test being positive (adjusted OD% > 10), p_i.

We further defined:

$$p_i = \rho_i * J_i \tag{7.2}$$

where J_i is an indictor variable representing infection status of the *i*th herd, and ρ_i is the seroprevalence conditional on the presence of infection. The term ρ therefore represents

the probability of finding infection in a randomly chosen pig from an infected herd. The latent variable J_i is distributed as:

$$J_i \sim Bernoulli(q_i) \tag{7.3}$$

where q_i is the probability of a herd being infected. This latent variable was modelled as:

$$\log(q_i/1 - q_i) = \alpha_0 + \alpha_1 x_{1i} + \ldots + \alpha_m x_{mi} + A_i \tag{7.4}$$

In Equation 7.4, the logit of the observed probability of the *i*th herd being infected, $logit(q_i)$, was modelled as a function of $m = 4$ farm-level explanatory variables (herd size, location, the number of sows present and health status) and a random effect term, A_i, which was normally distributed with a mean of zero and precision σ. For the ZIB model, the continuous variable herd size was categorised to facilitate model convergence. The categories chosen were the same as those used by the DSSCP (Alban et al. 2002).

The latent variable ρ_i was modelled as:

$$\log(\rho_i/1 - \rho_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_m x_{mi} + B_i \tag{7.5}$$

In Equation 7.5, the logit of the probability of observing infection in a randomly chosen pig from the *i*th infected farm was modelled as a function of the four farm-level explanatory variables defined earlier, and a random effect term for herd, B_i, which was normally distributed with a mean of zero and precision τ.

We set non-informed, normally distributed priors centred at zero and with a precision of 0.5 for each of the fixed effect terms, including the intercept. Sensitivity to these priors was evaluated by re-running the models with a precision of 1 and 0.2. For the precision of the random farm-level effects, σ and τ, we specified a precision of 1. Sensitivity to these priors was evaluated by re-running the models with a precision of 0.5 and 0.3.

Three chains were run and convergence was judged to have occurred on the basis of visual inspection of plots of the sampled values as a time series (Toft et al. 2007). The required number of iterations of the Gibbs sampler was determined by running sufficient iterations to ensure the Monte Carlo standard errors for each parameter were less than 5% of the

posterior standard deviations. A total of 30,060 iterations were run with a burn in of 1000 iterations.

We proposed using this model on 2003 data to inform sampling strategies for the subsequent year (2004). To check for consistency between years (2003 and 2004), we ran the model on both years of data separately and compared the magnitude and direction of the regression coefficients and the correlation between the random farm effects. A scatter plot of the median conditional seroprevalence q_i versus the median probability of infection ρ_i (Figure 7.3) was used to identify the cut-off for the two model derived risk-based sampling schemes A and B.

7.3.4 Comparison of the sampling schemes

The results from all four sampling schemes were compared by considering cost, the number of false negative farms (a measure of sensitivity), and the number of farms identified by the model with a within herd seroprevalence of ≥ 0.40.

Costs were compared by adding up the number of tests taken under each of the four sampling schemes. Only the costs of meat-juice testing were taken into account, with each meat-juice sample tested costing €2 . These costs are borne by the producers through levies on each pig slaughtered. There are follow-on tests once herds reach level 2 and 3, costing €200 with further costs if herds are found to be positive. Although important to affected producers, these follow-on tests were not considered in this study.

For each farm ($n = 8151$) there were 1020 iterations stored from the model and these were used to determine the false negative rate and the number of farms detected with a within-herd seroprevalence of ≥ 0.40 for each of the four sampling schemes.

The number of farms that were falsely reported as negative and the sensitivity for each of the four sampling schemes were determined using the following process:
(a) the J_i parameter, the indictor variable representing infection status of the *i*th herd, for 2004 was examined at each iteration. If it equalled one, then, for that iteration, the farm was considered infected. Otherwise for that iteration the farm was considered uninfected;
(b) ρ_i, the predicted within-herd seroprevalence given the herd was infected, for 2004 was determined for each iteration when the farm was infected. ρ_i was combined with the number of pigs sampled, using the binomial distribution to determine the number of positives

that would be detected at each iteration;
(c) a false negative iteration was defined as one where the farm was infected at the iteration, but no positives were detected at that iteration. The number of false negative iterations was summed and divided by the number of total iterations to give the number of false negative farms;
(d) this was expressed as the sensitivity of the sampling scheme by dividing the number of false negative farms by the total number of farms ($n = 8151$), and subtracting this fraction (the false negative fraction) from one.

The number of farms with a predicted seroprevalence of ≥ 0.40 for each of the four sampling schemes was determined using the following process:
(1) the number of positives detected at each iteration was determined as in steps (a) and (b) in the preceding paragraph;
(2) the number of positives was divided by the number sampled to give the detected seroprevalence at each iteration;
(3) the number of iterations that had detected seroprevalences at ≥ 0.40 was summed and divided by the number of total iterations to give the number of farms that had a detected seroprevalence at ≥ 0.40.

7.4 Results

7.4.1 Data sources

In 2003, there were 9735 herds in the programme. The median number of pigs finished per year was 2000 (IQR: 800–3700). In total, 5938 herds (61%) kept no sows, 1752 (18%) kept some and 2045 (21%) kept many. A total of 7107 herds (73%) were of conventional health status, 586 (6%) of SPF status and 2042 (21%) of SPF with *Mycoplasma*. Finally, 978 herds (10%) were from Sonderjylland.

7.4.2 Model development for the sampling schemes

All predictors were significant in the simple logistic regression model developed in R. The results of the Bayesian model using all these predictors are shown in Table 7.1. Compared

with pigs from conventional health status herds, pigs from SPF health status and SPF-*Mycoplasma* status herds had 0.69 (95% CI: 0.66–0.72) and 0.93 (95% CI: 0.91–0.96) times the odds of being *Salmonella* positive, respectively. Compared with herds having 1 to 125 sows, having none or more than 125 sows increased the odds of a pig being *Salmonella* positive by a factor of 1.33 (95% CI: 1.28–1.38) and 1.36 (95% CI: 1.32–1.41), respectively. Compared with farms located outside of Sonderjylland, the odds of pigs being *Salmonella* positive on farms within Sonderjylland was increased by a factor of 1.32 (95% CI: 1.28–1.36).

Estimated coefficients for the ZIB model are shown in Tables 7.2 and 7.3. Table 7.2 shows the factors included in the zero-inflated part of the model; these are interpreted as factors associated with the probability of a herd being infected. A herd producing less than 2000 (small), or greater than 5000 (large) pigs for slaughter per year had a 1.58 (95% CI: 1.18–2.11) and 2.08 (95% CI: 1.42–3.14) greater odds of infection with *Salmonella*, respectively, compared with herds producing between 2000 and 5000 (medium) pigs per year for slaughter. Compared with herds within farms located outside of Sonderjylland, the odds of a Sonderjylland herd being infected with *Salmonella* was decreased by a factor of 0.25 (95% CI: 0.19–0.33).

Table 7.3 shows the model results for the binomial part of the ZIB model; these are interpreted as variables associated with the level of seropositivity in a herd, given that the herd is infected. The odds of a pig being seropositive in an infected small or large herd was increased by a factor of 1.16 (95% CI 1.06-1.28) compared with a pig being seropositive in an infected medium herd. The remaining results were similar to those provided in Table 7.1 for the logistic regression model.

The ZIB model was insensitive to changes in the precision parameter of the prior distribution assigned to A_i and B_i. The zero-inflated part of the model showed a five-fold increase in the value of the posterior standard deviation when compared with the binomial part of the model.

As we planned to use this model, based on 2003 data, to predict the probability of infection and seropositivity in 2004, we checked for consistency between the two years. This was thought to be important, because substantial changes in pig- and herd-level risks for infection (arising from, for example, changes in herd size or changes in the price of feed) from one year to the next could reduce the ability of the 2003 model to predict herd-level

behaviour in 2004. The magnitude and sign of the regression coefficients for 2003 and 2004 were compared. There was no change in sign of the estimated regression coefficients for each year. The two alpha coefficients (for the variables SPF health status and many sows) showed minor changes in magnitude between years with overall conclusions remained unchanged. For example, the alpha coefficient for SPF health status changed from 0.56 in 2003 to 0.36 in 2004. The beta coefficients (for variables associated with seropositivity, given infection) were similar between years.

The 8151 random farm-level effects for the two years were compared using scatter-plots and the relationship between them quantified using Pearson's product-moment correlation coefficient. Figure 7.4 shows the random effect terms A_i from the zero-inflated part of the ZIB model from 2004 as a function of those from 2003. There are four points to note: (1) there is moderate positive correlation between years (cor = 0.18, 95% CI: 0.16–0.20); (2) there is a skew towards the bottom left of the plot indicating a larger variance in the negative than in the positive valued random farm-level effects; this is likely to be due to the positive value for the intercept, α_0 (2.36, 95% CI: 2.00–2.74); (3) the random effect terms are tightly clustered around zero indicating that that the contribution of herd-level variation to the outcome was small; and (4) the cruciate-shaped pattern allows one to visualise the change in the contribution of unmeasured herd-level factors over the two year period. For example, a point positioned in the bottom right quadrant of Figure 7.4 would represent a herd with a positive random effect term in 2003 changing to negative in 2004. Given the measured herd-level factors (i.e. the fixed effects) remained the same, this would reflect a change in the influence of unmeasured herd-level factors on the herd's probability of being infected with *Salmonella*.

Figure 7.5 shows the random effect terms B_i from the seropositivity part of the ZIB model. Here, there is stronger consistency between years compared with Figure 7.4 (cor = 0.52, 95% CI: 0.51–0.54), particularly in the herds with a seroprevalence lower than what the fixed effects part of the model alone would predict (negative random effects). Secondly, there is a skew towards the top right quadrant of the plot indicating a larger variance in the positive than in the negative-valued random farm effects. This is likely to be due to the negative value for the β_0 coefficient (-3.33, 95% CI: -3.41 – -3.24). Thirdly, the random effect terms are less tightly clustered around zero than they are in Figure 7.4. This indicates that that the contribution of unmeasured herd-level variation to the outcome was

large, relative to the fixed effects component of the model.

A scatter plot of the median conditional seroprevalence q_i as a function of the median probability of infection ρ_i for 2003 is shown in Figure 7.3. There is a partial distinction in predicted seroprevalence between herds that were detected as positive (red open circles) and those that were not (green open circles). This provided us with our cut-off threshold of 0.09 predicted seroprevalence for the sampling schemes. This plot suggests that many of the herds where the *Salmonella* has not been detected were actually infected but at a low prevalence.

7.4.3 Comparison of the sampling schemes

Table 7.4 shows the performance of each of the sampling schemes. The scheme with the lowest cost was MRBA; the one with the highest cost was OHS. The one with the lowest number of false negatives and highest sensitivity was OHS and the one with the highest number of false negatives and lowest sensitivity was MRBA and MRBB. The scheme that reported the largest number of high positive farms was MRBA and MRBB.

Table 7.1: Results of a logistic regression model showing factors associated with *Salmonella* seropositivity in 578,260 meat-juice ELISA results taken from 9735 Danish finisher herds in 2003 as a part of the national surveillance and control programme.

Variable	Level	Posterior Mean	Posterior SD	MC error	OR (95% CI)
Intercept	-	-2.88	0.01	$\leq$0.001	-
Herd size[a]	Continuous	$2.9*10^{-2}$	0	$\leq$0.001	1.02(1.01–1.03)
Health Status	Conventional	Reference	-	-	-
	SPF	-0.37	0.02	$\leq$0.001	0.69(0.66–0.72)[b]
	SPF (with *Mycoplasma*)	-0.07	0.01	$\leq$0.001	0.93(0.91–0.96)
Sow Status	No sows	0.31	0.02	$\leq$0.001	1.33(1.28–1.38)
	Some sows (1-125)	Reference	-	-	-
	Many sows (>125)	0.28	0.02	$\leq$0.001	1.36(1.32–1.41)
Sonderjylland	No	Reference	-	-	-
	Yes	0.28	0.02	$\leq$0.001	1.32(1.28–1.36)

Model Statistics: Intercept, -2.88; DIC, 8689.65.
SD: Standard deviation; CI: Bayesian credible interval; MC error: Monte Carlo standard error of the posterior mean; OR: odds ratio
[a] Number of finishers produced (rescaled by subtracting the minimum, then dividing by 1000).
[b] Interpretation: Once adjusted for herd size, number of sows, and location , a pig on a farm with SPF health status had 0.69 times the odds of being *Salmonella* positive compared with a pig on a farm with conventional health status (95%CI: 0.66-0.72).

Table 7.2: Zero-inflated binomial model output showing factors associated with *Salmonella* infection status in 8151 Danish finisher herds in 2003 as a part of the national surveillance and control programme.

Variable	Level	Posterior Mean	Posterior SD	MC error	OR (95% CI)
Intercept	-	2.36	0.19	0.008	-
Herd size	Small	0.46	0.15	0.004	1.58(1.18–2.11)
	Medium	Reference	-	-	-
	Large	0.73	0.21	0.005	2.08(1.42–3.14)
Health Status	Conventional	Reference	-	-	-
	SPF	0.56	0.46	0.011	1.67(0.85–5.02)
	SPF (with *Mycoplasma*)	-0.14	0.16	0.003	0.87(0.63–1.18)
Sow Status	None	0.14	0.21	0.008	1.15(0.75–1.71)
	Some	Reference	-	-	-
	Many	0.05	0.24	0.009	1.04(0.64–1.67)
Sonderjylland	No	Reference	-	-	-
	Yes	-1.38	0.14	0.003	0.25(0.19–0.33)[a]

SD: Standard deviation; CI: Bayesian credible interval; MC error: Monte Carlo standard error of the posterior mean; OR: odds ratio

[a] Interpretation: Once adjusted for herd size, number of sows, and herd health status, a farm located in Sonderjylland had 0.25 times the odds of being *Salmonella* positive compared with a farm located elsewhere (95%CI: 0.19-0.33).

Table 7.3: Zero-inflated binomial model output showing factors associated with *Salmonella* seropositivity in 8151 Danish finisher herds in 2003 as a part of the national surveillance and control programme.

Variable	Level	Posterior Mean	Posterior SD	MC error	OR (95% CI)
Intercept	-	-3.33	0.04	0.002	-
Herd size	Small	0.15	0.04	0.002	1.16(1.08–1.24)
	Medium	Reference	-	-	-
	Large	0.15	0.05	0.002	1.16(1.06–1.28)
Health Status	Conventional	Reference	-	-	-
	SPF	-0.37	0.07	0.002	0.69(0.61–0.78)
	SPF (with *Mycoplasma*)	-0.08	0.04	0.001	0.92(0.85–0.99)
Sow Status	None	0.28	0.05	0.003	1.32(1.20–1.45)
	Some	Reference	-	-	-
	Many	0.28	0.06	0.003	1.33(1.19–1.48)
Sonderjylland	No	Reference	-	-	-
	Yes	0.52	0.05	0.002	1.68(1.51–1.86)[a]

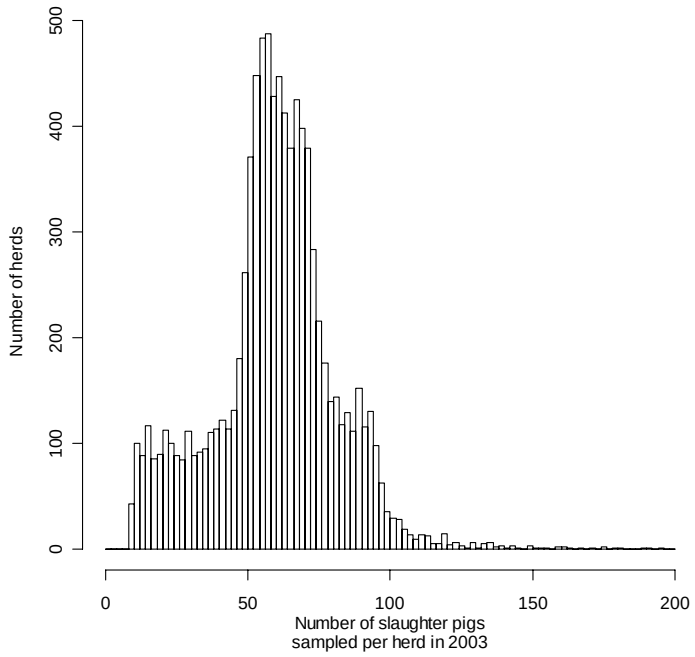
SD: Standard deviation; CI: Bayesian credible interval; MC error: Monte Carlo standard error of the posterior mean; OR: odds ratio

[a] Interpretation: Once adjusted for herd size, number of sows and herd health status , a pig on a farm located in Sonderjylland had 1.68 times the odds of being *Salmonella* positive compared with a pig on a farm located elsewhere (95%CI: 1.51–1.86).

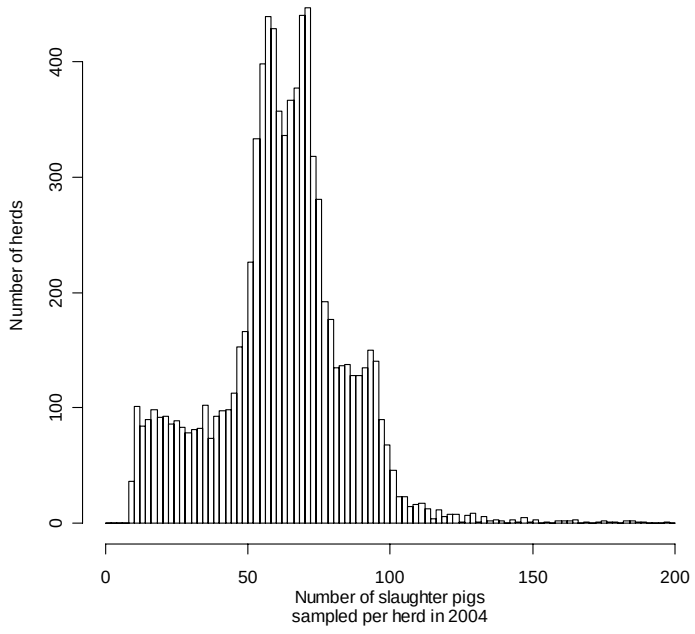
Table 7.4: Performance of four sampling schemes for surveillance for *Salmonella* in Danish finisher herds in 2004, $n = 8151$ herds

Sampling scheme	**OHS**	**DRB**	**MRBA**	**MRBB**
Number of false negative farms[a]	731	1186	3257	3251
Sensitivity	0.91	0.85	0.60	0.60
Number of high positive farms[b]	304	849	1148	1199
Cost of sampling scheme (€1000)	1,118	959	372	479

[a] Farms infected in 2004 with *Salmonella* but not detected by the sampling scheme
[b] Farms the sampling scheme has detected at a *Salmonella* seroprevalence of ≥ 0.40

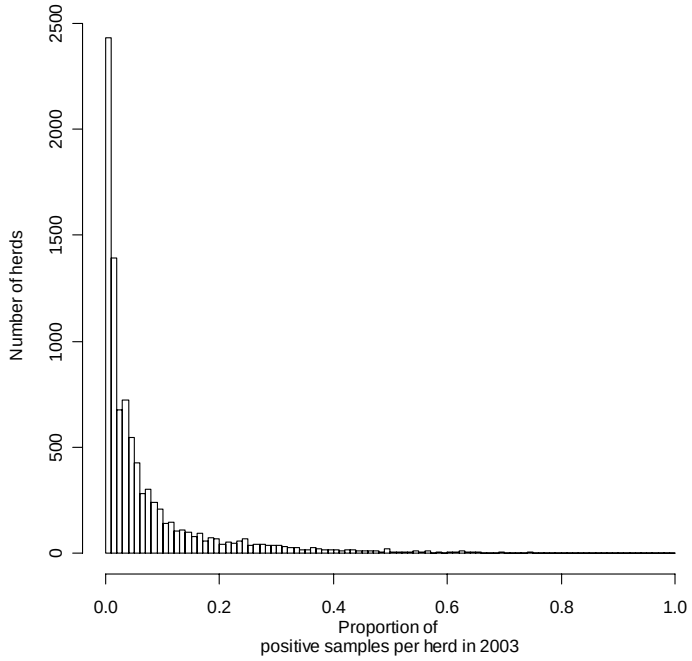


(a) 2003 histogram

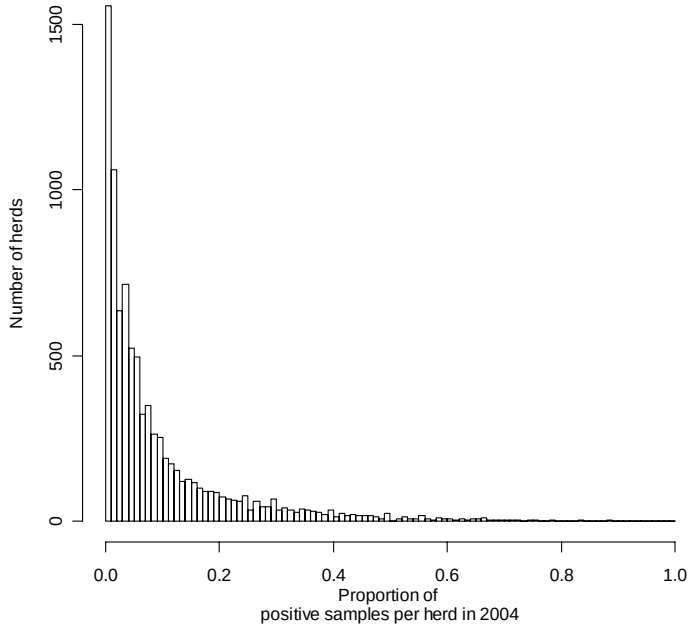


(b) 2004 histogram

Figure 7.1: The distribution of the number of pigs sampled for *Salmonella* per herd in (a) 2003 (five outliers removed) and (b) 2004 (seven outliers removed). Data are from the Danish swine *Salmonella* surveillance and control programme.



(a) 2003 histogram



(b) 2004 histogram

Figure 7.2: The distribution of the actual within-herd prevalence of *Salmonella* per herd in (a) 2003 and (b) 2004. Data are from the Danish swine *Salmonella* surveillance and control programme.

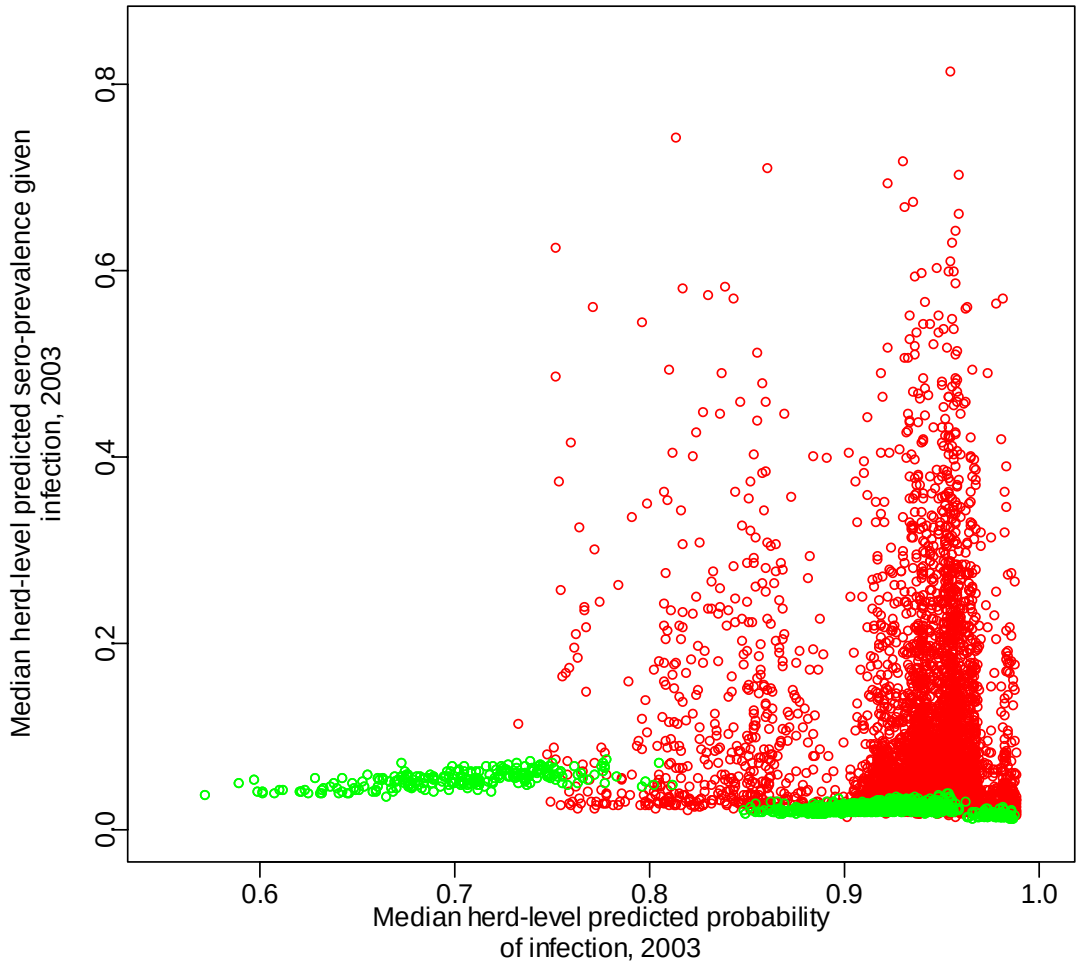


Figure 7.3: Scatter plot of median predicted seroprevalence as a function of median predicted probability of infection derived from a zero-inflated binomial model. Farms with at least one positive sample detected in are represented by the red open circles. Those with no positive samples detected are represented by the green open circles. Data are from 8151 farms sampled in the Danish swine *Salmonella* surveillance and control programme in 2003.

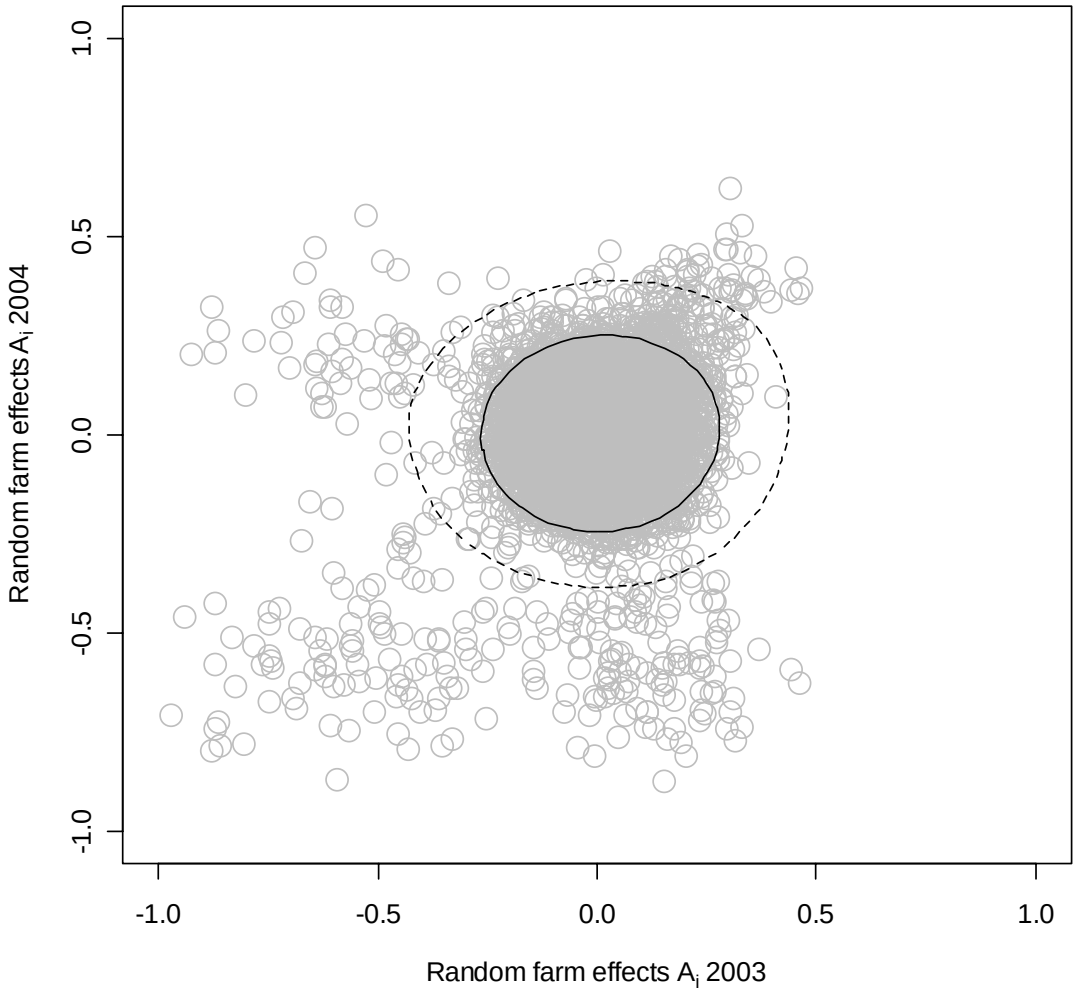


Figure 7.4: Scatter plot of the random farm effects A_i in 2003 as a function of those from 2004 (from the zero-inflated part of the ZIB model). Data are from 8151 farms sampled in the Danish swine *Salmonella* surveillance and control programme in 2003 and 2004. The solid and dashed lines represent the upper 10th and 25th percentiles of the smoothed density of the points.

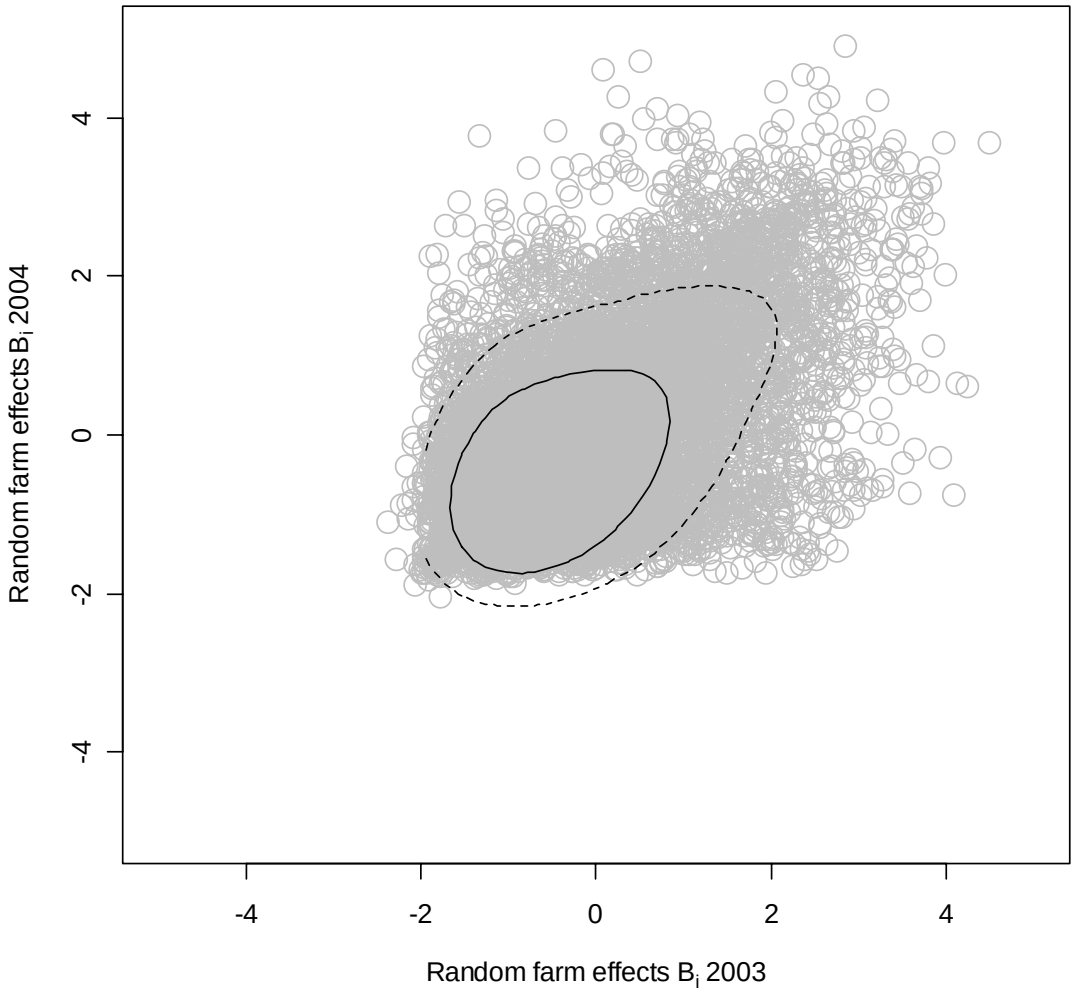


Figure 7.5: Scatter plot of the random farm effects B_i in 2003 as a function of those from 2004 (from the seropositivity part of the ZIB model). Data are from 8151 farms sampled in the Danish swine *Salmonella* surveillance and control programme in 2003 and 2004. The solid and dashed lines represent the upper 10th and 25th percentiles of the smoothed density of the points.

7.5 Discussion

We report on the use of a zero-inflated binomial model to investigate the performance of alternative sampling strategies for zoonotic *Salmonella*. To the best of our knowledge practical applications of this technique in the veterinary literature are scarce. Reports of the counterparts of these techniques for count data, zero-inflated Poisson and zero-inflated negative binomial models are more numerous. In Indonesia, Cheung (2006) used a ZIB model in the regression analysis of the cognitive function of Indonesian children. In an ecological context Martin et al. (2005) discuss the use of zero-inflated binomial and Poisson models in situations where both true and false zeroes occur. As far as we are aware, this is the first application of zero-inflated binomial modelling to endemic disease surveillance; there is potential to use this approach in the design of other surveillance systems.

Our model is based on our earlier work in Chapters 3 and 5 that identified farm- and herd-level risk factors for *Salmonella* status (Benschop et al. 2008*a*,*b*) in Danish pig herds. Our proposed sampling strategies were not based on season as we found this was not an associated risk factor based on work from Chapter 6 (Benschop et al. 2008*c*). However, it is important to consider seasonality in a targeted surveillance strategy, as time of the year typically has an effect on the incidence of infectious disease. For example, the pattern of human cases of salmonellosis consistently reports a late summer-early autumn peak in Denmark (Hald & Andersen 2001) and in Canada (Guerin et al. 2005*a*). This seasonality may be due to both direct and indirect effects of climate. The effect of seasonal effects has modified surveillance strategies for infectious diseases. In Denmark, for example, it has been recommended that sampling for infectious bovine rhinotracheitis (IBR) occurs primarily during the winter months of the year (Chriel et al. 2005). Human influenza surveillance in New Zealand using sentinel practices operates only during the winter months, from May to September (Population and Environmental Health Group 2008).

One variable, herd size, was a significant risk factor for both infection with *Salmonella* and subsequent seroprevalence. When compared with medium size herds (producing 2000 to 5000 pigs for slaughter per annum), both large (producing greater than 5000 pigs) and small (producing less than 2000 pigs) size herds were at greater risk. Keeping

none or greater than 125 sows was a risk factor when compared with keeping 1–125 in the seroprevalence model only. It is likely that the results for these two variables act through mechanisms such as buying in and mixing of pigs, number of visitors, biosecurity measures, feeding systems, and other management factors (Leontides et al. 2003, Farzan et al. 2006).

When compared with being located elsewhere, being located in Sonderjylland significantly decreased the odds of a herd being infected with *Salmonella* but significantly increased the odds of a pig from an infected herd being *Salmonella* positive. There is no convenient explanation for this seeming paradox, but it may be related to the herd demographics within this region. This region forms a border with Germany and there are two distinct types of farms present: the small family-owned and more traditional operations and larger modern premises, particularly on the island of Als. Earlier work reported in Chapter 4 has shown that farms in this region (compared with all other regions) showed the most variation in farm level prevalence of *Salmonella* (Benschop et al. 2006).

The use of the ZIB model has the potential to allow assessment of the extent of the similarities and differences between factors that affect herd infection status (introduction) and those that affect the seroprevalence in infected herds (persistence and spread). One can think of introduction as issues pertaining to with external biosecurity such as rodent control, number, and type of suppliers, and visitor policy, while persistence and spread falls under internal biosecurity such as type of partitions between pens (Lo Fo Wong et al. 2004*a*, Bollaerts et al. 2008) and use of an all-in-all-out production system vs. a continuous one (Belœil et al. 2004, Lurette et al. 2008). Figure 7.3 would suggest that many of the farms where the disease has not been observed are actually infected but with low prevalence. We predicted no farms had less than a 55% probability of infection.

We chose to use the same set of risk factors in both parts of the ZIB model and report all findings including those where the 95% Bayesian credible intervals overlapped the null value. We believe that our framework for zero-inflated modelling has provided a useful starting point for further exploration of the technique but at this stage we must be cautious not to over-interpret the results. With regard to risk factors for seroprevalence our results (Table 7.3) were very similar to those found when a logistic regression model alone (Table 7.1) was used. Although this may call into question our decision to use a ZIB model it is an interesting finding to be explored in the future.

7.5.1 A discussion of sampling

There is a a significant amount of literature that has preceded this work in the determination of sampling strategy as the DSSCP has evolved to where it is today (Mousing et al. 1997, Nielsen et al. 2001, Alban et al. 2002, Ekeroth et al. 2003, Enoe et al. 2003). The quandary remains that on the one hand if too few samples are taken then there is less chance of detecting a positive in an infected herd, however, if on the other hand, small numbers of samples may make it too easy for a herd to reach a cut-off proportion purely by chance. The current risk-based system seems a reasonable compromise in that only 12 samples a year are taken from herds that have been consistently negative (Ministry of Family and Consumer Affairs 2006). Once a positive is detected, herds return to the higher intensity herd-size based system with 60, 75 or 100 samples taken before a threshold criterion is applied. Our model supports that decision by providing reassurance that even if a non-detected farm is infected, it will most likely be infected at a low seroprevalence (Figure 7.3).

We present only a few sampling schemes, as this chapter's primary focus is the development of the concept rather than fine-tuning for the best sampling scheme. As expected, the scheme with the lowest cost (€372,000) was MRBA; the next lowest at almost one third as much again was MRBB at €479,000. It is important to reiterate that for each scenario these are the direct costs associated only with sampling meat-juice at the abattoir; no follow-on costs associated with on farm testing, hygienic slaughter or carcass downgrading are included. A full-cost benefit analysis is beyond the scope of this chapter but would be important groundwork prior to implementation of a change in the sampling regime.

The schemes with the lowest sensitivity were also MRBA and MRBB. We defined false negatives as farms that were infected in 2004 but had no positives detected by the sampling scheme. Both model base sampling schemes had 39% of farms falling into this category. If the aim of the scheme is to detect every infected herd then these schemes perform poorly. However the aim of the scheme is to identify herds with a high seroprevalence, and then enforce certain requirements on these herds, The detection of every infected herd is not an aim. This work would suggest that there are very few herds that are not infected and most herds have a predicted seroprevalence below 40% (see Figure 7.3).

We have fitted the model at one point in time, using accumulated data from 2003 to determine sampling for 2004. This model could be updated on a monthly basis allowing incorporation of the latest meat-juice results, and allowing for dynamic changes in covariates as herds increase in size, no longer keep sows, or enter an SPF health status, to take several examples. The development of model refinements, sampling schemes, and the methods to evaluate their performance are important areas for future work and would make the best use of this new tool. This should continue to occur in direct consultation with Danish authorities.

7.5.2 A discussion of future work

If we want to implement this type of targeted surveillance, we recommend gathering more data to inform these models. For example, the type of feed used has been found to be significantly associated with *Salmonella* status (Farzan et al. 2006, Hautekiet et al. 2008). This type of information may result in an increased sensitivity of the model based sampling strategy. Current registry databases such as the CHR should be advised to broaden the type of baseline information gathered to facilitate this. To assist in selecting the type of baseline information that would prove useful, a nested case-control study within the Danish cohort could be used to more precisely identify risk factors.

The incorporation of spatial autocorrelation into the model is a challenging and important area to consider. We have previously identified in Chapters 4 and 5 that there is spatial dependency between farms with regard to *Salmonella* status up to distances of approximately four km (Benschop et al. 2006, 2008*a*). A first step along this road would be to examine the model residuals for spatial autocorrelation. We propose that farms within this radius of level 2 or 3 herds be preferentially sampled. This would require careful consideration and be analytically complex, as the current allocation of herds into levels is continuous, i.e. a herd can be in level 2 one month then back to normal, level 1, in the next.

As well as the spatial dependency, we would like to extend the model further. We believe that it is likely a herd's infection probability and its seroprevalence are highly correlated. A way to express this is to use a multivariate normal distribution for the random effects A_i and B_i. It seems reasonable to assume that herds that are frequently infected have a

higher seroprevalence than herds that are infected only occasionally.

7.5.3 A discussion of bias, confounding, and chance

In 2003, the laboratory processing the meat-juice samples realised it was experiencing technical problems. These started in 2002, and were resolved in September 2003 (Bak et al. 2007). The cause was irregularities in the automatic microtitre-plate washing machine, which resulted in artificially low OD percentages in one corner of the microtitre plates. This resulted in fewer samples testing positive than were truly positive; a misclassification of the outcome of our model. However, this misclassification bias was almost certainly independent of the exposure and so any resulting bias would be towards the null (Sackett 1979). Another source of non-differential misclassification bias is that we used tests with imperfect sensitivity and specificity.

By using only herds that had 10 or more meat-juice samples taken, we excluded herds both new to the programme and those going out of production. It is feasible that herds in both these scenarios may be more likely to have a different *Salmonella* status than other herds resulting in differential selection bias.

Our use of random farm effects should have gone most of the way towards dealing with the issues of unmeasured variables such as the effect of the experience of the herd owner. The confounder of most importance is likely to be the effect of the penalty system on *Salmonella* status. Owners are highly motivated to avoid entry into levels 2 or 3, and if they do enter they try to leave as soon as possible as there are considerable costs associated with these classifications. This is one of the key reasons why predicting how a herd will perform based on the previous years' performance is complex.

CHAPTER 8

General Discussion

8.1 Introduction

This thesis has taken data from the first ten years of the Danish swine *Salmonella* surveillance and control programme (DSSCP) and used spatial, temporal, and risk factor analysis to develop methods for optimising the surveillance strategy. The first study (Chapter 3) developed a novel method of spatially adaptive smoothing to describe the spatial epidemiological features of the results from the first ten years of the programme (1995–2004). The conditional probability of a farm being a case was consistently high in the the south-west of Sonderjylland on the Jutland peninsula, identifying this area for further investigation and targeted surveillance. The identification of clustering of case farms informed the next study, described in Chapter 4.

Chapter 4 is an investigation of the patterns of spatial dependency in the data from 2003. K-function analyses provided evidence for aggregation of *Salmonella* case farms over that of all farms at distances of up to six kilometres. Semivariogram analyses of the random farm-level effects from a Bayesian logistic regression model (adjusted for herd size) of *Salmonella* seropositivity identified spatial dependency between pairs of farms up to a distance of four kilometres apart. The strength of the spatial dependency was positively associated with slaughter pig farm density. The study described how this might inform the surveillance programme by potentially targeting herds within a 4 kilometre radius of level 2 or 3 herds.

In the third study (Chapter 5), farm location details, routinely recorded surveillance information, and industry survey data were combined to build a Bayesian seroprevalence

model. This identified wet-feeding and specific pathogen free herd health status as protective factors for *Salmonella* seropositivity, while purchasing feed was a risk factor. After adjusting for these covariates, pockets of unexplained risk for *Salmonella* seropositivity were identified, and spatial dependency was found at distances of up to 6 km (95% CI: 2–35 km) between farms. A generalised linear spatial model was fitted to the Jutland data allowing formal estimation of the range of spatial correlation and a measure of the uncertainty around it. There was a large within-farm component to the variance, suggesting that gathering more farm level information would be advantageous if this approach was to be used to target surveillance strategy.

Chapter 6 again considers data from the whole study period, 1995 to 2004. A detailed temporal analysis of the data revealed there was no consistent seasonal pattern, and correspondingly no benefit in targeting sampling to particular times of the year. Spatio-temporal analyses suggested a local epidemic of increased seroprevalence occured in west Jutland in late 2000. Lorelogram analyses showed a defined period of statistically significant temporal dependency, suggesting that there is little value in sampling more frequently than every 10 weeks on the average farm.

The final study (Chapter 7) uses findings from the previous chapters to develop a zero-inflated binomial (ZIB) model to assist in the development of a risk-based sampling approach. The ZIB model is a useful tool when considering both risk factors for introduction of *Salmonella* between herds, and those for subsequent spread once introduced within a herd.

This thesis began with a literature review of change and emergence of diseases and how they present challenges for surveillance. There are two major changes within Denmark around the surveillance for zoonotic *Salmonella* that are worthy of mention. They will now be discussed in this concluding chapter of the thesis.

8.1.1 From on-farm interventions to interventions at the slaughterhouse

During the life of the Danish swine *Salmonella* surveillance and control programme there has been a change an increase in the application of interventions at the slaughterhouse

whilst maintaining those occurring on-farm. The focus is currently on post-harvest initiatives with increasing attention given to decontamination after slaughter and to surveillance cost-effectiveness.

A study by Alban & Stärk (2005) concluded that focussing solely on primary production, at the expense of the rest of the farm-to-fork continuum, would be an economically inefficient approach to reducing *Salmonella* in Danish pork further. Danish farmers have done much towards reducing *Salmonella* since the start of the programme such as increasing biosecurity and using wet or coarse feeding. The current system puts a constant pressure on the herds with highest *Salmonella* prevalence by the use of penalty fees (Nielsen et al. 2001). Other studies have also found that interventions at the slaughterhouse, rather than on farm, had the highest impact on the number of contaminated carcasses (van der Gaag et al. 2004*b*, Swanenburg et al. 2001*b*). The Danish Meat Association has an intensified *Salmonella* programme at slaughterhouses, the objective of which is to single out slaughterhouses that, over time, have a high *Salmonella* prevalence on individual carcasses.

A recent retrospective study of the Danish programme (Hurd et al. 2008) demonstrated that, except for the first few years, the on-farm programme had minimal impact in reducing the number of positive carcasses and pork attributed human cases. Prospective scenarios out to 2013 showed a similar result: on-farm efforts aimed at *Salmonella* reduction will not markedly improve public health. The Hurd et al. (2008) study showed that carcass decontamination was the most effective means of reducing human risk. Currently, hot-water decontamination is used by one Danish processing company for finishers from all level-3 *Salmonella* herds and those from herds positive for *Salmonella typhimurium* DT104.

8.1.2 The change in human cases of salmonellosis in Denmark

Throughout this thesis, I have made reference to the decline in estimated number of cases of salmonellosis in humans in Denmark attributable to pork consumption. These have decreased from 1444 in 1993 to 215 in 2005 (Nielsen et al. 2001, Ministry of Family and Consumer Affairs 2006). At the time of writing (December 2008) there is a large and sustained outbreak of human salmonellosis due to *Salmonella enterica* serotype Typhimurium phage type U292 occurring in Denmark. This was first detected in April 2008

(Ethelberg et al. 2008*a*), and a total of 1158 cases have been identified to 3rd December 2008, making it the largest outbreak recorded in Denmark since the present surveillance system became active in 1980 (Ethelberg et al. 2008*b*). Proposing of a further surveillance reduction strategy in the face of this epidemic, will be challenging.

The 2008 outbreak serves to remind us that we are dealing with a very dynamic system. There are many examples of *Salmonella* serotypes emerging to take up new niches (Callaway et al. 2008). New Zealand currently is in the midst of an outbreak of *Salmonella enterica* serotype Typhimurium phage type 42 (ProMED-mail 2008*c*). To January 14 2009 this outbreak has affected 68 people and has been linked to contaminated flour.

8.2 Lessons learnt

8.2.1 The value of multi-disciplinary collaboration

Working closely with a mathematician has led to the development of a predictive model for *Salmonella* surveillance in Denmark that could be applied to other surveillance systems (Chapter 7). Working closely with statisticians has led to the development of the novel adaptive smoothing technique (Chapter 3). This technique will soon be available as a package 'adsmooth', to be used within the software R (R Development Core Team 2007). Two recent Belgian publications on *Salmonella* in pig herds further reflect the benefits of similar inter-disciplinary collaborations. Hautekiet et al. (2008) have developed a sanitary risk index using weighted risk factors to classify high risk herds to form a basis for targeted sampling. Another Belgian group used semi-parametric quantile regression with P-splines to propose an alternative method to identify pig herds with a high *Salmonella* infection burden based on their previous serological data (Bollaerts et al. 2008).

Work within this thesis has also used techniques that have traditionally been used in other disciplines, and applied them to surveillance data. These include the inhomogeneous K-function in Chapter 4, the lorelogram in Chapter 6 and the zero-inflated binomial model in Chapter 7.

A highly technical approach can lead to impractical solutions, and it is important that theoretical advances are well grounded in biology, and are feasible. I have been fortunate in

working alongside highly-skilled people who deal with everyday problems and are keen to integrate their technical knowledge into practical situations. Furthermore, having a close working relationship with the veterinarians and scientists in the Danish Meat Association keeps the work centred on the need to stay practical.

There have been discussions as to whether Denmark would change from its current surveillance scheme to a targeted one, and it is most likely that Denmark will continue with the current programme. Despite some disadvantages, including expense, the current form of the DSSCP gives a direct measure of *Salmonella* in the herd, and Denmark is inclined towards having integrated surveillance schemes on a national scale. However, other countries and especially those new to the European Union that have no tradition for surveillance on a national scale may be more likely to embark on a risk-factor based sampling scheme. These countries are more likely to have fewer resources to commit to surveillance, and so a lower cost *Salmonella* surveillance programme could be appropriate.

8.2.2 The importance of data quality

I have been fortunate to have access to a very large data set that is well referenced in space and time. Every pig herd is required to register with the Danish Central Husbandry Register. This provides a unique identifier (the CHR number), with details of farm location given as the coordinates of the farm house. As the pig farms used in this work are all intensive production systems and small in land area, this use of the farm house location as a measure for the herd location is likely to be reasonable. However, this measure of location would be inappropriate if used to represent larger productive units such as high-country New Zealand sheep stations or cattle farms in the Australian Outback.

During the project, the Danish Meat Association (DMA) provided additional data for herds in 2003 and 2004. These included herd health status, number of sows, and herd size measured as the number of finishers produced by each herd in a year. These detailed data were used in Chapters 6 and 7. Metadata is information that describes the content, quality, condition, origin, and other characteristics of data, and is a key issue in the field of disease surveillance (Patridge & Namulanda 2008). Metadata about the DSSCP shows that the data is generally of high quality. For example, data on herd health status is from the Specific Pathogen Free Company, which is currently delivered through an on-line

connection. In 2003 and 2004, updates were monthly. Data for the number of finishers produced each year is delivered every 13 weeks in IBM compact disc format produced by the Zoonoses Register from the National Food Institute of Denmark. Data on the number of sows comes from a yearly report farmers make to the CHR. The sow numbers are not as precise as the other data, as farmers tend to use numbers from the previous year if they consider there has been no substantial change.

Chapter 5 used data from a questionnaire delivered to swine producers in 1995. There was missing covariate data in some of the responses to this questionnaire. This was compensated for by Bayesian imputation and comparing this with complete case analysis. In Chapters 3 and 5, there were missing spatial coordinates for some farms that were compensated for by randomly drawing easting and northing coordinates from within the boundaries of the farm's respective communes. The adjustment for herd size in the model used in Chapter 4 was based on the sample size taken, as at the time of that analysis the other data were not available.

In many other surveillance systems, data quality is poor. For example, Bisoffi (2008) reported that incidence rates of malaria in European travellers to Gambia are often difficult to determine. In this case there are both problems with the numerator and with the denominator. Many cases of malaria are not reported to authorities and it is difficult to enumerate the population at risk. This applies to both recognised groups of travellers to Gambia: the European who visit as tourists and also those visiting friends and relatives.

A recently complete PhD thesis within our research group has addressed issues around poor numerator and denominator data in poultry outbreaks of avian influenza in the Socialist Republic of Vietnam. In the case of the numerator, the surveillance and reporting intensity of outbreaks was thought to vary across the country (Lockhart 2008). A spatial zero-inflated Poisson model was used to assist in determining the risk factors for both outbreak detection and number of outbreaks at the commune level.

8.3 Future perspectives

8.3.1 Future work for these data

The studies presented in this thesis have identified a number of areas for future work.

An important area for further work would be to gather and analyse movement and contact information in this population. Social network analysis has its roots in human epidemiology and social sciences, but has recently been explored in a veterinary context (Westgarth et al. 2008, Brennan et al. 2008). The identification of high-risk farms through the measurement and analysis of contact networks may enable greater understanding of infection dynamics, and inform surveillance and control procedures (Christley et al. 2005). Movement of animals and animal product is a crucial control point in the spread of zoonotic disease, not only in the consideration of the recruitment of infected pigs into herds (Lo Fo Wong et al. 2004*a*) and their co-mingling during transport and lairage (Hurd et al. 2002), but also in the context of the partitioning and dissemination of infected food (Hodges & Kimball 2005).

The incorporation of spatial proximity as a risk factor to the predictive model is a challenging area for future work. Farms that perform poorly are currently penalised under the current system. The penalties are in place to encourage producers to make changes in herd management that will reduce *Salmonella* seroprevalence. Accordingly, the current allocation of herds into levels is continuous, i.e. a herd can be in level 2 one month and then be back to level 1 in the next. Targeting by spatial proximity may well be unacceptable to producers. For example, a farmer who has minimal risk factors and is performing well with respect to *Salmonella* would probably prefer to be tested directly under the current system than be penalised by location alone.

If we want to implement model derived risk-based surveillance it will be necessary to gather more data to inform these models. This type of information is likely to result in increased sensitivity, making these options more attractive to authorities and consumers. Current registry databases such as the CHR should be advised to broaden the type of baseline information gathered to facilitate this and provide incentives for producers to provide such data. To assist in selecting the type of baseline information that would prove useful, a nested case-control study within the Danish cohort could be used to more precisely identify risk factors.

8.3.2 Is risk-based sampling 'safe'?

The aim of risk-based surveillance is to provide the most sensitive means for disease surveillance in a cost effective manner. If we do apply risk-based sampling techniques, we need to be confident in the predictive value of the risk factors. It would be advisable to supplement targeted surveillance with standard surveys at a regular interval to maintain end user confidence. This should also assist in identifying new risk factors that may become important over time.

A current example of this is a concern that has recently emerged about variant Creutzfeldt-Jakob disease (vCJD) (ProMED-mail 2008*b*). Prior to December 2008, the risk factors for development of vCJD had included age, residence in the UK and methionine homozygosity (M/M) at codon 129 of the prion protein gene. Now a patient in a London hospital with a PRNP-129 methionine heterozygosity (M/V) genotype has been diagnosed clinically as a case of vCJD. If confirmed, this case may indicate that late onset vCJD is associated with this genotype, and that a new wave of vCJD may be imminent as a delayed consequence of the exposure of the UK population to BSE-contaminated meat.

8.3.3 Continual improvement of visualisation of surveillance data

In this thesis we developed a novel method of spatially adaptive smoothing and presented the results as static conditional probability surfaces. New ways of visualising surveillance data that present results in a readily-interpretable form are constantly being developed. For example, the use of movies, geovisual analytics and web-based mapping are becoming more widely used.

Vieira et al. (2008) created a movie to allow visualisation of the changes in magnitude and location of elevated breast cancer risk for the 40 years of residential history that was smoothed over space and time. This was achieved by application of a two-dimensional generalised additive model.

Geovisual analytics is a sub-area of the emerging research discipline of visual analytics, with specific focus on problems involving geographic phenomena (Keim et al. 2006). Geovisual analytics allows users to interactively explore visual representations of geographic information, tapping perceptual and cognitive abilities to recognise and process

patterns and outliers from a visual scene, link these patterns and outliers to existing knowledge bases, and arrive at an appropriate course of action given the visual input. This analytical approach has been applied to cervical cancer mortality data (Chen et al. 2008). Chen et al. apply their methods to cervical cancer mortality data for the United States between 2000 and 2004, and conclude that their proposed geovisual analytics approach compliments traditional statistical methods in cluster identification by enhancing the interpretation of identified clusters.

Electronic surveillance using web-based tools has proven to be of substantial value in reporting outbreaks of infectious disease. However, trying to pinpoint a potential outbreak and contain it before it spreads requires the constant surveillance of a continually growing number of disparate news sources and alert services. HealthMap, a surveillance tool developed by Clark Freifeld and John Brownstein, brings together disparate data sources to achieve a comprehensive view of the current global status of infectious diseases and their effect on human and animal health.[1]

8.3.4 Innovative surveillance

Surveillance for Johne's disease in New Zealand deer herds

An innovative surveillance system in is the Johne's Management Limited (JML) integrated system of feedback to the primary producer of information from the deer slaughter premises (Dr. Jaimie Glossop, personal communication). The information of predominant interest is lesion status, i.e. an enlarged mesenteric lymph node which has been found to be highly predictive of Johne's disease status. Animal and farm identifier, farm location, date of slaughter, and carcass weight information is also included. Every deer commercially slaughtered in New Zealand is included in the database, and it is expected that the system will be fully compatible with the recent National Animal Identification and Tracing System, to be introduced for deer in 2010. The deer industry in New Zealand is small, with approximately 500,000 deer slaughtered annually. This small size coupled with the strong producer board support are crucial to the success of this initiative. Monthly data is sent to producers facilitating strategic culling, breeding, and purchasing.

[1] http://www.healthmap.org/en

Plans are to analysis the data in space and time to produce reports of the prevalence of Johne's disease in slaughtered deer at the national, regional, and local levels.

Surveillance for 'one medicine'

More than 20 years ago, Calvin Schwabe coined the term 'one medicine', to focus attention on the similarity between human and veterinary health interests (Zinsstag et al. 2005). Today there is little doubt that veterinary medicine plays an essential role in protecting and promoting public health, especially in the prevention and control of zoonotic diseases (Sargeant 2008). Surveillance systems such as the global Salm-Surv project and ArboNet (Lemmings et al. 2006) include both human and animal disease, and the sampling of food, vectors, and environmental samples. Global issues such as the disruption of ecosystems, increased trade, and climate change will bring veterinarians, doctors, ecologists and social scientists together to respond to future challenges.

This complexity of natural phenomena is likely to result in increasing focus on the use of novel data sources. For example, the abundance and behaviour of wildlife may act as an early-warning system in the prediction of human disease. Pettersson et al. (2008) reports how the huddling behaviour of bank voles, influenced by lack of snow cover, may influence zoonotic disease transmission to humans in Finland. In a response to climatic conditions reservoir hosts may seek shelter or food closer to barns, houses, and other buildings, thereby increasing the exposure for the human population at risk.

Disease in pets or other animal 'sentinels' may also reflect disease in human populations. South Africa's human tuberculosis epidemic has jumped from humans to pets, zoo animals, and wildlife. The human strain of the disease has been found in springbok, mongooses, baboons, and chimpanzees (ProMED-mail 2008*d*). Worldwide, the human strain of the disease, *Mycobacterium tuberculosis*, is rare in animals, who are more commonly infected with *Mycobacterium bovis*.

A controversial area in future surveillance systems may be the use of humans as sentinels for animal or ecosystem disease. A study of avian influenza in Vietnam reports that human cases were more likely to be reported prior to outbreaks in poultry (Phan et al. 2008). The authors suggest this may have been due to delayed detection of clinical signs in poultry flocks, or enhanced detection and reporting of poultry outbreaks after the local emergence

of a human case. Similarly, Cook et al. (2004) suggest that human outbreak data can act as a pivotal warning system for ecosystem injury and to guide interventions to preserve both ecologic and human health. The outbreaks of hantaviruses in the Americas has acted as a bioindicator for the disruption of the local distribution of natural vegetation. The clearance and replacement of complex rainforest have encouraged a massive proliferation of small mammals that act as vectors for hantaviral diseases.

Molecular surveillance

Molecular technology continues to develop, and it is becoming increasingly inexpensive to determine the complete genome sequence of bacterial isolates. Genotypic methods have been developed which can distinguish numerous different strains of pathogenic bacteria, giving epidemiologists an unprecedented ability to differentiate between individual isolates. Typing and sub-typing are now essential tools for studying the microbiology and epidemiology of pathogen populations, and these tools are increasingly being used in infectious disease surveillance.

The fine typing of pathogens coupled with spatial information on cases has been used in the surveillance for meningococcal disease (Reinhardt et al. 2008). A consensus on molecular typing of meningococcal disease using variable regions of genes encoding immuno-dominant antigens has recently been reached in Europe (Jolley et al. 2007). Extending this tool to other countries with a functioning laboratory surveillance of meningococcal disease will be possible without changing typing attributes.

Work done in our research institute by combining typing information with spatial and epidemiological data has provided insight into transmission pathways for campylobacteriosis in the Manawatu region of New Zealand (French 2008). Multi-locus sequence typing was used. The poultry-strain associated human cases of campylobacteriosis were largely confined to urban dwellers. The ruminant-strain associated human cases were predominantly in rural dwellers, and in children and adults with an occupation that is likely to bring them into contact with ruminant faeces. This epidemiological information, when combined with the relative contamination levels on food products, suggests that poultry cases are likely to be acquired from food, whereas ruminant-associated cases are more likley to result from direct contact with animal faeces.

An important global health problem in the future and of today is the emergence of pathogens of heightened virulence, such as *Escherichia coli* O157 (Reid et al. 2000), and *Salmonella enteritidis* phage type-4, and the multidrug resistant phage type-DT104 (Callaway et al. 2008). There is evidence that this increase in virulence is through pathogens acquiring genome segments through lateral gene transfer that result in gain-of-function traits (Ochman et al. 2000). Stabler et al. (2008) report on the development and application of an Active Surveillance of Pathogens (ASP) microarray which represents known antibiotic resistance genes, virulence determinants, and pathogenicity traits from 151 bacteria species. Potential uses in surveillance include monitoring antimicrobial resistance, virulence profile, and gene flux in pathogens, along with potentially identifying gene acquisitions and new outbreak strains.

8.4 Conclusion

This thesis is concerned with the application of recently developed epidemiological and statistical tools to inform the optimisation of a national surveillance strategy of considerable importance to human health. Although data from the Danish *Salmonella* surveillance and control programme has been used in these investigations, the techniques may be readily applied to other surveillance data of similar quality. The challenges health professionals face in the future for zoonotic disease surveillance are likely to continue to expand as a result of a changing world. These changes may include increases in the disruption of ecosystems by development, globalisation of food and feed supply, changes in climate, and further disruption of human populations by conflict. Increased co-operation between veterinary and human health communities will be necessary to meet these challenges in public health. Inspiration can be taken from Calvin Schwabe, who in 1984 stated that *'Improved human health is the sole among veterinary medicine's several benefits to society that arises from virtually all of veterinarians' diverse activities.... There is now and always has been only one medicine'*, (cited by Sargeant (2008)).

Bibliography

Abrial, D., Calavas, D., Jarrige, N. & Ducrot, C. (2005), 'Spatial heterogeneity of the risk of BSE in France following the ban of meat and bone meal in cattle feed', *Preventive Veterinary Medicine* **67**(1), 69–82.

AIDS Epidemiology Group (2007), Unlinked anonymous study of HIV prevalence among attenders at sexual health clinics: 2005/06, report to the Ministry of Health, Technical report, University of Otago.

Alban, L. & Stärk, K. D. (2005), 'Where should the effort be put to reduce the *Salmonella* prevalence in the slaughtered swine carcass effectively?', *Preventive Veterinary Medicine* **68**(1), 63–79.

Alban, L., Boes, J., Kreiner, H., Petersen, J. V. & Willeberg, P. (2008), 'Towards a risk-based surveillance for *Trichinella* spp. in Danish pig production', *Preventive Veterinary Medicine* **87**(3-4), 340–357.

Alban, L., Stege, H. & Dahl, J. (2002), 'The new classification system for slaughter-pig herds in the Danish *Salmonella* surveillance-and-control program', *Preventive Veterinary Medicine* **53**(1-2), 133–146.

Allard, R. (1998), 'Use of time series analysis in infectious disease surveillance', *Bulletin of the World Health Organisation* **76**(4), 327–334.

Altizer, S., Dobson, A., Hosseini, P., Hudson, P., Pascual, M. & Rohani, P. (2006), 'Seasonality and the dynamics of infectious diseases', *Ecological Letters* **9**(4), 467–484.

Anonymous (2006), Annual report 2005, Technical report, National Committee for Pig Production, Copenhagen, Denmark.

Anyamba, A., Chretien, J. P., Small, J., Tucker, C. J. & Linthicum, K. J. (2006), 'Developing global climate anomalies suggest potential disease risks for 2006 - 2007', *International Journal of Health Geographics* **5**, 60.

Armstrong, D. (2003), 'Zoonoses Action Plan *Salmonella* monitoring programme update', *The Pig Journal*. Available as: `http://www.thepigsite.com/pigjournal/volume/52`. Accessed 22 December 2008.

Assunçao, R., Costa, M., Tavares, A. & Ferreira, S. (2006), 'Fast detection of arbitrarily shaped disease clusters', *Statistics in Medicine* **25**(5), 723–742.

Audigé, L., Doherr, M. G., Hauser, R. & Salman, M. D. (2001), 'Stochastic modelling as a tool for planning animal-health surveys and interpreting screening-test results', *Preventive Veterinary Medicine* **49**(1-2), 1–17.

Babin, S., Magruder, S., Hakre, S., Coberly, J. & Lombardo, J. (2007), Understanding the data: health indicators in disease surveillance, *in* J. Lombardo & D. Buckeridge, eds, 'Disease surveillance: a public health informatics approach', John Wiley and Sons, New Jersey, pp. 43–90.

Bach Knudsen, K. E. (2001), 'Development of antibiotic resistance and options to replace antimicrobials in animal diets', *Proceedings of the Nutritional Society* **60**(3), 291–299.

Baddeley, A. & Turner, R. (2005), 'Spatstat: an R package for analyzing spatial point patterns', *Journal of Statistical Software* **12**(6), 1–42.

Baddeley, A., Moller, J. & Waagepetersen, R. (2000), 'Non- and semi-parametric estimation of interaction in inhomogeneous point patterns', *Statistica Neerlandica* **54**(3), 329–350.

Baggesen, D. L. & Wegener, H. C. (1994), 'Phage types of *Salmonella enterica* serovar Typhimurium isolated from production animals and humans in Denmark', *Acta Veterinaria Scandinavica* **35**(4), 349–354.

Baggesen, D. L., Dahl, J., Wingstrand, A. & Nielsen, B. (1996*a*), Critical control points in pig herds in relation to subclinical *Salmonella* infection, *in* 'The 14th International Pig Veterinary Society Congress', Bologna, Italy.

Baggesen, D. L., Wegener, H. C., Bager, F., Stege, H. & Christensen, J. (1996*b*), 'Herd prevalence of *Salmonella enterica* infections in Danish slaughter pigs determined by microbiological testing', *Preventive Veterinary Medicine* **26**(3-4), 201–213.

Bak, H., Ekeroth, L. & Houe, H. (2007), 'Quality control using a multilevel logistic model for the Danish pig *Salmonella* surveillance antibody-ELISA programme', *Preventive Veterinary Medicine* **78**(2), 130–141.

Baker, M. G., Sneyd, E. & Wilson, N. A. (2007), 'Is the major increase in notified campylobacteriosis in New Zealand real?', *Epidemiology and Infection* **135**(1), 163–170.

Banerjee, S., Carlin, B. & Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Monographs on statistics and applied probability, Chapman and Hall/CRC, Florida.

Barber, D. A., Bahnson, P. B., Isaacson, R., Jones, C. J. & Weigel, R. M. (2002), 'Distribution of *Salmonella* in swine production ecosystems', *Journal of Food Protection* **65**(12), 1861–1868.

Baum, D., Harris, D., Nielsen, B. & Fedorka-Cray, P. (1998), Risk factors associated with increased seroprevalence of *Salmonella* in finishing swine, *in* 'The 15th International Pig Veterinary Society Congress', Birmingham, England.

Belœil, P., Fravalo, P., Fablet, C., Jolly, J., Eveno, E., Hascoet, Y., Chauvin, C., Salvat, G. & Madec, F. (2004), 'Risk factors for *Salmonella enterica* shedding by market-age pigs in French farrow-to-finish herds', *Preventive Veterinary Medicine* **63**(1-2), 103–120.

Benschop, J., Hazelton, M., Stevenson, M. A., Dahl, J., Morris, R. & French, N. P. (2008*a*), 'Descriptive spatial epidemiology of subclinical *Salmonella* infection in finisher pig herds: application of a novel method of spatially adaptive smoothing', *Veterinary Research* **39**, 02.

Benschop, J., Stevenson, M. A., Dahl, J. & French, N. P. (2008*b*), 'Towards incorporating spatial risk analysis for *Salmonella* seropositivity into the Danish swine surveillance programme', *Preventive Veterinary Medicine* **83**, 347–359.

Benschop, J., Stevenson, M. A., Dahl, J., Morris, R. & French, N. P. (2006), Descriptive spatio-temporal epidemiology of subclinical *Salmonella* infection in Danish finisher

pig herds, *in* '11th International Symposium on Veterinary Epidemiology and Economics', Cairns, Australia.

Benschop, J., Stevenson, M. A., Dahl, J., Morris, R. S. & French, N. P. (2008*c*), 'Temporal and longitudinal analysis of Danish swine salmonellosis control programme data: implications for surveillance', *Epidemiology and Infection* **136**, 1511–1520.

Berends, B. R., Urlings, H. A. P., Snijders, J. M. A. & Van Knapen, F. (1996), 'Identification and quantification of risk factors in animal management and transport regarding *Salmonella spp.* in pigs', *International Journal of Food Microbiology* **30**(1-2), 37–53.

Berends, B. R., Van Knapen, F., Mossel, D. A., Burt, S. A. & Snijders, J. M. (1998), 'Impact on human health of *Salmonella spp.* on pork in The Netherlands and the anticipated effects of some currently proposed control strategies', *International Journal of Food Microbiology* **44**(3), 219–229.

Besag, J. & Newell, J. (1991), 'The detection of clusters in rare diseases', *Journal of the Royal Statistical Society. Series A* **154**(1), 143–155.

Bi, P., Cameron, A. S., Zhang, Y. & Parton, K. A. (2008), 'Weather and notified *Campylobacter* infections in temperate and sub-tropical regions of Australia: an ecological study', *Journal of Infection* **57**(4), 317–323.

Biggeri, A., Bohning, D., Lesaffre, E., Viel, J., Lawson, A., Divino, F. & Frigessi, A. (1999), Introduction to spatial models in ecological analysis, *in* A. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J. Viel & R. Bertollini, eds, 'Disease Mapping and Risk Assessment for Public Health', John Wiley and Sons, Chichester, pp. 180–192.

Biggeri, A., Catelan, D., Rinaldi, L., Dreassi, E., Lagazio, C. & Cringoli, G. (2006*a*), 'Statistical methods for geographical surveillance in veterinary epidemiology', *Parassitologia* **48**(1-2), 73–76.

Biggeri, A., Dreassi, E., Catelan, D., Rinaldi, L., Lagazio, C. & Cringoli, G. (2006*b*), 'Disease mapping in veterinary epidemiology: a Bayesian geostatistical approach', *Statistical Methods in Medical Research* **15**(4), 337–352.

Bisoffi, Z. (2008), 'Malaria in travellers to Gambia', *Eurosurveillance* **13**(51), pii=19078.

Bithell, J. F. (1990), 'An application of density estimation to geographical epidemiology', *Statistics in Medicine* **9**(6), 691–701.

Bivand, R., Pebesma, E. & Gomez-Rubio, V. (2008), *Applied Spatial data analysis with R*, Use R!, Springer science and business media, New York.

Blaha, T. (2004), 'Up-to-date information from the German *QS Salmonella* monitoring and reduction programme', *Deutsche Tierarztliche Wochenschrifte* **111**(8), 324–326.

Bloom, M., Buckeridge, D. & Cheng, K. (2007), 'Finding leading indicators for disease outbreaks: Filtering, cross-correlation, and caveats', *Journal of the American Medical Informatics Association* **14**(1), 76–85.

Bohning, D. & Greiner, M. (2006), 'Evaluation of the cumulative evidence for freedom from BSE in birth cohorts', *European Journal of Epidemiology* **21**(1), 47–54.

Boklund, A., Alban, L., Mortensen, S. & Houe, H. (2004), 'Biosecurity in 116 Danish fattening swineherds: descriptive results and factor analysis', *Preventive Veterinary Medicine* **66**(1-4), 49–62.

Bollaerts, K., Aerts, M., Ribbens, S., Van der Stede, Y., Boone, I. & Mintiens, K. (2008), 'Identification of *Salmonella* high risk pig-herds in Belgium by using semiparametric quantile regression', *Journal Of The Royal Statistical Society Series A* **171**(2), 449–464.

Borch, E., Nesbakken, T. & Christensen, H. (1996), 'Hazard identification in swine slaughter with respect to foodborne bacteria', *International Journal of Food Microbiology* **30**(1-2), 9–25.

Botteldoorn, N., Heyndrickx, M., Rijpens, N., Grijspeerdt, K. & Herman, L. (2003), '*Salmonella* on pig carcasses: positive pigs and cross contamination in the slaughterhouse', *Journal of Applied Microbiology* **95**(5), 891–903.

Bowman, A. & Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis*, Oxford Statistical Sciences Series, Oxford University Press, Oxford.

Box, G. E. P., Jenkins, G. & Reinsel, G. (1994), *Time Series Analysis: Forecasting and Control*, third edn, Prentice Hall, Englewood Cliffs, New Jersey.

Bradley, R., Collee, J. G. & Liberski, P. P. (2006), 'Variant CJD (vCJD) and bovine spongiform encephalopathy (BSE): 10 and 20 years on: part 1', *Folia Neuropathologica* **44**(2), 93–101.

Brennan, M. L., Kemp, R. & Christley, R. M. (2008), 'Direct and indirect contacts between cattle farms in north-west England', *Preventive Veterinary Medicine* **84**(3-4), 242–260.

Broman, A. T., Shum, K., Munoz, B., Duncan, D. D. & West, S. K. (2006), 'Spatial clustering of ocular chlamydial infection over time following treatment, among households in a village in Tanzania', *Investigative Ophthalmology and Visual Science* **47**(1), 99–104.

Buckeridge, D. L. (2007), 'Outbreak detection through automated surveillance: a review of the determinants of detection', *Journal of Biomedical Informatics* **40**(4), 370–379.

Buehler, J. (2008), Surveillance, *in* K. Rothman, S. Greenland & T. Lash, eds, 'Modern Epidemiology', third edn, Lippincott Williams and Wilkins, Philadelphia, pp. 459–480.

Callaway, T. R., Edrington, T. S., Anderson, R. C., Byrd, J. A. & Nisbet, D. J. (2008), 'Gastrointestinal microbial ecology and the safety of our food supply as related to *Salmonella*', *Journal of Animal Science* **86**(14 Suppl), E163–E172.

Cambardella, C. A., Moorman, T. B., Novak, J. M., Parkin, T. B., Karlen, D. L., Turco, R. F. & Konopka, A. E. (1994), 'Field-scale variability of soil properties in central Iowa soils', *Soil Science Society of America Journal* **58**(5), 1501–1511.

Cameron, K. & Hunter, P. (2002), 'Using spatial models and kriging techniques to optimize long-term ground-water monitoring networks: a case study', *Environmetrics* **13**(5-6), 629–656.

Cannon, R. M. (2002), 'Demonstrating disease freedom—combining confidence levels', *Preventive Veterinary Medicine* **52**(3-4), 227–249.

Cardinal, M., Roy, R. & Lambert, J. (1999), 'On the application of integer-valued time series models for the analysis of disease incidence', *Statistics in Medicine* **18**(15), 2025–2039.

Carpenter, T. E. (2001), 'Methods to investigate spatial and temporal clustering in veterinary epidemiology', *Preventive Veterinary Medicine* **48**(4), 303–320.

Carstensen, B. & Christensen, J. (1998), 'Herd size and seroprevalence of *Salmonella enterica* in Danish swine herds: a random-effects model for register data', *Preventive Veterinary Medicine* **34**(2-3), 191–203.

Casey, P. G., Butler, D., Gardiner, G. E., Tangney, M., Simpson, P., Lawlor, P. G., Stanton, C., Ross, R. P., Hill, C. & Fitzgerald, G. F. (2004), '*Salmonella* carriage in an Irish pig herd: correlation between serological and bacteriological detection methods', *Journal of Food Protection* **67**(12), 2797–2800.

Centers for Disease Control and Prevention (2004), 'Syndromic surveillance: Reports from a national conference, 2003', *Morbidity and Mortality Weekly Report* **53**(Supplement), 1–268.

Centers for Disease Control and Prevention (2006), 'Ongoing multistate outbreak of *Escherichia coli* serotype O157:H7 infections associated with consumption of fresh spinach–United States, September 2006', *Morbidity and Mortality Weekly Report* **55**(38), 1045–1046.

Chambers, J., Cleveland, W., Kleiner, B. & Tukey, P. (1983), *Graphical methods for data analysis*, Wadsworth and L. Brooks, California.

Chatfield, C. (2004), *The Analysis of Time Series: An Introduction*, sixth edn, Chapman and Hall, London.

Checkley, W., Epstein, L. D., Gilman, R. H., Figueroa, D., Cama, R. I., Patz, J. A. & Black, R. E. (2000), 'Effect of El Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children', *Lancet* **355**(9202), 442–450.

Chen, D., Cane, M. A., Kaplan, A., Zebiak, S. E. & Huang, D. (2004), 'Predictability of El Niño over the past 148 years', *Nature* **428**(6984), 733–736.

Chen, J., Roth, R. E., Naito, A. T., Lengerich, E. J. & Maceachren, A. M. (2008), 'Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality', *International Journal of Health Geographics* **7**, 57.

Chetwynd, A. G. & Diggle, P. J. (1998), 'On estimating the reduced second moment measure of a stationary spatial point process', *Australian and New Zealand Journal of Statistics* **40**(1), 11–15.

Cheung, Y. B. (2006), 'Growth and cognitive function of Indonesian children: Zero-inflated proportion models', *Statistics in Medicine* **25**(17), 3011–3022.

Childs, D. Z., Cattadori, I. M., Suwonkerd, W., Prajakwong, S. & Boots, M. (2006), 'Spatiotemporal patterns of malaria incidence in northern Thailand', *Transactions of the Royal Society of Tropical Medicine and Hygiene* **100**(7), 623–631.

Choi, L., Dominici, F., Zeger, S. L. & Ouyang, P. (2005), 'Estimating treatment efficacy over time: a logistic regression model for binary longitudinal outcomes', *Statistics in Medicine* **24**(18), 2789–2805.

Chriel, M., Salman, M. D. & Wagner, B. A. (2005), 'Evaluation of surveillance and sample collection methods to document freedom from infectious bovine rhinotracheitis in cattle populations', *American Journal of Veterinary Research* **66**(12), 2149–2153.

Christensen, J. (2003), Danish swine salmonellosis control program: 1993 to 2001, *in* M. Salman, ed., 'Animal Disease Surveillance and Survey Systems', Iowa State Press, Iowa, pp. 185–207.

Christensen, J. & Rudemo, M. (1998), 'Multiple change-point analysis applied to the monitoring of *Salmonella* prevalence in Danish pigs and pork', *Preventive Veterinary Medicine* **36**(2), 131–143.

Christensen, O. & Ribeiro Jr., P. (2002), 'geoRglm: A package for generalised linear spatial models', *R-NEWS* **2**(2), 26–28.

Christley, R. M., Pinchbeck, G. L., Bowers, R. G., Clancy, D., French, N. P., Bennett, R. & Turner, J. (2005), 'Infection in social networks: using network analysis to identify high-risk individuals', *American Journal of Epidemiology* **162**(10), 1024–1031.

Clark, A. & Lawson, A. (2004), 'An evaluation of non-parametric relative risk estimators for disease maps', *Computational Statistics and Data Analysis* **47**(1), 63–78.

Clements, M. & Hendry, D. (2001), 'Forecasting with difference-stationary and trend-stationary models', *Econometrics Journal* **4**, S1–S19.

Cleveland, R., Cleveland, W., McRae, J. & Terpenning, I. (1990), 'STL: A seasonal-trend decomposition procedure based on loess', *Journal of Official Statistics* **6**(1), 3–33.

Cleveland, W. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Congdon, P. (2001), *Bayesian Statistical Modelling*, John Wiley and Sons, Chichester.

Cook, A., Jardine, A. & Weinstein, P. (2004), 'Using human disease outbreaks as a guide to multilevel ecosystem interventions', *Environmental Health Perspectives* **112**(11), 1143–1146.

Cooper, B. & Lipsitch, M. (2004), 'The analysis of hospital infection data using hidden Markov models', *Biostatistics* **5**(2), 223–237.

Corrigan, R., Waldner, C., Epp, T., Wright, J., Whitehead, S., Bangura, H., Young, E. & Townsend, H. (2006), 'Prediction of human cases of West Nile virus by equine cases, Saskatchewan, Canada, 2003', *Preventive Veterinary Medicine* **76**(3-4), 263–272.

Crighton, E. J., Moineddin, R., Mamdani, M. & Upshur, R. E. (2004), 'Influenza and pneumonia hospitalizations in Ontario: a time-series analysis', *Epidemiology and Infection* **132**(6), 1167–1174.

Crump, J. A., Murdoch, D. R. & Baker, M. G. (2001), 'Emerging infectious diseases in an island ecosystem: the New Zealand perspective', *Emerging Infectious Diseases* **7**(5), 767–772.

Cryer, J. & Chan, K. (2008), *Time series Analysis with applications in R*, Springer texts in statistics, second edn, Springer science and business media, New York.

Dahl, J. (1997), Cross-sectional epidemiological analysis of the relations between different herd factors and *Salmonella* seropositivity, *in* 'International symposium on Veterinary Epidemiology and Economics', AEEMA, Paris, pp. 04.23.1–04.23.3.

Daniell, P. (1946), 'Discussion of on the theoretical specifications and sampling properties of autocorrelated time-series', *Journal of the Royal Statistical Society* **8**, 88–90.

Dargatz, D. & Hill, G. (1996), 'Analysis of survey data', *Preventive Veterinary Medicine* **28**(4), 225–237.

De Gooijer, J. & Hyndman, R. (2006), 'Twenty-five years of time series forecasting', *International Journal of Forecasting* **22**(3), 443–473.

Deming, W. (1942), 'On a classification of the problems of statistical inference', *Journal of the American Statistical Association* **37**, 173–185.

Diggle, P. (1985), 'A kernel method for smoothing point process data', *Applied Statistics* **34**(2), 138–147.

Diggle, P. (1990), *Time Series: A Biostatistical Introduction*, Oxford Statistical Science Series, Oxford University Press, Oxford.

Diggle, P., Gomez-Rubio, V., Brown, P., Chetwynd, A. G. & Gooding, S. (2007), 'Second-order analysis of inhomogeneous spatial point processes using case-control data', *Biometrics* **63**(2), 550–557.

Diggle, P., Heagerty, P., Liang, K. & Zeger, S. (2002*a*), *The Analysis of Longitudinal Data*, second edn, Oxford University Press, Oxford.

Diggle, P. J. (2003), *Statistical analysis of spatial point patterns*, second edn, Arnold, London.

Diggle, P. J., Chetwynd, A. G., Haggkvist, R. & Morris, S. E. (1995), 'Second-order analysis of space-time clustering', *Statistical Methods in Medical Research* **4**(2), 124–136.

Diggle, P., Knorr-Held, L., Rowlingson, B., Su, T., Hawtine, P. & Bryant, T. (2004), On-line monitoring of public health surevillance data, *in* R. Brookmeyer & D. Stroup, eds, 'Monitoring the health of populations: statistical principles and methods for public health surveillance', Oxford University Press, New York, pp. 233–266.

Diggle, P., Moyeed, R., Rowlingson, B. & Thomson, M. (2002*b*), 'Childhood malaria in the Gambia: a case-study in model-based geostatistics', *Journal of the Royal Statistical Society Series C* **51**(4), 493–506.

Diggle, P., Rowlingson, B. & Su, T. (2005), 'Point process methodology for on-line spatio-temporal disease surveillance', *Environmetrics* **16**(5), 423–434.

Donders, A. R., van der Heijden, G. J., Stijnen, T. & Moons, K. G. (2006), 'Review: a gentle introduction to imputation of missing values', *Journal of Clinical Epidemiology* **59**(10), 1087–1091.

Dowdall, M., Lind, B., Gerland, S. & Rudjord, A. L. (2003), 'Geostatistical analysis as applied to two environmental radiometric time series', *Environmental Monitoring and Assessment* **83**(1), 1–16.

Duczmal, L. & Assunçao, R. (2004), 'A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters', *Computational Statistics and Data Analysis* **45**(2), 269–286.

Duczmal, L. & Buckeridge, D. (2006), 'A workflow spatial scan statistic', *Statistics in Medicine* **25**(5), 743–754.

Earnest, A., Chen, M. I., Ng, D. & Sin, L. Y. (2005), 'Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore', *BMC Health Services Research* **5**, 36.

Eddy, S. (2004), 'What is a hidden Markov Model?', *Nature Biotechnology* **22**(10), 1315–1316.

Ekeroth, L., Alban, L. & Feld, N. (2003), 'Single versus double testing of meat-juice samples for *Salmonella* antibodies, in the Danish pig-herd surveillance programme', *Preventive Veterinary Medicine* **60**(2), 155–165.

Endepols, S., Klemann, N., Pelz, H. J. & Ziebell, K. L. (2003), 'A scheme for the placement of rodenticide baits for rat eradication on confinement livestock farms', *Preventive Veterinary Medicine* **58**(3-4), 115–123.

Enoe, C., Wachmann, H. & Boes, J. (2003), Low intensity serological surveillance for *Salmonella enterica* in slaughter pigs from low prevalence herds in Denmark, *in* 'International Symposium of Veterinary Epidemiology and Economics', Vina del Mar, Chile.

Enserink, M. (2006), 'Emerging infectious diseases: During a hot summer, bluetongue virus invades northern Europe', *Science* **313**(5791), 1218–1219.

Ersbøll, A. K. & Nielsen, L. R. (2008), 'The range of influence between cattle herds is of importance for the local spread of *Salmonella Dublin* in Denmark', *Preventive Veterinary Medicine* **84**(3-4), 277–290.

Ethelberg, S., Wingstrand, A., Jensen, T., Sørensen, G., Müller, L., Nielsen, E. & Mølbak, K. (2008*a*), 'Large ongoing outbreak of infection with *Salmonella typhimurium* U292 in Denmark, February - July 2008', *Eurosurveillance* **13**(7), pii=18923.

Ethelberg, S., Wingstrand, A., Jensen, T., Sørensen, G., Müller, L., Nielsen, E. & Mølbak, K. (2008*b*), 'Large outbreaks of *Salmonella typhimurium* infection in Denmark in 2008', *Eurosurveillance* **13**(44), pii=19023.

European Food Safety Authority (2008), 'Report of the task force on zoonoses data collection on the analysis of the baseline survey on the prevalence of *Salmonella* in slaughter pigs, part A', *The EFSA Journal* **135**, 1–111.

Farzan, A., Friendship, R. M., Dewey, C. E., Warriner, K., Poppe, C. & Klotins, K. (2006), 'Prevalence of *Salmonella spp.* on Canadian pig farms using liquid or dry-feeding', *Preventive Veterinary Medicine* **73**(4), 241–254.

Fedorka-Cray, P. J., Gray, J. T. & C, W. (2000), *Salmonella* infections in pigs, *in* C.Wray & A.Wray, eds, '*Salmonella* in Domestic Animals', CABI Publishing, New York, pp. 191–207.

Fenton, S. E., Clough, H. E., Diggle, P. J., Evans, S. J., Davison, H. C., Vink, W. D. & French, N. P. (2008), 'Spatial and spatio-temporal analysis of *Salmonella* infection in dairy herds in England and Wales', *Epidemiology and Infection*. Aug 19, Epub ahead of print.

Fernandez-Perez, C., Tejada, J. & Carrasco, M. (1998), 'Multivariate time series analysis in nosocomial infection surveillance: a case study', *International Journal of Epidemiology* **27**(2), 282–288.

Fernando, S. (2008), PhD confirmation report, Technical report, Massey University.

Fevré, E. M., Bronsvoort, B. M., Hamilton, K. A. & Cleaveland, S. (2006), 'Animal movements and the spread of infectious diseases', *Trends in Microbiology* **14**(3), 125–131.

Fisker, N., Vinding, K., Mølbak, K. & Hornstrup, M. K. (2003), 'Clinical review of nontyphoid *Salmonella* infections from 1991 to 1999 in a Danish county', *Clinical Infectious Diseases* **37**(4), 47–52.

French, N. (2008), Enhancing surveillance of potentially foodborne enteric diseases in New Zealand: Human campylobacteriosis in the Manawatu, Technical report, New Zealand Food Safety Authority.

French, N. P., Berriatua, E., Wall, R., Smith, K. & Morgan, K. L. (1999), 'Sheep scab outbreaks in Great Britain between 1973 and 1992: spatial and temporal patterns', *Veterinary Parasitology* **83**(3-4), 187–200.

Fricker, R., Hegler, B. & Dunfee, D. (2008), 'Comparing syndromic surveillance detection methods: EARS versus a CUSUM-based methodology', *Statistics in Medicine* **27**, 3407–3429.

Funk, J., Davies, P. & Gebreyes, W. (2001), 'Risk factors associated with *Salmonella enterica* prevalence in three-site swine production systems in North Carolina, USA', *Berliner Und Munchener Tierarztliche Wochenschrift* **114**(9-10), 335–338.

German, R. (2000), 'Sensitivity and predictive value positive measurements for public health surveillance systems', *Epidemiology* **11**, 720–727.

Gibbens, J. C., Sharpe, C. E., Wilesmith, J. W., Mansley, L. M., Michalopoulou, E., Ryan, J. B. & Hudson, M. (2001), 'Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months', *Veterinary Record* **149**(24), 729–743.

Gilks, W., Thomas, A. & Spiegelhalter, D. (1994), 'A language and program for complex Bayesian modelling', *The Statistician* **43**, 169–178.

Giovannini, A., Savini, L., Conte, A. & Fiore, G. L. (2005), 'Comparison of BSE prevalence estimates from EU countries for the period July to December 2001 to the OIE and EU GBR classifications', *Journal of Veterinary Medicine Series B* **52**(6), 262–271.

Goldenberg, A., Shmueli, G., Caruana, R. A. & Fienberg, S. E. (2002), 'Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales', *Proceedings of the National Academy of Sciences of the United States of America* **99**(8), 5237–5240.

Gradel, K. O., Dethlefsen, C., H, C. S., Ejlertsen, T., H, T. S., Thomsen, R. W. & Nielsen, H. (2007), 'Severity of infection and seasonal variation of non-typhoid *Salmonella* occurrence in humans', *Epidemiology and Infection* **135**(1), 93–99.

Graham, S. L., Barling, K. S., Waghela, S., Scott, H. M. & Thompson, J. A. (2005), 'Spatial distribution of antibodies to *Salmonella enterica* serovar Typhimurium O antigens in bulk milk from Texas dairy herds', *Preventive Veterinary Medicine* **69**(1-2), 53–61.

Grassly, N. C. & Fraser, C. (2006), 'Seasonal infectious disease epidemiology', *Proceedings of the Biological Society* **273**(1600), 2541–2450.

Gray, J. T. & Fedorka-Cray, P. J. (2001), 'Survival and infectivity of *Salmonella choleraesuis* in swine feces', *Journal Of Food Protection* **64**(7), 945–949.

Green, C. G., Krause, D. & Wylie, J. (2006), 'Spatial analysis of *Campylobacter* infection in the Canadian province of Manitoba', *International Journal of Health Geographics* **5**, 2.

Guerin, M. T., Martin, S. W. & Darlington, G. A. (2005*a*), 'Temporal clusters of *Salmonella* serovars in humans in Alberta, 1990-2001', *Canadian Journal of Public Health* **96**(5), 390–395.

Guerin, M. T., Martin, S. W., Darlington, G. A. & Rajic, A. (2005*b*), 'A temporal study of *Salmonella* serovars in animals in Alberta between 1990 and 2001', *Canadian Journal of Veterinary Research* **69**(2), 88–99.

Ha, Y., Jung, K., Kim, J., Choi, C. & Chae, C. (2005), 'Outbreak of salmonellosis in pigs with postweaning multisystemic wasting syndrome', *Veterinary Record* **156**(18), 583–584.

Haine, D., Boelaert, F., Pfeiffer, D. U., Saegerman, C., Lonneux, J. F., Losson, B. & Mintiens, K. (2004), 'Herd-level seroprevalence and risk-mapping of bovine hypodermosis in Belgian cattle herds', *Preventive Veterinary Medicine* **65**(1-2), 93–104.

Haining, R. (2003), *Spatial Data Analysis: theory and practice*, Cambridge University Press, Cambridge.

Hald, T. & Andersen, J. S. (2001), 'Trends and seasonal variations in the occurrence of *Salmonella* in pigs, pork and humans in Denmark, 1995-2000', *Berliner und Muncheners Tierarztliche Wochenschrift* **114**(9-10), 346–349.

Hald, T., Vose, D., Wegener, H. C. & Koupeev, T. (2004), 'A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis', *Risk Analysis* **24**(1), 255–269.

Hald, T., Wegener, H., Borck, B., Lo Fo Wong, D. M., Baggesen, D., Madsen, M., Korsgaard, H., Ethelberg, S., Gerner-Smidt, P. & Mølbak, K. (2005), The intergrated surveillance of *Salmonella* in Denmark and the effect on public health., *in* F. Smulders & J. Collins, eds, 'Food safety assurance and veterinary public health. Volume 3. Risk management strategies: monitoring and surveillance.', Wageningen Academic Publishers, Wageningen, pp. 213–238.

Han, D., Rogerson, P. A., Nie, J., Bonner, M. R., Vena, J. E., Vito, D., Muti, P., Trevisan, M., Edge, S. B. & Freudenheim, J. L. (2004), 'Geographic clustering of residence in early life and subsequent risk of breast cancer (United States)', *Cancer Causes and Control* **15**(9), 921–929.

Hare, E. H. (1975), 'Season of birth in schizophrenia and neurosis', *American Journal of Psychiatry* **132**(11), 1168–1171.

Hashizume, M., Armstrong, B., Wagatsuma, Y., Faruque, A. S., Hayashi, T. & Sack, D. A. (2008), 'Rotavirus infections and climate variability in Dhaka, Bangladesh: a time-series analysis', *Epidemiology and Infection* **136**(9), 1281–1289.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Hauenstein, L., Wojcik, R., Loschen, W., Ashar, R., Sniegoski, C. & Tabernero, N. (2007), Putting it together: the biosurveillance information system, *in* J. Lombardo & D. Buckeridge, eds, 'Disease surveillance: a public health informatics approach', John Wiley and sons, New Jersey, pp. 193–261.

Hautekiet, V., Geert, V., Marc, V. & Rony, G. (2008), 'Development of a sanitary risk index for *Salmonella* seroprevalence in Belgian pig farms', *Preventive Veterinary Medicine* **86**(1-2), 75–92.

Hazelton, M. (2007), 'Bias reduction in kernel binary regression', *Computational Statistics and Data Analysis* **51**, 4393–4402.

Heagerty, P. & Zeger, S. (1998), 'Lorelogram: a regression approach to exploring dependence in longitudinal categorical responses', *Journal of the American Statistical Association* **93**(441), 150–162.

Hedemann, M. S., Mikkelsen, L. L., Naughton, P. J. & Jensen, B. B. (2005), 'Effect of feed particle size and feed processing on morphological characteristics in the small and large intestine of pigs and on adhesion of *Salmonella enterica* serovar Typhimurium DT12 in the ileum in vitro', *Journal of Animal Science* **83**(7), 1554–1562.

Held, L., Giusi, G., Frank, C. & Rue, H. (2006), 'Joint spatial analysis of gastrointestinal infectious diseases', *Statistical Methods in Medical Research* **15**(5), 465–80.

Held, L., Natario, I., Fenton, S. E., Rue, H. & Becker, N. (2005), 'Towards joint disease mapping', *Statistical Methods in Medical Research* **14**(1), 61–82.

Helfenstein, U. (1996), 'Box-Jenkins modeling in medical research', *Statistical Methods in Medical Research* **5**(1), 3–22.

Henderson, A. (2005), 'The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data', *Clinica Chimica Acta* **359**(1), 1–26.

Hodges, J. R. & Kimball, A. M. (2005), 'The global diet: trade and novel infections', *Global Health* **1**, 4.

Hook, E. & Regal, R. (2004), Completeness of reporting: capture-recapture methods in public health surveillance, *in* R. Brookmeyer & D. Stroup, eds, 'Monitoring the health of populations: statistical principles and methods for public health surveillance', Oxford University Press, New York, pp. 341–359.

Hosmer, D. & Lemeshow, S. (1989), *Applied Logistic Regression*, Wiley Series in Probablity and Mathematical Statistics, John Wiley and Sons, Inc., New York.

Houben, M., Coebergh, J., Birch, J., Tijssen, C., van Duijn, C. & McNally, R. (2005), 'Space-time clustering patterns of gliomas in The Netherlands suggest an infectious aetiology', *European Journal of Cancer* **41**(18), 2917–2923.

Hu, W., Mengersen, K., Bi, P. & Tong, S. (2006), 'Time-series analysis of the risk factors for haemorrhagic fever with renal syndrome: comparison of statistical models', *Epidemiology and Infection* **135**(2), 245–252.

Hu, W., Nicholls, N., Lindsay, M., Dale, P., McMichael, A. J., Mackenzie, J. S. & Tong, S. (2004), 'Development of a predictive model for Ross River virus disease in Brisbane, Australia', *American Journal of Tropical Medicine and Hygiene* **71**(2), 129–137.

Huang, L., Kulldorff, M. & Gregorio, D. (2007), 'A spatial scan statistic for survival data', *Biometrics* **63**, 109 – 118.

Hueston, W. & Bryant, C. (2005), 'Transmissible spongioform encephalopathies', *Journal of Food Science* **70**(5), R77–R87.

Hufnagel, L., Brockmann, D. & Geisel, T. (2004), 'Forecast and control of epidemics in a globalized world', *Proceedings of the National Academy of Sciences of the United States of America* **101**(42), 15124–15129.

Hunter, P. R., Chalmers, R. M., Syed, Q., Hughes, L. S., Woodhouse, S. & Swift, L. (2003), 'Foot and mouth disease and cryptosporidiosis: possible interaction between two emerging infectious diseases', *Emerging Infectious Diseases* **9**(1), 109–112.

Hurd, H. S., Enoe, C., Sørensen, L., Wachman, H., Corns, S. M., Bryden, K. M. & Grenier, M. (2008), 'Risk-based analysis of the Danish pork *Salmonella* program: past and future', *Risk Analysis* **28**(2), 341–351.

Hurd, H. S., McKean, J. D., Griffith, R. W., Wesley, I. V. & Rostagno, M. H. (2002), '*Salmonella enterica* infections in market swine with and without transport and holding', *Applied Environmental Microbiology* **68**(5), 2376–2381.

Hutwagner, L., Browne, T., Seeman, G. M. & Fleischauer, A. T. (2005), 'Comparing aberration detection methods with simulated data', *Emerging Infectious Diseases* **11**(2), 314–326.

Isaaks, E. & Srivastava, R. (1989), *An Introduction to Applied Geostatistics*, Oxford University Press, New York.

Isham, V. (1993), 'Stochastic models for epidemics with special reference to AIDS', *Annals of Applied Probability* **3**, 1–27.

Jajosky, R. A. & Groseclose, S. L. (2004), 'Evaluation of reporting timeliness of public health surveillance systems for infectious diseases', *BMC Public Health* **4**, 29.

Jay, M. T., Cooley, M., Carychao, D., Wiscomb, G. W., Sweitzer, R. A., Crawford-Miksza, L., Farrar, J. A., Lau, D. K., O'Connell, J., Millington, A., Asmundson, R. V., Atwill, E. R. & Mandrell, R. E. (2007), '*Escherichia coli* O157:H7 in feral swine near spinach fields and cattle, central California coast', *Emerging Infectious Diseases* **13**(12), 1908–1911.

Jelinek, T., Behrens, R., Bisoffi, Z., Bjorkmann, A., Andersen, P., Blaxhult, A., Gascon, J., Hellgren, U., Petersen, E. & Zoller, T. (2007), 'Recent cases of falciparum malaria imported to Europe from Goa, India, December 2006-January 2007', *Eurosurveillance* **12**(1), pii=70111.

Jensen, E. S., Lundbye-Christensen, S., Pedersen, L., Sorensen, H. T. & Schonheyder, H. C. (2003), 'Seasonal variation in meningococcal disease in Denmark: relation to age and meningococcal phenotype', *Scandinavian Journal of Infectious Diseases* **35**(4), 226–229.

Jernigan, J., Stephens, D. & Ashford, D. (2001), 'Bioterrorism-related inhalational anthrax: The first 10 cases reported in the United States', *Emerging Infectious Diseases* **7**(6), 933–944.

Johnson, G. D., Eidson, M., Schmit, K., Ellis, A. & Kulldorff, M. (2006), 'Geographic prediction of human onset of West Nile virus using dead crow clusters: An evaluation of year 2002 data in New York State', *American Journal of Epidemiology* **163**(2), 171–180.

Jolley, K., Brehony, C. & Maiden, M. (2007), 'Molecular typing of *Meningococci*: recommendations for target choice and nomenclature', *FEMS Microbiological Reviews* **31**, 89–96.

Jones, R., Liberatore, M., Fernandez, J. & Gerber, S. (2006), 'Use of a prospective space-time scan statistic to prioritize shigellosis case investigations in an urban jurisdiction', *Public Health Reports* **121**(2), 133–140.

Jung, I., Kulldorff, M. & Klassen, A. C. (2006), 'A spatial scan statistic for ordinal data', *Statistics in Medicine* **26**(7), 1594–1607.

Ka-Wai Hui, E. (2006), 'Reasons for the increase in emerging and re-emerging viral infectious diseases', *Microbes and Infection* **8**(3), 905–916.

Kadohira, M., Stevenson, M., Kanayama, T. & Morris, R. (2008), 'The epidemiology of bovine spongiform encephalopathy in Hokkaido, Japan, September 2001 to December 2006', *Veterinary Record* **163**, 709–713.

Kallio-Kokko, H., Uzcategui, N., Vapalahti, O. & Vaheri, A. (2005), 'Viral zoonoses in Europe', *FEMS Microbiology Reviews* **29**(5), 1051–1077.

Kao, R. R., Green, D. M., Johnson, J. & Kiss, I. Z. (2007), 'Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network of livestock movements in the UK', *Jornal of the Royal Society Interface* **4**(16), 907–916.

Kapel, C. (2005), 'Changes in the EU legislation on *Trichinella* inspection–new challenges in the epidemiology', *Veterinary Parasitology* **132**(1-2), 189–194.

Keenan, B. (2007), Leptospirosis: Reducing the impact on New Zealand workplaces, Technical report, Department of Labour.

Keim, D., Mansmann, F., Schneidewind, J. & Ziegler, H. (2006), Challenges in visual data analysis, *in* F. Mansmann, ed., 'Tenth International Conference on Information Visualization', London, UK, pp. 9–16.

Kelsall, J. & Diggle, P. (1995), 'Kernel estimation of relative risk', *Bernoulli* **1**(1/2), 3–16.

Kelsall, J. E. & Diggle, P. J. (1998), 'Spatial variation in risk of disease: a nonparametric binary regression approach', *Applied Statistics* **47**(4), 559–574.

Kim, C. H., Lee, C. G., Yoon, H. C., Nam, H. M., Park, C. K., Lee, J. C., Kang, M. I. & Wee, S. H. (2006), 'Rabies, an emerging disease in Korea', *Journal of Veterinary Medicine B: Infectoius Diseases and Veterinary Public Health* **53**(3), 111–115.

Kirk, S. F., Greenwood, D., Cade, J. E. & Pearman, A. D. (2002), 'Public perception of a range of potential food risks in the United Kingdom', *Appetite* **38**(3), 189–197.

Kleinman, K. P., Abrams, A. M., Kulldorff, M. & Platt, R. (2005), 'A model-adjusted space-time scan statistic with an application to syndromic surveillance', *Epidemiology and Infection* **133**(3), 409–419.

Knorr-Held, L. & Richardson, S. (2003), 'A hierarchical model for space-time surveillance data on meningococcal disease incidence', *Journal of the Royal Statistical Society Series C* **52**, 169–183.

Kovats, R. S., Bouma, M. J., Hajat, S., Worrall, E. & Haines, A. (2003), 'El Niño and health', *Lancet* **362**(9394), 1481–1489.

Kovats, R. S., Edwards, S. J., Charron, D., Cowden, J., D'Souza, R. M., Ebi, K. L., Gauci, C., Gerner-Smidt, P., Hajat, S., Hales, S., Hernandez Pezzi, G., Kriz, B., Kutsar, K., McKeown, P., Mellou, K., Menne, B., O'Brien, S., van Pelt, W. & Schmid, H. (2005), 'Climate variability and *Campylobacter* infection: an international study', *International Journal of Biometeorology* **49**(4), 207–214.

Kovats, R. S., Edwards, S. J., Hajat, S., Armstrong, B. G., Ebi, K. L. & Menne, B. (2004), 'The effect of temperature on food poisoning: a time-series analysis of salmonellosis in ten European countries', *Epidemiology and Infection* **132**(3), 443–453.

Kulldorff, M. (1997), 'A spatial scan statistic', *Communications in Statistics: Theory and Methods* **26**, 1481–1496.

Kulldorff, M. (2001), 'Prospective time periodic geographical disease surveillance using a scan statistic', *Journal of the Royal Statistical Society: Series A* **164**(1), 61–72.

Kulldorff, M., Athas, W., Feuer, E. J., Miller, B. A. & Key, C. (1998), 'Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico', *American Journal of Public Health* **88**(9), 1377–1380.

Kulldorff, M., Feuer, E. J., Miller, B. A. & Freedman, L. S. (1997), 'Breast cancer clusters in the northeast United States: a geographic analysis', *American Journal of Epidemiology* **146**(2), 161–170.

Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R. & Mostashari, F. (2005), 'A space-time permutation scan statistic for disease outbreak detection', *Public Library of Science Medicine* **2**(3), e59.

Kulldorff, M., Huang, L., Pickle, L. & Duczmal, L. (2006), 'An elliptic spatial scan statistic', *Statistics in Medicine* **25**(22), 3929–3943.

Langvad, B., Skov, M. N., Rattenborg, E. & Baggesen, D. L. (2003), 'Molecular epidemiology of a geographically localized *Salmonella typhimurium* DT104 outbreak in Danish cattle and pigs', *Acta Veterinaria Scandinavica* **44**(Supplement 1), 62.

Lawson, A. (2006), *Statistical methods in spatial epidemiology*, John Wiley and Sons, Chichester.

Lawson, A. (2009), *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*, CRC Press, Boca Raton.

Lawson, A. & Kleinman, K. (2005), Introduction: Spatial and syndromic surveillance for public health, *in* A. Lawson & K. Kleinman, eds, 'Spatial and syndromic surveillance for public health', John Wiley and sons, Chichester, pp. 1–10.

Lawson, A. & Williams, F. (1993), 'Application of extraction mapping in environmental epidemiology', *Statistics in Medicine* **12**, 1249–1258.

Lawson, A. B. & Williams, F. L. R. (1994), 'Armadale: A case-study in environmental epidemiology', *Journal of the Royal Statistical Society. Series A* **157**(2), 285–298.

Lawson, A. B., Biggeri, A. & Dreassi, E. (1999), Edge effects in disease mapping, *in* A. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J. Viel & R. Bertollini, eds, 'Disease Mapping and Risk Assessment for Public Health', John Wiley and Sons, Chichester, pp. 85–98.

Le Menach, A., Legrand, J., Grais, R. F., Viboud, C., Valleron, A. J. & Flahault, A. (2005), 'Modeling spatial and temporal transmission of foot-and-mouth disease in France: identification of high-risk areas', *Veterinary Research* **36**(5-6), 699–712.

Le Strat, Y. (2005), Overview of temporal surveillance, *in* A. Lawson & K. Kleinman, eds, 'Spatial and Syndromic Surveillance for Public Health', John Wiley & Sons Ltd., Chichester, pp. 13–28.

Le Strat, Y. & Carrat, F. (1999), 'Monitoring epidemiologic surveillance data using hidden Markov models', *Statistics in Medicine* **18**(24), 3463–3478.

Lemmings, J., Robinson, L., Hoffman, R., Mangione, E. & Humes, R. (2006), 'Assessing capacity for surveillance, prevention, and control of West Nile virus infection—United States, 1999 and 2004', *Morbidity and Mortality Weekly Report* **55**(6), 150–153.

Leontides, L., Grafanakis, E. & Genigeorgis, C. (2003), 'Factors associated with the serological prevalence of *Salmonella enterica* in Greek finishing swineherds', *Epidemiology And Infection* **131**(1), 599–606.

Letellier, A., Messier, S., Pare, J., Menard, J. & Quessy, S. (1999), 'Distribution of *Salmonella* in swine herds in Quebec', *Veterinary Microbiology* **67**(4), 299–306.

Ling, M., Rifai, H. S., Newell, C. J., Aziz, J. J. & Gonzales, J. R. (2003), 'Groundwater monitoring plans at small-scale sites—an innovative spatial and temporal methodology', *Journal of Environmental Monitoring* **5**(1), 126–134.

Ljung, G. M. & Box, G. E. P. (1978), 'On a measure of lack of fit in time series models', *Biometrika* **65**(2), 297–303.

Lo Fo Wong, D. M. A., Dahl, J., Stege, H., van der Wolf, P. J., Leontides, L., von Altrock, A. & Thorberg, B. (2004*a*), 'Herd-level risk factors for subclinical *Salmonella* infection in European finishing-pig herds', *Preventive Veterinary Medicine* **62**(4), 253–266.

Lo Fo Wong, D. M. A., Dahl, J., van der Wolf, P. J., Wingstrand, A., Leontides, L. & von Altrock, A. (2003), 'Recovery of *Salmonella enterica* from seropositive finishing pig herds', *Veterinary Microbiology* **97**(3-4), 201–214.

Lo Fo Wong, D. M. A., Dahl, J., Wingstrand, A., van der Wolf, P. J., von Altrock, A. & Thorberg, B. M. (2004*b*), 'A European longitudinal study in *Salmonella* seronegative- and seropositive-classified finishing pig herds', *Epidemiology and Infection* **132**(5), 903–914.

Lo Fo Wong, D. M. A., Hald, T., van der Wolf, P. J. & Swanenburg, M. (2002), 'Epidemiology and control measures for *Salmonella* in pigs and pork', *Livestock Production Science* **76**(3), 215–222.

Lockhart, C. (2008), Surveillance for Diseases of Poultry with Specific Reference to Avian Influenza, PhD thesis, Massey University.

Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Lewis, S., Loschen, W., Sari, J., Sniegoski, C., Wojcik, R. & Pavlin, J. (2003), 'A systems overview of the electronic surveillance system for the early notification of community-based epidemics (ESSENCE II)', *Journal of Urban Health* **80**(Supplement 1), 32–40.

Lopez-Lozano, J. M., Monnet, D. L., Yague, A., Burgos, A., Gonzalo, N., Campillos, P. & Saez, M. (2000), 'Modeling and forecasting antimicrobial resistance and its dynamic relationship to antimicrobial use: a time series analysis', *International Journal of Antimicrobial Agents* **14**(1), 21–31.

Louis, V. R., Gillespie, I. A., O'Brien, S. J., Russek-Cohen, E., Pearson, A. D. & Colwell, R. R. (2005), 'Temperature-driven *Campylobacter* seasonality in England and Wales', *Applied and Environmental Microbiology* **71**(1), 85–92.

Lurette, A., Belloc, C., Touzeau, S., Hoch, T., Ezanno, P., Seegers, H. & Fourichon, C. (2008), 'Modeling *Salmonella* spread within a farrow-to-finish pig herd', *Veterinary Research* **39**(5), 49.

Madigan, D. (2005), Bayesian data mining for health surveillance, *in* A. Lawson & K. Kleinman, eds, 'Spatial and syndromic surveillance for public health', John Wiley and sons, Chichester, pp. 203–221.

Majowicz, S. E., Edge, V. L., Fazil, A., McNab, W. B., Dore, K. A., Sockett, P. N., Flint, J. A., Middleton, D., McEwen, S. A. & Wilson, J. B. (2005), 'Estimating the under-

reporting rate for infectious gastrointestinal illness in Ontario', *Canadian Journal of Public Health* **96**(3), 178–181.

Mandl, K. D., Overhage, J., Wagner, M., Lober, W., Sebastiani, P., Mostashari, F., Pavlin, J., Gesteland, P., Treadwell, T., Koski, E., Hutwagner, L., Buckeridge, D., Aller, R. & Grannis, S. (2004), 'Implementing syndromic surveillance: a practical guide informed by early experience', *Journal of the American Medical Informatics Association* **11**(2), 141–150.

Marshall, J. & Hazelton, M. (2008), Boundary kernels for adaptive density estimators on regions with irregular boundaries, Technical report, Massey University.

Martin, P. A., Cameron, A. R. & Greiner, M. (2007*a*), 'Demonstrating freedom from disease using multiple complex data sources 1: A new methodology based on scenario trees', *Preventive Veterinary Medicine* **79**(2-4), 71–97.

Martin, P. A., Cameron, A. R., Barfod, K., Sergeant, E. S. & Greiner, M. (2007*b*), 'Demonstrating freedom from disease using multiple complex data sources 2: case study–classical swine fever in Denmark', *Preventive Veterinary Medicine* **79**(2-4), 98–115.

Martin, T., Wintle, B., Rhodes, J., Kuhnert, P., Field, S., Low-Choy, S., Tyre, A. & Possingham, H. (2005), 'Zero tolerance ecology: improving ecological inference by modeling the source of zero observations', *Ecology Letters* **8**(11), 1235–1246.

Martin, W. (2004), 'Measuring health and disease: progress in analytical approaches', *Preventive Veterinary Medicine* **62**(3), 165–175.

Matérn, B. (1960), *Spatial Variation*, Lecture Notes in Statistics, second edn, Springer, Berlin.

McLeod, K. (2000), 'Our sense of Snow: the myth of John Snow in medical geography', *Social Science and Medicine* **50**, 923–935.

McMichael, A. J., Woodruff, R. E. & Hales, S. (2006), 'Climate change and human health: present and future risks', *Lancet* **367**(9513), 859–869.

McNally, R., Alexander, F. & Bithell, J. F. (2006), 'Space-time clustering of childhood cancer in Great Britain: A national study, 1969-1993', *International Journal of Cancer* **118**(11), 2840–2846.

Merianos, A. (2007), 'Surveillance and response to disease emergence', *Current Topics in Microbiology, Immunology* **315**, 477–509.

Mikkelsen, L., Naughton, P., Hedemann, M. & Jensen, B. (2004), 'Effects of physical properties of feed on microbial ecology and survival of *Salmonella enterica* serovar Typhimurium in the pig gastrointestinal tract', *Applied and Environmental Microbiology* **70**(6), 3485–3492.

Miller, G., Dunn, G. M., Smith-Palmer, A., Ogden, I. D. & Strachan, N. J. (2004), 'Human campylobacteriosis in Scotland: seasonality, regional trends and bursts of infection', *Epidemiology and Infection* **132**(4), 585–593.

Miller, G. Y., Liu, X., McNamara, P. E. & Barber, D. A. (2005), 'Influence of *Salmonella* in pigs preharvest and during pork processing on human health costs and risks from pork', *Journal of Food Protection* **68**(9), 1788–1798.

Ministry of Family and Consumer Affairs (2005), Annual report on zoonoses in Denmark 2004, Technical report.

Ministry of Family and Consumer Affairs (2006), Annual report on zoonoses in Denmark 2005, Technical report.

Ministry of Family and Consumer Affairs (2007), Annual report on zoonoses in Denmark 2006, Technical report.

Mintiens, K., Laevens, H., Dewulf, J., Boelaert, F., Verloo, D. & Koenen, F. (2003), 'Risk analysis of the spread of classical swine fever virus through 'neighbourhood infections' for different regions in Belgium', *Preventive Veterinary Medicine* **60**(1), 27–36.

Moineddin, R., Upshur, R. E., Crighton, E. & Mamdani, M. (2003), 'Autoregression as a means of assessing the strength of seasonality in a time series', *BMC Population Health Metrics* **1**, 10.

Mousing, J., Jensen, P., Halgaard, C., Bager, F., Feld, N., Nielsen, B., Nielsen, J. & Bech-Nielsen, S. (1997), 'Nation-wide *Salmonella enterica* surveillance and control in Danish slaughter swine herds', *Preventive Veterinary Medicine* **29**(4), 247–261.

Murakami, S., Ogawa, A., Kinoshita, T., Matsumoto, A., Ito, N. & Nakane, T. (2006), 'Occurrence of swine salmonellosis in postweaning multisystemic wasting syndrome affected pigs concurrently infected with porcine reproduction and respiratory syndrome virus', *Journal of Veterinary Medical Science* **68**(4), 387–391.

Naumova, E. N., Jagai, J. S., Matyas, B., DeMaria, A., MacNeill, I. B. & Griffiths, J. K. (2007), 'Seasonality in six enterically transmitted diseases and ambient temperature', *Epidemiology and Infection* **135**, 281–292.

Newcombe, R. (1998), 'Interval estimation for the difference between independent proportions: comparison of eleven methods', *Statistics in Medicine* **17**(8), 873–890.

Nielsen, B. & Wegener, H. C. (1997), 'Public health and pork and pork products: regional perspectives of Denmark', *Revue Scientifique et Technique* **16**(2), 513–524.

Nielsen, B., Alban, L., Stege, H., Sørensen, L., Mogelmose, V., Bagger, J., Dahl, J. & Baggesen, D. (2001), 'A new *Salmonella* surveillance and control programme in Danish pig herds and slaughterhouses', *Berliner und Munchener Tierarztliche Wochenschrift* **114**(9-10), 323–326.

Nielsen, B., Baggesen, D., Bager, F., Haugegaard, J. & Lind, P. (1995), 'The serological response to *Salmonella* serovars Typhimurium and Infantis in experimentally infected pigs. The time course followed with an indirect anti-LPS ELISA and bacteriological examinations', *Veterinary Microbiology* **47**(3-4), 205–218.

Nielsen, B., Ekeroth, L., Bager, F. & Lind, P. (1998), 'Use of muscle fluid as a source of antibodies for serologic detection of *Salmonella* infection in slaughter pig herds', *Journal of Veterinary Diagnostic Investigation* **10**(2), 158–163.

Nobre, F. F., Monteiro, A. B., Telles, P. R. & Williamson, G. D. (2001), 'Dynamic linear models and SARIMA: a comparison of their forecasting performance in epidemiology', *Statistics in Medicine* **20**(20), 3051–3069.

Nollet, N., Maes, D., De Zutter, L., Duchateau, L., Houf, K., Huysmans, K., Imberechts, H., Geers, R., de Kruif, A. & Van Hoof, J. (2004), 'Risk factors for the herd-level bacteriologic prevalence of *Salmonella* in Belgian slaughter pigs', *Preventive Veterinary Medicine* **65**(1-2), 63–75.

Ochman, H., Lawrence, J. & Groisman, E. (2000), 'Lateral gene transfer and the nature of bacterial innovation', *Nature* **405**(6784), 299–304.

Oliveira, C. J., Carvalho, L. F. & Garcia, T. B. (2006), 'Experimental airborne transmission of *Salmonella Agona* and *Salmonella Typhimurium* in weaned pigs', *Epidemiology and Infection* **134**(1), 199–209.

Oliver, M. & Kharyat, A. (1999), Investigating the spatial variation of radon in soil geostatistically, *in* 'The IV International conference on Geocomputation', Fredericksberg, Virginia.

Patel, J. L. & Goyal, R. K. (2007), 'Applications of artificial neural networks in medical science', *Current Clinical Pharmacology* **2**(3), 217–226.

Patridge, J. & Namulanda, G. (2008), 'Describing environmental public health data: implementing a descriptive metadata standard on the environmental public health tracking network', *Journal of Public Health Management and Practice* **14**(6), 515–525.

Paul, M., Held, L. & Toschke, A. M. (2008), 'Multivariate modeling of infectious disease surveillance data', *Statistics in Medicine* **27**(29), 6250–6267.

Pearce, R. A., Bolton, D. J., Sheridan, J. J., McDowell, D. A., Blair, I. S. & Harrington, D. (2004), 'Studies to determine the critical control points in pork slaughter hazard analysis and critical control point systems', *International Journal of Food Microbiology* **90**(3), 331–339.

Penna, M. (2004), 'Use of an artificial neural network for detecting excess deaths due to cholera in Ceará, Brazil', *Revista de Saude Publica* **38**, 3.

Perez, A. M., Ward, M. P., Torres, P. & Ritacco, V. (2002), 'Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina', *Preventive Veterinary Medicine* **56**(1), 63–74.

Pettersson, L., Boman, J., Juto, P., Evander, M. & Ahlm, C. (2008), 'Outbreak of Puumala virus infection, Sweden', *Emerging Infectious Diseases* **14**(5), 808–810.

Pfeiffer, D., Robinson, T., Stevenson, M., Stevens, K., Rogers, D. & Clements, A. (2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

Phan, Q., Schauer, B., Stevenson, M., Jones, G., Morris, R. & Noble, A. (2008), Association between human cases and poultry outbreaks of highly pathogenic avian influenza in Vietnam from 2003 to 2007: A nationwide study, Technical report, Massey University.

Population and Environmental Health Group (2008), Notifiable and other disease in New Zealand. Annual Report 2007, Technical report, Institute of Environmental Science and Research Limited.

Porphyre, T. (2008), Factors Associated with the Transmission Dynamics of Bovine Tuberculosis in New Zealand, PhD thesis, Massey University.

ProMED-mail (2007), 'Rift valley fever - Kenya (North Eastern Province)', *ProMED-mail* p. 12 Jan: 20070101.0004. Available as: `http://www.promedmail.org`. Accessed 13 January 2007.

ProMED-mail (2008*a*), 'Dioxin contamination, pig meat—Ireland, Europe (03)', *ProMED-mail* p. 10 Dec:20081210.3883. Available as: `http://www.promedmail.org`. Accessed 24 December 2008.

ProMED-mail (2008*b*), 'Prion disease update 2008 (14): New vCJD wave imminent?', *ProMED-mail* p. 18 Dec:20081218.3980. Available as: `http://www.promedmail.org`. Accessed 22 December 2008.

ProMED-mail (2008*c*), 'Salmonellosis, serotype Typhimurium phage type 42 - New Zealand', *ProMED-mail* p. 4 Dec:20081204.3814. Available as: `http://www.promedmail.org`. Accessed 22 December 2008.

ProMED-mail (2008*d*), 'Tuberculosis—South Africa: Human to animal transmission', *ProMED-mail* p. 27 Aug:20080827.2680. Available as: `http://www.promedmail.org`. Accessed 22 December 2008.

Proux, K., Cariolet, R., Fravalo, P., Houdayer, C., Keranflech, A. & Madec, F. (2001), 'Contamination of pigs by nose-to-nose contact or airborne transmission of *Salmonella typhimurium*', *Veterinary Research* **32**(6), 591–600.

Qi, M. & Zhang, P. (2008), 'Trend time-series modeling and forecasting with neural networks', *IEEE Transactions on Neural Networks* **19**(5), 808–816.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L. (1989), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE* **77**, 257–286.

Radunz, B. (2006), 'Surveillance and risk management during the latter stages of eradication: Experiences from Australia', *Veterinary Microbiology* **112**(2-4), 283–290.

Rath, T., Carreras, M. & Sebastiani, P. (2003), Automated detection of Influenza epidemics with hidden Markov models, *in* M. Berthold, H. Lenz, E. Bradley, R. Kruse, C. Borgelt & P. Pfenning, eds, 'Advances in Intelligent data Aanlysis V', Springer-Verlag, Berlin, pp. 521–531.

Recuenco, S., Eidson, M., Kulldorff, M., Johnson, G. & Cherry, B. (2007), 'Spatial and temporal patterns of enzootic raccoon rabies adjusted for multiple covariates', *International Journal of Health Geographics* **6**, 14.

Reid, S., Herbelin, C., Bumbaugh, A., Selander, R. & Whittam, T. (2000), 'Parallel evolution of virulence in pathogenic *Escherichia coli*', *Nature* **406**(6791), 64–67.

Reinhardt, M., Elias, J., Albert, J., Frosch, M., Harmsen, D., Vogel, U., Elias, J., Reinhardt, M., Hautmann, W., Harms, I., Oppermann, H., Schroter, M., Hellenbrand, W., Oster, P., Kurzai, O., Taha, M. K., Nossal, R., Frosch, M. & Vogel, M. (2008), 'EpiScanGIS: an online geographic surveillance system for meningococcal disease [3rd Workshop on Epidemiology, Prevention and Treatment of Invasive Meningococcal Disease in Wurzburg 2006]', *International Journal of Health Geographics* **7**(4), 33.

Reis, B. Y. & Mandl, K. D. (2003), 'Time series modeling for syndromic surveillance', *BMC Medical Informatics and Decision Making* **3**, 2.

Ribeiro Jr., P. & Diggle, P. J. (2001), 'geoR: A package for geostatistical analysis', *R-News* **1**(2), 15–18.

Ripley, B. (1976), 'The second-order analysis of stationary point processes', *Journal of Applied Probability* **13**, 255–266.

Ripley, B. (1988), *Statistical inference for Spatial Processes*, Cambridge University Press, Cambridge.

Rogerson, P. (2005), Spatial surveillance and cumulative sum methods, *in* A. Lawson & K. Kleinman, eds, 'Spatial and Syndromic Surveillance for Public Health', John Wiley and Sons Ltd., Chichester, pp. 95–114.

Rolka, H., Burkom, H., Cooper, G. F., Kulldorff, M., Madigan, D. & Wong, W. K. (2007), 'Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs', *Statistics in Medicine* **26**, 1834–1856.

Rostagno, M. H., Hurd, H. S., McKean, J. D., Ziemer, C. J., Gailey, J. K. & Leite, R. C. (2003), 'Preslaughter holding environment in pork plants is highly contaminated with *Salmonella enterica*', *Applied and Environmental Microbiology* **69**(8), 4489–4494.

Rowlingson, B. & Diggle, P. J. (1993), 'Splancs: spatial point pattern analysis code in S-Plus.', *Computers and Geosciences* **19**, 627–655.

Sackett, D. L. (1979), 'Bias in analytic research', *Journal of Chronic Disease* **32**(1-2), 51–63.

Salvatori, V., Skidmore, A., Corsi, F. & Van der Meer, F. (1999), 'Estimating temporal independence of radio-telemetry data on animal activity', *Journal of Theoretical Biology* **198**(4), 567–574.

Sanchez, J., Stryhn, H., Flensburg, M., Ersbøll, A. K. & Dohoo, I. (2005), 'Temporal and spatial analysis of the 1999 outbreak of acute clinical infectious bursal disease in broiler flocks in Denmark', *Preventive Veterinary Medicine* **71**(3-4), 209–223.

Sanchez, S., Hofacre, C. L., Lee, M. D., Maurer, J. J. & Doyle, M. P. (2002), 'Animal sources of salmonellosis in humans', *Jornal of the American Veterinary Medical Association* **221**(4), 492–497.

Sargeant, J. M. (2008), 'The influence of veterinary epidemiology on public health: past, present and future', *Preventive Veterinary Medicine* **86**(3-4), 250–259.

Sawyer, M. (2000), 'Invited commentary: Artificial neural networks an introduction', *Surgery* **127**, 1–2.

Scallan, E. & Angulo, F. (2007), Surveillance for foodborne diseases, *in* N. M'ikanatha, R. Lynfield, C. van Beneden & H. de Valk, eds, 'Infectious Disease Surveillance', first edn, Blackwell Publishing, Massachusetts, pp. 57–68.

Schafer, J. L. & Graham, J. W. (2002), 'Missing data: our view of the state of the art', *Psychological Methods* **7**(2), 147–177.

Schlundt, J., Toyofuku, H., Jansen, J. & Herbst, S. A. (2004), 'Emerging food-borne zoonoses', *Revue Scientifique et Technique* **23**(2), 513–533.

Schwartz, J., Spix, C., Touloumi, G., Bacharova, L., Barumamdzadeh, T., le Tertre, A., Piekarksi, T., Ponce de Leon, A., Ponka, A., Rossi, G., Saez, M. & Schouten, J. P. (1996), 'Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions', *Journal of Epidemiology and Community Health* **50**(Suppl 1), S3–S11.

Sellers, R. F. (2006), 'Comparison of different control strategies for foot-and-mouth disease: a study of the epidemics in Canada in 1951/52, Hampshire in 1967 and Northumberland in 1966', *Veterinary Record* **158**(1), 9–16.

Shah, A. P., Smolensky, M. H., Burau, K. D., Cech, I. M. & Lai, D. (2006), 'Seasonality of primarily childhood and young adult infectious diseases in the United States', *Chronobiology International* **23**(5), 1065–1082.

Shewhart, W. (1931), *Economic control of quality of Manufactured Product*, Van Nostrand, New York.

Siqueira, J. B., J., Maciel, I. J., Barcellos, C., Souza, W. V., Carvalho, M. S., Nascimento, N. E., Oliveira, R. M., Morais, O. L. N. & Martelli, C. M. (2008), 'Spatial point analysis based on dengue surveys at household level in central Brazil', *BMC Public Health* **8**(1), 361.

Snow, J. (1854), *On the mode of communication of cholera*, Churchill Livingston, London.

Song, C. & Kulldorff, M. (2003), 'Power evaluation of disease clustering tests', *International Journal of Health Geographics* **2**, 9.

Spencer, S., Pirie, R. & French, N. (2008), The detection of spatially localised point source outbreaks in campylobacteriosis notification data, Technical report, Massey University.

Stabler, R. A., Dawson, L. F., Oyston, P. C., Titball, R. W., Wade, J., Hinds, J., Witney, A. A. & Wren, B. W. (2008), 'Development and application of the active surveillance of pathogens microarray to monitor bacterial gene flux', *BMC Microbiology* **8**(1), 177.

Stack, J. & Perrett, L. (2005), 'Brucellosis — Veterinary Laboratories Agency's (VLA) perspective on the outbreaks in Scotland 2003', *State Veterinary Journal* **15**(1), 9–12.

Stärk, K. D., Regula, G., Hernandez, J., Knopf, L., Fuchs, K., Morris, R. S. & Davies, P. (2006), 'Concepts for risk-based surveillance in the field of veterinary medicine and veterinary public health: Review of current approaches', *BMC Health Services Research* **6**(1), 20–28.

Stärk, K. D., Wingstrand, A., Dahl, J., Mogelmose, V. & Lo Fo Wong, D. M. A. (2002), 'Differences and similarities among experts' opinions on *Salmonella enterica* dynamics in swine pre-harvest', *Preventive Veterinary Medicine* **53**(1-2), 7–20.

Stege, H., Christensen, J., Nielsen, J. & Willeberg, P. (2001), 'Data-quality issues and alternative variable-screening methods in a questionnaire-based study on subclinical *Salmonella enterica* infection in Danish pig herds', *Preventive Veterinary Medicine* **48**(1), 35–54.

Steinbach, G. & Hartung, M. (1999), 'An attempt to estimate the share of human cases of salmonellosis attributable to *Salmonella* originating from pigs', *Berliner und Muncheners Tierarztliche Wochenschrift* **112**(8), 296–300.

Steinbach, G. & Kroell, U. (1999), '*Salmonella* infections in swine herds-epidemiology and importance for human diseases', *Deutsche Tierarztliche Wochenschrift* **106**(7), 282–288.

Stevenson, M. (2004), The spatio-temporal epidemiolgy of Bovine Spongiform Encephalopathy and Foot-And-Mouth disease in Great Britain, PhD thesis, Massey University.

Stevenson, M. A., Benard, H., Bolger, P. & Morris, R. S. (2005), 'Spatial epidemiology of the Asian honey bee mite (*Varroa destructor*) in the north island of New Zealand', *Preventive Veterinary Medicine* **71**(3-4), 241–252.

Stevenson, M. A., Wilesmith, J. W., Ryan, J. B., Morris, R. S., Lawson, A. B., Pfeiffer, D. U. & Lin, D. (2000), 'Descriptive spatial analysis of the epidemic of bovine spongiform encephalopathy in Great Britain to June 1997', *Veterinary Record* **147**(14), 379–384.

Stoumbos, Z., Reynolds, M., Ryan, T. & Woodall, W. (2000), 'The state of statistical process control as we proceed into the 21st Century', *Journal of the American Statistical Association* **95**(451), 992–998.

Swanenburg, M., Berends, B. R., Urlings, H. A., Snijders, J. M. & van Knapen, F. (2001*a*), 'Epidemiological investigations into the sources of *Salmonella* contamination of pork', *Berliner Und Munchener Tierarztliche Wochenschrift* **114**(9-10), 356–359.

Swanenburg, M., Urlings, H. A., Snijders, J. M., Keuzenkamp, D. A. & van Knapen, F. (2001*b*), '*Salmonella* in slaughter pigs: prevalence, serotypes and critical control points during slaughter in two slaughterhouses', *International Journal of Food Microbiology* **70**(3), 243–254.

Takahashi, K., Kulldorff, M., Tango, T. & Yih, K. (2008), 'A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring', *International Journal of Health Geographics* **7**, 14.

Tam, C. C., Rodrigues, L. C., O'Brien, S. J. & Hajat, S. (2006), 'Temperature dependence of reported *Campylobacter* infection in England, 1989-1999', *Epidemiology and Infection* **134**(1), 119–125.

Tango, T. (1999), Comparison of general tests for spatial clustering, *in* A. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J. Viel & R. Bertollini, eds, 'Disease mapping and risk assessment for public health', John Wiley and Sons, Chichester, pp. 111–117.

Tango, T. & Takahashi, K. (2005), 'A flexibly shaped spatial scan statistic for detecting clusters', *International Journal of Health Geographics* **4**, 11.

Taylor, L., Latham, S. & Woolhouse, M. E. (2001), 'Risk factors for human disease emergence', *Philosophical Transactions of the Royal Society B: Biological Sciences* **356**, 983–989.

Thacker, S., Parrish, R. & Trowbridge, F. (1988), 'A method for evaluating systems of epidemiological surveillance', *World Health Statistics Quarterly* **41**, 11–18.

Thornley, C. N., Baker, M. G., Weinstein, P. & Maas, E. W. (2002), 'Changing epidemiology of human leptospirosis in New Zealand', *Epidemiology and Infection* **128**(1), 29–36.

Thornley, C. N., Simmons, G. C., Callaghan, M. L., Nicol, C. M., Baker, M. G., Gilmore, K. S. & Garrett, N. K. (2003), 'First incursion of *Salmonella enterica* serotype Typhimurium DT160 into New Zealand', *Emerging Infectious Diseases* **9**(4), 493–495.

Tobias, A., Diaz, J., Saez, M. & Alberdi, J. C. (2001), 'Use of Poisson regression and Box-Jenkins models to evaluate the short-term effects of environmental noise levels on daily emergency admissions in Madrid, Spain', *European Journal of Epidemiology* **17**(8), 765–771.

Toft, N., Innocent, G. T., Gettinby, G. & Reid, S. W. (2007), 'Assessing the convergence of Markov Chain Monte Carlo methods: An example from evaluation of diagnostic tests in absence of a gold standard', *Preventive Veterinary Medicine* **79**(2-4), 244–256.

Tong, S. & Hu, W. (2002), 'Different responses of Ross River virus to climate variability between coastline and inland cities in Queensland, Australia', *Occupational and Environmental Medicine* **59**(11), 739–744.

Touloumi, G., Samoli, E., Pipikou, M., Le Tertre, A., Atkinson, R. & Katsouyanni, K. (2006), 'Seasonal confounding in air pollution and health time-series studies: effect on air pollution effect estimates', *Statistics in Medicine* **25**(24), 4164–4178.

Trottier, H., Philippe, P. & Roy, R. (2006), 'Stochastic modeling of empirical time series of childhood infectious diseases data before and after mass vaccination', *Emerging Themes in Epidemiology* **3**, 9.

Tukey, J. (1977), *Exploratory data analysis*, Addison-Wesley, Reading, Massachusetts.

Upshur, R. E., Moineddin, R., Crighton, E., Kiefer, L. & Mamdani, M. (2005), 'Simplicity within complexity: seasonality and predictability of hospital admissions in the province of Ontario 1988-2001, a population-based analysis', *BMC Health Services Research* **5**, 13.

van Beneden, C., Olsen, S., Skoff, T. & Lynfield, R. (2007), Active, population-based surveillance for infectious disease, *in* N. M'ikanatha, R. Lynfield, C. van Beneden & H. de Valk, eds, 'Infectious Disease Surveillance', first edn, Blackwell Publishing, Massachusetts, pp. 32–43.

van der Gaag, M. A., Saatkamp, H. W., Backus, G. B. C., van Beek, P. & Huirne, R. B. M. (2004*a*), 'Cost-effectiveness of controlling *Salmonella* in the pork chain', *Food Control* **15**(3), 173–180.

van der Gaag, M. A., Vos, F., Saatkamp, H. W., van Boven, M., van Beek, P. & Huirne, R. B. M. (2004*b*), 'A state-transition simulation model for the spread of *Salmonella* in the pork supply chain', *European Journal of Operational Research* **156**(3), 782–798.

van der Wolf, P., Wolbers, W., Elbers, A., van der Heijden, H., Koppen, J., Hunneman, W., van Schie, F. & Tielen, M. (2001), 'Herd level husbandry factors associated with the serological *Salmonella* prevalence in finishing pig herds in The Netherlands', *Veterinary Microbiology* **78**(3), 205–219.

van Pelt, W., Mevius, D., Stoelhorst, H. G., Kovats, S., van de Giessen, A. W., Wannet, W. & Duynhoven, Y. T. (2004), 'A large increase of *Salmonella* infections in 2003 in The Netherlands: hot summer or side effect of the avian influenza outbreak?', *Euro Surveillance* **9**(7), 17–19.

Vellinga, A. & Van Loock, F. (2002), 'The dioxin crisis as experiment to determine poultry-related *Campylobacter* enteritis', *Emerging Infectious Diseases* **8**(1), 19–22.

Verbeke, W., Frewer, L. J., Scholderer, J. & De Brabander, H. F. (2007), 'Why consumers behave as they do with respect to food safety and risk information', *Analytica Chimica Acta* **586**(1-2), 2–7.

Vieira, V. M., Webster, T. F., Weinberg, J. M. & Aschengrau, A. (2008), 'Spatial-temporal analysis of breast cancer in upper Cape Cod, Massachusetts', *International Journal of Health Geographics* **7**, 46.

Vigre, H., Baekbo, P., Jorsal, S., Bille-Hansen, V., Hassing, A., Enoe, C. & Botner, A. (2005), 'Spatial and temporal patterns of pig herds diagnosed with postweaning multisystemic wasting syndrome during the first two years of its occurrence in Denmark', *Veterinary Microbiology* **110**(1-2), 17–26.

Vorou, R. M., Papavassiliou, V. G. & Tsiodras, S. (2007), 'Emerging zoonoses and vector-borne infections affecting humans in Europe', *Epidemiology and Infection* **135**(8), 1231–1247.

Vourc'h, G., Bridges, V. E., Gibbens, J., De Groot, B. D., McIntyre, L., Poland, R. & Barnouin, J. (2006), 'Detecting emerging diseases in farm animals through clinical observations', *Emerging Infectious Diseases* **12**(2), 204–210.

Wakefield, J., Kelsall, J. & Morris, S. (2001), Clustering, cluster detection, and spatial variation in risk, *in* P. Elliott, J. Wakefield, N. Best & D. Briggs, eds, 'Spatial epidemiology: methods and application', Oxford University Press, Oxford, pp. 128–152.

Waller, L. A., Hill, E. G. & Rudd, R. A. (2006), 'The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations', *Statistics in Medicine* **25**(5), 853–865.

Ward, M. P. (2002), 'Clustering of reported cases of leptospirosis among dogs in the United States and Canada', *Preventive Veterinary Medicine* **56**(3), 215–226.

Ward, M. P. & Carpenter, T. E. (2000*a*), 'Analysis of time-space clustering in veterinary epidemiology', *Preventive Veterinary Medicine* **43**(4), 225–237.

Ward, M. P. & Carpenter, T. E. (2000*b*), 'Techniques for analysis of disease clustering in space and in time in veterinary epidemiology', *Preventive Veterinary Medicine* **45**(3-4), 257–284.

Wartenberg, D. (2001), 'Investigating disease clusters: why, when and how?', *Journal of the Royal Statistical Society: Series A* **164**(1), 13–22.

Watkins, R. E., Eagleson, S., Veenendaal, B., Wright, G. & Plant, A. J. (2008), 'Applying cusum-based methods for the detection of outbreaks of Ross River virus disease in Western Australia', *BMC Medical Informatics and Decision Making* **8**, 37.

Webster, S., Diggle, P. J., Clough, H., Green, R. & French, N. P. (2006), Strain-typing transmissible spongiform encephalopathies using replicated spatial data, *in* A. Baddeley, ed., 'Case Studies in Spatial Point Process Modeling', Springer, Berlin, p. 332.

Wegener, H. & Baggesen, D. (1996), 'Investigation of an outbreak of human salmonellosis caused by *Salmonella enterica ssp. enterica* serovar Infantis by use of pulsed field gel electrophoresis', *International Journal of Food Microbiology* **32**(1-2), 125–131.

Weinberger, M., Andorn, N., Agmon, V., Cohen, D., Shohat, T. & Pitlik, S. D. (2004), 'Blood invasiveness of *Salmonella enterica* as a function of age and serotype', *Epidemiology and Infection* **132**(6), 1023–1028.

West, H. & Harrison, P. (1997), *Bayesian forecasting and dynamic models*, second edn, Springer, New York.

Westgarth, C., Gaskell, R. M., Pinchbeck, G. L., Bradshaw, J. W., Dawson, S. & Christley, R. M. (2008), 'Walking the dog: exploration of the contact networks between dogs in a community', *Epidemiology and Infection*. Nov 19, Epub ahead of print.

Wheeler, D. C. (2007), 'A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003', *International Journal of Health Geographics* **6**, 13.

Wheeler, J. G., Sethi, D., Cowden, J. M., Wall, P. G., Rodrigues, L. C., Tompkins, D. S., Hudson, M. J. & Roderick, P. J. (1999), 'Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance', *British Medical Journal* **318**(7190), 1046–1050.

Wilesmith, J. W., Stevenson, M. A., King, C. B. & Morris, R. S. (2003), 'Spatio-temporal epidemiology of foot-and-mouth disease in two counties of Great Britain in 2001', *Preventive Veterinary Medicine* **61**(3), 157–170.

Willeberg, P. (2006), From Venn to now – science and application, *in* '11th International Symposium on Veterinary Epidemiology and Economics', Cairns, Australia.

Williamson, G. D. & Weatherby Hudson, G. (1999), 'A monitoring system for detecting aberrations in public health surveillance reports', *Statistics in Medicine* **18**(23), 3283–3298.

Wills, R., Gray, J., Fedorka-Cray, P., Yoon, K., Ladely, S. & Zimmerman, J. (2000), 'Synergism between porcine reproductive and respiratory syndrome virus and *Salmonella choleraesuis* in swine', *Veterinary Microbiology* **71**(3-4), 177–192.

Wilson, A. M., Salloway, J. C., Wake, C. P. & Kelly, T. (2004), 'Air pollution and the demand for hospital services: a review', *Environment International* **30**(8), 1109–1118.

Winfield, M. D. & Groisman, E. A. (2003), 'Role of nonhost environments in the lifestyles of *Salmonella* and *Escherichia coli*', *Applied Environmental Microbiology* **69**(7), 3687–3694.

Woodall, W. (2006), 'The use of control charts in health-care and public-health surveillance', *Journal of Quality Technology* **38**, 89–104.

Woolhouse, M. E. & Gowtage-Sequeria, S. (2005), 'Host range and emerging and reemerging pathogens', *Emerging Infectious Diseases* **11**(12), 1842–1847.

Yoon, H., Wee, S. H., Stevenson, M. A., O'Leary B, D., Morris, R. S., Hwang, I. J., Park, C. K. & Stern, M. W. (2006), 'Simulation analyses to evaluate alternative control strategies for the 2002 foot-and-mouth disease outbreak in the republic of Korea', *Preventive Veterinary Medicine* **74**(2-3), 212–225.

Yu, J. & Xi, L. (2009), 'A neural network ensemble-based model for on-line monitoring and diagnosis of out-of-control signals in multivariate manufacturing processes', *Expert Systems with Applications* **36**, 909–921.

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

Zepeda, C. & Salman, M. (2003), Planning survey, surveillance, and monitoring systems - roles and requirements, *in* M. Salman, ed., 'Animal Disease Surveillance and Survey Systems', Iowa State Press, Iowa, pp. 35–46.

Zhang, X., McEwen, B., Mann, E. & Martin, W. (2005), 'Detection of clusters of *Salmonella* in animals in Ontario from 1991 to 2001', *Canadian Veterinary Journal* **46**(6), 517–523.

Zhang, Y., Bi, P. & Hiller, J. (2008), 'Climate variations and salmonellosis transmission in Adelaide, South Australia: a comparison between regression models', *International Journal of Biometeorology* **52**, 179–187.

Zheng, P., Durr, P. A. & Diggle, P. J. (2004), Edge correction for spatial kernel smoothing methods - when is it neccessary?, *in* 'GisVET', Ontario.

Zinsstag, J., Schelling, E., Wyss, K. & Mahamat, M. B. (2005), 'Potential of co-operation between human and animal health to strengthen health systems', *Lancet* **366**(9503), 2142–2145.

APPENDIX A

Appendix 1

A.1 Introduction

This material about the handling of missing values appears as an appendix to Chapter 5. During the process of review for publication, this material was removed at the request of the reviewers and editors. They felt the inclusion of missing values pulled the manuscript in too many different directions. For completeness, I include this extra material which originally appeared in the unpublished version of Chapter 5.

A.2 Materials and methods

A.2.1 Data description and handling

A lot of data were missing. The variable feed type had the highest percentage of missing values at 29%. In other variables, missing values ranged from 3% to 11%. The variables of herd size and access to straw had no missing values and there were no missing serology results. We investigated how the missing covariate data related to seropositivity by testing the null hypothesis that the proportion of pigs positive was the same for farms both with (n = 1504) and without (n = 2280) missing covariate information (Newcombe 1998).

A.2.2 Risk factor analysis

To reduce the bias that might have been associated with complete case analysis, we imputed missing values in WinBUGS and re-ran the model using data from all 3784 farms.

As all missing data were binary, they were modelled by giving each missing value a number drawn from an arbitrary distribution: Bernoulli (0.5). Sensitivity to this was evaluated by re-running the models with missing covariate data drawn from extreme distributions: Bernoulli (0) and Bernoulli (1).

A.3 Results

The proportion of pigs positive was 10.4% for farms with missing covariate information, and 10.1% for farms with complete covariate information. The proportion of pigs positive was 10.7% for farms with missing feed type information, and 10% for farms with this information. The proportion of pigs positive was 8.7% for farms with missing health status information, and 10.3% for farms with this information. The proportion of pigs positive was 11.8% for farms with missing feed supply information, and 10% for farms with this information. All p-values were less than 0.03, indicating that the null hypothesis of no difference between these two groups (farms with and without missing covariate information) could be rejected.

Table A.1 shows the results using all 3784 farms with imputed data for the missing covariates. The direction of effect for each covariate was the same as in the complete case analysis (Table 5.3), but the magnitude of effect was reduced by approximately 5% for feed type, 10% for feed supply, and 20% for health status. There was no difference in the monitored parameters when the covariate priors were varied. However, the model was mildly sensitive to missing covariate data being drawn from the extreme distributions. With both Bernoulli (0) and (1) distributions, there were minor differences in the random farm effects and variance terms, with more substantial differences in the estimated regression coefficients. The direction of effect for each covariate was the same, although the magnitude of effect was reduced by approximately 10% for feed type, feed supply, and health status.

Table A.1: Factors associated with *Salmonella* seropositivity in 45,103 meat-juice ELISA results, taken from 3784 Danish finisher pig herds from 1st October to 31st December 1995, using values for the missing covariates drawn from a Bernoulli (0.5) distribution. Data originate from the Danish swine *Salmonella* surveillance and control programme.

Variable	Level	Posterior Mean	Posterior SD	MC error	OR(95% CI)
Herd size[1]	continuous	0.02	0.01	<0.01	1.02 (1.01–1.04)
Feed type	wet or mixed	-0.34	0.07	<0.01	0.71 (0.63–0.81)[2]
	dry	reference			
Feed supply	purchased	0.57	0.05	<0.01	1.77 (1.61–1.94)
	home mixed or both	reference			
Health status	SPF	-0.23	0.08	<0.01	0.80 (0.68–0.94)
	conventional	reference			

Model Statistics: Intercept, -2.86; DIC, 12407.78.
SD: Standard deviation; CI: Bayesian credible interval; MC error: Monte Carlo standard error of the posterior mean; OR: odds ratio
[1] Number of pen places for finishers (rescaled by subtracting the mean, then dividing by 100).
[2] *Interpretation*: Once adjusted for herd size, feed supply, and health status, a pig on a farm using wet-feeding had 0.71 times the odds of being *Salmonella* positive compared with a pig on a dry-feeding farm (95% CI: 0.63–0.81).

A.4 Discussion

Using only complete cases of data will introduce bias if the missing values are not completely at random (Schafer & Graham 2002). We overcame this issue and made use of data from all 3784 farms by imputing the missing values within a Bayesian framework. However, it is important to consider that our use of both imputed data and within-commune randomly generated coordinates may itself introduce bias or increased random error. We checked that this was not the case in our study by performing sensitivity analyses which compares the results both with and without the imputed data (Schafer & Graham 2002).