Nonlinear Time Series Analysis of Zoonoses

Lasse Engbo Christiansen

Kongens Lyngby 2004
IMM-PHD-2004-137

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Preface

This thesis was prepared at Informatics and Mathematical Modelling (IMM), the Technical University of Denmark (DTU) in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering. The project has been carried out in the period 15 August 2001 to 14 October 2004 in collaboration with the Danish Zoonosis Centre, now part of the Danish Institute for Food and Veterinary Research (DFVF), where half of the work has been done.

The thesis deals with applied statistics and mathematical modelling as tools for veterinary epidemiologists. The thesis consists of a summary report and a collection of four research papers written during the Ph.D. study and published elsewhere.

Lyngby, October 2004

Lasse Christiansen

Lasse Engbo Christiansen

Acknowledgements

First of all I would like to thank my supervisors. At IMM Prof. Henrik Madsen for his interest, support, and guidance during the last three years. At DFVF Prof. Henrik Caspar Wegener for making this collaboration possible, putting the research into a broader perspective, and his guidance. At DFVF Jens Strodl Andersen for many fruitful discussions and support throughout the project.

I am also grateful for the support from the Section for Antimicrobial Resistance at DFVF in particular Prof. Frank Møller Aarestrup for supporting the work on copper and Inge M. Hansen for doing the lab work.

Furthermore I wish to thank my fellow Ph.D. students and colleagues in the Mathematical Statistics section and my colleagues in the Department of Epidemiology and Risk Assessment at DFVF.

Finally, I am grateful to my wife, Astrid Christiansen, for her support and patience.

Summary

The present thesis consists of a summary report and four research papers. The subject of the thesis is temporal and spatial correlations in epidemiologic data.

The summary report consists of three parts. The first part is an application of locally weighted least squares, the second part suggests expansions of the spatial scan statistic, and the third part is on modelling using stochastic differential equations.

The first part is related to Paper B and the problem was how climatic variables predict *Campylobacter* in broiler flocks and humans. The problem was targeted using locally weighted least squares with zero to second order polynomials; cross validation was used to avoid overfitting the data. It was found that temperature is a good predictor for *Campylobacter* infections. For broiler flocks the highest proportion of positive flocks was seen in weeks with sustained high temperatures throughout a week, for humans warm days and more sunshine resulted in the highest incidence. The dependence was non-linear and the lag between high temperature and registred infection was 3 to 4 weeks. The ambient temperature is believed to drive something else that can transfer infections to broiler flocks.

The second part is related to Paper C and the question was if there was spatial clustering of the proportion of positive broiler flocks. The spatial scan statistic is able to both locate and test for significance of a cluster, the problem with the method is that previously only circular clusters were allowed. This work presents an algorithm for creation of the set of confocal ellipses which is a natural expansion of the set of concentric circles previously used. A new way of presenting the set of clusters is also presented. Clustering was found to occur;

Funen and the nearest part of Jutland is a low risk region whereas the Northwest part of Jutland is a high risk region for *Campylobacter* infected broiler flocks.

The third part is related to Paper D and E and the problem was to make a model for growth in a rich media including the dependence on the concentration of copper in the media. Bacterial growth involves many processes within the cells and it is shown why the continuous-discrete stochastic state space models have an advantage over the often used ordinary differential equations. The parameters of five candidate models have been estimated using data for one concentration of copper. With the given dataset the simpler models performed better than the more complicated models. It was chosen to implement two simple copper dependencies in three of the models. In all three models the multiplicative approach was better than the additive approach.

In all three parts suggestions for further work have been proposed.

Resumé

Denne afhandling består af en sammenfatning og fire videnskabelige artikler. Emnet for afhandlinger er tidslige og rumlige afhængigheder i epidemiologiske data.

Sammenfatningen består af tre dele. Den første del viser anvendelsen af en lokalt vægtet mindste kvadraters metode, den anden del præsenterer to udvidelser til en algoritme til at finde rumlige klynger af sygdomme, og den tredie del er en anvendelse af stokastiske differential ligninger.

Den første del er relateret til artikel B og omhandler brugen af vejrobservationer til prædiktion af *Campylobacter* infektioner i kyllingeflokke og mennesker. Lokalt vægtet mindste kvadrater med polynomier af nulte til anden orden blev brugt. Krydsvalidering blev anvendt for at undgå overfitning af data. Resultatet var at temperatur er en god prædiktor for *Campylobacter* infektioner. De højeste andele positive kyllingeflokke forekom når der var stabile høje temperaturer. For antallet af inficerede mennesker var risikoen størst når der havde været ekstra varme solskinsdage. Temperaturafhængigheden er for begge grupper ikkelineær, og det viste sig, at man opnåede den bedste prædiktion ved at benytte temperatur målinger 3 til 4 uger før infektionerne blev registreret. Den omgivende temperatur tænkes at påvirke et eller andet, som kan overføre infektioner til kyllinge flokke.

Den anden del er relateret til artikel C og spørgsmålet var om der var geografiske klyngedannelser i andelen af positive kyllingeflokke. Metoden spatial scan statistic kan både lokalisere og bestemme signifikans af klyngedannelse. Afhandlingen præsenterer en algoritme til at lave en familie af konfokale ellipser, som er en naturlig udvidelse af den normalt anvendte familie af koncentriske cirkler. Der

foreslås også en ny måde hvorpå man kan præsentere informationen i et sæt af klynger. Der var signifikant klyngedannelse i datasættet. Fyn og den nærmeste del af Jylland er et lavrisiko område og det nordvestlige Jylland er et højrisiko område for *Campylobacter* infektioner i kyllinge flokke.

Den tredie del er relateret til artiklerne D og E og omhandler modellering af bakterievækst i et rigt medie, og desuden hvordan væksten afhænger af koncentrationen af kobber i mediet. Der er mange processer involveret i mikrobiel vækst, og dette er med til at vise at kontinuere diskret observerede stokastiske tilstands modeller har fordele sammenlignet med de ofte anvendte ordinære differential ligninger. Parametrene i fem modeller er blevet estimeret på baggrund af data for én koncentration af kobber. Med det givne datasæt gav de simpleste modeller de bedste prædiktioner. Det blev besluttet at indbygge to simple kobber afhængigheder, en multiplikativ og en additiv, i de tre modeller. For alle tre modeller gav den multiplikative afhængighed en højere likelihood end den additive.

I alle tre dele er der forslag til videre arbejde.

Papers included in the thesis

[B] Mary Evans Patrick, Lasse Engbo Christiansen, Michael Wainø, Steen Ethelberg, Henrik Madsen, and Henrik Caspar Wegener. Effects of Climate on Incidence of Campylobacter spp. in Humans and Prevalence in Broiler Flocks in Denmark. *Applied and Environmental Microbiology*, 2004. Accepted for publication.

[C] Lasse Engbo Christiansen, Jens Strodl Andersen, Henrik Caspar Wegener, and Henrik Madsen. Spatial scan statistics using elliptic clusters. *J. of Agricultural, Biological, and Environmental Statistics*, 2004. Submitted.

[D] Lasse Engbo Christiansen, Jens Strodl Andersen, Henrik CasparWegener, and Henrik Madsen. Estimating the parameters of a stochastic differential equation model of bacterial growth. *Biometrics*, 2004. Submitted.

[E] Lasse Engbo Christiansen, Jens Strodl Andersen, Henrik CasparWegener, Frank Aarestrup, and Henrik Madsen. Modelling microbial growth in a rich medium and the effect of adding copper. *J. of Theoretical Biology*, 2004. Submitted.

Contents

Chapter 1

Introduction

1.1 Background

In modern epidemiology increasingly sophisticated methods are needed in order to utilize the increased amount of information on time, space, and hierachical structure in the data. The aim is to describe the correlations or dependencies in the data and it is no longer restricted to pure linear correlations - in many applications non-linear relations are needed.

The increased focus on food safety over the last decade is one of the driving forces in collecting more and more information. One of the means for collecting information is the surveillance programs; many of which are targeted at preventing human diseases.

A Danish example is Salmonella, that became a public issue in the late eighties. At that time the main source of human salmonellosis was *Salmonella* in broiler chickens. Therefore a control program was implemented in 1988. This was effective but the number of human infections went up again due to infected pork, this was targeted by another control program in 1993. Since 1994 shell eggs have been the largest single source of human incidence. This lead to a third surveillance based control program in 1997, and serologic and bactericologic tests are now used to direct eggs from infected flocks to heat treatment (Wegener et al., 2003). Eggs are still the largest source but the total number has come down. Until recent years Salmonella was responsible for the largest part of human infections with zoonotic[1] bacteria. Today *Campylobacter* has taken the lead. Again interventions have been implemented and a decrease in human cases has been observed.

Even though the number of cases is dropping there is still a high number of cases and it is known that only a small part of the actual human cases is registered. There is still room for improvements in the collection and usage of the surveillance data. The Danish Zoonosis Centre (DZC), now part of the Danish Institute for Food and Veterinary Research, is recieving test results from both veterinary, food, and human samples. The centre is responsible for prevention and control of zoonoses, hence there is a need for mathematical and statistical methods that can guide in this work.

The aim of this research is to demonstrate and develop tools that can help epidemiologists explain the observed variation corresponding to a description of the underlaying correlations and relations. It is the hope that ultimatively this thesis indirectly will help reducing the number of infections even further.

1.2 Overview of the thesis

The present chapter describes the background and motivation for the research and an outline of the thesis.

Chapter 2 provides a broader introduction to Paper B concerning the seasonal variations in *Campylobacter* infections in broiler flocks and humans. It also presents the used model in further detail and finally presents an outlook.

Chapter 3 is related to Paper C and contains a brief description of the model for spatial clustering and possible further enhancements of the model. This also includes further applications for the model.

Chapter 4 contains a description of the framework used to describe bacterial growth in a rich media. It also contains further background information and motivation than what is included in Paper D and E.

Chapter 5 is a discussion that summarizes the Chapters 2 trough 4 and presents a joint outlook.

[1]Zoonoses are infections transmissible from animals to humans.

Finally, Chapter 6 contains the overall conclusions.

Appendix B trough E contain the papers included in the thesis.

Chapter 2

Campylobacter and climate

The last few years *Campylobacter* has been the leading cause of enteric infections in Denmark. The same is observed in most developed and developing countries (Anonymous, 2000). Previously *Salmonella* was the predominant but implementation of control strategies has decreased the number of cases and in the meanwhile the number of human cases of campylobacteriosis has increased. The increase stopped two years ago and the reason is believed to be so-called logistic slaughtering, meaning that broilers from houses that usually deliver *Campylobacter* free flocks are used for production of fresh meat and the rest is frozen. Freezing meat infected with *Campylobacter* has been shown to reduce the count of bacteria with a factor 100 and thus lowering the risk of becoming infected when eating undercooked meat or cross-contaminated food. Furthermore *Campylobacter* differ from *Salmonella* and other bacteria as it in most cases doesn't grow at room temperature and below.

A pronounced seasonality in the number of infections with *Campylobacter* is observed, see Fig. 2.1. The ambition is to reduce the number of *Campylobacter* infections both in humans and in broilers. One important problem is how to reduce the observed seasonality? One solution would be to have winter all year round - but that is rather unlikely. One step on the way to an answer is to investigate if one or more climatic variables can explain the seasonality and especially trying to describe the part not explained directly by the climatic variables, i.e. are the residuals independent?

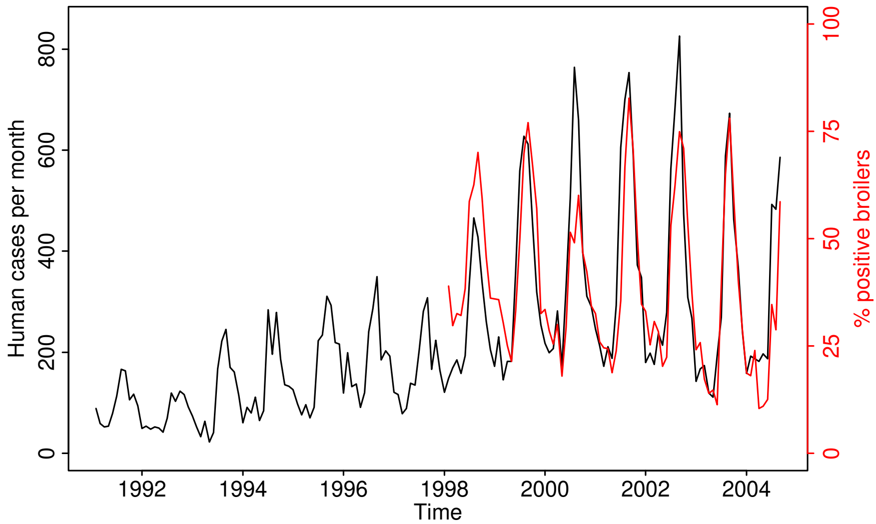

Figure 2.1: Seasonality of campylobacteriosis in humans and Campylobacter infections in broiler flocks in Denmark, 1991-2004

2.1 Data

The study period is 1998 through 2001. The reason for this time-span is that a national monitoring program for *Campylobacter* in Danish broiler flocks at slaughter was initiated January 1 1998 and the sampling plan was changed in the beginning of 2002. In the beginning of the slaughter line ten birds from each flock or batch were examined for *Campylobacter* by cloacal swabs. Each swab was tested individually and if one or more swabs from the same batch was positive the batch was labeled positive. One flock may be slaughtered in up to six batches, and in order to eliminate the correlation between batches only the first batch was included in the analysis. A total of 17,941 flocks were included in the analysis and 6,860 (38%) tested positive for *Campylobacter* spp., for further details see Paper B.

Human cases of campylobacteriosis are reported to Statens Serum Institut and after excluding cases from counties not reporting on a weekly basis, those with a known travel history, and those with missing county information a total of 12,008 human cases were included in the analysis. The background population was 4.4 million giving an average incidence of 68 cases per 100,000 population per year.

Both broiler flock status and human cases were aggregated on a weekly basis. Making the proportion of *Campylobacter* positive broiler flocks slaughtered within each week and the number of human cases of campylobacteriosis per week the two response variables.

Climate data was provided by the Danish Meteorological Institute (DMI). The variables used as explanatory variables were average ambient air temperature, maximum temperature, relative humidity, precipitation, and hours with sunlight. Again the variables are weekly country wide averages.

2.2 Mathematical model

The broiler data and the human data were modelled independently to see if they are predicted by the same climatic variables. The human count data was transformed by the logarithm prior to analysis and the proportion of infected broiler flocks was untransformed in the analysis. The outcome is denoted Y_t with t being the week number. Likewise $\mathbf{X}_t$ denotes the vector of climatic observations in week t. In the analysis one or two climatic variables are used at the same time, so the length of $\mathbf{X}_t$ is either 1 or 2. It is assumed that the weather of interest was in the preceding weeks, and hence a lag, $n > 0$, is introduced; i.e. $\mathbf{X}_{t-n}$ is the climate observations n weeks prior to t.

The model structure used is locally weighted polynomial regression. For each fitting point $\mathbf{x}$ the problem can be formulated as:

$$\arg\min_{\theta} \frac{1}{N} \sum_{t=1}^{N} w(\mathbf{X}_{t-n}, \mathbf{x})(Y_t - P(\mathbf{X}_{t-n}, \mathbf{x}, \theta))^2 \tag{2.1}$$

where $w(\mathbf{X}_s, \mathbf{x})$ is the weight function and $P(\mathbf{X}_s, \mathbf{x}, \theta)$ are the local polynomials around $\mathbf{x}$. The weight function was chosen to have a sperical kernel:

$$w(\mathbf{X}_s, \mathbf{x}) = W\left(\frac{\|\mathbf{X}_s - \mathbf{x}\|}{h(\mathbf{x})}\right)$$

The bandwidth, $h(\mathbf{x})$, is a nearest neighbour bandwidth meaning that $h(\mathbf{x})$ is chosen so that a certain fraction, α, of the data is within $h(\mathbf{x})$ from $\mathbf{x}$. The kernel function must be never increasing; it was chosen to use the tricube function:

$$W(u) = \begin{cases} (1-u^3)^3 & , u \in [0;1) \\ 0 & , u \in [1;\infty) \end{cases}$$

The local polynomials were chosen to be of zero, first, or second order. Either one or two climatic predictors were used at a time. With two predictors,

$\mathbf{x} = [x_1, x_2]^T$, and second order polynomials we have:

$$P(\mathbf{X}_s, \mathbf{x}, \theta) = f(\mathbf{X}_s - \mathbf{x}, \theta)$$

$$f(\mathbf{z}, \theta) = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \theta_{11} z_1^2 + \theta_{12} z_1 z_2 + \theta_{22} z_2^2$$

Lower order polynomials are obtained by removing the inappropriate terms.

It was chosen to fit the model in evenly distributed points within the range of the explanatory variables. This causes no problems in the one dimensional case but in the two dimensional case the observations will not span the full square given by the independent ranges. To get around this it was chosen only to estimate in fitting points within the convex hull spanned by the observations.

In order not to overfit the data it was chosen to use cross validation (CV) to select the optimal model. Four fold CV was used with folds over the years, i.e. all combinations of three years were used to estimate a model and then that model was used to predict the remaining year. The sum of squared prediction errors was then used to select the optimal model. The parameters to be selected are the lag of the predictors, $n \in [1, 5]$weeks, the order of the fitting polynomial, and the fraction, α, of the data used as bandwidth.

2.3 Summary of results

The nearest neighbour bandwidth was allowed to be within $[0.15; 0.50]$ with discrete jumps of 0.05. The lag was allowed to be 1 to 5 weeks and as mentioned above the order of the fitting polynomials ranged from 0 to 2. This resulted in 120 different models for each choice of one or two climatic variables. Based on the CV error one of those 120 models was selected for each choice of predictors and those are the models listed in Table B.1 and B.2 in Paper B.

As expected ambient air temperature was a very good predictor both for broilers and humans. Indeed, the maximum temperature with a lag of 4 weeks was the best single predictor of reported human cases of campylobacteriosis, 3 weeks was almost as good. For broiler flocks the average temperature 3 weeks prior to slaughter was the best single predictor. So when only one predictor is used humans and broilers seem to be linked somehow.

When using two predictors for human campylobacteriosis the combination of average temperature and hours of sunlight gave the lowest CV error. The only other pair having a CV error lower than when using one predictor was the combination of average and maximum temperature. For both models the lag was estimated to be 4 weeks. For broiler flocks the combination of average and maximum temperature lagged 3 weeks gave the best prediction. The combination of average temperature and hours with sunlight was better than using average temperature alone and again those two combinations are the only pairs that lowers the CV error.

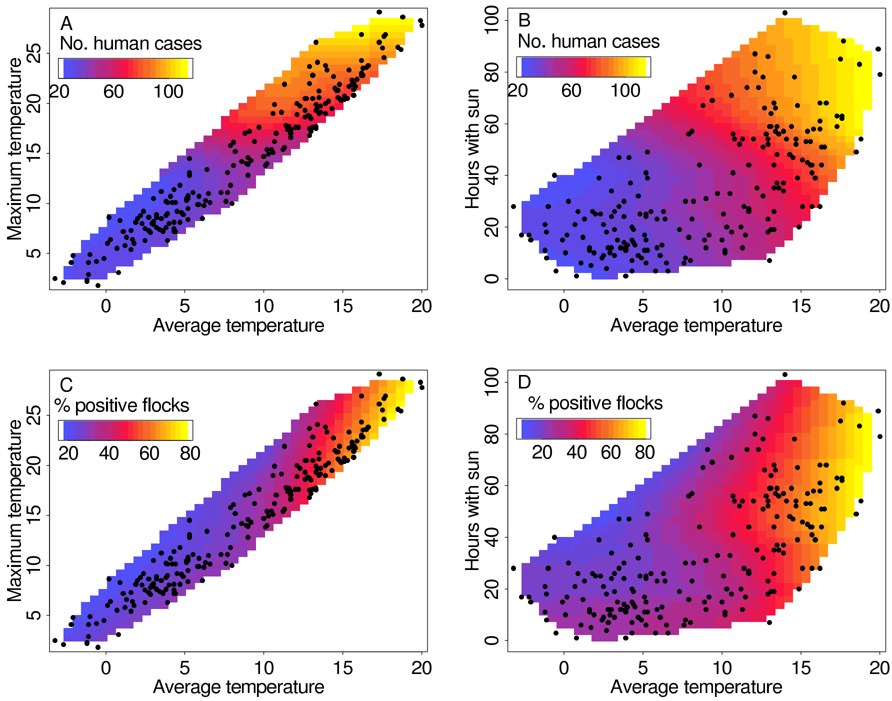

Figure 2.2: Effects of temperature and hours of sunlight on the incidence of campylobacteriosis in humans and the percent of Campylobacter infected broiler flocks at slaughter. Black dots are observations. (Copy of Fig. B.3.)

Fig. 2.2 shows the combination of average temperature with maximum temperature and hours with sunlight for both humans and broiler flocks. Comparing the dependence of human cases and % positive broiler flocks on average and maximum temperature it appears that the two have different links to temperature. For humans the largest dependence is on the maximum temperature and for a given average temperature the incidence increases as the maximum temperature increases. For broilers it appears to be quite different as for a given average temperature the expected proportion of positive flocks decreases when the maximum temperature increases. Also the dependence on sunlight is different - for humans an increase in the number of hours with sunlight neutralizes the protective effect of a decrease in average temperature.

Considering the residuals it becomes clear that the average temperature is not enough to explain the proportion of positive broiler flocks. Fig. B.5 shows the average temperature and the percentage of positive flocks over the winter and it is seen that the temperature follows a rather symmetric season whereas the broilers becomes less and less infected from autumn to late spring (Week No. 18) where a threshold is passed and the next summer peak begins.

2.4 Discussion and outlook

There is definitely a very pronounced seasonality, but based on this research the association to climate is not the same for humans and broilers. Regarding the broiler flocks it is seen that for a given average temperature a larger proportion of positive flocks is expected if the maximum temperature is relatively low. This corresponds to sustained high temperatures throughout the week. The lag of 3 weeks between high temperature and time of slaughter is in the middle of the growing period and as temperature (especially average temperature) is highly correlated between weeks, the lag of 3 weeks represents the growing period, which is about 39 days, quite well.

Modern broiler houses are climate controlled. The control objectives are that the indoor temperature has to follow a fixed curve during the growing period and that the relative humidity has allowed ranges during different stages of the growing period. The controller can adjust the ventilation and a heat source. The broilers produce a lot of energy so ventilation is always needed. When the surrounding temperature is high a larger air flow is needed to cool the house. During the warmest summer months in Denmark the broiler producers have fewer broilers in the houses in order to make it possible to cool down the house. During winter cold air is draged in and heated; the heating lowers the relative humidity so only a small air flow is needed to comply with the controller objectives. As the indoor climate is approximately the same during summer and winter it seems as if the inlet air should be targeted.

One study by Hald et al. (2004) was performed in the summer of 2003. Polyester nets were mounted over two wall inlets and a lot of flies were captured, speciated and counted. At the time of harvesting those flies were dead and dried. Therefore, tests for *Campylobacter* were performed on flies captured within 50m from the houses. It was possible to culture *Campylobacter jejuni* from 8.2% of the flies and using PCR methods 70.2% of the flies were positive for *Campylobacter*. Experimentally it has been shown that flies are able to transmit *C.*

jejuni to chickens (Shane et al., 1985). A new dataset with fly data is being collected over the summer of 2004 and will soon be analysed; it is a comparison between houses with and without permanent nets on all air inlets. Among other suggested sources are beetles living in the houses (Jacobs-Reitsma et al., 2003).

For humans it is different as for a given average temperature the incidence is predicted to increase with higher maximum temperatures which correlates with the number of hours with sunlight. One interpretation of this is that the human incidence is related to a good weather index. Hence, the focus should be on how warm and sunny days changes human behavior. Two obvious links are barbecuing and outdoor activities. Barbecuing is one of the recognized risk-factors for human campylobacteriosis (Neimann et al., 2003).

The estimated lag of 4 weeks from good weather to human infections corresponds quite well with the average time between onset of illness, visiting a healthcare facility, providing a culture and receiving a diagnosis which is 21 days (J. Neimann - personal communication). On top of the 21 days is the incubation period which usually is 1-7 days (Skirrow and Blaser, 2000). This fits quite well with a lag of about 4 weeks.

Some ideas for further research: The hypothesis that flies are involved in the summer peak observed in broiler flocks is being assessed. The amount of data that is being collected is not large and hence the study should be repeated on a larger scale. One suggestion would be to make a longitudinal analysis using ambient temperatures, indoor temperature, fan speed, fly count, and the *Campylobacter* status of the flocks from the houses and the flies.

As broiler meat is thought of as being the major source of *Campylobacter* bacteria leading to human campylobacteriosis the most efficient way will be to lower the concentration of bacteria on the meat in the supermarkets. This can be done in multiple ways: Lowering the *Campylobacter* counts in broilers entering the slaughterhouses either by producing negative flocks or by lowering the count of bacteria per bird. Improving the steps along the slaughter line so that the number of bacteria on infected carcasses is decreased. A large effort is put into finding solutions and hopefully it is only a matter of time before it is possible to reduce the number of human cases of campylobacteriosis caused by Danish poultry meat considerably. The restriction to Danish meat is that we are importing a lot of meat and the current legislation does not allow us to return infected meat as long as *Campylobacter* is found in meat produced in Denmark.

2.5 Conclusion

We have found a strong relationship between temperature and *Campylobacter* infections in both broilers and humans. On the other hand we have also shown that there are differences in the associations for the two. For the broilers the temperature is affecting the activity and presence of surrounding insects and wild life and the temperature also affects the ventilation of the houses and there by changing the exposure of the broilers. When temperature changes so does human behavior and therefore also the exposure, on top is the changes in the proportion of meat that is infected.

Chapter 3

Spatial clustering

The previous chapter described non-linear correlations between climate and *Campylobacter* infections. It is suggested that in future work local measurements of climatic variables could be used instead of country averages. Another approach to lowering the number of infected flocks is testing if geographical clustering of cases occurs and then use significant clusters in the search for causes of infections. When the cases are individual animals, flocks in houses, or human individuals the cases can be treated as a point process, i.e. all observations are made at points in space and time.

In the litterature many different algorithms and methods are described (Diggle and Tawn, 1998; Kulldorff, 1997, 1999). Most of the methods can either test if clustering occurs or find locations of possible clusters but only few methods are able to find both the location of a cluster and test the significance of the located cluster. One algorithm that is able to do both is the spatial scan statistic which originally was working in 1D (Naus, 1965b) and rectangles in 2D (Naus, 1965a) and now ND algorithms are available, at least in theory. Presently, the most well known implementation is in SaTScan (www.SaTScan.org). SaTScan is able to test for clustering in both space and time separately and combined.

3.1 Model description

The principle of the spatial scan statistic is relatively simple: The input is a spatial point process, i.e. records of cases and non-cases in the selected space, one such space could be a geographical region as a country and restricted to events in a given time window. The scanning is made by choosing a family of scanning windows, move, and scale the window as allowed, e.g. circles that can be moved and scaled. The statistic is obtained by using a likelihood ratio measure on all the scanning windows and choosing the window with the highest maximum likelihood.

In problems where the full population is observed the point process can be treated using a Bernoulli model. If the non-cases are only partly observed or if the risk of being a case is very low a Poisson model can be used. In this work it was decided only to work with the Bernoulli model as the datasets that were considered all contained the full population at risk.

The maximum likelihood is calculated by comparing the proportion of cases in the scanning window with the proportion of cases outside the scanning window. Let n_G be the size of the population and let c_G be the number of cases in the population. When searching for a high risk cluster a scanning window containing a population of n_Z with c_Z cases has the maximum likelihood:

$$L(Z) = \left(\tfrac{c_Z}{n_Z}\right)^{c_Z} \left(1 - \tfrac{c_Z}{n_Z}\right)^{n_Z - c_Z} \left(\tfrac{c_G - c_Z}{n_G - n_Z}\right)^{c_G - c_Z} \left(1 - \tfrac{c_G - c_Z}{n_G - n_Z}\right)^{n_G - c_G - (n_Z - c_Z)}$$

This has to be compared with the null hypothesis, that is that the whole region has the same risk:

$$L_0(G) = \left(\frac{c_G}{n_G}\right)^{c_G} \left(1 - \frac{c_G}{n_G}\right)^{n_G - c_G}$$

The likelihood ratio for the region Z is:

$$LR(Z) = \frac{L(Z)}{L_0(G)}$$

The most likely cluster is the region with the highest likelihood ratio. The analytical distribution of the likelihood ratio will in almost all cases be too complicated, and hence inference is done using Monte Carlo simulations to make a distribution of maximum likelihood ratios under the null hypothesis.

When interested in the significance of secondary clusters one should use the secondary clusters in the Monte Carlo simulated datasets to do inference. This would increase the time and memory used for each simulated dataset and instead an approximate test is performed by using the distribution of maximum likelihood ratios to test the secondary clusters. This implies that the true level of significance of secondary clusters is higher than or equal to the reported level of significance.

Besides using geographical coordinates to define regions, time can also be used as a variable. In combination with circular windows in space, time is added to create cylinders of various heights representing different time windows.

It was not possible to obtain the source code for SaTScan so I had to write my own code from scratch. In order to do the calculations on a computer the likelihood was transformed by taking the logarithm, the problem is handling the power of a number between zero and one raised to the total population in the dataset, the logarithmic transformation reduces this to a single product. My implementation, Episcan, is written in C++ and uses OpenMP to parallelize the Monte Carlo simulations. The implementation in C++ gives a favorable structure but to increase the performance there will be a new version written in C. The present version requires a large amount of memory and the calculations could be speeded up by changing the way the set of clusters is stored in memory. The present version contains a lot of pointers to objects with pointers to data and this two level structure will be reduced to one level.

3.2 Results

As a dataset consider the *Campylobacter* status of broiler flocks described in the previous chapter. There are information on which house every individual flock originates from and the locations of all farms are known, so the information needed in order to consider the data as a point process is available.

The first question is whether to include time in a clustering analysis of the selected dataset. In the Bernoulli model it is not possible to include explanatory variables and we know that there is a pronounced seasonality so it should be expected that the inclusion of time will return a summer peak as the most likely cluster. When running SaTScan with both geographical data and time data the most significant cluster is purely temporal and represents the summer with the highest proportion of positive flocks. This is as expected and doesn't bring any further information about the underlaying problem. Hence, time was omitted in the analysis.

Using the family of circles covering up to 50% of the points a set with 107,953 unique regions was generated. Based on 99,999 Monte Carlo simulations 21,324 clusters were significant at a 10^{-4} level, i.e. 21,324 clusters had a maximum likelihood ratio among the 9 highest Monte Carlo simulated maximum likelihood ratios.

The question is how to present the location of 21,324 clusters. It is expected that many of the clusters overlap, e.g. the clusters that only differ from the most likely cluster by one point will have maximum likelihood ratios similar to the most likely cluster. A simple and widely applied approach is not to allow any overlap. This is done by removing all clusters including points in the most likely cluster, finding the second most likely cluster and removing again, and repeating this untill a given level of significance is reached. This results in 9 non-overlapping clusters, see Fig. 3.1D. When interpreting the plot one has to keep in mind that the underlaying point process is stochastic. Therefore, the size of the clusters is stochastic - especially when having many locations with few observations at each point. When doing a strict reduction of any cluster that overlaps with a more likely cluster it becomes hard to judge how much the elimination would have changed if the most likely cluster had been slightly different. In the more recent versions of SaTScan different elimination schemes have been made available allowing for partial overlap with existing clusters, but still one has to keep in mind that the size of the clusters is stochastic.

In the present study a different approach for visualizing the information is proposed. Instead of doing an elimination based on overlap it is suggested that one selects a level of significance and reports the proportion of significant clusters that each location is part of. When searching for both high and low risk clusters proportions will be made for both, and the data should be visualized so that it is possible to see if a location is part of a low as well as a high risk cluster. An example can be seen in Fig. 3.1C.

The choice of scanning windows also affects the results. When using geographical data today's standard is the family of concentric circles, i.e. scaled circles from a list of centers. In many practical applications of the spatial scan statistic the shape of a potential cluster is not necessarily circular. The allowance of overlap is one way to identify other shapes of a cluster, another way is to propose another and more flexible family of clusters. A natural expansion of circles is a family of ellipses, the question is which? I propose the set of confocal ellipses, i.e. all combinations of two locations are used as foci, for each pair of foci a new potential cluster is created for each additional location when making the ellipse larger. A description of the proposed algorithm is found in Section C.3.

When using elliptic clusters and only reporting the proportion of high and low risk clusters that each location is part of (Fig. C.4D), the results look similar to those of the family of circles. The main difference is that the low prevalent area on Funen and the nearest part of Jutland becomes elliptical as the southern most locations in Jutland appear less frequent in the set of ellipses. When

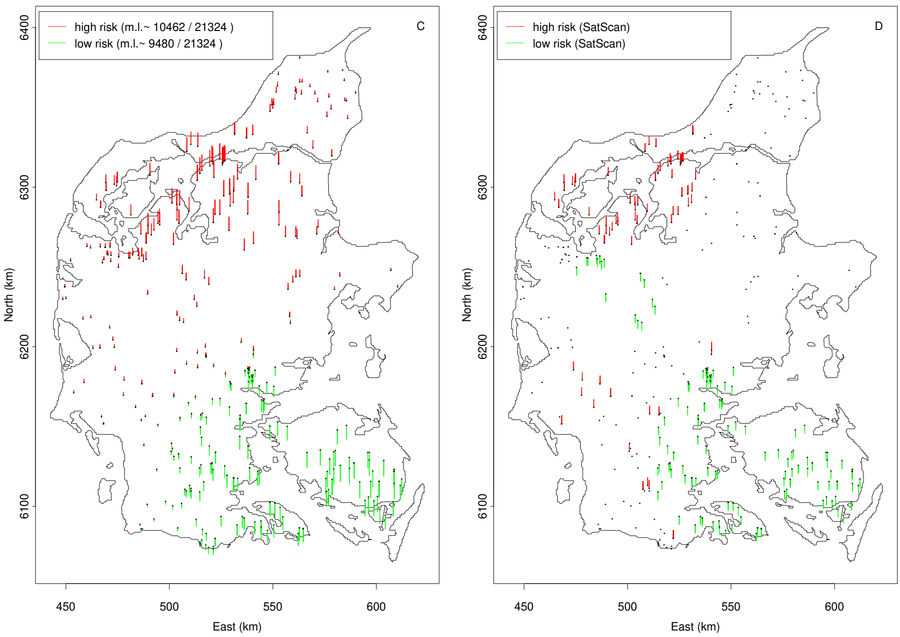

Figure 3.1: Plot C shows the geographical distribution of the 21,324 most significant circular clusters, the 21,324 clusters are the ones with a significance of 0.0001 or less. Green is low prevalent and red is high prevalent, the length of the markers indicates the proportion of clusters that a given house is in (see the legends). Plot D is made using the output from SaTScan. (Part of Fig. C.2.)

applying the elimination that doesn't allow any overlap Funen is preserved as a large low risk area (Fig. C.4B) whereas the high risk area in the Northwest part of Jutland splits in two small and very elongated clusters. Based on this observation one should consider only allowing ellipses where the ratio between the minor and major axis is larger than a certain threshold. This has not been implemented but will be an option in the next version of Episcan.

Based on this analysis the level of *Campylobacter* in broiler flocks on Funen is significantly lower than the average and there is a part of the Nortwest Jutland that has a significantly higher frequence of flocks carrying a *Campylobacter* infection.

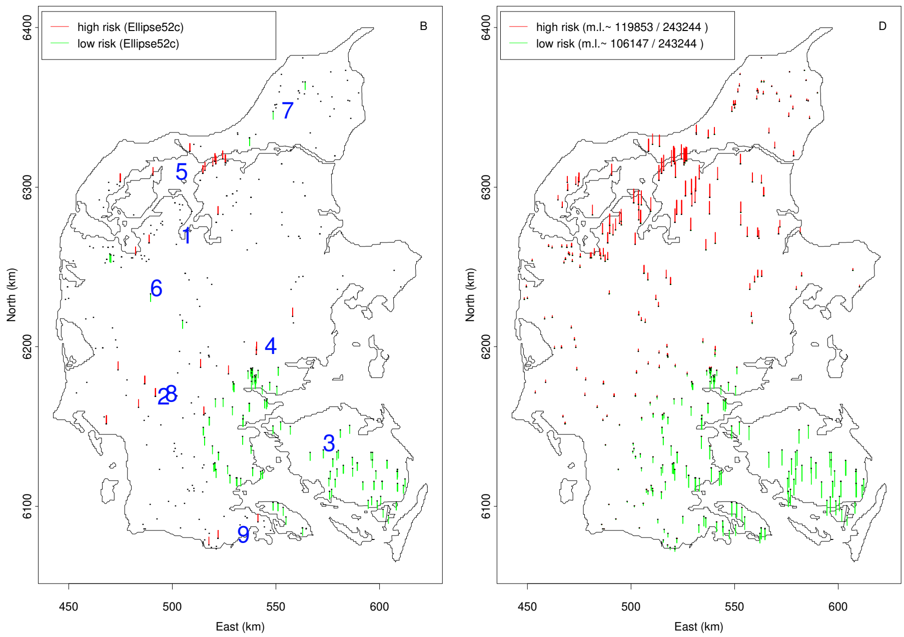

Figure 3.2: Plots presenting elliptic clusters that are significant at a 10^{-4} level. The two plots are made with $p_s = 0.2$. Plot B shows the result of SaTScan elimination and plot D shows all the significant clusters. (Part of Fig. C.4.)

3.3 Discussion and outlook

One of the limitations when using the Bernoulli model in combination with the spatial scan statistic is that it is not possible to use explanatory variables in the analysis. For the Poisson distribution it holds that the sum of two Poisson distributed variables again is a Poisson distributed variable:

$$Z = Z_1 + Z_2 \quad \wedge \quad Z_i \in P(\mu_i) \quad \Rightarrow \quad Z \in P(\mu_1 + \mu_2)$$

In contrast, the sum of two Binomial distributed variables is a Binomial distributed variable if and only if the probability parameter is the same. This can be assumed to be the case when no explanatory variables are used. If two points in the same cluster have different values of explanatory variables the probability of being a case is different and then the sum will no longer be a Binomial distributed variable because the variance is lower than expected from the Binomial distribution.

In principle every single observation may have its own risk, this works fine under the Poisson setup but not in the Bernoulli model. I suggest that it should be possible to make a pseudo likelihood method allowing for use of explanatory variables in combination with the Bernoulli model. One suggestion would be to multiply the likelihood function with the risk inside the cluster multiplied with the risk outside and normalize by the square of the overall risk. Further work is needed to justify this approach.

The next version of Episcan will most likely be made available as a package in R (www.r-project.org), this should make it easy to use and thereby diffuse it to many more applications. One application waiting for the next version is the human data from Paper B. The address is know for all Danes and every address has known coordinates yielding a very large completely observed point process. In practice downsampling to a regular grid will be performed. A run allowing overlapping circles using 3000 locations in a 4 by 4 km grid showed that the set of confocal ellipses would be appropriate but the new version is needed before it is possible with reasonable runtimes on the available high performance computer system at DTU.

Further work on the broiler dataset should focus on finding epidemiologically reasonable explanatory variables that differ between the high and low risk regions. Afterwards intervention experiments should be performed to test if a partial solution of the problem has been found. For elliptic clusters time can be added as a third dimension in the same way as for circular clusters.

In general the spatial scan statistic is a widely applicable tool. It does not provide many answers directly but it can be a very usefull guide when searching for the origin of a problem.

3.4 Conclusion

The use of the family of confocal ellipses has the potential to be a widely applicable tool in epidemiologic investigations. It is as natural an expansion from the set of circles as going from 1D to 2D datasets. Further work is needed to make the algorithm more accessible to the end user.

Chapter 4

Growth dynamics

In Denmark one of the major export articles is pork meat. During the last 15 years the production has increased almost every year and the increase is expected to continue. One of the factors driving this increase is more and more effective production systems including optimizing the diet to reduce the time to market and to increase the feed conversion rate. Previously antimicrobial growth promoters were used in the optimization but now these are banned. Another mean is the amount of micro minerals in the diet. It has been found that the growth rate increases when more copper is added to the diet, besides this copper and zinc are known to protect piglets against diarrhea. Copper has been tested at levels that if used for a few years would increase the amount of copper in the Danish soil to levels where it is illegal to grow anything in the fields (Personal communication with Jens Strodl Andersen).

One unknown consequence of the addition of copper is how it affects the intestinal flora of the pigs. This research is part of the foundation for studying competitive growth of bacteria under different concentrations of copper.

4.1 Growth in a rich environment

Before starting to model competitive growth in different concentrations of copper one has to select or propose a model for bacterial growth. To facilitate the choice growth experiments allow for the use of statistics. The ultimate would be in vivo growth experiments in the intestine of a pig but that setup limits the number of samples that can be taken. Therefore, it was decided to do in vitro experiments and liquid Müller Hinton broth was selected as it is a rich media (Primarily based on boiled meat).

Most of the growth studies that have been conducted to investigate growth have used synthetic media containing a limited number of substrates, starting with single substrate growth as described by Monod (1942). When using two or more substrates there has been a long debate on how the bacteria utilize the substrates (Lendenmann and Egli, 1995; Narang et al., 1997a; Kovárová-Kovar and Egli, 1998) - will they use the substrates simultaneously or sequentially? The present understanding is that it depends on the combination and concentration of substrates.

There are many ways to measure growth in experiments; two basic setups are continous and batch cultures. The latter is efficient when running several tests at the same time as automated equipment can be used to do optical density (OD) measurements on a tray with several wells at discrete time intervals. In other words bacterial growth is a continous time process that is observed in discrete time.

4.1.1 Additional data and linearization

The dataset used in Paper D and E is a subset of a larger dataset containing ten levels of copper. Fig. 4.1 shows the measured and the linearized data from the two repetitions of the ten levels of copper.

The rationale for the linearization is that when the concentration of bacteria, c, in a well gets higher some bacteria will be in the shade of other bacteria, the optical density is proportional to the absorbance, a, of ligth vertical through the well. The change in transmittance, t, when adding bacteria is given as:

$$\frac{dt}{dc} = -ut$$

where u is the area covered by one unit of bacteria. This means that when adding more bacteria the fraction t of the bacteria will absorb light and thereby

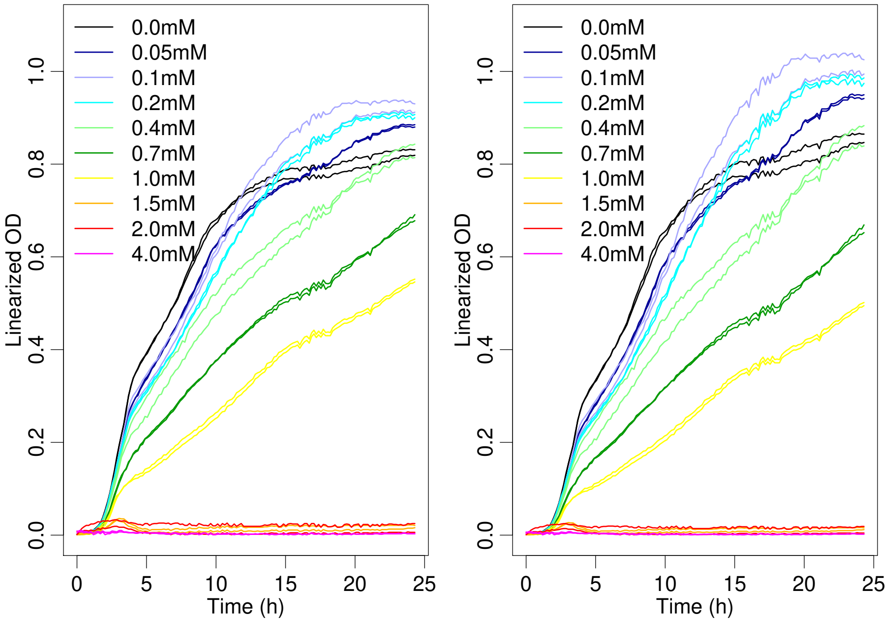

Figure 4.1: Growth curves for ten different levels of copper. The left panel shows the measured optical densities and the right panel shows the linearized optical densities.

reduce the transmittance. Solving for t:

$$t = \exp(-uc)$$

Finally, the absorbance is $1 - t$ resulting in:

$$a = 1 - \exp(-uc)$$

which is proportional to the optical density:

$$od = b(1 - \exp(-uc))$$

where b and u have to be estimated. This was done by making series of dilutions going to 3/4 of the previous concentration for each step; and using the results to get a linearized scale. The concentrations were not quantified so the scale is arbitrary.

4.2 Model structure

Most previously used models for growth on two or more substrates are non-linear dynamic models. One such example presenting an analogy between mixtures of substrates and competing species is presented by Narang (1998). Most of the models are based on ordinary differential equations (ODE) and parameters are mostly estimated from single substrate growth or other studies (Narang et al., 1997a). The parameters in single substrate models are often estimated by minimizing the sum of squares of residuals of simulations. The problem with this approach is that the use of ODEs implicitly makes the assumption that the model is correct and describes the behavior. It is well known that a large number of metabolic processes are involved in bacterial growth and that most of these are not fully understod. Hence, ODEs should be used with care and instead continuous-discrete stochastic state space models (Paper D) should be considered.

In the general case, the continuous-discrete stochastic state space model is a model that consists of a set of non-linear discretely, partially observed Itô stochastic differential equations (SDE) with measurement noise, i.e.:

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \tag{4.1}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{4.2}$$

where $t \in \mathbb{R}$ is the time variable; $\boldsymbol{x}_t \in \mathcal{X} \subset \mathbb{R}^n$ is a vector of state variables; $\boldsymbol{u}_t \in \mathcal{U} \subset \mathbb{R}^m$ is a vector of input variables; $\boldsymbol{y}_k \in \mathcal{Y} \subset \mathbb{R}^l$ is a vector of output variables; $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ is a vector of (possibly unknown) parameters; $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ are known non-linear functions; $\{\boldsymbol{\omega}_t\}$ is an n-dimensional standard Wiener process and $\{\boldsymbol{e}_k\}$ is an l-dimensional white noise process with $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta}))$. The diffusion term in (4.1) is assumed to be independent of the state variables, i.e. only using additive noise, as this makes parameter estimation more feasible.

In practice a model formulated as a system of ODEs is easily converted to a continuous-discrete stochastic state space model by multiplying the ODEs with dt and adding Wiener processes to the state equations.

The advantage of using SDEs over ODEs is that several statistical tests and tools become available. If an extended Kalman filter (EKF) is used with the state space model described by Eq. 4.1 and 4.2 an exact maximum likelihood procedure can be achieved as described in Paper D. The one-step prediction errors as defined by the EKF can be used to test whether the noise is white and whether it is Gaussian as described by Box and Jenkins (1976). If the noise is non-Gaussian the method is no longer exact maximum likelihood but a more general prediction error (PE) procedure (Kristensen et al., 2004). If the one-step prediction errors are non-white noise the errors can be used to assist when building a better model. It is also possible to test if it can be assumed that the model is the correct model (Bak et al., 1999). Finally it is possible to test for significance of the individual parameters trough t-tests.

4.3 Model selection

Five candidate models proposed by Venkatesh et al. (1997), Narang et al. (1997a), and Lendenmann and Egli (1998) were selected. Two versions of the model by Venkatesh, a simplified version of the model by Narang (The reduction can be seen in App. A), and two different models by Lendenmann were entered into CTSM 2.3 (www.imm.dtu.dk/ctsm) in order to estimate the parameters. It was decided to use three substrates in all the models. The parameters were estimated using two repetitions of an *E. coli* growth experiment with 0.1 mM copper, as described in Paper E. In the models with many parameters some of the parameters were highly correlated and in order to get around this some of the parameters were fixed. If the parameters are identifiable the way to get less correlated estimates is to use datasets with more variation fx. by adding more substrate at random intervals during the experiment. This is not feasible with a batch setup but could be done with a continuous culture. Based on the summary of the results in Table E.1 and the ability to represent the data by pure simulation as seen in Fig. E.1 it was decided to model the copper dependence using the simplified Venkatesh model and the two models due to Lendenmann.

The model given by Eq. 16 in Lendenmann and Egli (1998) will be used as an example in the remaining part of this chapter. The model is given by:

$$dc = c\frac{\sum_{i=1}^{3} V_i s_i/K_i}{1+\sum_{i=1}^{3} s_i/K_i}dt - \sigma_s \sum_{i=1}^{3} d\omega_{t,i} \tag{4.3}$$

$$ds_i = -c\frac{V_i s_i/K_i}{1+\sum_{j=1}^{3} s_j/K_j}dt + \sigma_s d\omega_{t,i} \tag{4.4}$$

where index i is over the three substrates. The s_i are the concentration of the three substrates, V_i and K_i are the standard Monod parameters representing the maximum growth rate and the concentration of substrate yielding half of the maximum growth rate. c is the concentration of bacteria. σ_s is the standard deviation of the increments feed by by the Wiener process in unit time intervals. The Wiener processes are added to the substrate equations and subtracted from the concentration of bacteria in order preserve mass balance.

4.4 Modelling the copper dependence

Before starting to model the copper dependence it is important to know a little about the expected effect of copper. Copper is an essential mineral for many living organisms including bacteria. On the other hand copper-ions are also toxic to the cells when the concentration gets to high, the mechanism of toxicity is not fully understood as described by Nies (1999) and Rosen (2002). Hence copper has to be present for bacteria to grow and low concentrations of copper will enhance growth but at higher levels of copper growth will be limited and at even higher levels growth will be inhibited.

The ten different concentrations of copper span the expected behaviors (Fig. 4.1). In the wells with 1.5mM or more copper the bacteria never really started growing, though it seems as if some of them gave it a try during the first four hours but had to give in as the levels of copper were too high. For those that did grow the first growth phase lasted about the first four hours no matter the concentration of copper. The dependence on copper is seen as the concentration of bacteria is quite dependent on copper after four hours. When no copper was added the growth seemed superior during the first ten hours after which it started to slow down. When a little copper was added (0.05mM) the growth was slower, but it continued growing. With a concentration of copper on 0.1mM the highest OD values were optained, adding more copper only reduced the growth rate and for concentrations above 0.4mM the cultures did not finish growing within the 24 hours that the experiment lasted.

To start describing a part of the copper dependence it was decided to use a dataset with the five concentrations 0.1mM to 1.0mM. Those data describe the reduction of the growth rate for concentrations of copper higher than the optimal one, thought of as being the one that reaches the highest OD.

When wanting to model a system where the underlaying dynamics are unknown or partly unknown it is generally a good aproach to start with simple solutions incorporating the knowledge that one has. In the present case it is known that the overall growth is reduced when the concentration of copper is increased. But where should the copper dependence be put into the models? Several papers describing heavy metal ion toxicity describe a non specific uptake system and different efflux systems for removing excess copper from the cells (Dameron and Harrison, 1998; Nies, 1999; Petersen and Møller, 2000; Rosen, 2002; Banci et al., 2003; Rensing and Grass, 2003). Assuming that the main part of the energy related to copper is used for efflux two simple dependencies are proposed.

The first is additive and represents the assumption that besides using energy for growing every bacteria uses energy to transport copper out of the cell:

$$dc = c\frac{\sum_{i=1}^{3} V_i s_i/K_i}{1+\sum_{i=1}^{3} s_i/K_i}dt - c(a\cdot cu + b)dt - \sigma_s \sum_{i=1}^{3} d\omega_{t,i} \quad (4.5)$$

$$ds_i = -c\frac{V_i s_i/K_i}{1+\sum_{j=1}^{3} s_j/K_j}dt + \sigma_s d\omega_{t,i} \quad (4.6)$$

where cu is the concentration of copper; b is the basic maintenance and a is the coefficient for the additional maintenance due to copper efflux.

The second version is multiplicative and it reflects the assumption that a fixed proportion of the energy is used for efflux of copper:

$$dc = \left(c\frac{\sum_{i=1}^{3} V_i s_i/K_i}{1+\sum_{i=1}^{3} s_i/K_i}dt - \sigma_s \sum_{i=1}^{3} d\omega_{t,i}\right)(1-a(cu-0.1)) \quad (4.7)$$

$$ds_i = -c\frac{V_i s_i/K_i}{1+\sum_{j=1}^{3} s_j/K_j}dt + \sigma_s d\omega_{t,i} \quad (4.8)$$

again a is the dependence on the concentration of copper. The concentration of copper was subtracted 0.1mM to make a linearization around the concentration where the candidate models were estimated.

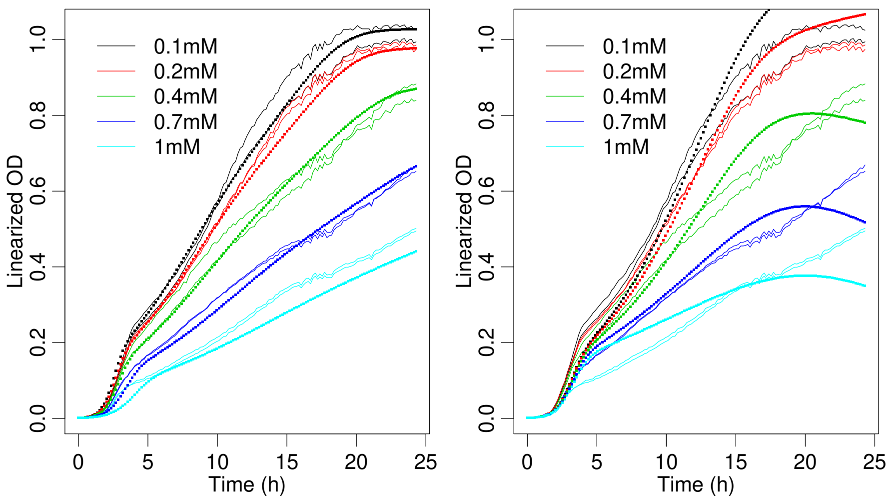

Figure 4.2: The ten time series with 0.1 mM Cu and simulations based on the estimated parameters of the multiplicative (left part) and additive (right part) version of the first Lendenmann model. (Copy of Fig. E.2)

The log likelihoods for the two different copper dependencies were 5,582 for the additive model and 5,633 for the multiplicative model. In Fig. 4.2 pure simulations of the multiplicative and additive models can be seen. Overall the multiplicative version is best at reproducing the copper dependence. The main problem with the multiplicative model is during the first growth phase where the simulations split before the data. The additive model is completely different during the first growth phase as it doesn't split until the end; it also has a too pronounced death phase during the last five to ten hours. A death phase is expected some time after bacteria have stopped growing but here it is over emphasized in order to obtain the separation between the different levels of copper.

Also for the two other set of models the multiplicative models had higher likelihoods than the additive models (See Table E.2).

The next step is to target the first growth phase and one solution would be to combine the additive and multiplicative models. When doing this the estimated model has a log likelihood slightly lower than the pure multiplicative model. In fact the parameter estimates are almost the same as for the pure multiplicative model. The basic maintenance term in the additive model is different from zero but the copper dependence of the additive part was not different from zero. Having a super model containing the two models to be compared is an efficient way to judge which of the models gives the best description, especially when the estimated parameters of the super model reduce the super model to one of the models to be compared.

4.5 Discussion

The main difference between the two models where the copper dependence was not implemented and the other three is that the two have a higher complexity - especially the Narang model which originally had five levels that were reduced to four prior to the analysis was very difficult to estimate with the available data. On top of the two layers that all models have concentration of bacteria and substrates, the narang model has a precursor between the level of substrate and the concentration of bacteria as well as simplified enzyme kinetics controling the usage of the different substrates. If interested in estimating all parameters in such complex models it is necessary to aquire a dataset containing more dynamic information, i.e. in order to estimate embedded parameters related to dynamics at different time scales the system being studied has to be excited in a way that reflects the different time scales. This could be done using a continuous culture where the inflow of media is turned on and off randomly to create a white noise signal, e.g. using a so called PRBS signal.

Given the present dataset the multiplicative implementation of the copper dependence resulted in the highest likelihood. But there is still room for improvement - especially during the first growth phase. If the multiplicative is accepted then one possible explanation for the behavior during the first four hours is that the uptake of the first substrate and its conversion into a precursor is faster than what is needed for the maximum growth rate; hence there is enough energy to be used for efflux of copper while maintaining the maximum growth rate until the first substrate is used. This could be done by adding a state for a precursor as in the Narang model.

Since the ban for use of growth promoters, i.e. antibiotics used at subtherapeutic levels, there has been a great interest in other substances with antimicrobial effects. After the stop for growth promoters the main problem has been diarrhea in piglets (Poulsen and Carlson, 2004), the diarrhea reduces growth and a large effort has been put into finding ways to avoid the problem. As seen in this study copper works as an antimicrobial when the concentration is high enough. One recognized problem with increasing the allowed amount of copper in the feed is the toxic effect to plants - if the industry continues feeding the current level of copper then in about 100 years no fields in Denmark can be used for cultivation of plants (Kærgaard et al., 2002).

Another potential problem feeding animals high levels of copper is changes in the intestinal flora, some bacteria are less sensitive to copper than others, e.g. *Salmonella* Typhimurium tolerates higher concentrations of copper than *E. coli* (Seen in the remaining part of the present dataset and not yet published). Those are two bacteria that are widely known, but the intestine contains numerous bacteria that are not even named, and it is unknown how copper affects their growth.

4.6 Conclusion

The use of continuous-discrete stochastic state space models have been motivated and examples of usage have been given. The growth medium used in the experiments was rich but using three substrates it was possible to give a reasonably good description of the behavior.

Two different copper dependencies and a combination of the two have been implemented and the embedded parameters were estimated. The result was that the multiplicative approach was the most promissing and a suggestion for an improvement has been given.

Chapter 5

Discussion

The methodology used in this research spans over several topics in applied statistics. It is all about using statistics as a guide for researchers in other fields; the guidance is performed by describing and modelling linear correlations and indeed non-linear relations in time and space.

The choice of veterinary epidemiology as the target was based on the fact that more and more information becomes available for research in this field. The data structure in veterinary datasets is often very hierarchical, with levels including slaughterhouse, farm, house on farm, section within the house, and animal. On top of these levels are temporal dynamics and spatial correlations and relations between houses, production systems, and farms. This means that a large part of the classical statistic methods will not be applicable as the observations are neither independently nor identically distributed. In many cases new methods that describe these typically non-linear correlations have been described analytically, but there is also a wide range of cases where it is infeasible to make an analytical distribution for doing inference. In many of those cases it is possible to simulate under the null hypothesis and then Monte Carlo simulations may be used to generate a distribution that can be used to do inference.

In Denmark we have a large production of animals and since the late eighties there has been an increasing consumer concern about diseases that are transfered from animals to humans - the so called zoonosis. The concern has been targeted by implementing several surveillance based control programs creating systematic collected data for a very large part of the Danish animal production. The veterinary epidemiologists at the Danish Zoonosis Centre do have access to those data and via collaboration with the Statens Serum Institut they also have access to the result of human isolates that have been tested. A large part of the information is only used in summaries but those data contain a vast potential for data mining. The problem or challenge is that the complicated structure has to be included in the model in order to perform the analysis but this is often not feasible and therefore many new tools are being developed.

Danish trees do not grow into the sky; a lot of information is reported and stored in databases but as a large part of the data is collected as part of surveillance systems the sampling is not always good enough to use many of the statistical tools at hand or in the pipeline. One of the important reasons for this is that every sample has a price and cost benefit analysis are often used more or less directly when making control programs. One example is the Danish surveillance of *Salmonella Enteritidis* in shell eggs. The program started in 1997 and is based on a combination of serologic and bacteriologic test. Every nine weeks the yolk from 60 eggs from each flock are tested serologically and two so called sock samples or pooled feacal droppings are tested bacteriologically. Skov et al. (2002) have experimentially infected a group of day old chickens and followed up with both kinds of tests. The proportion of bacteriologic positive tests declined within 8 weeks and increased again at the onset of egg production. The serologic test on serum started to increase after 3 weeks, peaked after 9 weeks, and after a year some tests were still positive. The study started with 100 birds and two thirds died during the study. This has to be compared with commercial producers often having 10,000 birds or more. Apparently there is a lag of about 3 weeks in the dynamics and a time constant in the order of 8 weeks, when sampling every 9th week it is not possible to use these samples to make a model as the sampling frequency is far to low considering the Nyquist frequency. No studies have been performed on the infection dynamics in a real house.

A different example is the VETSTAT database containing information on all veterinary medicine usage. The purpose is to monitor the medicine usage. One application of the database is to do inference on the risk of bacteria becoming resistant when using a drug, but in many cases the number of farms that can go into analysis is limited due to insufficient sampling.

After all there are many large and good datasets collected in Denmark, some are used in this work and many other await to be analyzed.

In the first part of this research climate variables were used as predictors. It was not the direct cause of infection but driving other sources of infection. Indeed bacterial communities are highly affected by the surrounding environment and ambient air temperature is one of the important factors. The last part of this research is a pilot study that is going to be used for competition studies between different bacteria under different environments occuring along the farm to fork chain. The aim of that work is to develop predictive models describing the importance of using various antimicrobial compounds for the spread of zoonotic bacteria along the entire food chain.

One thing to be aware of is not to repeat the story of the antibiotics - they were very efficient for a period of time but then resistance started to become a problem and now resistance is a major problem especially in countries where growth promoters are still used, e.g. in the US a large proportion of the broiler meat in the stores contain bacteria resistant to ciprofloxacin (personal communication with Lance B. Price from John Hopkins School of Public Health).

One of the problems that needs to be targeted within near future is how to handle outbreaks or epidemics of exotic diseases in Denmark. The United Kingdom has learned a lot about foot and mouth disease (FMD) the hard way; avian influenza outbreaks have been seen in the Netherlands in 2003 and in several Asian countries this year. So far Denmark has been lucky but the risk is there, and as was seen with SARS an infection can easily get on an airplane and spread worldwide. It is necessary that studies use the existing cases to gather information in order to say we do have a level of preparedness and modelling will be a helpful tool trying to stop these diseases. Unfortunately the trainings of preparedness within DFVF have only concerned the laboratories, no training of modelling has been done to be prepared if an outbreak occurs. It should be a goal to be able to use all the database information in an early warning system.

Most of the modelling currently done in Denmark only uses a small part of the potential for dynamic models that exists in the data, and knowing that many processes are dynamic and happen in continuous time should make it obvious to use continuous time dynamic models. It is advocated to use stochastic models in most cases as complete knowledge of the system and its inputs is unrealizable.

One of the challenges in todays epidemiology is modelling. Many different tools are available; each has its pros and cons, e.g. at DFVF there is an upcoming Epi-Lab project comparing three different models of FMD under a Danish setting. Many problems can be handled by various statistical programs and packages but other have to be programmed at some level. The highlevel statistical languages S-plus and R are usefull in many cases but for some problems they are too slow. In the cases where the interpreting languages are too slow one solution is to link and call precompiled code from within S-plus and R. The external code can be written in C, C++, or Fortran. I have an example related to a Baysian selection of a set of biochemical tests where the calculation of the objective function was 1500 times faster in C than in R!

When the calculation time becomes an issue it is important to understand the structure of the language where the calculations are made in order to make it efficient. It is often possible to make calculations in say Splus 10-20 times faster by using array based function calls. If a large part of the calculations are independent it may be quite easy to parallelize that part of the calculations and get a speedup if multiple processors are available. In fact several chip producers are about to make multi-threading CPUs, i.e. one chip with two or more processors on it. An example is Sun Microsystem's UltraSPARC IV having two UltraSPARC III processors on each chip. When having a shared memory parallel computer, and those will be on the desktop in a couple of years, an easy way to parallelize is by using OpenMP (www.openmp.org). Often Monte Carlo simulations can be parallelized by adding a single statement to the original source code.

CHAPTER 6

Conclusion

It has been shown that there is a non-linear correlation between temperature and the occurrence of *Campylobacter* in broilers and humans. The correlation is slightly different for the two and possible links between temperature and infection has been described.

The spatial scan statistic has been further developed by adding the set of confocal ellipses and discussing the way the information in a set of clusters is visualized. Using the expansion of the spatial scan statistic significant clustering of *Campylobacter* occurrence in broiler flocks has been found. The algorithm should be made more accessible and a way to allow explanatory variables in combination with the Bernoulli model has been suggested but needs further analysis.

Continuous-discrete stochastic state space models have been used to model growth of *E. coli* in a rich media. Two different implementations of copper dependence have been suggested and implemented in three models. The multiplicative implementation was the better and suggestions for improvements have been proposed.

The aim was to demonstrate and develop statistical tools that can help epidemiologists and this has been achieved.

Appendix A

Reduction of the Narang model

The full model described by Narang et al. (1997a) with n substrates is given by the following set of ordinary differential equations:

$$\frac{dc}{dt} = V_g \frac{p}{K_g + p} c \tag{A.1}$$

$$\frac{dp}{dt} = \sum_{j=1}^{n} Y_j V_{x,j} \frac{x_j}{K_{x,j} + x_j} - V_g \frac{p}{K_g + p} - V_g \frac{p}{K_g + p} p \tag{A.2}$$

$$\frac{dx_i}{dt} = V_{s,i} e_i \frac{s_i}{K_{s,i} + s_i} - V_{x,i} \frac{x_i}{K_{x,i} + x_i} - V_g \frac{p}{K_g + p} x_i \tag{A.3}$$

$$\frac{ds_i}{dt} = -V_{s,i} e_i \frac{s_i}{K_{s,i} + s_i} c \tag{A.4}$$

$$\frac{de_i}{dt} = V_{e,i} \frac{x_i}{K_{e,i}(p) + x_i} + k^*_{e,i} - k_{d,i} e_i - V_g \frac{p}{K_g + p} e_i \tag{A.5}$$

where $K_{e,i}(p) = K^0_{e,i} \left(1 + \frac{p}{K_{inh,i}}\right)$, the dependence on p is intended to reflect the effect of catabolite repression.

The five states are the concentration of bacteria, c, the concentration of precursor per cell, p, the concentration of inducer per cell for the i'th enzyme, x_i, the i'th substrate, and the conentration of the i'th enzyme per cell, e_i. For a description of the individual parameters see the original paper.

The model has five states of which only one is observed so in order to improve the chances of fitting the model with the provided datasets it is advice to reduce the model complexity.

I decided to eliminate the state for the inducer by assuming that it is quasi-stationary, i.e. that dx_i/dt equals zero almost all the time. This is in agreement with Narang et al. (1997a): "We begin the process of simplifying these equations by expressing mathematically the fact that the fast variables x_i and p rapidly achieve quasisteadystate." I tried eliminating p as well but this resulted in a much more complex right hand sides of the differential equations and made the estimation harder.

The equation to be solved is:

$$V_{s,i}e_i\frac{s_i}{K_{s,i}+s_i} - V_{x,i}\frac{x_i}{K_{x,i}+x_i} - V_g\frac{p}{K_g+p}x_i = 0$$

The three fractions are all of the order of one. Using the information on parameter values in the original paper the order of magnitude of the three terms are $1630 \cdot 2 \cdot 10^{-4} \approx 0.3$, 10, and $1.0 \cdot 10^{-4}$, respectively. hence the last term vanishes and we are left with:

$$V_{x,i}\frac{x_i}{K_{x,i}+x_i} = V_{s,i}e_i\frac{s_i}{K_{s,i}+s_i}$$

this can be directly inserted into Eq. A.2. To solve for x_i it is easier to start with the following:

$$\begin{aligned}
\frac{x}{K+x} &= a \quad \Leftrightarrow \\
x &= aK + ax \quad \Leftrightarrow \\
x(1-a) &= Ka \quad \Leftrightarrow \\
x &= K\frac{a}{1-a}
\end{aligned}$$

In order to see that:

$$x_i = K_{x,i}\frac{\frac{V_{s,i}}{V_{x,i}}e_i\frac{s_i}{K_{s,i}+s_i}}{1-\frac{V_{s,i}}{V_{x,i}}e_i\frac{s_i}{K_{s,i}+s_i}} \tag{A.6}$$

$e_iV_{s,i}/V_{x,i}$ is of the order of 0.03. Hence the denominator if is almost unity and therefore:

$$x_i \approx \frac{K_{x,i}V_{s,i}}{V_{x,i}}e_i\frac{s_i}{K_{s,i}+s_i}$$

In Eq. A.5 x_i is to be inserted into:

$$
\begin{aligned}
\frac{x_i}{K_{e,i}(p)+x_i} &= \frac{\dfrac{K_{x,i}V_{s,i}}{V_{x,i}}e_i\dfrac{s_i}{K_{s,i}+s_i}}{K_{e,i}(p)+\dfrac{K_{x,i}V_{s,i}}{V_{x,i}}e_i\dfrac{s_i}{K_{s,i}+s_i}} \\
&= \frac{\dfrac{e_is_i}{K_{s,i}+s_i}}{\dfrac{V_{x,i}K_{e,i}(p)}{V_{s,i}K_{x,i}}+\dfrac{e_is_i}{K_{s,i}+s_i}}
\end{aligned}
$$

Redefining $K_{e,i}$ to include $V_{x,i}/(V_{s,i}K_{x,i})$ and ignoring the dependence on p:

$$\frac{dc}{dt} = V_g\frac{p}{K_g+p}c \tag{A.7}$$

$$\frac{dp}{dt} = \sum_{j=1}^{n}Y_jV_{s,j}\frac{e_js_j}{K_{s,j}+s_j}-V_g\frac{p}{K_g+p}-V_g\frac{p}{K_g+p}p \tag{A.8}$$

$$\frac{ds_i}{dt} = -V_{s,i}\frac{e_is_i}{K_{s,i}+s_i}c \tag{A.9}$$

$$\frac{de_i}{dt} = V_{e,i}\frac{\dfrac{e_is_i}{K_{s,i}+s_i}}{K_{e,i}+\dfrac{e_is_i}{K_{s,i}+s_i}}+k^*_{e,i}-k_{d,i}e_i-V_g\frac{p}{K_g+p}e_i \tag{A.10}$$

The only remaining difference to the model used for analysis is that the parameter Y_i is fixed to unity and therefore omitted. The justification for doing that is that the data used for analysis does not include information on the initial concentration of the substrates this can be shown to give one free parameter per substrate so in order to estimate the parameters one parameter has to be fixed. By defining $Y_i \equiv 1$ one unit of substrate is converted to one unit of bacteria, this makes it easy to link the growth phases with the use of the substrates.

Appendix B

Effect of Climate on Incidence of Campylobacter spp.

Accepted for publication in *Applied and Environmental Microbiology.*

Effects of Climate on Incidence of *Campylobacter* spp. in Humans and Prevalence in Broiler Flocks in Denmark.

Mary Evans Patrick[1], Lasse Engbo Christiansen[1,3], Michael Wainø[2], Steen Ethelberg[4], Henrik Madsen[3], Henrik Caspar Wegener[1*]

[1]Danish Institute for Food and Veterinary Research, DK-2860 Søborg, Denmark;
[2]Danish Institute for Food and Veterinary Research, DK-8200 Århus N;
[3]Technical University of Denmark, DK-2800 Kgs. Lyngby;
[4]Statens Serum Institut, DK-2300 Copenhagen S

B.1 Abstract

Campylobacter infections are increasing and pose a serious public health problem in Denmark. Infections in humans and broiler flocks show a similar seasonality suggesting that climate may play a role in infection. We examine the effect of temperature, precipitation, relative humidity and hours of sunlight on *Campylobacter* infections in humans and broiler flocks by using lag dependence functions, locally fitted linear models, and cross validation methods. In humans, the best model includes average temperature and sunlight 4 weeks prior to infection; maximum temperature lagged at 4 weeks was the best single predictor. In broilers, average and maximum temperature 3 weeks prior to slaughter gave the best estimate; average temperature lagged at 3 weeks was the best single predictor. The combined effect of temperature and sunlight or the combined effect of temperature and relative humidity predicted incidence in humans equally well. On broiler flock incidence these factors explained considerably less. Potential explanations are discussed and future research is suggested.

B.2 Background

Campylobacter has become the leading cause of enteric zoonotic infection in most developed and developing countries and the incidence appears to be increasing (Anonymous, 2000). Typical symptoms include diarrhea, which can be bloody, abdominal pain, fever, nausea, headache and vomiting. Illness is typically self-limiting and usually resolves in about one week, however severe illness and long-term complications such as arthritis, septicemia and Guillan-Barre Syndrome sometimes occur. Incidence of human campylobacteriosis in Denmark has risen steadily since 1992, reaching 4,620 cases or 86 per 100,000 population in 2001

(Anonymous, 2001). The majority of cases occur sporadically and outbreaks are rare. It is estimated that approximately 80% of cases are domestically acquired (Anonymous, 2001). *Campylobacter* is commonly found in domesticated animals including poultry, cattle, pigs, sheep, cats and dogs (Hald and Madsen, 1997) and migratory birds such as ducks (Cabrita et al., 1992; Luechtfeld et al., 1980) and seagulls (Glunder et al., 1992). *Campylobacter* has also been isolated from rodents (Cabrita et al., 1992), beetles (Jacobs-Reitsma et al., 1995), houseflies (Rosef and Kapperud, 1983), and surface water sources (Brennhovd et al., 1992; Dackowska-Kozon and Brzostek-Nowakowska, 2001).

Many risk factors for *Campylobacter* infection have been identified including consumption of raw or undercooked poultry (Kapperud et al., 1992; Eberhart-Phillips et al., 1997; Neimann et al., 2003; Studahl and Andersson, 2000) handling raw poultry (Hopkins and Scott, 1983; Neal and Slack, 1995), contact with chicken or hens (Studahl and Andersson, 2000) consumption of unpasteurized or contaminated milk (Studahl and Andersson, 2000; Friedman et al., 2000) or untreated or contaminated drinking water (Engberg et al., 1998; Melby et al., 1990), barbecuing (Neimann et al., 2003; Studahl and Andersson, 2000) contact with pets (Eberhart-Phillips et al., 1997; Kapperud et al., 1992; Tenkate and Stafford, 2001) and international travel (Eberhart-Phillips et al., 1997; Rodrigues et al., 2001). In Denmark, consumption of poultry is thought to be the major risk factor for *Campylobacter* infections (Neimann et al., 2003). A national monitoring program estimated the average prevalence of *Campylobacter* in Danish broiler flocks to be 42% in 2001 (Anonymous, 2001).

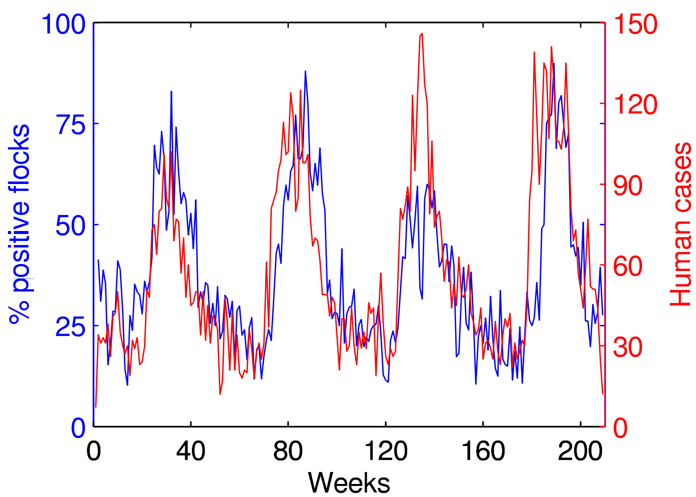

Figure B.1: Seasonality of campylobacteriosis in humans and Campylobacter infections in broiler flocks in Denmark, 1998-2001

In Denmark, human *Campylobacter* infections and percentage of infected broiler flocks show a similar seasonal trend with peak infections occurring in July and August (weeks 27-35) (Fig. B.1). Seasonal trends in human infection have also been observed in several other countries (Nylen et al., 2002; Rautelin and Hänninen, 2000). In Norway, cases show an increased summer peak corresponding to increased latitude (Kapperud and Aasen, 1992). Seasonal variations in *Campylobacter* infection in broilers (Berndtson et al., 1996; Jacobs-Reitsma et al., 1994; Kapperud et al., 1993; Wallace et al., 1997) and cattle (Stanley et al., 1998) have been reported from several countries.

These seasonal trends suggest that climatic factors may play a role in infection. Studies have documented an effect of climate, particularly temperature, on human disease (Checkley et al., 2000; Speelmon et al., 2000; Tong et al., 1998). Other studies have suggested associations between environmental factors such as temperature, humidity and sunlight and *Campylobacter* carriage in broilers (Wallace et al., 1997; Doyle, 1984). Our study further examines the role of climate on *Campylobacter* infections by comparing the effects of temperature, precipitation, relative humidity and sunlight on the number of human cases in Denmark and the prevalence of *Campylobacter* in Danish broiler flocks at slaughter.

B.3 Methods

Data collection. Human *Campylobacter* cases reported to the Statens Serum Institute between 1998 and 2001 were summarized by week based on the date that the specimen was received in the laboratory. Our analysis is based on the counties of Århus, Bornholm, Frederiksborg, Fyn, Nordjylland, Ribe, Roskilde, Storstrøms, Sønderjylland, Vejle, Vestsjælland and Viborg and Copenhagen and Frederiksberg municipalities which consistently reported cases by week between 1998 and 2001. Copenhagen county and Ringkøbing county were excluded from analysis because of missing weekly information during 1998. This excluded 0.9 million out of 5.3 million population. Greenland and the Faroe Islands were also excluded from analysis. Cases which reported international travel in the week preceding *Campylobacter* infection were excluded from analysis, as were cases for which county name within Denmark was missing. It was assumed that the remaining cases were exposed within Denmark.

Under a national monitoring program for Danish broiler flocks, ten birds from each flock or batch[1] are examined for *Campylobacter* by cloacal swabs at slaughter. The percent of flocks testing positive by week were summarized between

[1] One flock may be slaughtered in up to six different batches.

1998 and 2001. Only results from the first batch were included in our analysis because slaughtering a flock in several batches may increase the risk for contamination, e.g. introduction from the machine used to depopulate the house, and due to the dependence between batches from the same flock. If the house on a farm from which the slaughtered flock came was unknown, the sample was excluded from analysis to ensure validity of the data.

Between 1998 and 2001, a total of 16,305 human campylobacteriosis cases were reported in Denmark; 94% were *C.* jejuni. Of these, 2,658 were excluded because they were reported from Copenhagen or Ringkøbing county, 1,554 cases were excluded because they had a known history of international travel and 85 cases were excluded because the county name was not reported. During this same time, 23,279 broiler samples were collected; 5,338 were excluded from analysis either because they were not from the first batch slaughtered or were from an unknown house. Of the remaining 17,941 samples, 6,860 (38%) tested positive for *Campylobacter* spp.; 85% of these samples were *C.* jejuni.

Climate data were provided by the Danish Meteorological Institute (DMI). Average temperature, maximum temperature, relative humidity, precipitation measured in mm, and amount of sunlight were averaged for the entire country by week between 1998 and 2001. Some of the climatic factors were found to be highly correlated. For example, relative humidity was negatively correlated with hours of sunlight and average temperature was generally 5 degrees lower than maximum temperature.

Data analysis. Weekly measures of the number of human cases and percent of infected broiler flocks were fitted using univariate and bivariate models with climate variables as the regressors. Locally fitted linear models (Madsen, 1997; Nielsen and Madsen, 2001) (The S-plus library LFLM can be downloaded from www.imm.dtu.dk/˜han/software.html) with nearest neighbor bandwidths were used with zero, first and second order polynomials. Each of the regressors was lagged 0 to 6 weeks. For each climatic variable a four fold cross validation (Hastie et al., 2001; Madsen, 1997) was used to select the optimal lag, degree of the fitting polynomial and bandwidth. The bivariate models were estimated independent of the univariate models. The four fold cross validation was done using all combinations of three years data to estimate a model used to predict the fourth year, the reported cross validation error is the sum of squared prediction errors for the four folds (years). The model with the lowest cross validation error (CV) was determined to be the best model. The squared multiple correlation coefficient (R^2) for each model was also included; however, a higher R^2 did not necessarily correspond to the best model. The human count data was transformed by the logarithm prior to analysis and the proportion of infected broiler flocks was untransformed in the analysis.

B.4 Results

Humans and climate. Maximum temperature four weeks prior to reported infection was the best single predictor of campylobacteriosis incidence explaining 68% of the variation (Table B.1). Maximum temperatures between 13-20°C corresponded to the steepest increase in incidence while a smaller increase was seen at temperatures above 20°C (Fig. B.2b). Average temperature three weeks prior to reported infections was nearly as good a predictor as maximum temperature (Table B.1). The steepest increase in incidence of human cases occurred at average temperatures between 8-13°C, with a smaller increase above 13°C (Fig. B.2a). Hours of sunlight lagged at five weeks explained slightly more than half of the variation while relative humidity also lagged at five weeks explained 44%. Precipitation was not at all important in predicting human incidence, as it explained only 6% of the variation.

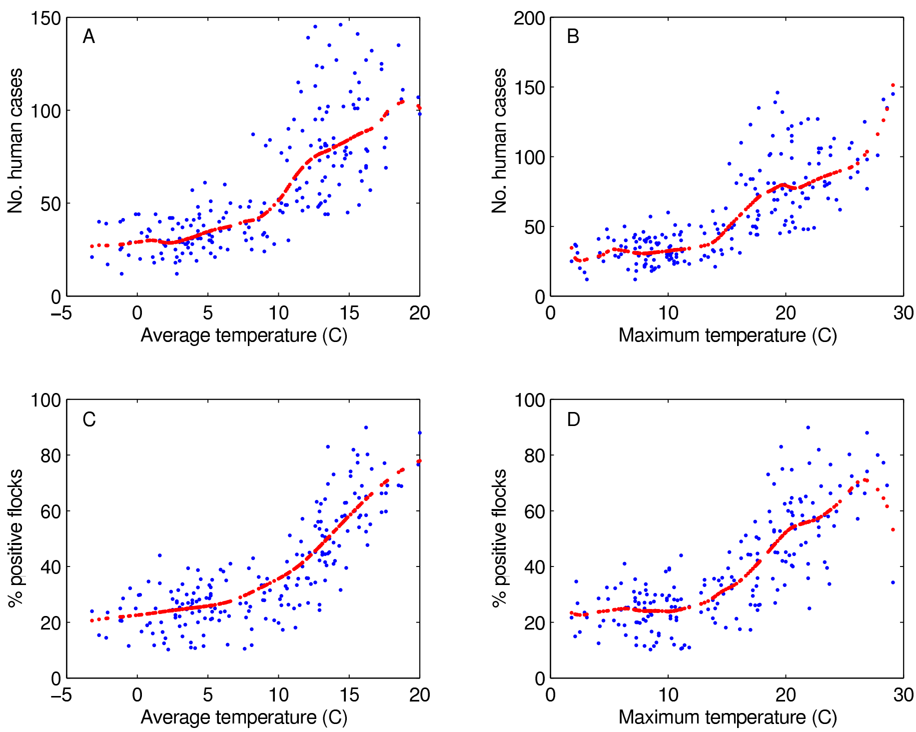

Figure B.2: Impact of temperature on human campylobacteriosis incidence and percent of Campylobacter infected broiler flocks at slaughter. Blue dots are observations and red dots are estimations.

Table B.1: Climatic Factors Used To Predict Human Campylobacter Infections

	Lag (weeks)	Band-width	Order	CV*	R^{2**}
Human one predictor					
Ave. temp.	3	0.40	0	22.441	0.6662
Max. temp.	4	0.50	1	21.746	0.6776
Hours of sunlight	5	0.50	2	29.860	0.5748
Rel. humidity	5	0.60	1	38.067	0.4398
Precipitation	3	0.20	1	62.758	0.0632
Human two predictors					
Ave. temp. & max. temp.	4	0.45	1	21.549	0.7062
Ave. temp. & hours of sunlight	4	0.40	1	20.585	0.7159
Ave. temp. & rel. humidity	3	0.55	1	21.838	0.6907
Ave. temp. & precipitation	1	0.50	1	22.811	0.6687
Max. temp. & hours of sunlight	4	0.45	1	21.967	0.6955
Max. temp. & rel. humidity	4	0.40	1	22.730	0.6948
Max. temp. & precipitation	4	0.25	1	22.343	0.7044

* Cross-validation error ** Explanation of variation

Using multiple climatic factors, the best model for predicting human campylobacteriosis incidence (i.e. the model with the lowest cross validation error) included average temperature and hours of sunlight four weeks prior to reported infection (Table B.1). The combined effect of high average temperatures with many hours of sunlight resulted in the greatest number of human campylobacteriosis cases (Fig. B.3b). For temperatures below 17°C lower average temperatures and more hours of sunlight predicts a similar number of cases as higher average temperatures with less hours of sunlight. For example, at 10°C and 60 hours of sunlight the predicted number of human cases was approximately the same as at 15°C and 30 hours of sunlight. Average temperature and maximum temperature lagged at four weeks was the next best model for predicting human cases. When maximum temperature was above 15°C, increases in maximum temperature resulted in a larger change in incidence than corresponding increases in average temperature (Fig. B.3a). Overall, these models explained 72% and 71% of the variation, respectively, and were each better predictors than using either maximum or average temperature alone.

Using average temperature and relative humidity three weeks prior to reported infection was also an important predictor of human incidence (Table B.1). The greatest effect existed at high temperatures and relatively low humidity. A slight gradient was seen with low average temperatures and low humidity exerting a similar effect on the number of cases as higher temperatures with high humidity (Fig. B.4). For example, the estimated number of cases at 10°C and 60% relative

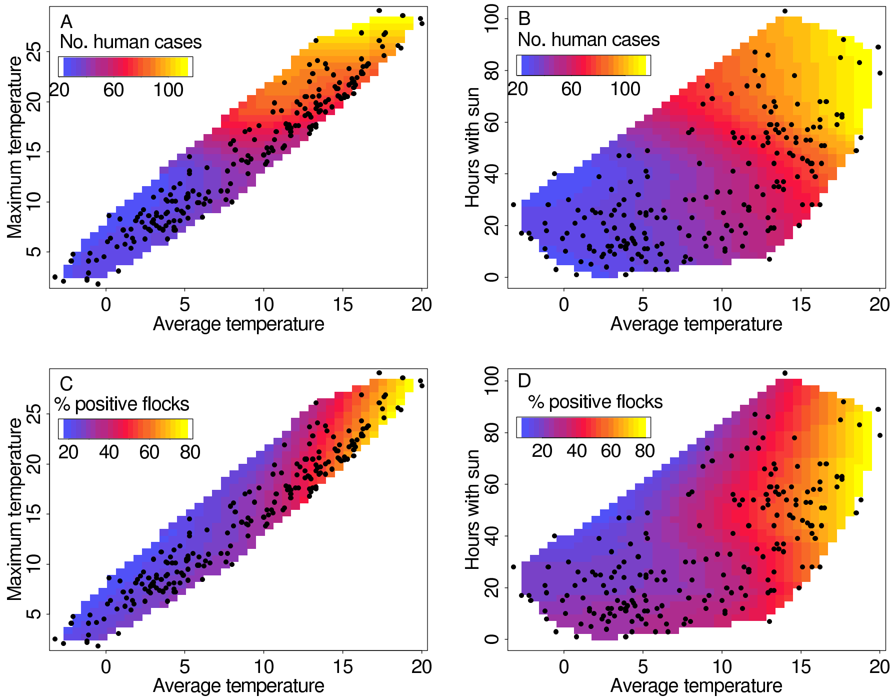

Figure B.3: Effects of temperature and hours of sunlight on incidence of campylobacteriosis in humans and percent of Campylobacter infected broiler flocks at slaughter. Black dots are observations.

humidity was similar to that at 13°C and 85% relative humidity. This was not unexpected as relative humidity is inversely correlated with hours of sunlight.

Models including average temperature and precipitation, or maximum temperature and either hours of sunlight, relative humidity, or precipitation did not predict any better than using maximum temperature alone.

Broilers and climate. Average temperature three weeks prior to slaughter explained 74% of the variation and was the single best predictor of broiler prevalence (Table B.2). A lag of two and four weeks produced similar CV values. At average temperatures above 8°C, there was a large increase in broiler prevalence with increasing temperature (Fig. B.2c). Maximum temperature lagged at four weeks was the next best single predictor of broiler prevalence, explaining 66% of the variation (Fig. B.2d). Similar to what was shown with human cases, a large increase in incidence was seen at temperatures between 13-20°C, followed

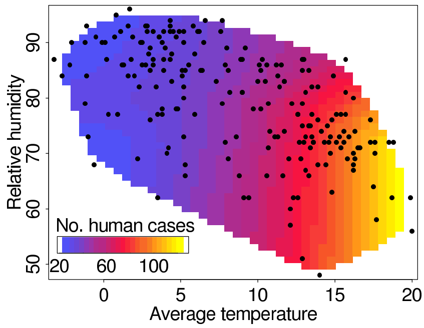

Figure B.4: Effect of average temperature and relative humidity on incidence of human campylobacteriosis. Black dots are observations.

by a smaller increase at higher temperatures. Although temperatures were similar during autumn and spring, a higher percentage of broilers were positive for *Campylobacter* in the autumn (Fig. B.5). During the fall, the average temperature and the percentage of positive broilers decreased at similar rates, however, in the spring, temperatures began to increase before the percentage of infected broiler flocks. There appeared to be a threshold at the end of spring after which the broiler incidence increased dramatically. This effect was still visible even after accounting for a temperature lag of 3 weeks. As single predictors, hours of sunlight and relative humidity lagged at five weeks explained 50% and 38% of the variation, respectively, while precipitation lagged at one week explained only 17%.

Using multiple climatic factors, average temperature and maximum temperature three weeks prior to slaughter provided the best estimate (i.e. the lowest cross validation error) of broiler prevalence explaining 76% of the variation and was better than any of the single predictors alone (Table B.2). In contrast to humans, the greatest prevalence occurred when there was a small difference between average and maximum temperature (i.e. high average temperature and relatively low maximum temperature) (Fig. B.3c). A model including average temperature and hours of sunlight was the next best predictor explaining 77% of the variation and was better than a model with average temperature alone (Table B.2). Interestingly, hours of sunlight had a more pronounced effect at temperatures below 8°C, with fewer hours of sunlight corresponding to a higher prevalence (Fig. B.3d). At temperatures above 13°C, the effect was less apparent with the highest prevalence corresponding to 50-60 hours of sunlight. However, the overall contribution of hours of sunlight to increased prevalence at these temperatures was not as important. Models including either relative humidity or precipitation with average temperature were not any better than a model including average temperature alone.

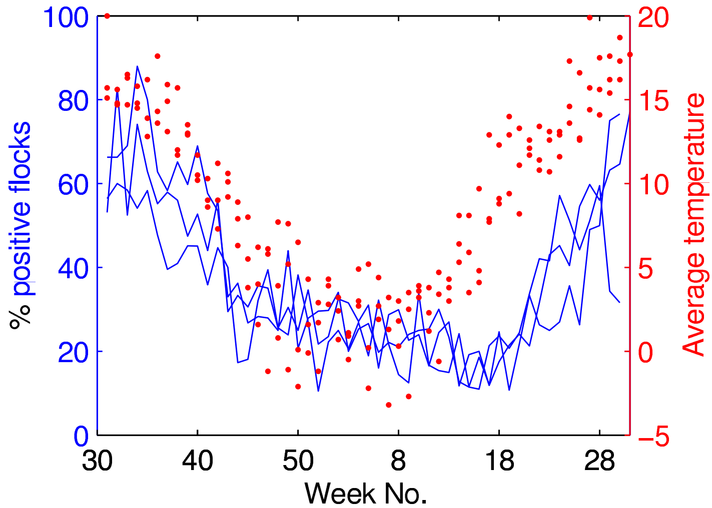

Figure B.5: Effects of cold season on percentage of Campylobacter infected broiler flocks at slaughter. Blue line represent percent positive broiler flocks. Each line represents one year from summer to the next summer (e.g. summer 1998-summer 1999). Red dots represent average temperature. The week no. on the x-axis are calendar week no.

B.5 Discussion

Our study found that as the temperature increases, there is a corresponding increase in the number of *Campylobacter* cases in humans and percent of infected broiler flocks at slaughter. A threshold average temperature above 8°C and maximum temperature above 13°C was seen with largest increase in incidence at temperatures up to 20°C, and a smaller increase above 20°C. Temperature had its greatest impact 3-4 weeks before reported infection for both broilers and humans. The combination of high average and relatively low maximum temperature (i.e. a small variation in temperature within one week) corresponded to the second highest incidence of human cases and highest percent of infected broilers, indicating that it may be the effect of sustained high temperatures rather than intermittent high temperature that is the key factor in predicting infection. Climatic factors explained a larger proportion of the variance in *Campylobacter* prevalence in broilers than in humans.

In humans, models including average temperature and either maximum temperature or hours of sunlight provided the best estimates of *Campylobacter* incidence. At a lag of four weeks, a model including average and maximum

Table B.2: Climatic Factors Used To Predict Campylobacter Carriage in Broiler Flocks

	Lag (weeks)	Band-width	Order	CV*	R^{2**}
Broiler one predictor					
Ave. temp.	3	0.50	1	2.0958	0.7439
Max. temp.	4	0.30	0	2.8193	0.6637
Hours of sunlight	5	0.45	1	3.7508	0.5035
Rel. humidity	5	0.40	0	4.7387	0.3770
Precipitation	1	0.50	0	7.1459	0.1706
Broiler two predictors					
Ave. temp. & max. temp.	3	0.50	1	1.9973	0.7623
Ave. temp. & hours of sunlight	3	0.50	2	2.0824	0.7702
Ave. temp. & rel. humidity	2	0.50	2	2.1220	0.7616
Ave. temp. & precipitation	3	0.30	1	2.1064	0.7601
Max. temp. & hours of sunlight	4	0.50	2	2.7474	0.7066
Max. temp. & rel. humidity	4	0.50	2	2.7228	0.7245
Max. temp. & precipitation	3	0.50	2	2.7458	0.7027

* Cross-validation error ** Explanation of variation

temperature was nearly consistent with a recent study of *S.* Typhimurium infections that found a positive association with air temperature 2-3 weeks before notification (Hald and Andersen, 2001). In a recent case-control study of sporadic *Campylobacter* cases in Denmark, the average time between onset of illness, visiting a healthcare facility, providing a culture and receiving a diagnosis was 21 days (J. Neimann - personal communication). Thus, sustained high temperatures four weeks prior to reported infection roughly corresponds to the time of exposure. This could mean either that temperature plays a direct role in exposure, or more likely that temperature is a marker for behavioral factors, which tend to occur more often in summer months. At a lag of four weeks, a model including average temperature and hours of sunlight also fits with a hypothesis of behavioral factors. In Denmark, as in other northern countries, the amount of sunlight coupled with warmer temperatures during summer months greatly impacts human behavior and may encourage outdoor activities, such as barbecuing, that increase the risk for *Campylobacter* infection. Other outdoor activities such as picnicking, camping or swimming may increase exposure to *Campylobacter* from the environment and natural water sources such as lakes and rivers. Additionally, some studies have suggested that exposure to UV radiation (sunlight) may result in increased susceptibility to infectious diseases (Norval, 2001). Although the mechanisms involved are highly complex and have not been fully investigated, this could serve as a possible partial explanation for our findings.

Our study also showed an association between average temperature and relative humidity lagged at three weeks. This is consistent with a study in Peru that found the number of hospital admissions for diarrheal illness to be related to increased ambient temperature and decreased humidity during an El Niño year (Checkley et al., 2000). However, the mechanisms behind this association remain uncertain and more work is needed to support these findings.

The degree to which temperature and sunlight influence *Campylobacter* incidence in humans should be balanced with the contribution of campylobacteriosis that is attributed to consumption of broiler meat. In Denmark, consumption of poultry is a major risk factor for campylobacteriosis. While it is possible that consumption patterns vary by season, this variation is not enough on its own to explain the seasonal trend.

In broilers, average and maximum temperature three weeks prior to slaughter resulted in the most accurate estimation of prevalence, while a model of average temperature and hours of sunlight was better than any single predictor. This is consistent with another study, which found a positive correlation between minimum temperature, hours of sunshine and *Campylobacter* carriage in poultry (Wallace et al., 1997). As suggested by Jacobs-Reitsma et. al, it may be that temperature has a direct role in *Campylobacter* carriage and/or that sources of contamination within the farm environment such as migratory birds, beetles or rodents are temperature-dependent (Jacobs-Reitsma et al., 1994). Interestingly, temperature did not seem to have as great an effect on prevalence during colder months as it did during warmer months indicating that factors other than temperature play a role during this time or that a threshold temperature is needed. The fact that the incidence was higher during autumn could mean that the effect of higher temperatures or increased prevalence in reservoirs in the summer impacts prevalence in the fall. The higher prevalence seen at colder temperatures and few hours of sunlight fits with the biology of *Campylobacter* which is know to tolerate cold (Chan et al., 2001) and exhibit sensitivity to light (Obiri-Danso et al., 2001).

As *Campylobacter* is sensitive to dry conditions, high relative humidity and/or precipitation contributes to the survival of *Campylobacter* in the environment. Although it was observed that elevated temperature and high humidity coincided with increased *Campylobacter* carriage in poultry in one study (Doyle, 1984), relative humidity and precipitation were not important predictors of broiler prevalence in our study. This lack of association may be due to the fact that measurements were taken and summarized nationally and did not account for regional variations. Temperature and humidity levels measured within the broiler houses may provide a more accurate assessment of their true impact.

Complex interactions exist between climatic factors making elucidation of the effects of a single factor extremely difficult to describe and this should be kept in mind when interpreting our results. Data describing human cases, broiler prevalence and climate were analyzed nationally and the effects of regional climate variation were not examined. Our analysis did not account for factors other than climate that may influence *Campylobacter* prevalence. Although our study excluded persons with a known history of international travel, not all cases were interviewed about travel history, which may bias our results.

The used model is a static time series model. This was chosen to investigate the direct influences by the climatic factors and rapid secondary dynamics. The observed difference between spring and autumn could not be described by the chosen model and this indicates that there most likely are some dynamics involved in the infection of broiler flocks that have time constants higher than a few weeks. The next step is to make a dynamic model, but the choice of model structure will affect the results so some knowledge or assumption of the route of infection is needed to assure useful results.

B.6 Conclusion

We found a strong relationship between temperature and *Campylobacter* infections in both broilers and humans, with the greatest effects at maximum weekly temperatures above 13°C. The combined effect of temperature with hours of sunlight, and to a lesser extent relative humidity, seemed to predict infection in humans, but their role in predicting broiler prevalence was less apparent and more work is needed to support these findings. Future research should focus on elements within the broiler environment that may be susceptible to climate as well as the interaction of micro-climatic factors on and around broiler farms. There is a need to quantify the contribution of broilers as a source of *Campylobacter* in human infections and further examine the effect of temperature on human cases after accounting for this contribution. Investigations should be conducted into food consumption and preparation practices and poultry sales that may vary by season. Finally, the role of climatic factors on *Campylobacter* incidence in other countries should be explored.

References

Anonymous (2000). The increasing incidence of human campylobacteriosis. Technical report, Report and proceedings of a WHO Consultation of Experts. Copenhagen, Denmark 21-25 November 2000, empty.

Anonymous (2001). Annual report on zoonosis in Denmark 2001. Technical report, Danish Zoonosis Centre, Copenhagen.

Berndtson, E., Emanuelson, U., Engvall, A., and Danielsson-Tham, M. (1996). A 1-year study of Campylobacters in 18 Swedish chicken farms. *Preventive Veterinary Medicine*, **26**, 167–85.

Brennhovd, O., Kapperud, G., and Langeland, G. (1992). Survey of thermotolerant *Campylobacter* spp. and *Yersinia* spp. in three surface water sources in norway. *International Journal of Food Microbiology*, **15**, 327–338.

Cabrita, J., Rodrigues, J., Braganca, F., Morgado, C., and Goncalves, A. P. (1992). Prevalence, biotypes, plasmid profile and antimicrobial resistance of *Campylobacter* isolated from wild and domestic animals from northeast Portugal. *Journal of Applied Bacteriology*, **73**, 279–285.

Chan, K. F., Tran, H. L., Kanenaka, R. Y., and Kathariou, S. (2001). Survival of clinical and poultry-derived isolates of *Campylobacter jejuni* at a low temperature (4°c). *Applied Environmental Microbiology*, **67**, 4186–4191.

Checkley, W., Epstein, L. D., Gilman, R. H., Figueroa, D., Cama, R. I., Patz, J. A., and Black, R. E. (2000). Effects of El Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children. *The Lancet*, **355**, 424–450.

Dackowska-Kozon, E. and Brzostek-Nowakowska, J. (2001). *Campylobacter* spp. in waters of three main western pomerania water bodies. *Int. J. Hyg. Environ. Health*, **203**, 435–443.

Doyle, M. P. (1984). Association of *Campylobacter jejuni* with laying hens and eggs. *Applied Environmental Microbiology*, **47**, 533–536.

Eberhart-Phillips, J., Walker, N., Garrett, N., Bell, D., Sinclair, D., Rainger, W., and Bates, M. (1997). Campylobacteriosis in New Zealand: results of a case-control study. *J. Epidemiol. Community Health*, **51**, 686–691.

Engberg, J., Gerner-Smidt, P., Scheutz, F., Moller-Nielsen, E., On, S. L., and Molbak, K. (1998). Water-borne *campylobacter jejuni* infection in a danish town—a 6-week continuous source outbreak. *Clin. Microbiol. Infect.*, **4**, 648–656.

Friedman, C. R., Neimann, J., Wegener, H. C., and Tauxe, R. V. (2000). Epidemiology of *Campylobacter jejuni* infections in the United States and other industrialized nations. In Nachamkin, I. and Blaser, M., editors, Campylobacter *, 2nd ed.* American Society for Microbiology, Washington, D.C.

Glunder, G., Neumann, U., and Braune., S. (1992). Occurrence of *Campylobacter* spp. in young gulls, duration of *Campylobacter* infection and reinfection by contact. *Zentralbl Veterinarmed [B]*, **39**, 119–122.

Hald, B. and Madsen, M. (1997). Healthy puppies and kittens as carriers of *Campylobacter* spp. with special reference to *Campylobacter upsaliensis*. *Journal of Clinical Microbiology*, **35**, 3351–3352.

Hald, T. and Andersen, J. S. (2001). Trends and seasonal variations in the occurrence of Salmonella in pigs, pork and humans in Denmark, 1995-2000. *Berl. Münich Tierärztl Wschr.*, **114**, 346–349.

Hastie, T., Tibshirani, R., and Friedman., J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, Berlin.

Hopkins, R. S. and Scott, A. S. (1983). Handling raw chicken as a source of sporadic *Campylobacter jejuni* infections [letter]. *J. Infect. Dis.*, **148**, 770.

Jacobs-Reitsma, W. F., Bolder, N. M., and Mulder, R. W. A. W. (1994). Cecal carriage of Campylobacter and Salmonella in dutch broiler flocks at slaughter; a one year study. *Poultry Science*, **73**, 1260–66.

Jacobs-Reitsma, W. F., van de Giessen, A. W., Bolder, N. M., and Mulder, R. W. (1995). Epidemiology of *Campylobacter* spp. at two Dutch broiler farms. *Epidemiol. Infect.*, **114**, 413–421.

Kapperud, G. and Aasen, S. (1992). Descriptive epidemiology of infections due to thermotolerant Campylobacter spp. in Norway, 1979-1988. *APMIS*, **100**, 883–90.

Kapperud, G., Skjerve, E., Bean, N. H., Ostroff, S. M., and Lassen, J. (1992). Risk factors for sporadic *Campylobacter* infections: results of a case-control study in southeastern Norway. *J. Clin. Microbiol.*, **30**, 3117–3121.

Kapperud, G., Skjerve, E., Vik, L., Hauge, K., Lysaker, A., Aalmen, I., Ostroff, S. M., and Potter, M. (1993). Epidemiological investigation of risk factors for *Campylobacter* colonization in norwegian broiler flocks. *Epidemiology and Infection*, **111**(2), 245–55.

Luechtfeld, N. A., Blaser, M. J., Reller, L. B., and Wang, W. L. (1980). Isolation of *Campylobacter* fetus subsp. jejuni from migratory waterfowl. *J. Clin. Microbiol.*, **12**, 406–408.

Madsen, H. (1997). *Statistics in finance.* Technical University of Denmark, Informatics and Mathematical Modelling, Lyngby.

Melby, K., Dahl, O. P., and Penner, J. L. (1990). Clinical and serological manifestations in patients during a waterborne epidemic due to *Campylobacter jejuni*. *J. Infect.*, **21**, 309–316.

Neal, K. R. and Slack, R. C. (1995). The autumn peak in *Campylobacter* gastroenteritis. are the risk factors the same for travel- and uk-acquired *Campylobacter* infections? *J. Public Health Med.*, **17**, 98–102.

Neimann, J., Engberg, J., Mølbak, K., and Wegener, H. C. (2003). A case-control study of risk factors for sporadic *Campylobacter* infections in Denmark. *Epidemiology and Infection*, **130**(3), 353–66.

Nielsen, H. A. and Madsen, H. (2001). A generalization of some classical time series tools. *Computational Statistics and Data Analysis*, **37**, 13–31.

Norval, M. (2001). Effects of solar radiation on the human immune system. *Journal of Photochemistry and Photobiology*, **63**, 28–40.

Nylen, G., Dunstan, F., Palmer, S. R., Anderson, Y., Bager, F., Cowden, J., Feierl, G., Galloway, Y., Kapperud, G., Megraud, F., Molbak, K., Petersen, L. R., and Ruutu, P. (2002). The seasonal distribution of campylobacter infection in nine european countries and New Zealand. *Epidemiology and Infection*, **128**(3), 383–90.

Obiri-Danso, K., Paul, N., and Jones, K. (2001). The effects of UVB and temperature on the survival of natural populations and pure cultures of *Campylobacter jejuni*, *Camp. coli*, *Camp. lari* and urease-positive thermophilic *Campylobacters* (uptc) in surface waters. *J. Appl. Microbiol.*, **90**, 256–267.

Rautelin, H. and Hänninen, M.-L. (2000). Campylobacters: the most common bacterial enteropathogens in the Nordic countries. *Ann. Med.*, **32**, 440–5.

Rodrigues, L. C., Cowden, J. M., Wheeler, J. G., Sethi, D., Wall, P. G., Cumberland, P., Tompkins, D. S., Hudson, M. J., Roberts, J. A., and Roderick, P. J. (2001). The study of infectious intestinal disease in England: risk factors for cases of infectious intestinal disease with *Campylobacter jejuni* infection. *Epidemiol. Infect.*, **127**, 185–193.

Rosef, O. and Kapperud, G. (1983). House flies (musca domestica) as possible vectors of *Campylobacter fetus subsp. jejuni*. *Appl. Environ. Microbiol.*, **45**, 381–383.

Speelmon, E. C., Checkley, W., Gilman, R. H., Patz, J., Calderon, M., and Manga, S. (2000). Cholera incidence and El Niño-related higher ambient temperature. *JAMA*, **283**, 3072–3074.

Stanley, K. N., Wallace, J. S., Currie, J. E., Diggle, P. J., and Jones, K. (1998). The seasonal variation of thermophilic campylobacters in beef cattle, dairy cattle and calves. *Journal of Applied Microbiology*, **85**, 472–80.

Studahl, A. and Andersson, Y. (2000). Risk factors for indigenous *Campylobacter* infection: a swedish case-control study. *Epidemiol. Infect.*, **125**, 269–275.

Tenkate, T. D. and Stafford, R. J. (2001). Risk factors for *Campylobacter* infection in infants and young children: a matched case-control study. *Epidemiol. Infect.*, **127**, 399–404.

Tong, S., Bi, P., Parton, K., Hobbs, J., and McMichael, A. (1998). Climate variability and transmission of epidemic polyarthritis. *Lancet*, **351**, 1100.

Wallace, J. S., Stanley, K. N., Currie, J. E., Diggle, P. J., and Jones, K. (1997). Seasonality of thermophilic *Campylobacter* populations in chickens. *Journal of Applied Microbiology*, **82**, 219–24.

Appendix C

Spatial scan statistics using elliptic clusters

Submitted to *J. of Agricultural, Biological, and Environmental Statistics.*

Spatial scan statistics using elliptic clusters

L. E. Christiansen[1,2,*], J. S. Andersen[2], H. C. Wegener[2], and H. Madsen[1]

1: Informatics and Mathematical Modelling, The Technical University of Denmark, Denmark.
2: Department of Epidemiology and Risk Assessment, Institute for Danish Food and Veterinary Research, Denmark.
Corresponding author: Lasse E. Christiansen, Building 321, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. E-mail: lec@imm.dtu.dk, Phone: +45 45253372, Fax: +45 45882673

Abstract

The spatial scan statistic is widely used to search for clusters. This paper shows that the usually applied elimination of secondary clusters as implemented in SatScan is sensitive to smooth changes in the shape of the clusters. We present an algorithm for generation of set of confocal elliptic clusters and propose a new way to present the information when a spatial point process is considered. This method gives smooth changes for smooth expansions of the set of clusters. The proposed method for presenting the information i a set of clusters contains more information. We suggest that where possible, more than one illustration method is used to avoid possible limitations of the chosen method.

Keywords: Overlapping clusters; Campylobacter; broilers; likelihood ratio; Monte Carlo.

C.1 Introduction

In epidemiology the spatial scan statistic is widely used to search for clusters in spatial data one popular implementation is SaTScan (www.SaTScan.org). The software has been used for many applications including infectious diseases (Sauders et al., 2003), cancer (Roche et al., 2002), and veterinary epidemiology (Enemark et al., 2002; Hoar et al., 2003).

SaTScan is used in cases with few spatial locations as well as in cases with many locations.

In the case where secondary clusters are of interest SaTScan eliminates all clusters overlapping with the most significant cluster; this gives a large reduction in the number of significant clusters - especially if a case with hundreds of spatial locations is considered. The question is how much information is lost during this reduction?

It is a fact that not all naturally occuring phenomena are circular and hence there is a need for methods that can find clusters with other shapes. In this paper we suggest a method leading to set of confocal ellipses.

C.2 Model description

The spatial scan statistic is a log likelihood ratio statistic for spatial point processes. Originally the spatial scan statistic was a 1D statistic but it has been developed to manage both 2D and 3D problems (Kulldorff, 1997), (Kulldorff, 1999) . The 1D version was designed for temporal data, the higher dimensional designs were made for spatial and for combinations of spatial and temporal data. The implementation in SaTScan includes methods for both Bernoulli and Poisson point processes with circular windows of varying size. Improvements of the algorithm have been suggested by several researchers, eg. Boscoe et al. (2003) have presented a scheme that does the elimination on 10 levels of significance and Gangnon and Clayton (2001) worked with a weighted average likelihood ratio test.

The present work only concerns the Bernoulli model but the comments apply to the Poisson model as well. Let a geographical region G be the area of interest. Within this region there are a number of observations or a population n_G including a number of cases c_G. Any subset of points in G is a potential cluster. Let $Z \subset G$ contain n_Z observations including c_Z cases.

Let $\mathcal{Z}$ be the set of subsets of a given family, eg. circular subsets with varying centre and radius. (A description of the generation is postponed until Section C.3.)

At first high risk clusters are considered. The null hypothesis is that the incidence rate is the same all over the study region and the alternative hypothesis is that the rate is higher in Z. Under the null hypothesis the maximum likelihood is given by:

$$L_0(G) = \left(\frac{c_G}{n_G}\right)^{c_G} \left(1 - \frac{c_G}{n_G}\right)^{n_G - c_G}$$

Under the alternative hypothesis the maximum likelihood for the subset Z being a high risk cluster is:

$$L(Z) = \left(\tfrac{c_Z}{n_Z}\right)^{c_Z} \left(1 - \tfrac{c_Z}{n_Z}\right)^{n_Z - c_Z} \left(\tfrac{c_G - c_Z}{n_G - n_Z}\right)^{c_G - c_Z} \left(1 - \tfrac{c_G - c_Z}{n_G - n_Z}\right)^{n_G - c_G - (n_Z - c_Z)}$$

if $c_Z/n_Z > (c_G - c_Z)/(n_G - n_Z)$ and

$$L(Z) = \left(\frac{c_G}{n_G}\right)^{c_G} \left(1 - \frac{c_G}{n_G}\right)^{n_G - c_G}$$

otherwise. If low risk clusters are of interest one uses non-cases instead of cases as the observations. If the interest is in both high and low risk clusters simultaneously, $c_Z/n_Z > (c_G - c_Z)/(n_G - n_Z)$ determines if a subset is to be considered a high risk subset (true) or a low risk subset (false).

Next, we find the subset with the highest maximum likelihood:

$$\hat{Z} = \{Z : \ L(Z) \geq L(Z') \ \ \forall \, Z' \in \mathcal{Z}\}$$

This is the most likely cluster and to do statistical inference we find the likelihood ratio (LR) as:

$$LR = \frac{L(\hat{Z})}{L_0(G)}$$

The distribution of the likelihood ratio is not available in a closed analytical form. Instead Monte Carlo simulation is used to generate a distribution of the likelihood ratios under the null hypothesis. See Section C.4 for further details on the Monte Carlo simulations.

C.3 Generation of subsets

In SaTScan concentric circles is the only choice for the construction of a set of subsets. In this paper we present a new algorithm for generation of confocal ellipses leading to a larger set of subsets and a set that includes the set of concentric circles as a true subset. The motivation for this larger and more flexible set is that not all naturally occuring phenomena can be assumed to be circular.

First we will present the algorithm used in SaTScan to generate circular subsets. In principle it is possible to make an infinite set of circles in the study region but in practice only different subsets of the locations in the point process are of interest. Thus for a given center an increase in radius is only of interest if a new point enters the circle. Having n locations and setting the maximum cluster size to 50% leads to a maximum of $0.5 \cdot n$ subsets per center. Centers that are too close give rise to the same subsets and are not of interest. The question is how to choose the centers. In SaTScan one can either specify a list of centers or use the locations of the points in the point process as centers. We have chosen the later.

An algorithm for generation of concentric circles including up to a pre specified proportion of the locations, say 0.50 is as follows:

1: Select one point which will be the center and calculate the distance from this point to all other points.

2: Sort the points according to the distance to the selected center.

3: Make the point at the selected center the first subset.

4: Include the nearest point and make it the next subset.

5: Keep including the next nearest point, until the pre specified proportion is included, making a new subset for every inclusion.

6: Repeat 1 trough 5 making all points a center one by one.

7: Remove identical subsets.

We propose to make a larger family of subsets consisting of concentric ellipses that include the above set of concentric circles as a subset. The family of ellipses is a larger and more flexible family than the circles but still the set is going to be finite as only unique subsets are of interest. We have decided that the foci are to be chosen among the locations in the point process. This makes the maximum length of the constructed set of subsets $n^3 p_m p_s$, where p_m is the proportion of the total number of locations in the largest cluster, and p_s is the proportion of the total number of locations that are used as the secondary foci.

An algorithm for generation of confocal ellipses including up to a pre specified proportion of the locations, say 0.50 is as follows:

1: Select one point, which will be the first focal and calculate the distance from this point to all other points.

2: Sort the points according to the distance to the first focal.

3: Choose the first focal as the second focal in order to start with the set of circles. Include the next nearest point, until the pre specified proportion is included, making a new subset for every inclusion.

4: *Make the next nearest point the second focal and calculate the distance from this point to all other points.*

5: *Sort the points according to the sum of the distances to the two foci.*

6: *Include the nearest point and make it the next subset.*

7: *Keep including the next nearest point, until the pre specified porportion is included, making a new subset for every inclusion.*

8: *Repeat 4 trough 7 until the pre specified proportion of secondary foci has been used.*

9: *Repeat 1 trough 8 making all points the first focus, one at a time.*

10: *Remove identical subsets.*

The removal of identical subsets can be optimized by adding a Step *4b: If the two foci were previously used as foci go to Step 4.* This reduces the number of generated subsets in Step 1 through 9 by up to 50% dependent on the mutual distances, ie. dependent on changes in the density of locations.

C.4 Making inference

In the previous section two different algorithms were described for the generation of a set of subsets. Furthermore an expression for the likelihood ratio associated with each subset has been derived. Finding the most likely subset is an easy task as it is just the one with the highest likelihood ratio.

In the general setting it is infeasible, if at all possible, to write an expression for the distribution of likelihood ratios in closed analytical form. Hence, in order to test if the null hypothesis holds we use Monte Carlo simulations. One Monte Carlo simulation is made by generating a new dataset under the null hypothesis, ie. a point process with the same number of observations at the same locations and with the same risk of being a case as the over all risk in the original dataset. Then the most likely cluster is found using the simulated data and its likelihood ratio is recorded. In order to make inference usually one should make at least 999 such simulations - the number depends on the considered confidence limit. In the case of a 1% significance level and 9,999 simulations the most likely subset is significant if its likelihood ratio is among the 99 highest Monte Carlo maximum likelihood ratios (MCMLR).

If the null hypothesis is rejected for the most likely subset the question is how about all the other subsets? The subsets having one additional point or lacking one point when compared with the most likely subset will in most cases have likelihood ratios just below the most likely one; this is to remind us that the underlying cluster may be of a different size. It is also of interest if there are other significant subsets representing totally different clusters. When testing other clusters than the most likely do keep in mind that the Monte Carlo generated distribution is made of the most likely cluster in the simulated datasets. The second most likely cluster should be tested against the second most likely in the simulated datasets, and so forth. This is not feasible, but the fact that the second most likely cluster has a lower likelihood ratio than the most likely cluster makes it possible to make an approximate test. When using the distribution for the most likely cluster to test secondary clusters the true significance level is greater than or equal to the reported significance level.

In a case with many subsets having a likelihood ratio above the significance level it becomes difficult to present the information. There are several possible solutions to this problem, the one implemented in SaTScan starts by eliminating all subsets having a non empty intersection with the most likely subset. Then all subsets having a non empty intersection with the second most likely subset are eliminated and so forth. This makes a dramatic reduction in the number of subsets, but is it always a reasonable way to do it? We propose an alternative: For each location the proportion of significant subsets that it belongs to is used as a measure (it could be weighted using the likelihood ratios). In the case where both high and low risk areas are of interest these are treated separately and care should be taken to make it possible to see if some locations are both included in high and low risk subsets. This gives a more detailed picture of the underlying cluster(s).

C.5 Examples

A dataset with *Campylobacter* infections in Danish broilers is used to illustrate the differences in acquired information using two different schemes to present significant clusters as well as the differences between circular and elliptic clusters.

C.5.1 Dataset: *Campylobacter* in Danish broilers

The dataset is from a national monitoring program for Danish broiler flocks examined for *Campylobacter* by cloacal swabs at slaughter. All poultry flocks slaughtered in Denmark between 1998 and 2001 were in the program. During this period 23,279 broiler flocks were sampled; 7,741 flocks were excluded from the dataset either because they were reported to come from a farm not registred for broiler production or were reported to come from an unknown house. Another 7,217 flocks were excluded because they were not from the first batch. The reason for only including samples from the first batch is that the risk increases for the following batches and to avoid dependence between batches. This leaves 8321 flocks in the dataset.

Only flocks originating in Jutland or at Funen are included; this reduces the number of houses from 828 to 794 and the number of flocks from 8321 to 8056. Houses on the same farm are considered lying on a row with 10 meters between houses. In total 3080 of the flocks were positive corresponding to a 38.2% risk that a flock is infected.

The sampling was performed at the slaughterhouse at the entrance to the slaughterline by collecting one cloacal swab from ten animals. If at least one of these ten swabs tested positive for *Campylobacter* spp. the flock was labeled positive.

The dataset was collected under a national monitoring program and has also been used to find climatic predictors for the pronounced seasonality observed in *Campylobacter* infections in broiler flocks (Paper B).

C.5.2 Presenting significant clusters

The geographical locations in the dataset described above were used to generate circular clusters including up to 50% of the locations. Before removing duplets the full set of clusters contained 315,218 clusters. That set was reduced to 107,953 unique clusters. The next step was calculating the maximum likelihood ratios based on the dataset. We chose to use the 10,000 most likely clusters in our analysis, ie. the 10,000 highest maximum likelihood ratios.

In this case the most likely cluster is low prevalent, but why this cluster and not the one including one additional location or one less? The left part of Figure C.1 shows the likelihood ratio for the circles that have the same center as the most likely cluster, it is seen that the clusters containing just about the same number of locations have about the same likelihood ratio. Therefore one has to be aware that the underlying cluster may be different from the most likely one. It is also noted that a range of clusters less than half the size have relatively high likelihood ratios. When looking in the right part one can see that the proportion of positive flocks is much lower in the region with about 100 locations when compared to the proportion of the most likely cluster. This indicates that there is a core of the most likely cluster that is accountable and hence those locations may be of greater interest than the most likely cluster. It is also seen that the likelihood ratio drops fast when the circles becomes larger than the most likely cluster.

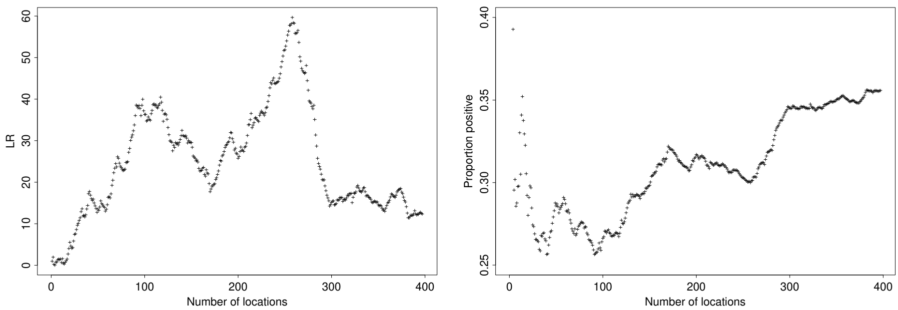

Figure C.1: The likelihood ratio (left) and proportion of positive flocks (right) as a function of the number of locations in the concentric circles that has the same center as the most likely circle.

Let us take a look at the clusters. Figure C.2 shows four maps of the region of interest with a dot at every broiler house and red and green bars showing the occurence of locations in high and low risk clusters, respectively. The lower right plot shows the result of a run with SaTScan showing the clusters that are significant at a 5% level. The upper left plot is made using the 100 most significant clusters without eliminating any clusters besides the duplets. The length of the bars indicate how large a proportion of the high and low risk clusters a location belongs to. The highest occurence of high and low is indicated in the legend. This is the same for the second and third plots, except that they are made using 1,000 and 21,324 clusters, respectively. The 21,324 clusters are the circular clusters that are significant at a 10^{-4} level.

It is seen that the 100 most likely clusters all define the same low risk area and this was also expected. When including 1,000 clusters some high risk clusters are included, and these define an area in the North-Western part of Jutland and a few covers the middle part of Jutland. So far the areas covered by clusters are in agreement with the two most likely clusters found using SaTScan. This agreement fails when the 21,324 most significant clusters are used, as can be seen in the lower left subfigure. The low risk area is almost stable but the high risk clusters cover more and more. Some locations in the Northern part of the low risk area are included both in low and high risk clusters. When comparing with SaTScan one notes that SaTScan finds a low risk area just south of the first high risk cluster and some small high risk clusters surrounding the most likely cluster. Those small high risk areas appear as most other clusters are removed due to an overlap with the most likely cluster. So we see that the selection done by SaTScan is in agreement with our representation as long as the clusters are far appart.

C.5.3 Elliptic clusters

The next step is a comparison of circular and elliptic clusters. First take a look at Figure C.3. The yellow distribution is made using circles including up to 50% of the locations. The other distributions are made using ellipses also limited to 50% of the locations and including up to the 0.3%, 1%, 20%, and 50% nearest locations as the second focus, ie. $p_m = 0.5$ in all cases and $p_s \in [0, 0.003, 0.01, 0.20, 0.50]$ in the specified order. In both parts of the figure a smooth transition is seen - more flexible shapes result in higher likelihood ratios. This is as expected.

The removal of dublets makes a large reduction in the number of clusters where the likelihood ratio has to be evaluated. Table C.1 shows that the reduction is between a factor of 3 and 10, and that a larger reduction is seen for more flexible clusters, ie. ellipses with a high p_s. The predicted maximum length for the largest case with $p_m = p_s = 0.5$ is above 125 mio. clusters. This number reduces to less than 8 mio. clusters due to the removal of clusters thus speeding up the calculation by a factor of 15. $p_s = 0.003$ was chosen as this only allows the nearest neighbor as a secondary focus and this only results in a small increase in the number of unique clusters. In fact the most likely cluster is the same when $p_s < 0.025$. From the table it is also seen that the highest likelihood ratio is about 3-4 times higher than the highest MCMLR indicating that the clusters are highly significant in the present dataset. To emphasize the significance another test was preformed with 9,999,999 Monte Carlo simulations, and in this case the highest MCMLR was 23.4981, so these clusters are extremely significant with the given test.

In order to investigate how the elimination implemented in SaTScan works, when applied to a more flexible set of clusters, all the clusters having a p-value below 10^{-4} were selected. The two plots in the top of Figure C.4 show the clusters found with $p_s = 0.01$(left) and 0.2(right). These should be compared with Figure C.2D. It is seen that the change from circles to ellipses with $p_s = 0.01$ primarily changes the shape of the secondary clusters a little. When compared with the set of ellipses with $p_s = 0.2$ the picture is changed quite dramatically. The low risk cluster on Funen and surroundings is almost stable but the Southern most locations in Jutland that used to be part of the cluster have now joined a

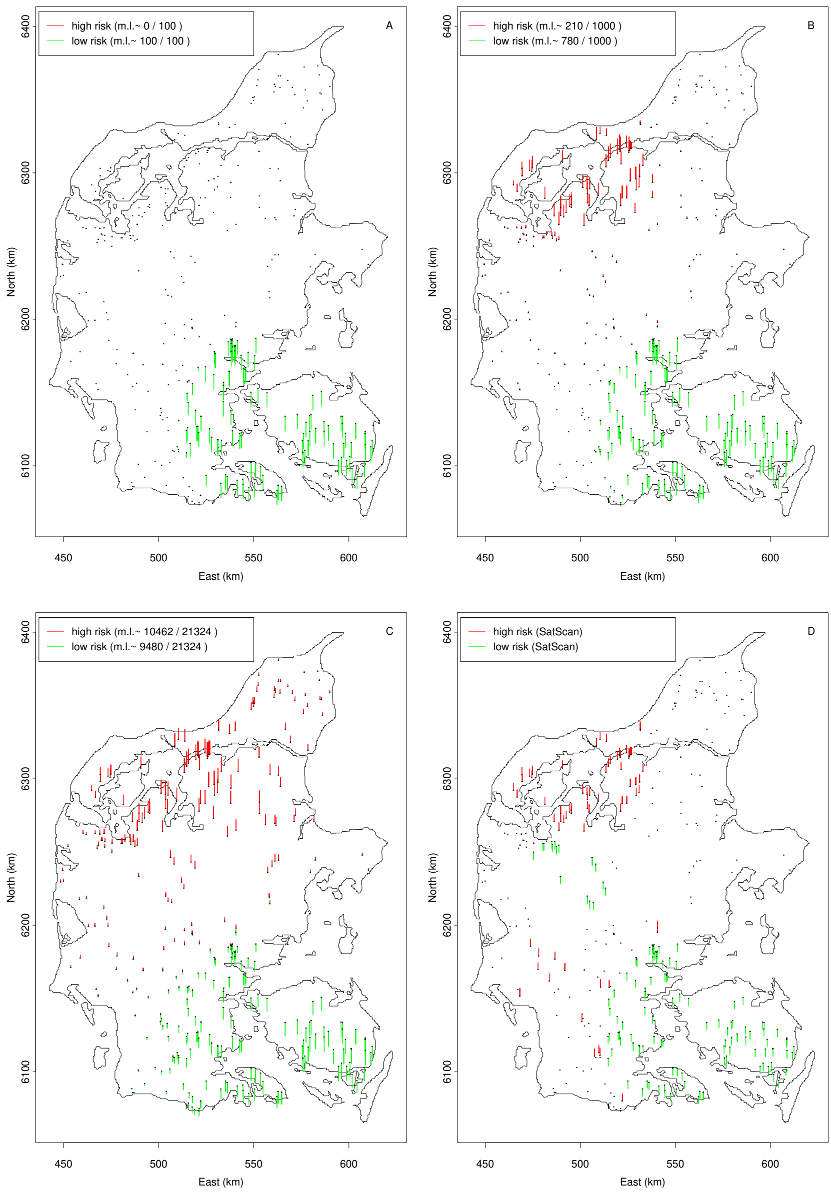

Figure C.2: Plots A trough C showing the geographical distribution of the 100, 1,000, and 21,324 most significant circular clusters, the 21,324 clusters are the ones with a significance of 0.0001 or less. Green is low prevalent and red is high prevalent, the length of the markers indicates the proportion of clusters that a given house is in (see the legends). The lower right plot is made using the output from SaTScan.

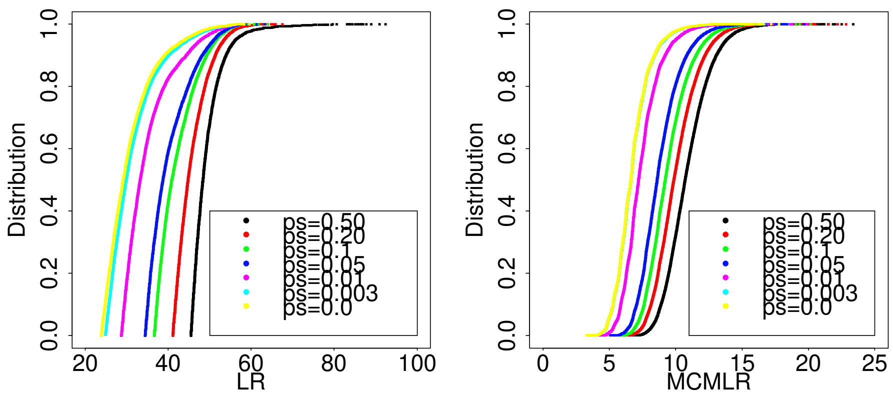

Figure C.3: Distributions of the 10,000 highest likelihood ratios (left) and the 99,999 MCMLRs (right).

p_m	p_s	# clusters with duplets	# clusters without duplets	Highest LR	Highest MCMLR
0.50	0.0	315,218	107,953	59.628	17.7875
0.50	0.003	537,374	126,708	59.628	16.689
0.50	0.01	1,525,790	223,623	59.628	19.9197
0.50	0.02	3,080,090	351,406	59.628	19.3834
0.50	0.05	7,662,602	717,338	63.1764	19.8396
0.50	0.10	15,134,726	1,308,741	64.0019	20.1673
0.50	0.20	29,791,478	2,655,066	67.4917	22.8339
0.50	0.50	70,288,418	7,939,616	92.3859	23.3868

Table C.1: Summary of results for 8 runs with fixed p_m and varying p_s. For all runs 99,999 Monte Carlo simulations are made, and thus a cluster is significant at a 10^{-5} level if its likelihood ratio is above the highest MCMLR.

high risk cluster. The high risk area in the North-Western part of Jutland is now split into two small high risk clusters leaving the impression that it may be a localized phenomenon. The two bottom plots show the geographical distribution of all the clusters that are significant at a 10^{-4} level with $p_s = 0.01$(left) and 0.2(right). When compared with Figure C.2C only small changes are seen, so the overall changes in the most likely clusters are small. In the case with $p_s = 0.2$ a total of 243,244 clusters were significant at a 10^{-4} level and the SaTScan elimination reduces this to 9 clusters of which only one appears to be almost stable as the flexibility is increased.

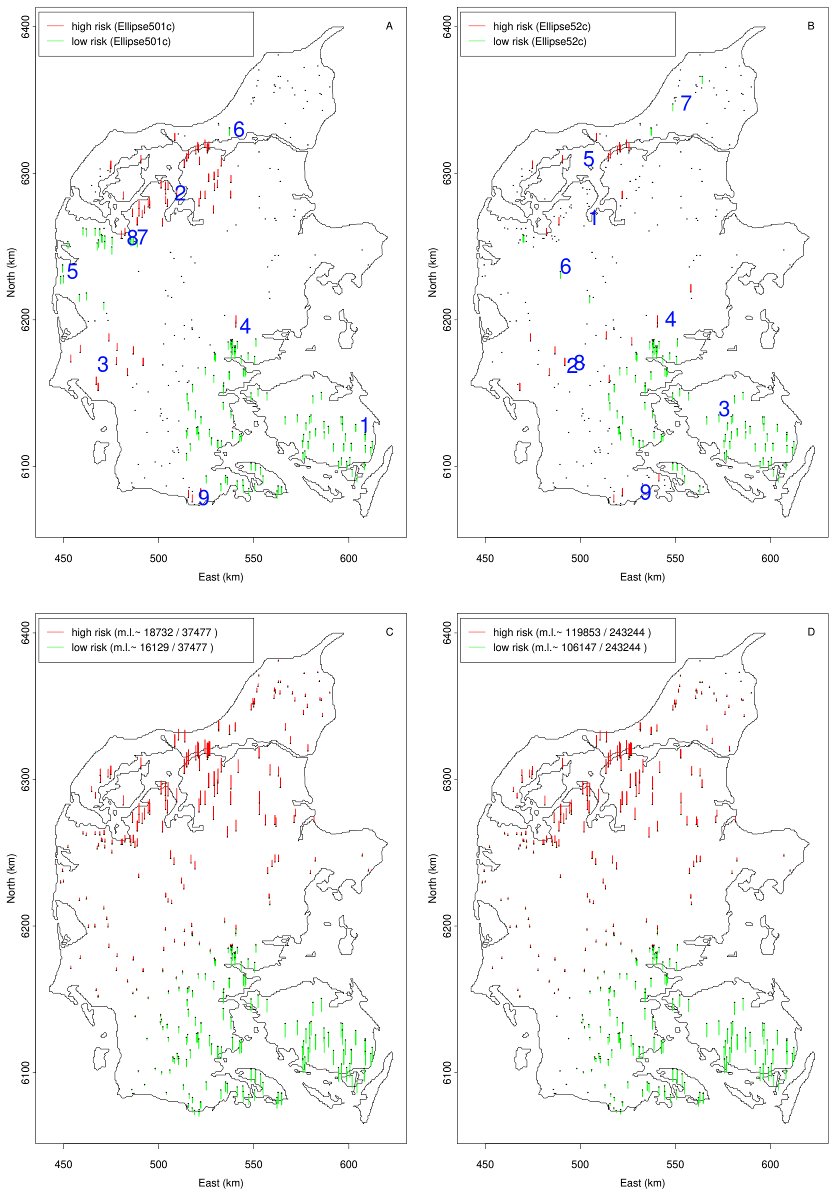

Figure C.4: Plots presenting elliptic clusters that are significant at a 10^{-4} level. The two to the left are made with $p_s = 0.01$ and the two to the right with $p_s = 0.2$. The two at top show the result of SaTScan elimination and the two at the bottom show all the significant clusters.

C.6 Discussion

As more and more computational power becomes available it becomes feasible to run larger computations. This is most likely the reason why Geographic Information Systems (GIS) and analysis of geographic data are increasingly used. Following the increase in computer power we tend to focus on gathering more detailed datasets to be analysed. When applying methods that originally were tested for use on more simple datasets we must be aware that some possible artifacts may come out.

It is well defined how to find the most likely cluster, but in cases with small populations or few observations on each geographical location the size is highly dependent on the outcome of the underlying process at the locations that are in the other rim and just outside the most likely cluster. Hence, the size is stochastic and this should be kept in mind when interpreting the most likely (and other) clusters. The effect may be more pronounced when looking at secondary clusters with the SaTScan elimination scheme. As long as the secondary clusters are found far away from more significant clusters it seems to be reasonable, but the elimination fails to find the true underlying clusters if the shape of those clusters are not in the chosen set of clusters; and also if the chosen set of shapes forces a cluster to cover more than the true underlying cluster and hence eliminates the neighboring underlying clustes. These effects become most pronounced when high and low risk clusters are geographically close under the given set of clusters.

One way to circumvent some of these problems is to create a more flexible set of clusters, and we have suggested to create clusters built up from confocal ellipses as they contain the set of concentric circles as a true subset. When a large proportion of the locations are included as secondary foci this gives rise to a very flexible set of clusters. We find that this flexibility is needed to avoid the problems with underlying clusters having odd shapes lying relatively close. The problem with such a flexible set is how to present the acquired information. In this study, 243,244 clusters were significant at a 10^{-4} level, they were reduced to 9 when the SaTScan elimination scheme was applied. Those 9 clusters make it hard to judge where the underlying problems are. This information, which must be the most important, is more easily accessible when all the highly significant clusters are represented. The disadvantage is that the borders of the areas of interest become less well defined. One could choose to select those locations that are part of more than a certain proportion of the highly significant clusters (in the present representation this corresponds to only selecting those locations having a bar longer than a selected threshold). Again one should keep in mind that the underlying process is stochastic and hence all defined borders of clusters should be interpreted with some flexibility.

When evaluating the additional flexibility provided by elliptic clusters one should keep in mind that rotating and scaling the coordinate system can transform an elliptic cluster into a circular cluster. Hence one should use an evaluation that is quasi-stable for smooth changes in the flexibility of the set of clusters under inspection. We have examplified that this is not the case when using the elimination scheme implemented in SaTScan, ie. non overlapping clusters.

The way we have chosen to evaluate and present the information may not be the best in all circumstances. We propose that for a given case one should try to present the results of the spatial scan statistic in several ways to get a better understanding of the underlying process. This could also include the use of the likelihood ratios to calculate a weighted proportion of clusters that a given location belongs to.

This work is based on an all new implementation of the algorithms. It is written in C++ and OpenMP has been used to make a parallel version so that a multiprocessor computer can be used to speed up the calculations. We have used up to 20 CPUs on a SunFire 15K in the calculations and more could easily have been used. The scalability is good as all the Monte Carlo smulations are independent and this is the most time consuming part of the calculations.

C.7 Conclusion

We have shown that set of confocal ellipses creates a smooth expansion (with discrete steps due to the point process) of concentric circles. Furthermore, we have provided examples of the problems that can occur when the SaTScan elimination scheme is applied, in particular rotations, scalings and clusters in close proximity. A new way to present the information and other alternatives have been suggested. It is found to be important to use more than one way to illustrate the information to avoid possible pitfalls.

C.8 Acknowledgement

We acknowledge the Danish Center for Scientific Computing (DCSC) for the support.

References

Boscoe, F. P., McLaughlin, C., Schymura, M. J., and Kielb, C. L. (2003). Visualization of the spatial scan statistic using nested circles. *Health and Place*, **9**, 273–277.

Enemark, H. L., Ahrens, P., Juel, C. D., Petersen, E., Petersen, R. F., Andersen, J. S., Lind, P., and Thamsborg, S. M. (2002). Molecular characterization of Danish cryptosporidium parvum isolates. *Parasitology*, **125**, 331–341.

Gangnon, R. E. and Clayton, M. K. (2001). A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine*, **20**, 2977–2987.

Hoar, B. R., Chomel, B. B., Rolfe, D. L., Chang, C. C., Fritz, C. L., Sacks, B. N., and Carpenter, T. E. (2003). Spatial analysis of Yersinia pestis and bartonella vinsonii subsp berkhoffii seroprevalence in California coyotes (canis latrans). *Preventive Veterinary Medicine*, **56**, 299–311.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

Kulldorff, M. (1999). *Recent Advances on Scan Statistics and Applications*, chapter Spatial scan statistics: Models, calculations and applications, pp. 303–322. Birkhauser, Boston.

Patrick, M. E., Christiansen, L. E., Wainø, M., Ethelberg, S., Madsen, H., and Wegener, H. C. (2004). Effects of climate on incidence of *Campylobacter* spp. in humans and prevalence in broiler flocks in Denmark. *Accepted by Applied and Environmental Microbiology*, **x**, xx.

Roche, L. M., Skinner, R., and Weinstein, R. B. (2002). Use of a geographic information system to identify and characterize areas with high proportions of distant stage breast cancer. *Journal of Public Health Management and Practice*, **8**, 26–32.

Sauders, B. D., Fortes, E. D., Morse, D. L., Dumas, N., Kiehlbauch, J. A., Schukken, Y., Hibbs, J. R., and Wiedmann, M. (2003). Molecular subtyping to detect human listeriosis clusters. *Emerging Infectious Diseases*, **9**, 672–680.

Appendix D

Estimating the parameters of a stochastic differential equation model of bacterial growth

Submitted to *Biometrics*.

Estimating the parameters of a stochastic differential equation model of bacterial growth.

Lasse Engbo Christiansen[1,2,*], Jens Strodl Andersen[2], Henrik Caspar Wegener[2], and Henrik Madsen[1]

1: Informatics and Mathematical Modelling, The Technical University of Denmark, Denmark.
2: Department of Epidemiology and Risk Assessment, Institute for Danish Food and Veterinary Research, Denmark.
Corresponding author: Lasse E. Christiansen, Building 321, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. E-mail: lec@imm.dtu.dk, Phone: +45 45253372, Fax: +45 45882673

D.1 Summary

Non-linear dynamic systems in continuous time that are partially observed in discrete time, such as bacterial growth, are estimated and modelled using continuous-discrete stochastic state space models where the dynamics are described by stochastic differential equations. This class of models provides a decomposition of the total error into system noise, as Wiener processes, and measurement noise, described as a Gaussian white noise sequence. The presented estimation scheme features maximum likelihood (ML) and maximum a posteriori (MAP) estimation of the unknown parameters. The method gives less biased results in the presence of significant process noise than the more commonly used output error (OE) method, where a minimization of the simulation error is considered.

E. coli growth data under different concentrations of copper are used to illustrate some of the advantages of the ML method over the OE method.

Keywords: Stochastic Differential Equations; bacteria; dynamics

D.2 Introduction

A large effort is made for understanding and describing the mechanisms involved in bacterial growth. One of the questions to be answered is how a bacteria selects between several nutrients available in the surrounding media.

In the litterature many matematical models have been suggested with different motivations. Most of the models are formulated as a set of ordinary differential equations (ODE) containing states for different substrates and bacteria. The conversion from substrate to bacteria is modeled by Monod kinetics multiplied by some activation factor. This activation factor is either found from some algebraic equations or by adding additional states to the set of differential equations. One motivation for the later is that these states can describe the involved enzyme kinetics.

From biochemistry it is known that many intracellular mechanisms are involved in the regulation of the cell, and for many purposes it is infeasible to establish a model that incorporates all these as there are too many parameters to be estimated and the complexity becomes too high to simulate on todays computers. Thus one has to consider a reduced model and the question is how to take the processes that were not explicitly included in the model into account.

When a model based on ODEs is used all processes that are not explicitly included are fully neglected and this will in many cases result in unlikely parameter estimates and a need for a more complex model. The complex model might however be non operational for the intended application. Applying stochastic differential equations (SDE) (Øksendal, 1989) instead of ODEs allows for use of noise processes to describe the variation not explicitly described by the model.

In practice observations are made in discrete time with some measurement error. Using ODEs as part of a modelling procedure, an output error (OE) method minimizing the sum of squared residuals from simulations is implicitly considered. In this ODE/OE case it is not possible to separate the measurement error from the model errors. Using SDEs this is possible as system noise takes care of the model incorrectness and a separate measurement error can be estimated. These models are entitled continuous-discrete stochastic state space models.

D.3 Methods

This section contains a condensed outline of a maximum likelihood (ML) estimation scheme for estimating parameters in continuous-discrete stochastic state space models. A detailed description can be found in the documentation for CTSM (www.imm.dtu.dk/ctsm); the program used for the modeling.

The last subsection contains a description of the dataset used to illustrate the difference between OE and ML methods.

D.3.1 Model structure

In the general case, the continuous-discrete stochastic state space model is a model that consists of a set of non-linear discretely, partially observed Itô SDE's with measurement noise, i.e.:

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \tag{D.1}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{D.2}$$

where $t \in \mathbb{R}$ is the time variable; $\boldsymbol{x}_t \in \mathcal{X} \subset \mathbb{R}^n$ is a vector of state variables; $\boldsymbol{u}_t \in \mathcal{U} \subset \mathbb{R}^m$ is a vector of input variables; $\boldsymbol{y}_k \in \mathcal{Y} \subset \mathbb{R}^l$ is a vector of output variables; $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ is a vector of (possibly unknown) parameters; $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ are known non-linear functions; $\{\boldsymbol{\omega}_t\}$ is an n-dimensional standard Wiener process and $\{\boldsymbol{e}_k\}$ is an l-dimensional white noise process with $\boldsymbol{e}_k \in N\left(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta})\right)$.

The diffusion term in (D.1) is assumed to be independent of the state variables, as this makes parameter estimation more feasible.

D.3.2 Parameter estimation

The solution to (D.1) is a Markov process and an estimation scheme based on probabilistic methods can therefore be applied to estimate the unknown parameters of the model in (D.1)-(D.2), e.g. *maximum likelihood* (ML) or *maximum a posteriori* (MAP), where the latter can be applied if prior information about the parameters is available. Let:

$$\mathbf{Y} = [\mathcal{Y}_{N_1}^1, \mathcal{Y}_{N_2}^2, \dots, \mathcal{Y}_{N_i}^i, \dots, \mathcal{Y}_{N_S}^S] \tag{D.3}$$

be a set of S stochastically independent sequences of consecutive measurements, where each sequence is:

$$\mathcal{Y}_{N_i}^i = [\boldsymbol{y}_{N_i}^i, \dots, \boldsymbol{y}_k^i, \dots, \boldsymbol{y}_1^i, \boldsymbol{y}_0^i] \tag{D.4}$$

Let $p(\boldsymbol{\theta})$ be a prior probability density function for the parameters. In the general case, point estimates of the parameters in (D.1)-(D.2) can then be found as the parameters $\boldsymbol{\theta}$ that maximize the posterior probability density function:

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto \left(\prod_{i=1}^{S} p(\mathcal{Y}_{N_i}^i|\boldsymbol{\theta}) \right) p(\boldsymbol{\theta}) \tag{D.5}$$

or equivalently:

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto \left(\prod_{i=1}^{S} \left(\prod_{k=1}^{N_i} p(\boldsymbol{y}_k^i|\mathcal{Y}_{k-1}^i, \boldsymbol{\theta}) \right) p(\boldsymbol{y}_0^i|\boldsymbol{\theta}) \right) p(\boldsymbol{\theta}) \tag{D.6}$$

where the rule $P(A \cap B) = P(A|B)P(B)$ has been applied in a successive manner to form products of conditional probability density functions.

In order to obtain an exact evaluation of the posterior probability density function in (D.6), the initial probability density functions $p(\boldsymbol{y}_0^i|\boldsymbol{\theta})$, $i = 1, \ldots, S$, must be known and all subsequent conditional probability density functions must be determined by successive prediction based on Kolmogorov's forward equation and updating based on Bayes' rule (Jazwinski, 1970). In practice this approach is computationally infeasible and an alternative is therefore needed. Methods based on approximate non-linear filters provide a computationally feasible (Nielsen et al., 2000) solution of the models of the type in (D.1)-(D.2). However, since the diffusion term in (D.1) has been assumed to be independent of the state variables, a much simpler alternative can be applied here.

More specifically, since the SDE's in (D.1) are driven by a Wiener process, and since increments of a Wiener process are Gaussian, it is reasonable to assume, under some regularity conditions, that the conditional probability density functions can be adequately approximated by Gaussian densities, which means that a method based on the extended Kalman filter (EKF) can be applied for calculating the conditional mean and covariance. The Gaussian density is completely characterized by its mean and covariance, so by introducing:

$$\hat{\boldsymbol{y}}_{k|k-1}^i = E\{\boldsymbol{y}_k^i|\mathcal{Y}_{k-1}^i, \boldsymbol{\theta}\} \tag{D.7}$$

$$\boldsymbol{R}_{k|k-1}^i = V\{\boldsymbol{y}_k^i|\mathcal{Y}_{k-1}^i, \boldsymbol{\theta}\} \tag{D.8}$$

$$\boldsymbol{\epsilon}_k^i = \boldsymbol{y}_k^i - \hat{\boldsymbol{y}}_{k|k-1}^i \tag{D.9}$$

and by further assuming that the prior probability density function for the parameters is Gaussian as well and therefore also introducing:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = E\{\boldsymbol{\theta}\} \tag{D.10}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = V\{\boldsymbol{\theta}\} \tag{D.11}$$

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}} \tag{D.12}$$

the posterior probability density function becomes:

$$\begin{aligned} &p(\boldsymbol{\theta}|\mathbf{Y}) \propto \\ &\left(\prod_{i=1}^{S} \left(\prod_{k=1}^{N_i} \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\epsilon}_k^i)^T (\boldsymbol{R}_{k|k-1}^i)^{-1} \boldsymbol{\epsilon}_k^i\right)}{\sqrt{\det\left(\boldsymbol{R}_{k|k-1}^i\right)}\left(\sqrt{2\pi}\right)^l} \right) \times p(\boldsymbol{y}_0^i|\boldsymbol{\theta}) \right) \\ &\times \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\right)}{\sqrt{\det\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)}\left(\sqrt{2\pi}\right)^p} \end{aligned} \tag{D.13}$$

The parameter estimates can now be determined by further conditioning on the initial conditions:

$$\mathbf{y_0} = [\boldsymbol{y}_0^1, \boldsymbol{y}_0^2, \dots, \boldsymbol{y}_0^i, \dots, \boldsymbol{y}_0^S] \tag{D.14}$$

and applying non-linear optimisation to find the maximum of the logarithm of the resulting posterior probability density function, i.e.:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \{\ln(p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{y_0}))\} \tag{D.15}$$

In the general case the estimation method implied by (D.15) is a MAP procedure, but, if no prior information about the parameters is available, it reduces to a ML procedure.

In either case, the innovations $\boldsymbol{\epsilon}_k^i$ and their covariances $\boldsymbol{R}_{k|k-1}^i$ can be computed recursively by means of the EKF for each set of parameters $\boldsymbol{\theta}$ in the optimisation. More specifically, the EKF consists of the output *prediction* equations:

$$\hat{\boldsymbol{y}}_{k|k-1}^i = \boldsymbol{h}(\hat{\boldsymbol{x}}_{k|k-1}^i, \boldsymbol{u}_k^i, t_k^i, \boldsymbol{\theta}) \tag{D.16}$$

$$\boldsymbol{R}_{k|k-1}^i = \boldsymbol{C}\boldsymbol{P}_{k|k-1}^i\boldsymbol{C}^T + \boldsymbol{S} \tag{D.17}$$

the *innovation* equation:

$$\boldsymbol{\epsilon}_k^i = \boldsymbol{y}_k^i - \hat{\boldsymbol{y}}_{k|k-1}^i \tag{D.18}$$

the Kalman *gain* equation:

$$\boldsymbol{K}_k^i = \boldsymbol{P}_{k|k-1}^i\boldsymbol{C}^T(\boldsymbol{R}_{k|k-1}^i)^{-1} \tag{D.19}$$

the *updating* equations:

$$\hat{\boldsymbol{x}}_{k|k}^i = \hat{\boldsymbol{x}}_{k|k-1}^i + \boldsymbol{K}_k^i\boldsymbol{\epsilon}_k^i \tag{D.20}$$

$$\boldsymbol{P}_{k|k}^i = \boldsymbol{P}_{k|k-1}^i - \boldsymbol{K}_k^i\boldsymbol{R}_{k|k-1}^i(\boldsymbol{K}_k^i)^T \tag{D.21}$$

and the state *prediction* equations:

$$\frac{d\hat{\boldsymbol{x}}_{t|k}^i}{dt} = \boldsymbol{f}(\hat{\boldsymbol{x}}_{t|k}^i, \boldsymbol{u}_t^i, t, \boldsymbol{\theta}) \tag{D.22}$$

$$\frac{d\boldsymbol{P}_{t|k}^i}{dt} = \boldsymbol{A}\boldsymbol{P}_{t|k}^i + \boldsymbol{P}_{t|k}^i\boldsymbol{A}^T + \boldsymbol{\sigma}\boldsymbol{\sigma}^T \tag{D.23}$$

which are solved for $t \in [t_k^i, t_{k+1}^i[$. In the above equations the following notation has been applied:

$$\boldsymbol{A} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}_t}|_{\hat{\boldsymbol{x}}_{k|k-1}^i, \boldsymbol{u}_k^i, t_k^i} \;, \; \boldsymbol{C} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}_t}|_{\hat{\boldsymbol{x}}_{k|k-1}^i, \boldsymbol{u}_k^i, t_k^i}$$
$$\boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{u}_k^i, t_k^i, \boldsymbol{\theta}) \;, \; \boldsymbol{S} = \boldsymbol{S}(\boldsymbol{u}_k^i, t_k^i, \boldsymbol{\theta})$$

Initial conditions for the EKF are $\hat{\boldsymbol{x}}^i_{t|t_0} = \boldsymbol{x}^i_0$ and $\boldsymbol{P}^i_{t|t_0} = \boldsymbol{P}^i_0$, which can either be prespecified or estimated as a part of the overall problem.

The prediction equations (D.22)-(D.23) can be solved in several ways. The most direct is to integrate the linearized system over the time interval $[t^i_k, t^i_{k+1}[$, but the EKF algorithm is sensitive to non-linear effects so this should only be done for linear or almost linear systems. For non-linear systems and especially stiff systems one should consider using a numerical ODE solution preferably with an intelligent update of $\boldsymbol{A}$. The covariance of the state, $\boldsymbol{P}$, is symmetric so the number of states in the ODE's to be solved are $n + n \cdot (n+1)/2$.

D.3.3 Uncertainty of parameter estimates

Any resonable statistical parameter estimation scheme should include an assessment of the uncertainty of the estimates. An estimate of the uncertainty of the parameter estimates can be obtained by using the fact that by the central limit theorem the estimator in (D.15) is asymptotically Gaussian with mean $\boldsymbol{\theta}$ and covariance matrix:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{H}^{-1} \tag{D.24}$$

where the matrix $\boldsymbol{H}$ is given by the elements:

$$h_{ij} = -E\left\{\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln\left(p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{y_0})\right)\right\}$$

for $i, j = 1, \ldots, p$. An approximation to $\boldsymbol{H}$ can be obtained from the elements:

$$h_{ij} \approx -\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln\left(p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{y_0})\right)\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

for $i, j = 1, \ldots, p$, which is the Hessian at the minimum of the objective function. A measure of the uncertainty of the individual estimates can be obtained by decomposing the covariance matrix:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{\sigma}_{\hat{\boldsymbol{\theta}}} \boldsymbol{R} \boldsymbol{\sigma}_{\hat{\boldsymbol{\theta}}} \tag{D.25}$$

into $\boldsymbol{\sigma}_{\hat{\boldsymbol{\theta}}}$, which is a diagonal matrix of the standard deviations of the parameter estimates, and $\boldsymbol{R}$, which is the corresponding correlation matrix.

D.3.4 Model validation

The uncertainty information thus obtained can subsequently be applied to perform tests of various hypotheses, e.g. to determine the significance of the individual parameters through t-tests.

If the model is correct then the sequence of one-step prediction errors is a white noise sequence, this can be tested. Another test is the likelihood ratio test that can be used to compare two models where one is a subset of the other. It is also possible to test for model correctness, i.e. to test if one can assume that the model is the correct model (Bak et al., 1999).

D.3.5 Different estimation schemes

As mentioned above the described method is a MAP procedure that reduces to a ML procedure if no prior information about the parameters is available.

Above it has been assumed that the noise is Gaussian. If that assumption fails the method is no longer an exact ML procedure but a prediction error (PE) procedure (Kristensen et al., 2004). Using one-step prediction errors it is possible to test whether the noise is white and whether it is Gaussian (Box and Jenkins, 1976). On top there are tools that can be used to assist when building a new model og revising an exsisting model, e.g. adding a state for a parameter to see how it changes over time.

In previous studies parameters have often been estimated by minimizing the sum of squared errors of simulations of ODEs; this is a so called output error (OE) procedure as it minimizes the sum of squares of residuals of simulations. The advantage of the continuous-discrete stochastic state space models described in the present paper is that the framework provides tools for statistical tests of model assumptions, and these tests cannot be made using OE procedures. The conversion from discrete observed ODEs to continuous-discrete stochastic state space models is simple as it is a matter of multiplying by dt on both sides of the equations describing the dynamics and adding Wiener processes to the states.

D.3.6 Case data

The differences between OE and PE methods will be illustrated using data from an ongoing study of the influence of copper on the growth dynamics of fecal bacteria. The motivation for the study is that large amounts of copper is added to feed for production animals. The question is if *Salmonella* Typhimurium is more tolerant than E. coli.

Optical density (OD) measurements of bacterial growth in Müeller Hinton liquid broth (MH) was made using Microbiology Reader Bioscreen C that features automated measurement of growth in 200 wells in each run. The data used for this paper comes from 10 wells with dublicates of 5 concentrations of copper: 0.1 mM, 0.2 mM, 0.4 mM, 0.7 mM, and 1.0 mM. The medium consists of 1/3 MH and 2/3 cellular saltwater. The media was titrated to get pH=6.0 and autoclaved before use. $5\mu L$ McFarlan corresponding to 5e5 CFU, was inoculated into each well before starting the experiment. OD measurements were taken every 10 minutes during 24 hours resulting in 145 measurements per well.

Optical density measurements are not linearly related to the actual bacterial count as the probability of having bacteria in the shade of other bacteria increases with the number of bacteria. For low ODs one can assume linearity but instead measurements made on series of dilutions were used to estimate the non-linearity. Based on the assumption that all bacteria blocks the same area the following parametric expression:

$$\mathtt{OD} = a \cdot (1 - exp(-b \cdot \mathtt{CFU})) \tag{D.26}$$

is obtained. This expression was fitted and used to calculate linearized OD values, ie. values proportional to the CFU.

D.4 Results

There are many proposals for models of bacterial growth on a mixture of substrates (Lendenmann and Egli, 1998; Narang et al., 1997a; Venkatesh et al., 1997). These models are based on ordinary differential equations that easily can be converted/extended to stochastic differential equations by adding Wiener processes to some or all of the equations. In an other paper ([E]) some of the different models are compared within the scope of SDEs. In this paper the model described by Eq. 16 in Lendenmann and Egli (1998) with three substrates was selected for illustration. The growth medium in this study contains many different substrates but it is infeasible to estimate too many parameters so a model with 3 substrates was selected allowing for the system noise to describe the infeasibility due to the remaining substrates. The model is extended with a term describing the effect of increased copper in the media and it is given by the following equations:

$$dc = \left(c\sum_i \frac{\mu_i(s_i/K_i)}{1+\sum_j(s_j/K_j)}dt - \sigma_s\sum_i d\omega_{t,i}\right)\left(1-a\left(\frac{cu}{0.1}-1\right)\right) \quad \text{(D.27)}$$

$$ds_1 = -c\frac{\mu_1(s_1/K_1)}{1+\sum_j(s_j/K_j)}dt + \sigma_s d\omega_{t,1} \quad \text{(D.28)}$$

$$ds_2 = -c\frac{\mu_2(s_2/K_2)}{1+\sum_j(s_j/K_j)}dt + \sigma_s d\omega_{t,2} \quad \text{(D.29)}$$

$$ds_3 = -c\frac{\mu_3(s_3/K_3)}{1+\sum_j(s_j/K_j)}dt + \sigma_s d\omega_{t,3} \quad \text{(D.30)}$$

where c is the concentration of bacteria, subscript $i \in [1,2,3]$, s_i are the substrates, cu is the concentration of copper, μ_i and K_i are the Monod growth constants, σ_s is the standard deviation of the Wiener processes $d\omega_{t,i}$, and a is the dependency on cu.

In the measurement setup only the concentration of bacteria can be measured leading to the following discrete observation equation:

$$OD_k = c_k + e_k \quad \text{(D.31)}$$

where OD_k is the linearized OD (using Eq. D.26), c_k is the concentration of bacteria, and e_k is the measurement error, $e_k \in N(0,\sigma_e^2)$, all at time k.

The above is the equations for the PE method. To obtain an OE method one chooses $\sigma_s = 0$. In practice a very small value for σ_s is chosen to avoid differences due to an other integrator and optimization algorithm. The parameter estimates found by the described procedure is shown in Table D.1. For both estimations the initial concentration was set to 0.001 and K_3 was set to 0.1. The value of the objective function, the negative logarithm of the posterior probability density function, for the PE and OE estimates were 5671.7 and 3230.0, respectively.

When looking at the actual estimates one notice that the two methods lead to almost the same estimate for parameters, especially when the estimated standard deviations are taken into account. As expected the most significant differences is found in the noise estimates. σ_e is estimated almost ten times higher for the OE when compared with the PE. The PE estimated system noise is having almost the same standard deviation as the measurement noise, emphasizing that the allowance of system noise does make a difference. Figure D.1 shows the ten timeseries with the linearized OD measurements, colors are used to indicate the concentration of copper. It was chosen to include the simulations based on the OE estimated parameters in the figure because the PE based one step predictions would almost coincide with the data. Recalling that the PE log likelihood

Table D.1: The estimated parameter values using the PE and OE procedure with the respective standard deviations (s.d.) of the parameter estimates. A not applicable (N/A) s.d. means that the parameter was fixed.

Parameter	PE estimate	PE s.d.	OE estimate	OE s.d.
c_0	0.0010	N/A	0.0010	N/A
$s_{1,0}$	0.15985	1.4198E-02	0.17303	3.6194E-03
$s_{2,0}$	0.053067	1.0407E-02	0.28123	9.4832E-03
$s_{3,0}$	0.58589	6.0465E-02	0.53936	1.0337E-02
K_1	0.079974	1.3173E-02	0.087955	1.1433E-03
K_2	0.015460	4.9427E-04	0.078863	1.6041E-03
K_3	0.1	N/A	0.1	N/A
a	0.046535	3.6966E-03	0.046395	1.8005E-04
μ_1	11.996	1.6990E-02	11.474	2.5830E-01
μ_2	0.53846	9.0491E-02	0.45053	1.9638E-02
μ_3	0.089478	1.1607E-02	0.079137	1.8102E-03
$log(\sigma_s)$	-5.4670	6.0270E-02	-21.0	N/A
$log(\sigma_e^2)$	-11.733	6.6584E-02	-7.3240	3.7110E-02

is almost twice the OE log likelihood provides further insight to the fact that the PE method gives a better description than the OE method. The bacteria did not grow in a 2mM solution. It is seen that the model captures the separation due to copper quite well and over all the shape of the simulated growth curves is describing the observed data. Taking a closer look one sees that the simulated initial growth phase (0-5 hours) differs especially for the two highest concentrations of copper and there is also some dynamics after about 15 hours that isn't captured.

D.5 Discussion

The results show that the PE method gives a better description of the data and it provides tools for model validation and tools for investigating model deficiencies that can be used as guides when developing a new model from scratch, modifing an existing model and when comparing different models.

Around 16 to 18 hours after the beginning the OD measurements show an apparent increase in the measurement noise. A closer look at the data in Figure D.1 shows that this additional noise appears to be correlated between the wells. One

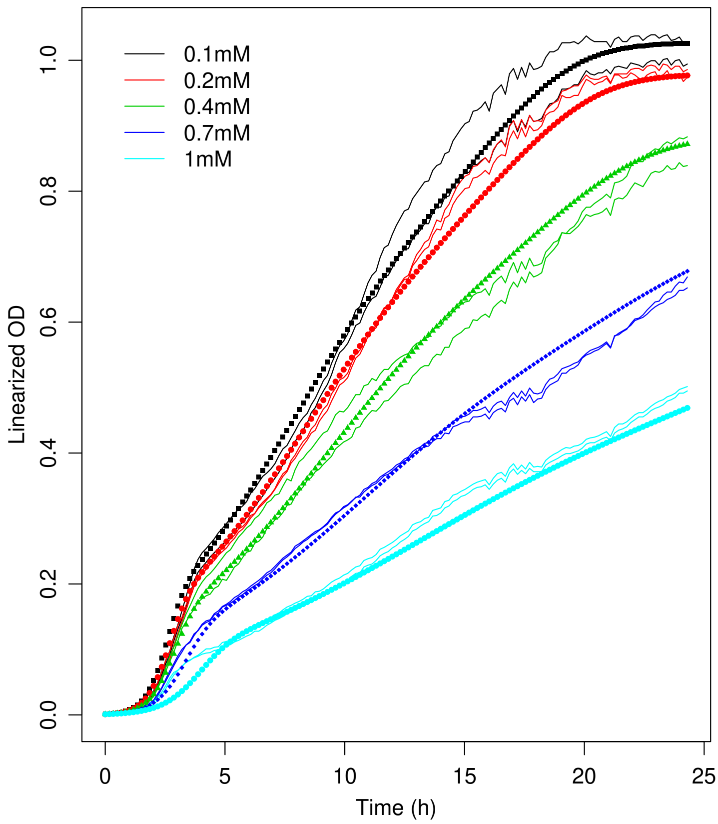

Figure D.1: The observed ten timeseries and simulations based on the OE estimated parameters.

explanation may be that the calibration of the OD measurement is somewhat unstable in that period. One way to try to adjust would be to include wells with different but fixed OD during the experiment and use these to correct the measurements.

The lack of fit after a bit less than 4 hours is certainly due to a model deficiency and this is one target for possible future improvements of the model. The fact that the estimated OE model fit the later parts (after 6 hours) of the experiments quite well indicates that the parameters related to this part of the growth curves are biased to correct for the model deficiencies in the first part. This is one reason why one should not expect the same parameter estimates with the two methods - at least not for all parameters.

The log likelihood obtained with the PE method will in practice always be higher than what can be optained with the OE method, though models with infinitesimal estimates of the system noise, i.e. a model that fully describe the data, will result in approximately the same log likelihood.

D.6 Conclusion

Discrete time observations of continuous time systems have been modelled using continuous-discrete stochastic state space models. Indeed the proposed MAP/PE framework provides several attractive features; in this paper the advantages of system noise have been presented including a more liable estimate of the measurement noise and less biased parameter estimates when compared with OE methods.

An estimation scheme has been presented including a free software implementation. The advantages of PE methods over OE methods has been illustrated with an example based on real data.

References

Bak, J., Nielsen, H. A., and Madsen, H. (1999). Goodness of fit of stochastic differential equations. In Holm, P. L. . A., editor, *21. Symposium i anvendt statistik*, Handelshøjskolen, København.

Box, G. and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden Day, San Fransisco.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press, New York, USA.

Kristensen, N., Madsen, H., and Jorgensen, S. (2004). Parameter estimation in stochastic grey-box models. *Automatica*, **40**(2), 225–237.

Lendenmann, U. and Egli, T. (1998). Kinetic models for the growth of *Escherichia coli* with mixtures of sugars under carbon-limited conditions. *Biotechnology and Bioengineering,*, **59**, 99–107.

Narang, A., Konopka, A., and Ramkrishna, D. (1997). The dynamics of microbial growth on mixtures of substrates in batch reactors. *Journal of Theoretical*

Biology, **184**, 301–17.

Nielsen, J. N., Madsen, H., and Young, P. C. (2000). Parameter estimation in stochastic differential equations: An overview. *Annual Reviews in Control*, **24**, 83–94.

Øksendal, B. (1989). *Stochastic Differential Equations - An introduction with Applications*. Springer Verlag, Berlin.

Venkatesh, K., Doshi, P., and Rengaswamy, R. (1997). Optimal strategy to model microbial growth in a multiple substrate environment. *Biotechnology and Bioengineering*, **56**(6), 635–44.

Appendix E

Modelling microbial growth in a rich medium and the effect of adding copper

Submitted to *J. of Theoretical Biology*.

Modelling microbial growth in a rich medium and the effect of adding copper

Lasse Engbo Christiansen[1,2,*], Jens Strodl Andersen[2], Henrik Caspar Wegener[2], Frank Møller Aarestrup[2], and Henrik Madsen[1]

1: Informatics and Mathematical Modelling, The Technical University of Denmark, Denmark.

2: Department of Epidemiology and Risk Assessment, Institute for Danish Food and Veterinary Research, Denmark.

Corresponding author: Lasse E. Christiansen, Building 321, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. E-mail: lec@imm.dtu.dk, Phone: +45 45253372, Fax: +45 45882673

E.1 Abstract

Bacterial growth in natural environments will often be on mixtures of substrates. In this work Müeller Hinton liquid broth was used as a rich growth medium for *Escherichia coli*. The parameters of five multi substrate models have been estimated using continuous-discrete stochastic state space models in the form of stochastic differential equations. Discrete time measurements of the optical density of batch cultures were used for the estimations.

Three models were selected for implementing additive as well as multiplicative dependencies on copper in the growth media resulting in six models. A linear dependence on the copper concentrations in the range $[0.1; 1]mM$ a was fitted for all models. The multiplicative implementation of the copper dependence resulted in the highest likelihood.

Keywords: Stochastic Differential Equations; maximum likelihood; mixed substrate growth; E. coli; copper

E.2 Introduction

Microbial growth on mixtures of substrates has been modelled in many ways for different purposes and under different settings (Kovárová-Kovar and Egli, 1998). *Escherichia coli* (*E. coli*) has been the bacteria of interest in many studies and the dynamic phase of bacterial growth is one of the most important factors for a large number of metabolic processes known to be involved in the growth of *E. coli*. Estimation of the parameteres in all those processes is often impossible due to correlation between parameters and lack of information in the estimation data, i.e. the data doesn't span the possible dynamics represented by the model. One approach is to separate into groups of sub processes and then later combine into a full model. This approach adds the uncertainties of the sub processes and interactions between sub processes may be excluded in the estimation. If the focus is to use the model for prediction the likelihood that the model will perform bad is relatively high. Thus most mathematical models are simplifications accounting only for the major processes (Gombert and Nielsen, 2000; Nielsen and Joergensen, 1996). Since Monod's model (Monod, 1942) of growth on a single substrate many other parameterisations have been made of single substrate growth. Many of the single substrate models can be solved analytically allowing for easy fitting with output error methods, i.e. minimization of squared simulation errors (Impe et al., 1995).

When modelling growth on two substrates the bacteria can either grow sequentially or simultaneously on the substrates. One assumption is that bacteria always will try to optimize their growth given the surrounding substrates and other compounds. This implies that presence of large concentrations of one optimal substrate will lead to sequential growth whereas low concentrations of substrates or combinations of less optimal substrates lead to simultaneous growth (Lendenmann and Egli, 1995).

Models fulfilling the above assumptions include degration of di-saccharides to mono-saccharides, serving as a common intermediate, when the level of the latter is low (Paul et al., 1998). Other models use simplified enzyme kinetics to control the utilization of the available substrates, e.g. combined with linear optimization (Venkatesh et al., 1997) or an intermediate precursor (Narang et al., 1997b). These models try to describe some of the dynamics involved in growth. Even simpler models have been made only including sums of modified Monod terms that reduce to the Monod equations in the case of only one substrate (Lendenmann and Egli, 1998). In all cases the growth is modeled using one or more Monod or Monod like terms with some activation mechanism.

Many metals are required for microbial growth, but even at relatively low concentrations highly toxic effects are reported (Nies, 1999). One of the required metals is copper and a large effort has been put into understanding the mechanisms involed in copper trafficking in bacteria (Banci et al., 2003; Dameron and Harrison, 1998; Jun et al., 2003; Rosen, 2002). A lot of this work has been carried out on *E. coli* (Grass and Rensing, 2001; Gupta et al., 1997; Petersen and Møller, 2000; Rensing and Grass, 2003). Only few studies are concerned with the changes in growth kinetics when exposed to copper (Cabrero and Mirada, 1998). Copper in various concentrations are frequently present where bacterial communities grow and a better understanding of how even low concentrations affect growth rates help us to understand the bacterial ecology better.

When running experiments to measure growth dynamics there are basically two ways: batch and continuous cultures. Using batch cultures information can be acquired at all growth stages, however many models focus only on the exponential phase including the models in this paper. Continuous cultures, on the other hand, give good estimates of growth rates in stable equilibriums. These can be thought of as estimates of the growth rate in points of a growth curve but this is not generally true when multiple substrates are available due to the history of the batch culture and the fact that there is no stable equilibrium.

The present study focuses on modelling growth kinetics of *E. coli* and how the effect of adding copper to the growth medium affects growth.

E.3 Methods

E.3.1 Growth measurements

Bacterial growth was measured using optical density (OD) measurements in a systematic setting by using a Microbiology Reader Bioscreen C allowing for measuring up to 200 wells in each batch. OD measurements were taken at $492nm$ every 10 minutes during 24 hours resulting in 145 measurements per well.

The growth media was 1/3 Müeller Hinton liquid broth (MH) and 2/3 cellular saltwater (0.9%). Five different concentrations of copper (0.1 mM, 0.2 mM, 0.4 mM, 0.7 mM, and 1.0 mM) were made by adding $CuSO_4$. Afterwards each of the five media was titrated to pH=6.0 and autoclaved before use. For each of the five concentrations two wells were prepared with $250\mu L$ broth and $5\mu L$ McFarlan which corresponds to 5e5 CFU were inoculated into each well before starting the experiment. Pilot experiments showed that adding Na_2SO_4 to concentrations of 1% did not inhibit bacterial growth (Data not shown).

Optical density measurements are not linearly related to the actual bacterial count as the probability of having bacteria in the shade of other bacteria increases with the number of bacteria. For low ODs one can assume linearity, but to allow higher OD measurements it was chosen to make series of dilutions to estimate the nonlinearity. A parametric expression for the relationship is suggested: Assuming that all bacteria covers one unit area then the next bacteria will have a probability not to overlap with the bacteria already present. This is proportional to the measured OD and it can be shown that: $\mathtt{OD} = a \cdot (1 - exp(-b \cdot \mathtt{CFU}))$. This was fitted and used to calculate linearized OD values, ie. values proportional to the CFU.

E.3.2 Modelling growth

The OD measurements are discrete in time but the batch cultures are continuous in time. It is known that the candidate models for describing bacterial growth are approximations hence system noise is introduced into the dynamical description in order to describe all the variations in the data. This leads to the class of continuous-discrete stochastic state space models that also facilitates a separation of system noise and measurement noise.

In the general case, the continuous-discrete stochastic state space model is a model that consists of a set of nonlinear discretely, partially observed Itô SDE's (Øksendal, 1989) with measurement noise, i.e.:

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta})dt + \boldsymbol{\sigma} d\boldsymbol{\omega}_t \tag{E.1}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{E.2}$$

where $t \in \mathbb{R}$ is the time variable; $\boldsymbol{x}_t \in \mathcal{X} \subset \mathbb{R}^n$ is a vector of state variables; $\boldsymbol{u}_t \in \mathcal{U} \subset \mathbb{R}^m$ is a vector of input variables; $\boldsymbol{y}_k \in \mathcal{Y} \subset \mathbb{R}^l$ is a vector of output variables; $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ is a vector of (possibly unknown) parameters; $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\boldsymbol{\theta}) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ are known nonlinear functions; $\{\boldsymbol{\omega}_t\}$ is an n-dimensional standard Wiener process and $\{\boldsymbol{e}_k\}$ is an l-dimensional white noise process with $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{\theta}))$. The diffusion term in (E.1) is assumed to be independent of the state variables, as this makes parameter estimation more feasible. CTSM 2.3 (www.imm.dtu.dk/ctsm) was used to perform maximum likelihood (ML) estimates of the parameters in this class of models. The method can be seen in Christiansen et al.[D], for a more thorough description see the documentation that comes with the software.

Five candidate models for growth on a rich substrate are presented. They represent different approaches to the description of growth. All models have states for the concentration of bacteria, c, and three different substrates: s_1, s_2, s_3. One unit of substrate is assumed converted to one unit bacteria. The first and third model have additional states handling the activation of growth on the substrates, one non-parametric and one based on enzyme dynamics. The last two candidates handle the activation of growth on the different substrates by normalizations of the Monod like terms. In all models the parameters V_i and K_i are the Monod parameters representing the maximum growth rate and the concentration of the substrates at half maximum growth rate, respectively. In all models system noise have been included by adding Wiener processes (Øksendal, 1989). The noise added to the substrate equations has been subtracted from the concentration of bacteria to maintain the mass balance.

The first model was presented by Venkatesh et al. (1997). In the original formulation the activation is handled by algebraic equations. This would be possible too in the selected framework, but in order to test whether there is near instantaneous adjustment or a delay in the adjustment of the activations the last two equations were added:

$$
\begin{aligned}
dc &= c\left(\sum_{i=1}^{3}\frac{V_i s_i}{K_i+s_i}\right)dt+\sigma_c d\omega_{t,0}-\sigma_s\sum_{i=1}^{3}d\omega_{t,i} && \text{(E.3)}\\
ds_1 &= -c\frac{V_1 s_1}{K_1+s_1}dt+\sigma_s d\omega_{t,1} && \text{(E.4)}\\
ds_2 &= -c\frac{V_2 s_2}{K_2+s_2}a_2 dt+\sigma_s d\omega_{t,2} && \text{(E.5)}\\
ds_3 &= -c\frac{V_3 s_3}{K_3+s_3}a_3 dt+\sigma_s d\omega_{t,3} && \text{(E.6)}\\
da_2 &= 4(1-a_2)a_2/t_2\left(V_1-\frac{V_1 s_1}{K_1+s_1}\right)dt+\sigma_a d\omega_{t,4} && \text{(E.7)}\\
da_3 &= 4(1-a_3)a_3/t_3\left(V_1-\frac{V_1 s_1}{K_1+s_1}-\frac{V_2 s_2}{K_2+s_2}a_2\right)dt+\sigma_a d\omega_{t,5} && \text{(E.8)}
\end{aligned}
$$

Where a_2 and a_3 are the states describing activation of utilization of the second and third substrates; t_2 and t_3 are time constants for the delay in usage of the last two substrates. The Wiener process $\sigma_c d\omega_{t,0}$ is included to improve the numerical stability and is fixed to have a very small variance.

If the two time constants t_2 and t_3 are estimated to be small, a second model which is a simplified version of the Venkatesh model will be included. The simplification results in the original Venkatesh model is made by removing the equations for da_2 and da_3 and replacing a_2 with $(1-s_1/(K_1-s_1))$ and a_3 with $(1-s_1/(K_1+s_1)-s_2/(K_2+s_2))$ in Eq.E.5 and E.6, respectively.

The third model is a reduced version of the model presented by Narang et al. (1997a). The additional states are a pool of biosynthetic precursors, p, and enzymes for the different substrates, e_i. The original model also included equations for inducers, x_i, for the enzymes. The full model contains many parameters and it is infeasible, if at all possible, to estimate these based solely on OD measurements of growth curves. As the dynamics for x_i are assumed to be relatively fast. A reduction of the model is obtained by assuming that x_i are quasi-stationary, i.e. assuming $dx_i/dt = 0$ almost all the time. Solving the equations for x_i and inserting them into the other equations gives the proposed model. Besides process noise on the substrates Wiener processes have been added to the precursor, again securing the mass balance, and to the enzyme levels. The resulting model is:

$$
\begin{aligned}
dc &= \frac{V_g p}{K_g + p} c dt - \sigma_p d\omega_{t,1} && \text{(E.9)}\\
dp &= \left(\sum_{i=1}^{3} Y_i \frac{V_i e_i s_i}{K_i + s_i} - (1+p) \frac{V_g p}{K_g + p} \right) dt + \sigma_p d\omega_{t,1} - \sigma_s \sum_{i=2}^{4} d\omega_{t,i} && \text{(E.10)}\\
ds_1 &= \frac{V_1 e_1 s_1}{K_1 + s_1} c dt + \sigma_s d\omega_{t,2} && \text{(E.11)}\\
ds_2 &= \frac{V_2 e_2 s_2}{K_1 + s_2} c dt + \sigma_s d\omega_{t,3} && \text{(E.12)}\\
ds_3 &= \frac{V_3 e_3 s_3}{K_1 + s_3} c dt + \sigma_s d\omega_{t,4} && \text{(E.13)}\\
de_1 &= \frac{V_{e1} e_1 s_1 / (K_1 + s_1)}{K_{e1} + e_1 s_1 / (K_1 + s_1)} + k_a - k_d e_1 - \frac{V_g p}{K_g + p} e_1 && \text{(E.14)}\\
de_2 &= \frac{V_{e2} e_2 s_2 / (K_2 + s_2)}{K_{e2} + e_2 s_2 / (K_2 + s_2)} + k_a - k_d e_2 - \frac{V_g p}{K_g + p} e_2 && \text{(E.15)}\\
de_3 &= \frac{V_{e3} e_3 s_3 / (K_3 + s_3)}{K_{e3} + e_3 s_3 / (K_3 + s_3)} + k_a - k_d e_3 - \frac{V_g p}{K_g + p} e_3 && \text{(E.16)}
\end{aligned}
$$

The fourth and fifth model are simpler than the previous as they use a normalization to control the usage of the different substrates instead of using additional equations. They are both suggested by Lendenmann and Egli (1998). The fourth model is described by Eq. 16 in the paper and $1/K_i$ are used for normalization. This model will be labeled Lendenmann 1. Note that in the case of one substrate the model reduces to the ordinary Monod equation. The motivation for using K_i is that this constant gives the substrate concentration where the growth rate is 50% of μ_i and hence it is a reasonable scale for the substrate concentration. The fourth model is formulated as:

$$
\begin{aligned}
dc &= c \frac{\sum_{i=1}^{3} V_i s_i / K_i}{1 + \sum_{i=1}^{3} s_i / K_i} dt - \sigma_s \sum_{i=1}^{3} d\omega_{t,i} && \text{(E.17)}\\
ds_1 &= -c \frac{V_1 s_1 / K_1}{1 + \sum_{i=1}^{3} s_i / K_i} dt + \sigma_s d\omega_{t,1} && \text{(E.18)}\\
ds_2 &= -c \frac{V_2 s_2 / K_2}{1 + \sum_{i=1}^{3} s_i / K_i} dt + \sigma_s d\omega_{t,2} && \text{(E.19)}\\
ds_3 &= -c \frac{V_3 s_3 / K_3}{1 + \sum_{i=1}^{3} s_i / K_i} dt + \sigma_s d\omega_{t,3} && \text{(E.20)}
\end{aligned}
$$

The fifth model is described by Eq. 17 in the paper by Lendenmann et al..

Here μ_i/K_i are used as the normalization constants. This model will be labeled Lendenmann 2. The motivation for this is parallel to the specific affinity in enzyme kinetics defined as ν_{max}/K_m.

$$dc = \frac{V_1 \sum_{i=1}^{3}(V_i/K_i)s_i}{V_1 + \sum_{i=1}^{3}(V_i/K_i)s_i} dt + \sigma_c d\omega_{t,0} - \sigma_s \sum_{i=1}^{3} d\omega_{t,i} \tag{E.21}$$

$$ds_1 = \frac{V_1(V_1/K_1)s_1}{V_1 + \sum_{i=1}^{3}(V_i/K_i)s_i} dt + \sigma_s d\omega_{t,1} \tag{E.22}$$

$$ds_2 = \frac{V_1(V_2/K_2)s_2}{V_1 + \sum_{i=1}^{3}(V_i/K_i)s_i} dt + \sigma_s d\omega_{t,2} \tag{E.23}$$

$$ds_3 = \frac{V_1(V_3/K_3)s_3}{V_1 + \sum_{i=1}^{3}(V_i/K_i)s_i} dt + \sigma_s d\omega_{t,3} \tag{E.24}$$

At the estimation V_i/K_i have been estimated as a single parameter. $\sigma_c d\omega_{t,0}$ is a Wiener process with a very small variance, which is introduced only for numerical stability.

E.3.3 Modelling dependence on Cu

A multiplicative and an additive dependence on copper is proposed. The multiplicative models are obtained by multiplying the whole right hand side of Eq. E.3, E.9, E.17, and E.21 by $(1 - a(cu - 0.1))$, where a is to be estimated using the ten time series with the five different concentrations of copper. It was chosen to subtract "0.1" from the copper concentration to make a linearization around the lowest concentration of copper, which is the same that is used for the five candidate models. The motivation for the multiplicative copper model is that efflux of copper consumes energy that can not be used for growth implying that each cell division takes more substrate when the concentration of copper is higher.

The additive models are made by subtracting $c(a \cdot cu + b)dt$ from the right hand side of Eq. E.3, E.9, E.17, and E.21. Where b is the basic maintenance and a is the copper dependence of the maintainance. Again this should be thought of as a linearization of the copper dependence in the range $[0.1; 1.0]mM$. The motivation for the additive model is the assumption that there is an increased maintenance when the concentration of copper is increased. Again this is due to increased energy used for efflux when the concentration increases and since the influx of copper is believed to be diffusive (Rensing and Grass, 2003).

E.4 Results

The parameters of the five candidate models for growth on a rich substrate were estimated using CTSM 2.3. Three of the models are selected for implementation of copper dependence. The second subsection focuses on the two different strategies for describing the copper dependence.

E.4.1 Selecting models

To find suitable candidates for modelling the Cu dependence the two time series with 0.1mM Cu were used. The parameters of all five models were estimated using CTSM 2.3. For the models with many parameters some of the parameters were fixed because of correlation with other parameters. This was especially the case with the Narang model as the parameters related to the enzyme kinetics are quite well hidden when only the concentration of bacteria is measured.

The parameters for the delay in usage of the second and third substrates, t_2 and t_3 were estimated to be small. Hence the parameters of the Simplified Venkatesh (Simplified V.) model were estimated. The Simplified V. model also had problems with very high correlations between the parameter estimates. There are several ways to come around this problem: One would be to measure more variables another would be to excite the system eg. by adding more substrate during the experiment - not a constant flow but for instant on/off in a random fashion. Both options require more complicated setups making them more time consuming. On top it may not be possible to run as many experiments at a time.

Table E.1: Summary of the estimation results from the five candidate models.

Model	No. parameters fixed	No. parameters estimated	Log likelihood
Venkatesh	9	9	1064
Simplified V.	3	10	1081
Narang	20	13	440
Lendenmann 1	2	10	1083
Lendenmann 2	2	9	1075

A summary of the estimation results can be seen in Table E.1. Based on the log likelihood the Narang model is found to be inferior to the other four models. Further, it is also the model with the most parameters which is a further

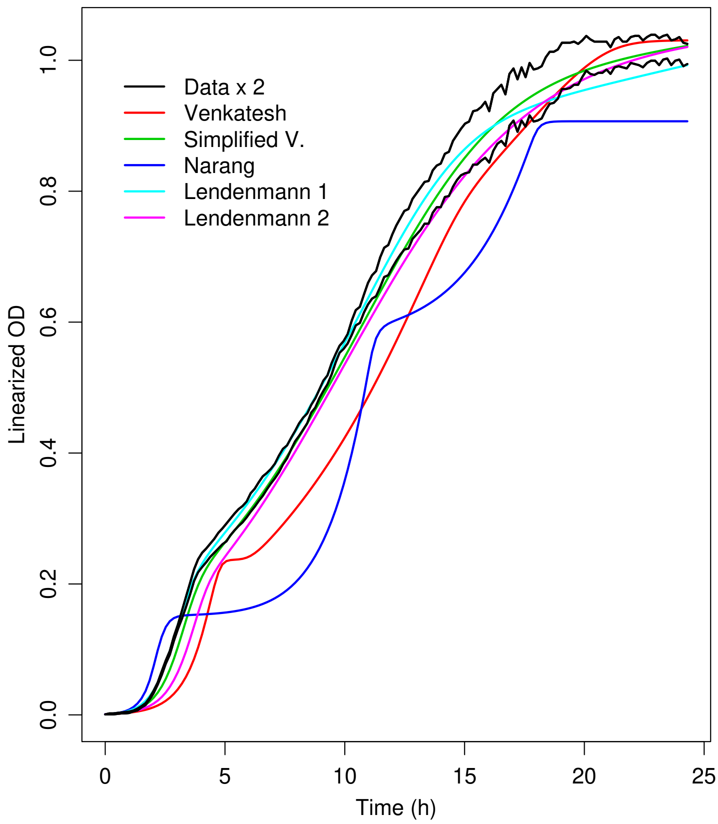

Figure E.1: The two time series with 0.1 mM Cu and simulations based on the the estimated parameters of the five models.

disadvantage due to the increased complexity of the estimation task. The time constants of the Venkatesh model, t_2 and t_3, where fixed at a small value in the final run because they where estimated to be very small leading to a very stiff set of equations causing numerical problems even though the BDF integration scheme for stiff systems was applied (Hindmarsh, 1983).

Pure simulations of the five models using the estimated parameters can be seen in Fig. E.1. It is seen that the Simplified V. model and both of the models due to Lendenmann resembles the variation seen in the data. In particular the Lendenmann 1 model seems to capture most of the dynamics. Both the Venkatesh model and the Narang model make too sharp a break when the first substrate is depleted after about four hours.

Table E.2: Summary of the estimation results from the six models used to describe the copper dependence.

Model	Cu dependence	Log likelihood	Highest corr.	Est. $\log(\sigma_s)$	Est. $\log(\sigma_{obs.})$
Simplified V.	add.	5490	0.86	-4.92	-7.00
Simplified V.	mul.	5600	0.999	-4.81	-6.30
Lendenmann 1	add.	5582	0.95	-5.55	-5.85
Lendenmann 1	mul	5633	0.78	-5.01	-6.00
Lendenmann 2	add.	5551	0.86	-5.17	-6.03
Lendenmann 2	mul.	5640	0.91	-4.99	-6.13
Lendenmann 1	mul. + add.	5628	0.96	-5.02	-6.00
Lendenmann 2	mul. + add.	5641	0.95	-5.00	-6.12

Based on these observations it was chosen only to continue with the Simplified Venkatesh and the two models due to Lendenmann.

E.4.2 Modelling dependence on Cu

The multiplicative and the additive dependence on copper were introduced for each of the three models. This results in estimation of parameters in six models. The number of parameters to be estimated is one additional parameter for the multiplicative models and two additional parameters for the additive models. The number of fixed parameters is the same as in the previous section. A summary of the estimations can be seen in Table E.2. The log likelihoods of all the multiplicative models are higher than all the additive models. The Simplified V. model caused problems with high correlations between the parameter estimates. Attempts were made to make the optimization algorithm select another path resulting in a lower correlation but those were unsuccessful.

When looking at the estimated noise parameters σ_s and σ_{obs} it is notable that the additive models have lower estimates of the system noise than the multiplicative models. Note also that the Simplified V. models both have higher system noise than the other models, and that the system noise is estimated to be lower for the Lendenmann 1 model.

Pure simulations were made based on the Lendenmann 1 model to illustrate the differences between the multiplicative and additive models. The simulations can be seen in Fig. E.2. It is seen that the multiplicative model seems to give the best overall description. During the initial 4 hours none of the models seem to be able

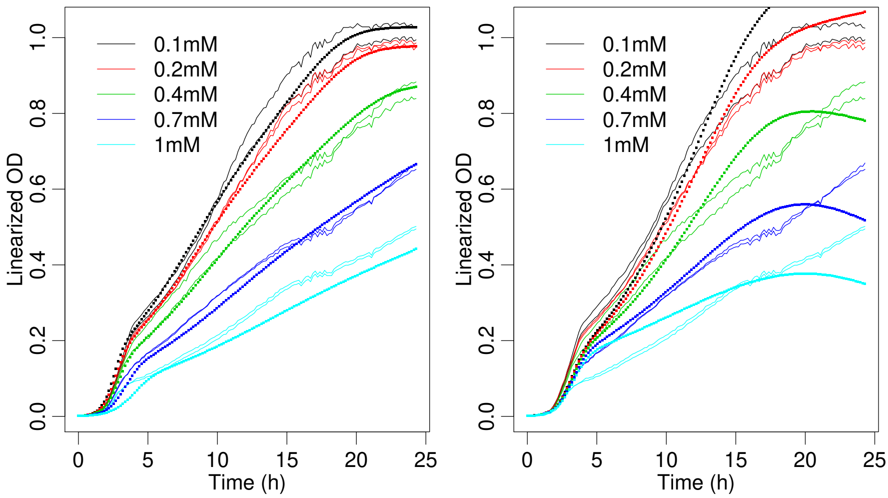

Figure E.2: The ten time series with 0.1 mM Cu and simulations based on the estimated parameters of the multiplicative (left part) and additive (right part) version of the first Lendenmann model.

to reproduce the copper dependence - the multiplicative model splits too early and the additive model doesn't split until the first substrate is completely used. This indicates that a combination of the models can be a solution. From looking at the simulation with the additive model it is clear that the copper dependence is estimated to be too strong as the simulation shows a death process that is not seen in the data. It should be noted that a similar death proces has been observed in other datasets than the one used in this paper.

Based on these observations it seems as if a combination of the additive and multiplicative model could be an advancement. Therefore, the multiplicative versions of the Lendenmann models have been extended by subtracting $c(a_s \cdot cu + b_s)$ from the equations for the concentration of bacteria, c. As can be seen in Table E.2 the inclusion of two additional parameters did not increase the log likelihoods when compared with the multipicative models. On top the parameter estimates are more correlated and there is no significant decrease in the system noise meaning that the combined models doesn't improve the description. In fact both estimates of a_s are approximately zero meaning that it is only the multiplicative part that is used to describe the copper dependence.

E.5 Discussion

The five candidate models represent different levels of complexity both when considering the number of states and the number of parameters to be estimated. The more complicated models includes representations of several layers of intracellular mechanisms. In the present setup the models are only partly observed as only the concentration of bacteria is measured. The reduced Narang model used in this paper contains 4 levels, which is one less than the original model, and as only the top level is observed and the system is unperturbed there is only little information available for estimating such complicated models. The three simple models that were selected for modelling the copper dependence only includes 2 layers explicitly and the assumption that one unit of substrate is converted to one unit of bacteria results in models with relatively few parameters.

The multiplicative version of the models appear to be better at describing the copper dependence than the additive models. The same answer was obtained from combining the two copper dependencies in one model. Still the multiplicative model is not perfect as it doesn't reproduce the copper dependence in the first growth phase (the first 4 hours). In the first growth phase the growth rate is the same for all concentrations of copper but it seems as if the substrate is used faster when the concentration of copper is high. The multiplicative models show a growth rate proportional to the copper concentration. This observation could be the basis for development of a new model featuring growth at the same rate for a range of copper concentrations in a way so that the consumption of substrate is dependent on the copper concentration. One idea inspired by the Narang model is to include a state for a precursor in a way where the first substrate can be converted faster to the precursor than the precursor can be used for growth.

In all models the measurement noise were estimated to be small. The Bioscreen C has a resolution of 0.001 and the standard deviations of the measurement noise was estimated to be in the range $[0.0009; 0.003]$. Thus it seems as if the Bioscreen C is quite accurate when compared the resolution of the measurements. This is an indication that the applied methodology was able to separate the system noise and the measurement noise.

E.6 Conclusion

Five models have been modeled using continuous-discrete stochastic state space models. The estimation of the parameters in the models was made based on discrete observations of the optical density of batch cultures. The growth medium was a rich medium and from simulations it was seen that 3 substrates was enough to capture the observed dynamics.

Two different copper dependencies were implemented in three of the models and they showed that the multiplicative approach was giving a higher likelihood than the additive approach. There is still room for improvements of the model and a suggestion has been proposed.

Acknowledgments

We acknowledge Inge M. Hansen from the Danish Institute for Food and Veterinary Research for doing the lab work and FELFO for financial support.

References

Gombert, A. K. and Nielsen, J. (2000). Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, **11**, 180–6.

Grass, G. and Rensing, C. (2001). Genes involved in copper homeostasis in *Escherichia coli*. *Journal of Bacteriology,*, **183**(6), 2145–2147.

Gupta, S. D., Wu, H. C., and Rick, P. D. (1997). A *Salmonella* typhimurium genetic locus which confers copper tolerance on copper-sensitive mutants of *Escherichia coli*. *Journal of Bacteriology,*, **179**(6), 4977–4984.

Hindmarsh, A. C. (1983). ODEPACK, a systematized collection of ODE solvers. In Stepleman, R. S., editor, *Scientific Computing (IMACS Transactions on Scientific Computation, Vol. 1)*, pp. 55–64. North-Holland, Amsterdam.

Impe, J. F. V., Nicolai, B. M., Schellekens, M., Martens, T., and Baerdemaeker, J. D. (1995). Predictive microbiology in a dynamic environment: a system theory approach. *International Journal of Food Microbiology*, **25**, 227–49.

Jun, Y., Yi, L., Yong, T., Jianben, L., Xiong, C., Qin, Z., Jiaxin, D., Songsheng, Q., and Ziniu, Y. (2003). Action of cu/sup 2+/ on bacillus thuringiensis growth investigated by microcalorimetry. *Applied Biochemistry and Microbiology*, **39**(6), 576–580.

Kovárová-Kovar, K. and Egli, T. (1998). Growth kinetics of suspended microbial cells: From single-substrate-controlled growth to mixed-substrate kinetics. *Microbiology and Molecular Biology Reviews*, **62**(3), 646–66.

Lendenmann, U. and Egli, T. (1995). Is *Escherichia coli* growing in glucose-limited chemostat culture able to utilize other sugars without lag? *Microbiology - Reading*, **141**(1), 71–78.

Lendenmann, U. and Egli, T. (1998). Kinetic models for the growth of *Escherichia coli* with mixtures of sugars under carbon-limited conditions. *Biotechnology and Bioengineering,*, **59**, 99–107.

Monod, J. (1942). *Recherches sur la croissance des cultures bactériennes*. Hermann et Cie, Paris, France.

Narang, A., Konopka, A., and Ramkrishna, D. (1997a). The dynamics of microbial growth on mixtures of substrates in batch reactors. *Journal of Theoretical Biology*, **184**, 301–17.

Narang, A., Konopka, A., and Ramkrishna, D. (1997b). New patterns of mixed-substrate utilization during batch growth of *Escherichia coli* K12. *Biotechnology and Bioengineering,*, **55**(5), 747–57.

Nielsen, J. and Joergensen, H. (1996). A kinetic model for the penicillin biosynthetic pathway in penicillium chrysogenum. *Control Engineering Practice*, **4**(6), 765–771.

Nies, D. H. (1999). Microbial heavy-metal resistance. *Appl Microbiol Biotechnol*, **51**, 730–750.

Paul, G., Syddall, M., Kent, C., and Thomas, C. (1998). A structured model for penicillin production on mixed substrates. *Biochemical Engineering Journal*, **2**(1), 11–21.

Petersen, C. and Møller, L. B. (2000). Control of copper homeostasis in escherichia coli by a p-type atpase, copa, and a merr-like transcriptional activator, copr. *Gene*, **261**, 289–298.

Rensing, C. and Grass, G. (2003). *Escherichia coli* mechanisms of copper homeostasis in a changing environment. *FEMS Microbiology Reviews*, **27**, 197–213.

Rosen, P. B. (2002). Transport and detoxification systems for transition metals, heavy metals and metalloids in eukaryotic and prokaryotic microbes. *Compar-*

ative Biochemistry and Physiology, **133**, 689–693.

Venkatesh, K., Doshi, P., and Rengaswamy, R. (1997). Optimal strategy to model microbial growth in a multiple substrate environment. *Biotechnology and Bioengineering*, **56**(6), 635–44.

Øksendal, B. (1989). *Stochastic Differential Equations - An introduction with Applications*. Springer Verlag, Berlin.

Bibliography

Anonymous (2000). The increasing incidence of human campylobacteriosis. Technical report, Report and proceedings of a WHO Consultation of Experts. Copenhagen, Denmark 21-25 November 2000, empty.

Anonymous (2001). Annual report on zoonosis in Denmark 2001. Technical report, Danish Zoonosis Centre, Copenhagen.

Bak, J., Nielsen, H. A., and Madsen, H. (1999). Goodness of fit of stochastic differential equations. In Holm, P. L. . A., editor, *21. Symposium i anvendt statistik*, Handelshøjskolen, København.

Banci, L., Bertini, I., Ciofi-Baffoni, S., Conte, R. D., and Gonnelli, L. (2003). Understanding copper trafficking in bacteria: Interaction between the copper transport protein copz and the n-terminal domain of the copper atpase copa from bacillus subtilis. *Biochemistry*, **42**, 1939–1949.

Berndtson, E., Emanuelson, U., Engvall, A., and Danielsson-Tham, M. (1996). A 1-year study of Campylobacters in 18 Swedish chicken farms. *Preventive Veterinary Medicine*, **26**, 167–85.

Boscoe, F. P., McLaughlin, C., Schymura, M. J., and Kielb, C. L. (2003). Visualization of the spatial scan statistic using nested circles. *Health and Place*, **9**, 273–277.

Box, G. and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden Day, San Fransisco.

Brennhovd, O., Kapperud, G., and Langeland, G. (1992). Survey of thermotolerant *Campylobacter* spp. and *yersinia* spp. in three surface water sources in Norway. *International Journal of Food Microbiology*, **15**, 327–338.

Cabrero, A. and Mirada, F. (1998). Effects of copper and zinc on the activated sludge bacteria growth kinetics. *Water Research*, **32**(5), 1355–1362.

Cabrita, J., Rodrigues, J., Braganca, F., Morgado, C., and Goncalves, A. P. (1992). Prevalence, biotypes, plasmid profile and antimicrobial resistance of *Campylobacter* isolated from wild and domestic animals from northeast portugal. *Journal of Applied Bacteriology*, **73**, 279–285.

Chan, K. F., Tran, H. L., Kanenaka, R. Y., and Kathariou, S. (2001). Survival of clinical and poultry-derived isolates of *campylobacter jejuni* at a low temperature (4°c). *Applied Environmental Microbiology*, **67**, 4186–4191.

Checkley, W., Epstein, L. D., Gilman, R. H., Figueroa, D., Cama, R. I., Patz, J. A., and Black, R. E. (2000). Effects of El Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children. *The Lancet*, **355**, 424–450.

Dackowska-Kozon, E. and Brzostek-Nowakowska, J. (2001). *Campylobacter* spp. in waters of three main western Pomerania water bodies. *Int. J. Hyg. Environ. Health*, **203**, 435–443.

Dameron, C. T. and Harrison, M. D. (1998). Mechanisms for protection against copper toxicity. *American Journal of Clinical Nutrition*, **67**, 1091S–1097S.

Diggle, P. J. and Tawn, J. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied statistics)*, **47**(3), 299–350.

Doyle, M. P. (1984). Association of *campylobacter jejuni* with laying hens and eggs. *Applied Environmental Microbiology*, **47**, 533–536.

Eberhart-Phillips, J., Walker, N., Garrett, N., Bell, D., Sinclair, D., Rainger, W., and Bates, M. (1997). Campylobacteriosis in New Zealand: results of a case-control study. *J. Epidemiol. Community Health*, **51**, 686–691.

Enemark, H. L., Ahrens, P., Juel, C. D., Petersen, E., Petersen, R. F., Andersen, J. S., Lind, P., and Thamsborg, S. M. (2002). Molecular characterization of danish cryptosporidium parvum isolates. *Parasitology*, **125**, 331–341.

Engberg, J., Gerner-Smidt, P., Scheutz, F., Moller-Nielsen, E., On, S. L., and Molbak, K. (1998). Water-borne *campylobacter jejuni* infection in a Danish town—a 6-week continuous source outbreak. *Clin. Microbiol. Infect.*, **4**, 648–656.

Friedman, C. R., Neimann, J., Wegener, H. C., and Tauxe, R. V. (2000). Epidemiology of *campylobacter jejuni* infections in the United States and other industrialized nations. In Nachamkin, I. and Blaser, M., editors, Campylobacter *, 2nd ed.* American Society for Microbiology, Washington, D.C.

Gangnon, R. E. and Clayton, M. K. (2001). A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine*, **20**, 2977–2987.

Glunder, G., Neumann, U., and Braune., S. (1992). Occurrence of *Campylobacter* spp. in young gulls, duration of *Campylobacter* infection and reinfection by contact. *Zentralbl Veterinarmed [B]*, **39**, 119–122.

Gombert, A. K. and Nielsen, J. (2000). Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, **11**, 180–6.

Grass, G. and Rensing, C. (2001). Genes involved in copper homeostasis in Escherichia coli. *Journal of Bacteriology,*, **183**(6), 2145–2147.

Gupta, S. D., Wu, H. C., and Rick, P. D. (1997). A salmonella typhimurium genetic locus which confers copper tolerance on copper-sensitive mutants of escherichia coli. *Journal of Bacteriology,*, **179**(6), 4977–4984.

Hald, B. and Madsen, M. (1997). Healthy puppies and kittens as carriers of *Campylobacter* spp. with special reference to *campylobacter upsaliensis*. *Journal of Clinical Microbiology*, **35**, 3351–3352.

Hald, B., Skovgård, H., Bang, D. D., Pedersen, K., Dybdahl, J., Jespersen, J. B., and Madsen, M. (2004). Flies and *campylobacter* infection of broiler flocks. *Emerging Infectious Diseases*.

Hald, T. and Andersen, J. S. (2001). Trends and seasonal variations in the occurrence of Salmonella in pigs, pork and humans in Denmark, 1995-2000. *Berl. Münich Tierärztl Wschr.*, **114**, 346–349.

Hastie, T., Tibshirani, R., and Friedman., J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, Berlin.

Hindmarsh, A. C. (1983). ODEPACK, a systematized collection of ODE solvers. In Stepleman, R. S., editor, *Scientific Computing (IMACS Transactions on Scientific Computation, Vol. 1)*, pp. 55–64. North-Holland, Amsterdam.

Hoar, B. R., Chomel, B. B., Rolfe, D. L., Chang, C. C., Fritz, C. L., Sacks, B. N., and Carpenter, T. E. (2003). Spatial analysis of yersinia pestis and bartonella vinsonii subsp berkhoffii seroprevalence in California coyotes (canis latrans). *Preventive Veterinary Medicine*, **56**, 299–311.

Hopkins, R. S. and Scott, A. S. (1983). Handling raw chicken as a source of sporadic *campylobacter jejuni* infections [letter]. *J. Infect. Dis.*, **148**, 770.

Impe, J. F. V., Nicolai, B. M., Schellekens, M., Martens, T., and Baerdemaeker, J. D. (1995). Predictive microbiology in a dynamic environment: a system theory approach. *International Journal of Food Microbiology*, **25**, 227–49.

Jacobs-Reitsma, W., Fiddelaers, J., Bolder, N., Beumer, R., Wagenaar, J., and Hazeleger, W. (2003). Darkling beetles and larvae as vectors for Campylobacter and Salmonella transmission to broilers. CHRO2003, available from `http://www.rivm.nl/carma/verwantonderzoek/index_eng.html`.

Jacobs-Reitsma, W. F., Bolder, N. M., and Mulder, R. W. A. W. (1994). Cecal carriage of Campylobacter and Salmonella in dutch broiler flocks at slaughter; a one year study. *Poultry Science*, **73**, 1260–66.

Jacobs-Reitsma, W. F., van de Giessen, A. W., Bolder, N. M., and Mulder, R. W. (1995). Epidemiology of *Campylobacter* spp. at two dutch broiler farms. *Epidemiol. Infect.*, **114**, 413–421.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory.* Academic Press, New York, USA.

Jun, Y., Yi, L., Yong, T., Jianben, L., Xiong, C., Qin, Z., Jiaxin, D., Songsheng, Q., and Ziniu, Y. (2003). Action of cu/sup 2+/ on bacillus thuringiensis growth investigated by microcalorimetry. *Applied Biochemistry and Microbiology*, **39**(6), 576–580.

Kærgaard, N., Bredahl, L., Frandsen, S. E., Hermansen, J. E., Ovesen, L., Sandøe, P., Wegener, H., and Walter-Jørgensen, A. (2002). *Den danske svinesektors samfundsmæssige betydning*, chapter Sundhed, miljø og dyrevelfærd, pp. 79–80. Jordbrugsforlaget, Det Kgl. Danske Landhusholdningsselskab, Greve, Denmark. In Danish.

Kapperud, G. and Aasen, S. (1992). Descriptive epidemiology of infections due to thermotolerant Campylobacter spp. in Norway, 1979-1988. *APMIS*, **100**, 883–90.

Kapperud, G., Skjerve, E., Bean, N. H., Ostroff, S. M., and Lassen, J. (1992). Risk factors for sporadic *Campylobacter* infections: results of a case-control study in southeastern Norway. *J. Clin. Microbiol.*, **30**, 3117–3121.

Kapperud, G., Skjerve, E., Vik, L., Hauge, K., Lysaker, A., Aalmen, I., Ostroff, S. M., and Potter, M. (1993). Epidemiological investigation of risk factors for campylobacter colonization in norwegian broiler flocks. *Epidemiology and Infection*, **111**(2), 245–55.

Kovárová-Kovar, K. and Egli, T. (1998). Growth kinetics of suspended microbial cells: From single-substrate-controlled growth to mixed-substrate kinetics. *Microbiology and Molecular Biology Reviews*, **62**(3), 646–66.

Kristensen, N., Madsen, H., and Jorgensen, S. (2004). Parameter estimation in stochastic grey-box models. *Automatica*, **40**(2), 225–237.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

Kulldorff, M. (1999). *Recent Advances on Scan Statistics and Applications*, chapter Spatial scan statistics: Models, calculations and applications, pp. 303–322. Birkhauser, Boston.

Lendenmann, U. and Egli, T. (1995). Is escherichia coli growing in glucose-limited chemostat culture able to utilize other sugars without lag? *Microbiology - Reading*, **141**(1), 71–78.

Lendenmann, U. and Egli, T. (1998). Kinetic models for the growth of escherichia coli with mixtures of sugars under carbon-limited conditions. *Biotechnology and Bioengineering,*, **59**, 99–107.

Luechtfeld, N. A., Blaser, M. J., Reller, L. B., and Wang, W. L. (1980). Isolation of Campylobacter fetus subsp. jejuni from migratory waterfowl. *J. Clin. Microbiol.*, **12**, 406–408.

Madsen, H. (1997). *Statistics in finance*. Technical University of Denmark, Informatics and Mathematical Modelling, Lyngby.

Melby, K., Dahl, O. P., and Penner, J. L. (1990). Clinical and serological manifestations in patients during a waterborne epidemic due to *campylobacter jejuni*. *J. Infect.*, **21**, 309–316.

Monod, J. (1942). *Recherches sur la croissance des cultures bactériennes*. Hermann et Cie, Paris, France.

Narang, A. (1998). The dynamical analogy between microbial growth on mixtures of substrates and population growth of competing species. *Biotechnology and Bioengineering*, **59**(1), 116–121.

Narang, A., Konopka, A., and Ramkrishna, D. (1997a). The dynamics of microbial growth on mixtures of substrates in batch reactors. *Journal of Theoretical Biology*, **184**, 301–17.

Narang, A., Konopka, A., and Ramkrishna, D. (1997b). New patterns of mixed-substrate utilization during batch growth of Escherichia coli k12. *Biotechnology and Bioengineering,*, **55**(5), 747–57.

Naus, J. I. (1965a). Clustering of random points in two dimensions. *Biometrika*, **52**(1/2), 263–267.

Naus, J. I. (1965b). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, **60**(310), 532–538.

Neal, K. R. and Slack, R. C. (1995). The autumn peak in *Campylobacter* gastroenteritis. are the risk factors the same for travel- and UK-acquired *Campylobacter* infections? *J. Public Health Med.*, **17**, 98–102.

Neimann, J., Engberg, J., Mølbak, K., and Wegener, H. C. (2003). A case-control study of risk factors for sporadic Campylobacter infections in Denmark. *Epidemiology and Infection*, **130**(3), 353–66.

Nielsen, H. A. and Madsen, H. (2001). A generalization of some classical time series tools. *Computational Statistics and Data Analysis*, **37**, 13–31.

Nielsen, J. and Joergensen, H. (1996). A kinetic model for the penicillin biosynthetic pathway in penicillium chrysogenum. *Control Engineering Practice*, **4**(6), 765–771.

Nielsen, J. N., Madsen, H., and Young, P. C. (2000). Parameter estimation in stochastic differential equations: An overview. *Annual Reviews in Control*, **24**, 83–94.

Nies, D. H. (1999). Microbial heavy-metal resistance. *Appl Microbiol Biotechnol*, **51**, 730–750.

Norval, M. (2001). Effects of solar radiation on the human immune system. *Journal of Photochemistry and Photobiology*, **63**, 28–40.

Nylen, G., Dunstan, F., Palmer, S. R., Anderson, Y., Bager, F., Cowden, J., Feierl, G., Galloway, Y., Kapperud, G., Megraud, F., Molbak, K., Petersen, L. R., and Ruutu, P. (2002). The seasonal distribution of campylobacter infection in nine European countries and New Zealand. *Epidemiology and Infection*, **128**(3), 383–90.

Obiri-Danso, K., Paul, N., and Jones, K. (2001). The effects of uvb and temperature on the survival of natural populations and pure cultures of campylobacter jejuni, camp. coli, camp. lari and urease-positive thermophilic campylobacters (uptc) in surface waters. *J. Appl. Microbiol.*, **90**, 256–267.

Øksendal, B. (1989). *Stochastic Differential Equations - An introduction with Applications*. Springer Verlag, Berlin.

Paul, G., Syddall, M., Kent, C., and Thomas, C. (1998). A structured model for penicillin production on mixed substrates. *Biochemical Engineering Journal*, **2**(1), 11–21.

Petersen, C. and Møller, L. B. (2000). Control of copper homeostasis in escherichia coli by a p-type atpase, copa, and a merr-like transcriptional activator, copr. *Gene*, **261**, 289–298.

Poulsen, H. D. and Carlson, D. (2004). Environmental effects of the termination of AGP use in Denmark - in relation to zinc and copper. In *Beyond Antimicrobial Growth Promoters in Food Animal Production*, pp. 91–99. Danish Institute of Agricultural Sciences.

Rautelin, H. and Hänninen, M.-L. (2000). Campylobacters: the most common bacterial enteropathogens in the Nordic countries. *Ann. Med.*, **32**, 440–5.

Rensing, C. and Grass, G. (2003). Escherichia coli mechanisms of copper homeostasis in a changing environment. *FEMS Microbiology Reviews*, **27**, 197–213.

Roche, L. M., Skinner, R., and Weinstein, R. B. (2002). Use of a geographic information system to identify and characterize areas with high proportions of distant stage breast cancer. *Journal of Public Health Management and Practice*, **8**, 26–32.

Rodrigues, L. C., Cowden, J. M., Wheeler, J. G., Sethi, D., Wall, P. G., Cumberland, P., Tompkins, D. S., Hudson, M. J., Roberts, J. A., and Roderick, P. J. (2001). The study of infectious intestinal disease in England: risk factors for cases of infectious intestinal disease with *campylobacter jejuni* infection. *Epidemiol. Infect.*, **127**, 185–193.

Rosef, O. and Kapperud, G. (1983). House flies (musca domestica) as possible vectors of Campylobacter fetus subsp. jejuni. *Appl. Environ. Microbiol.*, **45**, 381–383.

Rosen, B. P. (2002). Transport and detoxification systems for transition metals,heavy metals and metalloids in eukaryotic and prokaryotic microbes. *Comparative Biochemistry and Physiology*, **133**, 689–693.

Sauders, B. D., Fortes, E. D., Morse, D. L., Dumas, N., Kiehlbauch, J. A., Schukken, Y., Hibbs, J. R., and Wiedmann, M. (2003). Molecular subtyping to detect human listeriosis clusters. *Emerging Infectious Diseases*, **9**, 672–680.

Shane, S. M., Montrose, M. S., and Harrington, K. S. (1985). Transmission of *campylobacter jejuni* by the housefly (*musca domestica*). *Applied Environmental Microbiology*, **29**, 384–391.

Skirrow, M. B. and Blaser, M. J. (2000). Clinical aspects of *Campylobacter* infection. In Nachamkin, I. and Blaser, M. J., editors, Campylobacter , pp. 69–88. ASM Press, Washington DC.

Skov, M. N., Feld, N. C., Carstensen, B., and Madsen, M. (2002). The serologic response to Salmonella enteritidis and Salmonella typhimurium in experimentally infected chickens, followed by an indirect lipopolysaccharide enzyme-linked immunosorbent assay and bacteriologic examinations through a one-year period. *Avian Diseases*, **46**, 265–73.

Speelmon, E. C., Checkley, W., Gilman, R. H., Patz, J., Calderon, M., and Manga, S. (2000). Cholera incidence and El Niño-related higher ambient temperature. *JAMA*, **283**, 3072–3074.

Stanley, K. N., Wallace, J. S., Currie, J. E., Diggle, P. J., and Jones, K. (1998). The seasonal variation of thermophilic campylobacters in beef cattle, dairy cattle and calves. *Journal of Applied Microbiology*, **85**, 472–80.

Studahl, A. and Andersson, Y. (2000). Risk factors for indigenous *Campylobacter* infection: a Swedish case-control study. *Epidemiol. Infect.*, **125**, 269–275.

Tenkate, T. D. and Stafford, R. J. (2001). Risk factors for campylobacter infection in infants and young children: a matched case-control study. *Epidemiol. Infect.*, **127**, 399–404.

Tong, S., Bi, P., Parton, K., Hobbs, J., and McMichael, A. (1998). Climate variability and transmission of epidemic polyarthritis. *Lancet*, **351**, 1100.

Venkatesh, K., Doshi, P., and Rengaswamy, R. (1997). Optimal strategy to model microbial growth in a multiple substrate environment. *Biotechnology and Bioengineering*, **56**(6), 635–44.

Wallace, J. S., Stanley, K. N., Currie, J. E., Diggle, P. J., and Jones, K. (1997). Seasonality of thermophilic campylobacter populations in chickens. *Journal of Applied Microbiology*, **82**, 219–24.

Wegener, H. C., Hald, T., Wong, D. L. F., Madsen, M., Korsgaard, H., Bager, F., Gerner-Smidt, P., and Mølbak, K. (2003). *Salmonella* control programs in Denmark. *Emerging Infectious Diseases*, **9**(7), 774–780.